Princess Sumaya University for Technology
جامعة الأميرة سميّة للتكنولوجيا

صندوق دعم البحث العلمي والابتكار
Scientific Research and Innovation Support Fund

# Jordanian Journal of Computers and Information Technology

June 2025 VOLUME 11 NUMBER 02 ISSN 2415 - 1076 (Online)
ISSN 2413 - 9351 (Print)

JJCIT

www.jjcit.org jjcit@psut.edu.jo

**An International Peer-Reviewed Scientific Journal Financed by the Scientific Research and Innovation Support Fund**

## Jordanian Journal of Computers and Information Technology (JJCIT)

The Jordanian Journal of Computers and Information Technology (JJCIT) is an international journal that publishes original, high-quality and cutting edge research papers on all aspects and technologies in ICT fields.

JJCIT is hosted and published by Princess Sumaya University for Technology (PSUT) and supported by the Scientific Research Support Fund in Jordan. Researchers have the right to read, print, distribute, search, download, copy or link to the full text of articles. JJCIT permits reproduction as long as the source is acknowledged.

### Aims and Scope

The JJCIT aims to publish the most current developments in the form of original articles as well as review articles in all areas of Telecommunications, Computer Engineering and Information Technology and make them available to researchers worldwide. The JJCIT focuses on topics including, but not limited to: Computer Engineering & Communication Networks, Computer Science & Information Systems and Information Technology and Applications.

### Indexing

JJCIT is indexed in:

ELSEVIER Scopus

INDEXED IN EMERGING SOURCES CITATION (Web of Science) CLARIVATE ANALYTICS

DOAJ DIRECTORY OF OPEN ACCESS JOURNALS

crossref Text and Data Mining

OCLC WorldCat

BASE

ACADEMIA

Google scholar

### Editorial Board Support Team

### JJCIT Address

Website: www.jjcit.org
Email: jjcit@psut.edu.jo
Address: Princess Sumaya University for Technology, Khalil Saket Street, Al-Jubaiha
B.O. Box: 1438 Amman 11941 Jordan
Telephone: +962-6-5359949
Fax: +962-6-7295534

## EDITORIAL BOARD

## INTERNATIONAL ADVISORY BOARD

"Opinions or views expressed in papers published in this journal are those of the author(s) and do not necessarily reflect those of the Editorial Board, the host university or the policy of the Scientific Research Support Fund".

"ما ورد في هذه المجلة يعبر عن آراء الباحثين ولا يعكس بالضرورة آراء هيئة التحرير أو الجامعة أو سياسة صندوق دعم البحث العلمي والابتكار".

# INTRUSION DETECTION SYSTEM FOR INTERNET OF MEDICAL THINGS USING GRU WITH ATTENTION MECHANISM-BASED HYBRID DEEP LEARNING TECHNIQUE

Naveen Saran and Nishtha Kesswani

*(Received: 6-Sep.-2024, Revised: 8-Nov.-2024, Accepted: 21-Nov.-2024)*

## *ABSTRACT*

*The proliferation of Internet of Things devices in healthcare, specifically the Internet of Medical Things, has revolutionized patient care and health-monitoring systems. Integrating these interconnected medical devices introduces unprecedented security challenges, necessitating robust Intrusion Detection Systems (IDSs) to safeguard patient data and healthcare infrastructure. To protect the IoMT devices from numerous malicious attacks, researchers have developed numerous Intrusion Detection Systems, but the development of an effective and real-time IDS remains a challenge. Our proposed IDS addresses this gap and surpasses state-of-the-art IDS techniques for IoMT networks. In this research paper, we have proposed a novel IDS approach for IoMT, leveraging a Hybrid Deep Learning technique to enhance detection accuracy and efficiency. By combining the strengths of Gated Recurrent Unit (GRU) and Attention Mechanism, the proposed IDS achieves superior performance in detecting anomalous activities in medical networks. We evaluated the proposed IDS model on two publicly available benchmark intrusion datasets and achieved 99.99 % accuracy on the ICU Healthcare Dataset and 98.94 % on the NF-TON-IoT Dataset. Precision, Recall, F1-score metrics and ROC-AUC for the proposed model are promising. We also added Noise to the features to show how effectively the model performed in noisy environments. Moreover, we used the K-Fold Cross Validation Technique to cross-validate the model's performance on both datasets, ensuring the reliability and applicability of the suggested IDS model for IoMT networks.*

## *KEYWORDS*

*Internet of medical things, Intrusion-detection system, Principal-component analysis, Deep learning, Gated recurrent unit, Attention mechanism, K-fold cross validation.*

## 1. INTRODUCTION

The Internet of Things (IoT) innovation has changed various industries, such as health care. 560M wearable are expected to ship by 2024, which will track and visualize real-time healthcare data [1] and this number could increase. The Internet of Medical Things (IoMT) has introduced several devices, including wearable health monitors, implantable medical devices and smart hospital equipment, that have brought a revolution in the way patients are treated, diagnosed and cared for by enabling remote monitoring, real-time health tracking, personalized medicine [2] and ultimately become an integral part of Smart Healthcare Systems. However, the widespread adoption of IoMT devices has introduced unprecedented security challenges, particularly concerning the protection of sensitive patient data produced by these Internet of Medical Things (IoMT) apps for remote healthcare monitoring from vital signs and other signals like Electro-Cardio-Gram (ECG) and Electro-Encephalo-Gram (EEG) and the integrity of healthcare infrastructure. Nonetheless, instances of cyber-attacks targeting sensitive medical information present a severe risk. The aim is to protect health data against multiple intrusion attacks [3]. Thus, an attacker tampering with this data can cause severe medical problems, including misdiagnosis, thereby resulting in delays in emergency care or causing death [4]. Consequently, the research examines the safety of IoMT-generated health data from the perspective of Smart Healthcare Systems.

Attackers can remotely control IoMT devices to build IoT-based botnets, since they are simple to hack and attack. These assaults result in violations, infringements and disclosures of sensitive data inside the wider IoT-enabled system. Common attacks on Internet of Things (IoT)-based healthcare devices

N. Saran is with Department of Computer Science and N. Kesswani is with Department of Data Science & Analytics, Central University of Rajasthan, Bandarsindri, Kishangarh, Ajmer, Rajasthan, India. Emails: naveen.saran90@gmail.com and nishtha@curaj.ac.in

include denial-of-service (DOS), ransomware, distributed denial-of-service (DDoS) and botnet attacks [5]. Figure 1 illustrates the cyber-attacks on IoMT network communication targeting vulnerable sensor devices towards known, unknown and zero-day intrusion attacks by intruders. The inter-connectivity of these devices, coupled with their susceptibility to cyber-threats, significantly expose patients' data safety.

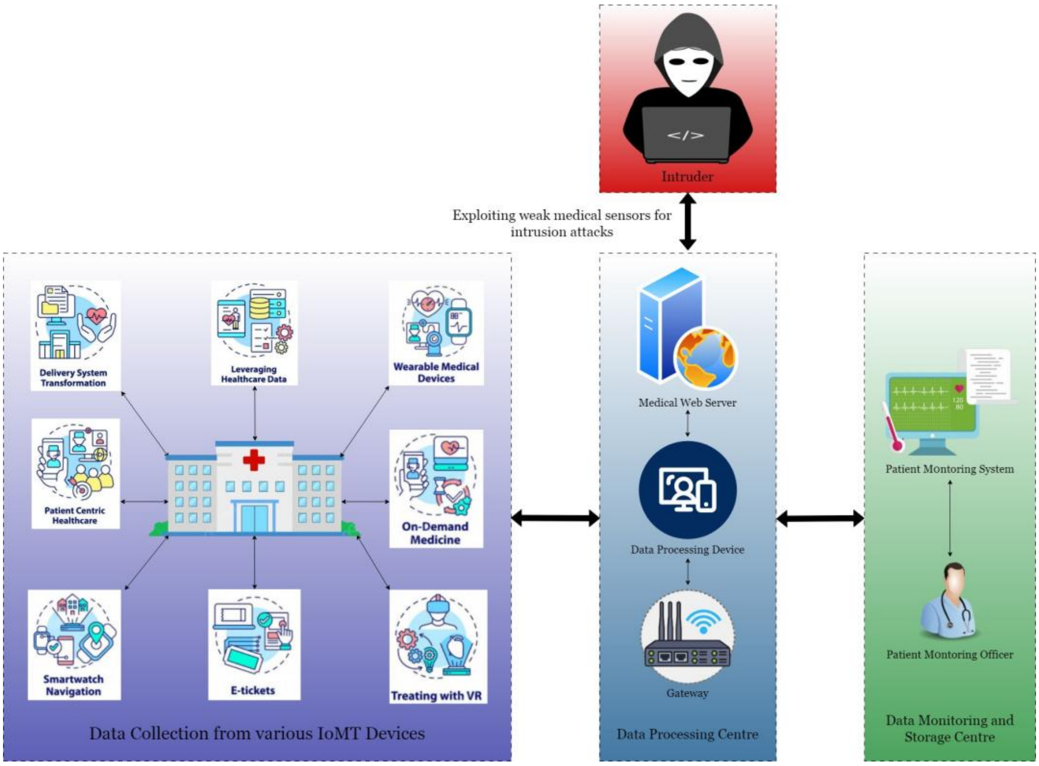

Intruder
Exploiting weak medical sensors for intrusion attacks
Delivery System Transformation
Leveraging Healthcare Data
Wearable Medical Devices
Patient Centric Healthcare
On-Demand Medicine
Smartwatch Navigation
E-tickets
Treating with VR
Data Collection from various IoMT Devices
Medical Web Server
Data Processing Device
Gateway
Data Processing Centre
Patient Montoring System
Patient Montoring Officer
Data Monitoring and Storage Centre

Figure 1. Internet of medical things vulnerability and attacks.

To address such emerging cyber-threats, there has been a need for robust Intrusion Detection Systems (IDSs) specifically designed for the Internet of Medical Things [6]-[8]. An IDS acts as the primary defence line against cyber-attacks by continuously monitoring network traffic, identifying any abnormal activity and alerting healthcare providers about possible security breaches instantly. However, traditional IDS solutions may not be suitable in an IoMT environment due to their unique nature where resources are limited, devices have different architectures and they operate dynamically [9]. To address these challenges, this research proposes a novel approach for intrusion detection in IoMT, leveraging a Hybrid Deep Learning technique to enhance detection accuracy and efficiency. The novelty of the proposed approach is the integration of a Gated Recurrent Unit (GRU) with an Attention Mechanism for Intrusion Detection and Classification and Principal Component Analysis (PCA) for Feature Engineering. The proposed Intrusion Detection System (IDS) seeks to offer all-encompassing coverage of anomalous behaviors and cyber-threats within medical networks. Also, we have used K-Fold Cross-validation procedures, which assess the model's performance on every fold, to ensure the model's robustness.

This paper presents a detailed analysis of the proposed IDS architecture, its implementation and its performance evaluation using real-time scenario-based publicly available benchmark datasets like ICU Healthcare Dataset and NF-TON-IoT Dataset. In summary, this research contributes to advancing security in healthcare by developing a specialized IDS solution tailored for the Internet of Medical Things. By leveraging the power of hybrid Deep Learning techniques, the proposed IDS offers enhanced known, unknown and zero-day intrusion attack-detection capabilities in real-time scenarios, safeguarding patient data and ensuring the integrity of IoMT infrastructure. The main contributions of this paper are as follows:

- **Comprehensive Data Pre-processing:** This has been achieved by standardizing numerical features, reducing Dimensions using PCA and handling class imbalance with SMOTE.

- **GRU with Attention Mechanism:** The incorporation of a custom attention layer captures temporal dependencies and important features.
- **Hyper-parameter Tuning:** This has been carried out using keras-tuner to optimize model parameters.
- **Robust Model Training:** This has been ensured by early stopping to prevent over-fitting.
- **Robustness to Noise:** The model performance has been evaluated with added noise.
- **Cross-validation (CV) for Stability:** A stratified K-Fold CV was carried out for stability assessment.
- **Scalability and Real-time Potential:** The proposed model is designed for real-time intrusion-detection scenarios and the scalable architecture is adaptable to new threats.

Our research presents an innovative approach to enhancing Intrusion-detection Systems for the Internet of Medical Things by implementing a sophisticated GRU model with an attention mechanism. This advanced architecture effectively captures temporal dependencies and highlights significant features in IoMT data, improving the detection of intricate cyber-attacks. Comprehensive data pre-processing steps, including standardization, dimensionality reduction using PCA and handling class imbalance with SMOTE, ensure the model's effectiveness. Hyper-parameter tuning using a keras-tuner optimizes model parameters and early stopping prevents over-fitting during robust model training. We tackle the issue of imbalanced data by utilizing the Synthetic Minority Over-sampling Technique (SMOTE) and ensure the model's robustness through K-Fold Cross-validation. Our IDS is designed for real-time processing, enabling swift detection and response to emerging threats. It offers high scalability and adaptability, accommodating new threats and processing diverse IoMT data. The model's robustness to noise is verified by evaluating performance with added noise. Evaluated on comprehensive ICU Healthcare and NF-TON-IoT Datasets, our model accurately identifies known and novel attacks, significantly enhancing security in real-time IoMT environments. Our evaluation results demonstrate that this proposed method outperforms existing techniques, which lack these advanced capabilities.

The rest of the research article is organized as follows: The paper's second section discusses the related research works based on machine-learning and deep-learning techniques used to implement the IDS model for the IoMT network. The third section discusses the preliminary concepts used in the proposed IDS model. The fourth section consists of the proposed research methodology. The fifth section contains complete information about the experimental setting required to operate and build an effective IDS model for IoMT. The sixth section comprises the experimental results and a detailed analytical report of our proposed IDS model. At last, section seven discusses the conclusion and the future scope of the proposed research work.

## 2. Related Works

In this work, we analyze the literature on intrusion detection in the Internet of Medical Things networks, focusing on applying various Machine Learning (ML) and Deep Learning(DL) techniques researchers employ to develop their respective IDS models. We explore the use of traditional ML algorithms such as Decision Trees (DTs) [10], Random Forests (RFs) and Ensemble Learning (EL) [11] for anomaly detection and classification tasks in IoMT networks. Additionally, we examine advanced DL approaches, including Convolutional Neural Networks (CNNs) [12], Recurrent Neural Networks (RNNs) [13] and others that offer data-driven and automated solutions for identifying complex patterns and anomalies in network-traffic data. Through this comprehensive review, we aim to highlight the strengths and drawbacks of ML and DL approaches in the context of IoMT, providing insights into their efficacy, scalability and adaptability to evolving cyber-threats in healthcare environments. Tables 1 and 2 emphasize the critical facts of state-of-the-art machine-learning and deep-learning Techniques proposed by researchers to construct an IDS model for IoMT networks.

### 2.1 Machine Learning-based IDS for IoMT

Using a fog-cloud-based architecture, Kumaret al. [14] suggested Ensemble Learning (EL) based IDS for IoMT environments. To identify attackers in the edge-centric IoMT framework, Nandy et al. [15] suggested an Empirical Intelligent Agent (EIA) based on a novel Swarm-Neural Network (Swarm-NN) technique. To secure the data of IoMT applications, Singh et al. [16] presented a Dew-Cloud-based model employing Hierarchical Federated Learning (HFL). Wagan et al. [17] described the Duo-Secure IoMT framework, which distinguishes between routine IoMT data and attack patterns using multi-modal sensory signals data. Khan et al. [18] examined the suggested technique for IoMT cyber-attacks by employing ensemble- based techniques and fog-cloud infrastructure. Using a meta-learning strategy, Zukaib et al. [19] developed a Meta-IDS model that improves the detection of known and zero-day intrusions in the IoMT environment.

Table 1. State-of-the-art machine-learning techniques.

| References | Methodology | Dataset | Accuracy | Limitations |
|---|---|---|---|---|
| Kumar et al. [14] | Ensemble learning using fog-cloud architecture. | TON-IoT | 96.35% | To detect intrusions in IoMT networks, the suggested model is not sufficiently compared in the paper to other cutting-edge ensemble or DL models. |
| Nandy et al. [15] | Using a Swarm-Neural Network approach based on Empirical Intelligent Agents (EIAs) to detect threats and evaluate the effectiveness of health data. | TON-IoT | 99.50% | The paper does not compare the proposed Swarm-NN approach and existing intrusion-detection techniques in the IoMT frameworks. |
| Singh et al. [16] | For data privacy in IoMT, the Hierarchical Long- Short Term Memory model is implemented at dispersed Dew servers with a cloud computing-supported backend. | TON-IoT | 99.31% | The paper's comparison of the HFL-HLSTM model with existing intrusion-detection techniques is limited. |
| Wagan et al. [17] | Dynamic Fuzzy C-Means clustering with Bi-LSTM technique to identify attack patterns within the IoMT network. | WUSTL EHMS 2020 | 89.67% | Dataset holds very few records. |
| Khan et al. [18] | Fog-cloud architecture and Ensemble Learning to handle IoMT security concerns. | TON-IoT | 98.56% | The paper fails to identify emerging or unknown attacks in future IoMT environments. |

Table 2. State-of-the-art deep-learning techniques.

| References | Methodology | Dataset | Accuracy | Limitations |
|---|---|---|---|---|
| RM et al. [20] | PCA with Grey-wolf Optimization for feature engineering and DNN for attack classification. | KDD99, UNSW-B15 | 99.99%, 89.13% | Lacks real-time implementation in live IoMT environment with sensitive medical data. |
| Khan et al. [21] | SDN enabled hybrid deep learning-based IDS for IoMT. | IoT | 99.83% | The paper lacks a comprehensive comparison with other state-of-the-art malware-detection methods on the same dataset. |
| Awotunde et al. [22] | Swarm-Neural Network-based IDS model for IoMT devices. | NF-TON-IoT | 89.00% | The paper fails to specify which existing models or techniques were used to compare the proposed model's performance on the same dataset. |
| Saran et al. [23] | S-RNN, LSTM and GRU-based deep-learning technique to build IDS model for IoMT devices. | ICU Healthcare | 99.00% | The paper fails to handle imbalanced datasets and model's robustness in noisy conditions. |
| Saheed et al. [24] | Deep Recurrent Neural Networks and Supervised machine-learning models were used to develop an IDS for IoMT environment. | NSL-KDD | 99.76% | The dataset lacks the latest real-world IoMT attacks and traffic. |
| Changanti et al. [25] | PSO with DNN is used to implement an effective and accurate IDS in IoMT. | WUSTLE HMS 2020 | 96.00% | Dataset holds very few records. |
| Alzubi et al. [26] | Blended DL framework leveraging the CNN-LSTM to recognize the latest intrusion attacks and defend the healthcare data. | CIC-IDS 2018 | 98.53% | Data imbalance and bias-mitigation strategies are not investigated. |

### 2.2 Deep-learning Techniques for IDS in IoMT

An IDS based on Deep Neural Networks (DNNs) was proposed by RM et al. [20] to identify and anticipate unknown threats in an IoMT context. Using a combination of CNN and LSTM techniques, Khan et al. [21] provided an SDN-enabled hybrid Deep Learning-based IDS for IoMT. To identify intrusions in the data-centric IoMT system, Awotunde et al. [22] proposed a Swarm-Neural Network-based intrusion detection system (IDS) model. Saran et al. [23] discussed S-RNN, LSTM and GRU-based Deep-learning Techniques to identify intrusion attacks in the IoMT environment. In Saheed et al. [24], it was shown how an efficient and effective IDS for classifying and forecasting unforeseen cyber-threats in the IoMT environment can be developed using a Deep Recurrent Neural Network (DRNN) and Supervised Machine Learning (SML) models (Random Forest, Decision Tree, KNN and Ridge Classifier). To enhance the performance of IDS in IoMT, Changanti et al. [25] introduced the DNN-based DL model and the PSO feature-selection technique. Alzubi et al. [26] presented a hybrid deep-learning framework that combines the advantages of Long Short-Term Memory (LSTM) and Convolutional Neural Network (CNN) to effectively detect the most recent intrusion attacks and safeguard medical data.

## 3. PRELIMINARIES

Preliminary concepts that are used to build a robust, secure and effective IDS solution are discussed in the following sub-section.

### 3.1 Gated Recurrent Unit (GRU) with Attention Mechanism

By enabling the model to focus on the most essential parts of the input sequence during prediction, the integration of an Attention Mechanism [27] with a Gated Recurrent Unit (GRU) [28] enhances the performance of the GRU. Unlike the general LSTM, CNN and GRU hybrid techniques, this technique efficiently prioritises essential temporal dependencies for the model, leading to enhanced accuracy in various IoT/IoMT applications. The GRU effectively prevents information loss by using update and reset gates to regulate information flow and refresh the hidden state. To determine what information should be prioritized, the Attention Mechanism computes a context vector by adding the weighted total of the hidden states. Alignment scores are used to calculate the hidden-state weights. By combining these two methods, the model's capacity to manage long-range dependencies in sequential data is greatly enhanced, which makes it very helpful for complicated-pattern identification in IoMT contexts. The core equations for the GRU with Attention Mechanism are as follows:

**GRU Equations:**

$$\mathrm{z}_t = \sigma(\mathrm{W}_z\mathrm{x}_t + \mathrm{U}_z\mathrm{h}_{t-1} + \mathrm{b}_z) \tag{1}$$

$$\mathrm{r}_t = \sigma(\mathrm{W}_r\mathrm{x}_t + \mathrm{U}_r\mathrm{h}_{t-1} + \mathrm{b}_r) \tag{2}$$

$$\tilde{\mathrm{h}}\boldsymbol{t} = \tanh(\mathrm{W}_h\mathrm{x}_t + \mathrm{U}_h(\mathrm{r}_t \odot \mathrm{h}_{t-1}) + \mathrm{b}_h) \tag{3}$$

$$\mathrm{h}_t = (1 - \mathrm{z}_t) \odot \mathrm{h}_{t-1} + \mathrm{z}_t \odot \tilde{\mathrm{h}}\boldsymbol{t} \tag{4}$$

where $\mathrm{x}_t$ is the input, zt is the update gate, $\mathrm{r}_t$ is the reset gate, $\mathrm{h}\tilde{}_t$ is the candidate's hidden state and $\mathrm{h}_t$ is the hidden state at time step *t*. The functions σ and tanh represent the sigmoid and hyperbolic tangent activation functions, respectively. ⊙ is the element-wise multiplication function [29].

**Attention-mechanism Equations:**

$$\boldsymbol{e_{t,i}} = \mathrm{V}^{\mathrm{T}}\, \mathbf{tanh}(\mathbf{W}_{\boldsymbol{e}}\boldsymbol{h_t} + \mathbf{U}_{\boldsymbol{e}}\mathbf{h}_{\boldsymbol{i}} + \mathbf{b}_{\boldsymbol{e}}) \tag{5}$$

$$\boldsymbol{\alpha_{t,i}} = \frac{\mathbf{exp}(\boldsymbol{e_{t,i}})}{\sum_{\boldsymbol{j=1}}^{\boldsymbol{T}} \mathbf{exp}(\boldsymbol{e_{t,j}})} \tag{6}$$

$$\mathrm{c}_t = \sum_{i=1}^{T} \alpha_{t,i}\mathrm{h}_i \tag{7}$$

where the context vector is $\mathrm{c}_t$, the weight vector is v and the hidden state at time step *t* is represented by $\mathrm{e}_{t,i}$, the alignment score is $\alpha_{t,i}$ and the total number of time steps in the input sequence is T. The hyperbolic tangent activation function for the attention mechanism is represented as tanh. Additionally, the weight matrices for the respective gates and the attention mechanism are $\mathrm{W}_z$, $\mathrm{W}_r$, Wh, $\mathrm{W}_e$; the weight matrices for the hidden state and the attention mechanism are $\mathrm{U}_z$, $\mathrm{U}_r$, $\mathrm{U}_h$, $\mathrm{U}_e$; and the bias vectors for the respective gates and the attention mechanism are $\mathrm{b}_z$, $\mathrm{b}_r$, $\mathrm{b}_h$, $\mathrm{b}_e$.

## 4. Proposed Research Methodology

The proposed IDS model to detect known, unknown and zero-day attacks in IoMT leverages a GRU with an Attention Mechanism-based DL technique. This comprehensive approach includes data pre-processing, feature extraction using PCA and handling class imbalance with SMOTE. The model is rigorously trained and evaluated, ensuring robustness and reliability through K-Fold cross-validation and hyper-parameter tuning with Keras Tuner, which allows defining an optimized attention function that directly helps priorities the most important features without huge extra computational costs, unlike all other existing attention-based IDS models accessible for IoT and IoMT applications. The following sub-sections provide a detailed breakdown and analysis of the critical components of the operational methodology.

### 4.1 Proposed IDS-model Architecture

Figure 2 illustrates the proposed IDS model architecture utilizing a GRU with Attention Mechanism-based Hybrid DL technique. This architecture is designed to effectively capture temporal dependencies and highlight significant features in IoMT data. Figures 3 and 4 depict the detailed GRU with Attention Mechanism model layer architecture applied to the ICU Healthcare and NF-TON-IoT datasets, respectively, demonstrating the multiple layers (Input, GRU, Attention and Dense or Output Layers) to construct the robust and secure IDS model for IoMT network.

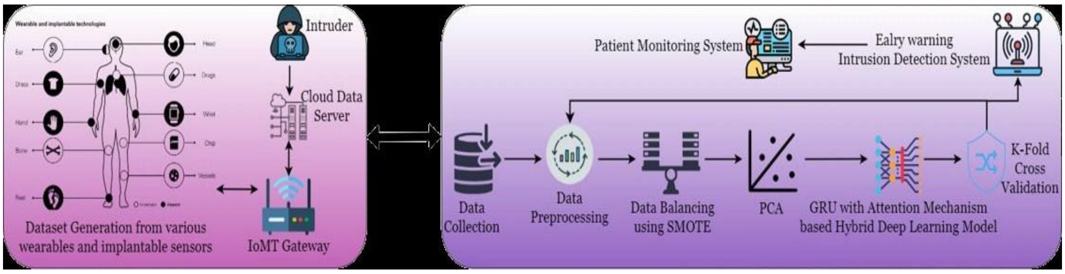
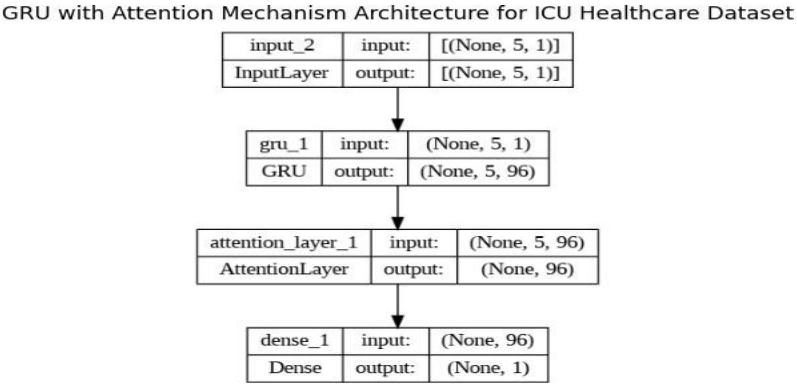

Wearable and implantable technologies
Intruder
Cloud Data Server
Dataset Generation from various wearables and implantable sensors
IoMT Gateway
Patient Monitoring System
Ealry warning Intrusion Detection System
Data Collection
Data Preprocessing
Data Balancing using SMOTE
PCA
GRU with Attention Mechanism based Hybrid Deep Learning Model
K-Fold Cross Validation

Figure 2. Intrusion-detection system's model architecture for IoMT network.


GRU with Attention Mechanism Architecture for ICU Healthcare Dataset

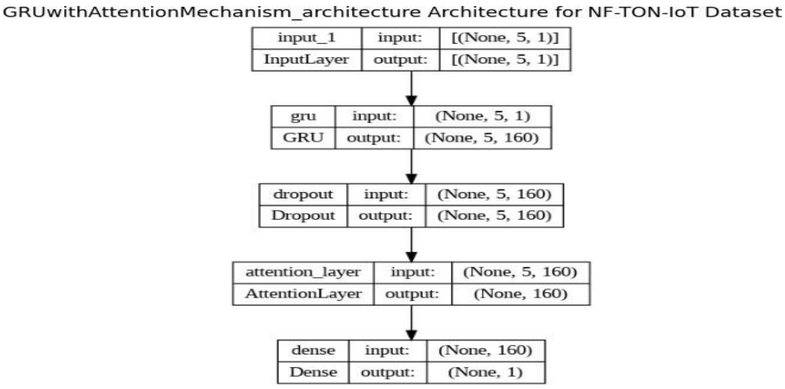
input_2 | input: | [(None, 5, 1)]
InputLayer | output: | [(None, 5, 1)]

gru_1 | input: | (None, 5, 1)
GRU | output: | (None, 5, 96)

attention_layer_1 | input: | (None, 5, 96)
AttentionLayer | output: | (None, 96)

dense_1 | input: | (None, 96)
Dense | output: | (None, 1)

Figure 3. GRU with attention mechanism model layer architecture for ICU healthcare dataset.


GRUwithAttentionMechanism_architecture Architecture for NF-TON-IoT Dataset

input_1 | input: | [(None, 5, 1)]
InputLayer | output: | [(None, 5, 1)]

gru | input: | (None, 5, 1)
GRU | output: | (None, 5, 160)

dropout | input: | (None, 5, 160)
Dropout | output: | (None, 5, 160)

attention_layer | input: | (None, 5, 160)
AttentionLayer | output: | (None, 160)

dense | input: | (None, 160)
Dense | output: | (None, 1)

Figure 4. GRU with attention mechanism model layer architecture for NF-TON-IoT dataset.

## 4.2 Proposed IDS Framework

This GRU with an Attention Mechanism-based Hybrid Deep Learning IDS framework presents a comprehensive workflow for addressing known, unknown and zero-day intrusion attacks in real-time IoMT scenarios. It includes data pre-processing, model building and evaluation, feature reduction and performance visualization, providing a detailed approach for enhancing model performance in this critical field. The framework utilizes the ICU Healthcare and NF-TON-IoT datasets to evaluate the effectiveness of the proposed IDS model, ensuring the robustness of the proposed Intrusion Detection Systems in the IoMT network environment. The step-by-step operational phases to construct our proposed secure IDS model for the IoMT network are as follows.

1) **Data Pre-processing**
    a) Cleaning the data by handling infinite, missing and duplicate values and encoding categorical target labels in numerical format.
    b) Balancing the dataset using SMOTE (Synthetic Minority Over-sampling Technique) to address the class-imbalance issue.
    c) Splitting the data into training and testing sets and performing feature scaling.

2) **Feature Extraction Using Principal Component Analysis (PCA)**

    PCA extracts features, lowering the dataset's dimensionality while keeping the essential features for efficient and accurate model training.

3) **Training the GRU with Attention Mechanism-based Hybrid DL Model for Classification**
    a) **Model Architecture:** Constructs and trains the GRU with Attention Mechanism-based hybrid DL model aiming to classify network-traffic data effectively into different categories of network attacks or normal behavior. This model combines:
        i. GRU layers for temporal feature extraction.
        ii. The Attention Mechanism is trained to learn where to pay attention at each time step by aligning the relevant contextual information with the hidden states of GRU.
        iii. Dense layers for final classification.

    b) **Hyper-parameter Tuning**
        i. It utilizes Keras Tuner for hyper-parameter optimization, finding the best configuration for GRU units and dropout rate.
        ii. It uses early stopping, which stops the training process when the model's performance on a validation set reaches an unacceptable level, to prevent overfitting.

    c) **Model Training**
        i. The model is trained on the pre-processed and feature-extracted dataset.
        ii. The model's performance is assessed in depth by utilizing a variety of metrics, including Accuracy, ROC curves, Precision, Recall, F1-Score and Confusion Matrix.

4) **Cross-validation:** The K-Fold cross-validation technique is utilized to rigorously evaluate the model's performance across five different splits of the datasets, ensuring robustness and generalizability.

5) **Handling Noisy Data**
    a) We use Gaussian noise to demonstrate the robustness of the model under noisy conditions with an average of zero and a standard deviation of 0.1 introduced to the features, with the effect of emulating interference related to surroundings and communication networks in IoT systems such as fluctuations in network signals and transmission errors.
    b) A Random Forest Classifier is trained on noisy data to compare performance.
    c) We found the accuracy for instances with the added noise, proving its robustness based on the ICU Healthcare dataset as 99.98% and 99.93% for the NF-TON-IoT dataset.

6) **Performance Visualization**

Algorithm 1 presents a detailed methodology for constructing our advanced Hybrid Deep Learning IDS model incorporating all the above processes tailored specifically for IoMT networks.

**Algorithm 1:** Algorithm for Intrusion Detection in IoMT

```
Require: Dataset $D$ with features $X \in \mathbf{R}^{m.n}$ and labels $y \in \mathbf{Z}^{m}$ (where $m$ is the number of samples and
    $n$ is the number of features).
Ensure: Trained GRU with Attention Mechanism model
Load and Preprocess Data:
Load dataset D from the CSV file
Separate features X and target y
Select only numeric features. $\mathbf{X}_{\mathbf{numeric}}$
Standardize numerical features using StandardScaler:
    $\mathbf{X}_{\mathbf{scaled}}$ = scaler.fit_transform($X_{\text{numeric}}$)
Dimensionality Reduction with PCA: $\mathbf{X}_{\mathbf{pca}}$ = pca.fit_transform($X_{\text{scaled}}$)
Encode Target Variable using LabelEncoder:
    $\mathbf{y}_{\mathbf{encoded}}$ = label_encoder.fit_transform($y$)
Split Data into Training and Testing Sets: $\mathbf{X}_{\mathbf{train}}$, $\mathbf{X}_{\mathbf{test}}$, $\mathbf{y}_{\mathbf{train}}$, $\mathbf{y}_{\mathbf{test}}$ =
    train_test_split($X_{\text{pca}}$, $y_{\text{encoded}}$, test_size = 0.2, random_state = 42)
Handle Class Imbalance with SMOTE:
    $X_{\mathbf{smote}}$, $y_{\mathbf{smote}}$ = smote.fit_resample($X_{\text{train}}$, $y_{\text{train}}$)
Define and Train GRU with Attention Model:
Define AttentionLayer class
Convert data to NumPy array and reshape for GRU model: $\mathbf{X}_{\mathbf{train\_np}}$, $\mathbf{X}_{\mathbf{test\_np}}$
Define a model-building function for hyperparameter tuning using Keras Tuner to find the
    best model
Train the best model on $\mathbf{X}_{\mathbf{train\_np}}$, $\mathbf{y}_{\mathbf{smote}}$ with EarlyStopping callback
Save and load the trained model
Evaluate the Model on Test Data: $\mathbf{X}_{\mathbf{test\_np}}$, $\mathbf{y}_{\mathbf{test}}$
Calculate and print accuracy, precision, recall, F1-score and AUC-ROC
Visualize loss and accuracy graphs and confusion matrix
Evaluate Model with Noise and Class Imbalance:
Add Random Noise to features and split data into train and test sets
Train and evaluate the accuracy of the Random Forest classifier on noisy data
K-Fold Cross-Validation:
Perform and visualize the K-fold cross-validation scores on five folds to assess model
    generalization
```

# 5. EXPERIMENTAL SETTING

## 5.1 Experimental Setup

We utilized a high-performance computing environment to establish a robust experimental setup for our Intrusion Detection System tailored for Internet of Medical Things (IoMT) network communication. This setup featured a 1TB hard drive operating on an Intel Core i7-6700 CPU clocked at 3.40GHz with 8GB RAM, running Ubuntu 20.04.4 LTS for stability and advanced networking capabilities. Leveraging Google Colab, a scalable cloud-based platform renowned for its computational power, facilitated our IDS model's development and simulation phases, accommodating extensive computations and large datasets. We chose Python for its rich ecosystem of libraries supporting machine-learning models, which is crucial for our testing and validation processes. To rigorously assess the effectiveness and accuracy of our IDS model, we conducted experiments using two prominent publicly available benchmark datasets, the ICU Healthcare dataset and the NF-TON-IoT dataset. These datasets are well-regarded for their comprehensive coverage of cyber-attacks across the IoMT network environment, providing critical benchmarks for evaluating our system.

## 5.2 Dataset Description

### 5.2.1 ICU Healthcare Dataset

The ICU Healthcare Intrusion Dataset [30] offers a rich source of network-traffic data captured from a simulated ICU healthcare environment, providing valuable insights into cyber-security threats and vulnerabilities in healthcare settings. The dataset consists of the records under three types of network traffic classes: Attack class, Environment-monitoring class and Patient-monitoring class, as depicted in Figure 5. In Figure 6, we have demonstrated the distribution of balanced and unbalanced labeled classes in the ICU Healthcare dataset using SMOTE. The detailed specification of the ICU Healthcare dataset is mentioned below.

1. **Size:** The dataset comprises 1,88,694 network-traffic data records captured over a specific period, providing a comprehensive representation of various network activities within an ICU healthcare environment.
2. **Features:** Each data record includes detailed information about network-traffic attributes in 52 categories such as source IP address, destination IP address, protocol type, timestamp, packet size and network port.
3. **Traffic Composition:** The dataset holds three types of traffic classes: a.) Attack class, b.) Environment-monitoring class and c.) Patient-monitoring class. The Normal class (Environment-monitoring class and Patient-monitoring class) and Attack class of the IoT healthcare dataset hold 1,08,568 and 80,126 records, respectively.
4. **Labels:** The dataset includes labeled instances as 0 (Normal) and 1 (Attack), indicating the presence or absence of cyber-security threats or anomalies in the IoMT-enabled heathcare network traffic. The dataset contains four distinct attack types: Brute Force, MQTT Publish Flood, MQTT Distributed Denial-of-Service (DDoS) and SlowITE [31].

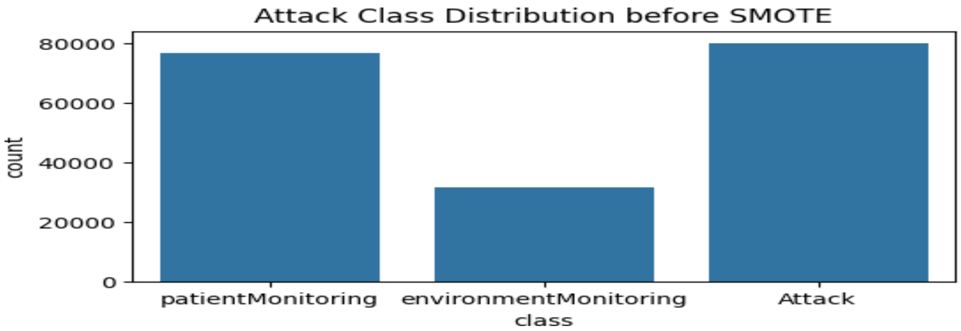


Figure 5. ICU healthcare dataset attack categories.

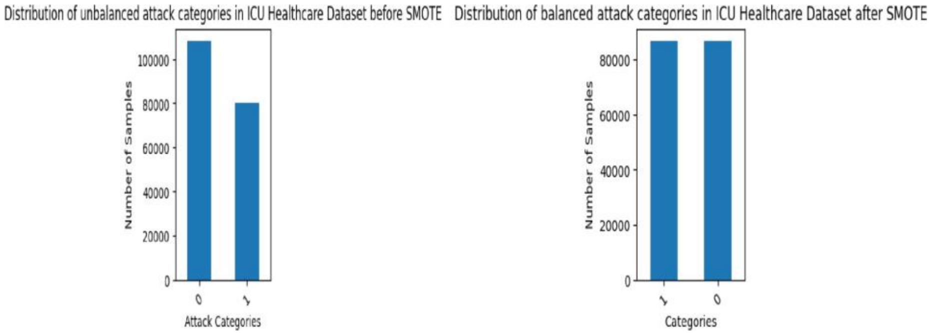


Figure 6. ICU healthcare dataset attack categories, before and after SMOTE.

### 5.2.2 NF-TON-IoT Dataset

The NF-ToN-IoT dataset offers a comprehensive source of network-traffic data specifically captured from IoT environments. This dataset is essential for understanding cyber-security threats and vulnerabilities within IoT networks. Its availability supports research, development and evaluation of Intrusion Detection Systems (IDSs), Machine-learning models and security measures tailored for safeguarding IoT infrastructure. The dataset comprises records categorized under different types of network-traffic classes, as depicted in Figure 7, providing a valuable resource for anomaly detection and traffic analysis. Figure 8 demonstrates balancing unbalanced distributed labeled classes in the NF-TON-IoT Dataset using SMOTE. A thorough inspection of the NF-TON-IoT dataset is as follows:

1. **Size:** The dataset comprises 1,379,274 network traffic-data records, offering a comprehensive representation of various network activities within IoT environments.
2. **Features:** Each data record includes detailed information about network traffic attributes in 14

categories, such as source IP address, destination IP address, protocol type, packet size, network port, among others.

3. **Traffic Composition:** The dataset includes two types of traffic classes:
   a) **Attack class:** Represents malicious network traffic consisting of DoS, Injection, DDoS, Scanning, Password, MITM, XSS, Backdoor and Ransomware.
   b) **Benign class:** Represents non-malicious, regular network traffic.
4. **Labels:** Labels help in identifying the presence or absence of cyber-security threats or attacks in IoT network traffic:
   a) **0 (Normal):** Indicates benign network traffic.
   b) **1 (Attack):** Indicates malicious network traffic.

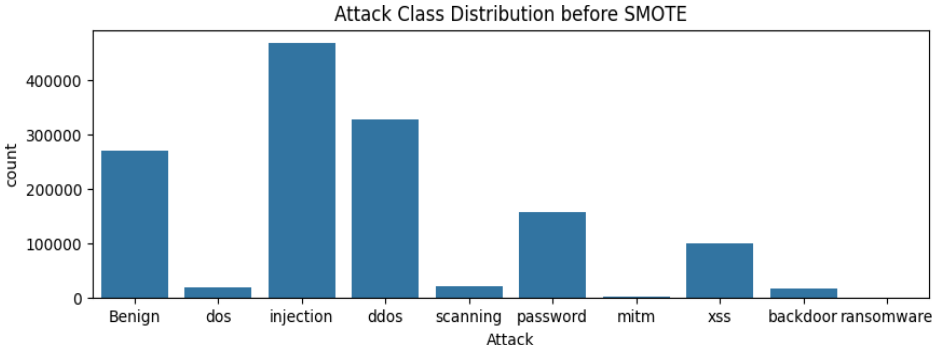


Figure 7. NF-TON-IoT dataset attack categories.

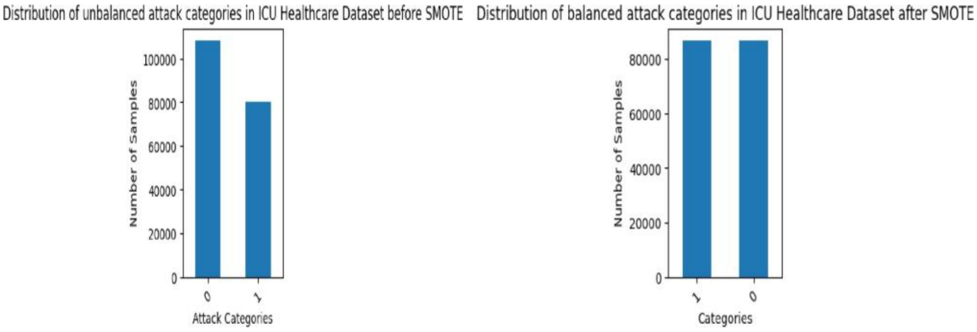


Figure 8. NF-TON-IoT dataset attack categories, before and after SMOTE.

## 5.3 Performance Evaluation

### 5.3.1 Performance Metrics

Performance metrics serve as fundamental measures for evaluating the effectiveness of classification models, offering critical insights into the capability of the proposed IDS model to classify instances and detect anomalies accurately. These metrics collectively provide a robust evaluation framework for assessing its real-world efficacy. The key performance metrics include:

1) **Precision (%):** Precision measures how many accurate positive predictions there are among all the positive predictions the model makes. This can be computed as follows:

$$\text{Precision } \% = \left(\frac{TP}{TP+FP}\right) \times 100 \tag{8}$$

2) **Recall (%):** The ratio of true positive predictions to all real positive instances in the dataset is known as recall and it may be calculated as follows:

$$\text{Recall } \% = \left(\frac{TP}{TP+FN}\right) \times 100 \tag{9}$$

3) **F1-score (%):** The F1-score, calculated as follows, is the harmonic mean of precision and recall and offers a fair evaluation of the model’s performance.

$$\text{F1} - \text{score } \% = \left(\frac{2TP}{2TP+FP+FN}\right) \times 100 \tag{10}$$

4) **Accuracy (%):** Accuracy reflects the proportion of correctly classified instances among all instances in the dataset, expressed as:

$$\text{Accuracy}\ \% = \left(\frac{TP+TN}{TP+TN+FP+FN}\right) \times 100 \tag{11}$$

5) **Accuracy with noise (%):** The algorithm for calculating model accuracy when noisy data is included is the same as for conventional accuracy calculation, but it is applied to the predictions produced on the noisy dataset and can be expressed as follows:

$$\text{Accuracy with noise}\ \% = \left(\frac{TP_{noisy}+TN_{noisy}}{TP_{noisy}+TN_{noisy}+FP_{noisy}+FN_{noisy}}\right) \times 100 \tag{12}$$

6) **Mean K-Fold Cross-validation (CV) Accuracy (%):** The average accuracy ratings from each cross-validation fold are the mean K-Fold cross-validation accuracy. This offers a broader gauge of the model's effectiveness, denoted by:

$$\text{Mean K} - \text{Fold CV Accuracy}\ \% = \left(\frac{\sum_{i=1}^{K}\left(\frac{TP_i+TN_i}{TP_i+TN_i+FP_i+FN_i}\right)}{K}\right) \times 100 \tag{13}$$

### Confusion Matrix

We assess the effectiveness of our suggested classification model using a confusion matrix. By comparing predicted labels with actual labels, it classifies predictions into True Positives (TP), True Negatives (TN), False Positives (FP) and False Negatives (FN). Figures 9 and 10 show the true and predicted values derived from our proposed Hybrid DL IDS model in confusion-matrix format on ICU Healthcare and NF-TON-IoT datasets, respectively.

## 6. Experiments and Result Analysis

In real-time Internet of Medical Things (IoMT) contexts, the proposed hybrid Deep Learning IDS model based on Attention Mechanism and GRU is designed to identify known, unknown and zero-day attacks. Evaluation of the IDS model utilized the benchmark ICU Healthcare and NF-TON-IoT Datasets. In the operational phase, raw datasets underwent comprehensive pre-processing steps, including data cleaning, normalization and dimensionality reduction using PCA to enhance computational efficiency. Subsequently, the processed data was fed into the GRU with Attention Mechanism for sequence modeling, leveraging its ability to capture intricate temporal dependencies in healthcare-network traffic.

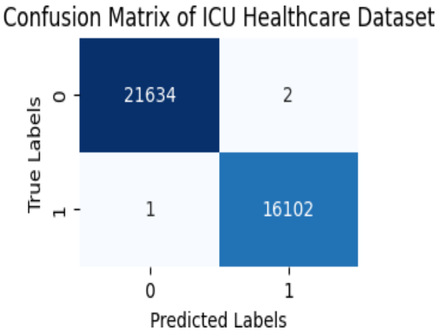


Figure 9. Confusion matrix of ICU healthcare dataset.

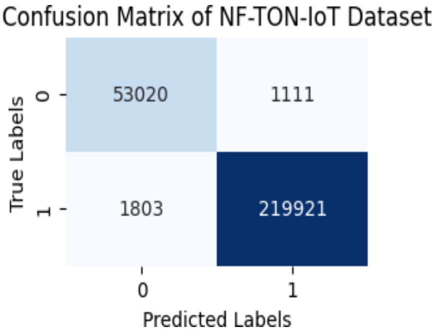


Figure 10. Confusion matrix of NF-TON-IoT dataset.

Table 3. Classification report of the proposed IDS model on both the given datasets.

| Dataset | Precision (%) | Recall (%) | F1-score (%) | Accuracy (%) | Accuracy with Noise (%) | Mean K-Fold CV Accuracy (%) | Support |
|---|---|---|---|---|---|---|---|
| ICU Healthcare Dataset | 99.98 | 99.99 | 99.99 | 99.99 | 99.98 | 99.99 | 37739 |
| NF-TON-IoT Dataset | 99.50 | 99.17 | 99.34 | 98.94 | 99.93 | 97.30 | 275855 |

The evaluation of the proposed IDS model demonstrates robust performance across multiple metrics and datasets. The accuracy achieved by the ICU Healthcare and NF-TON-IoT datasets reached 99.99% and 98.94%, respectively. Precision, Recall, F1-score, Accuracy with Noise and Mean K-fold cross-validation accuracy results are summarized in the Classification Report as shown in Table 3. The GRU with Attention Mechanism-based Hybrid DL IDS model was trained for 20 epochs on both datasets, maintaining consistent loss and accuracy trends throughout each epoch. Figures 11 and 12 illustrate dataset loss and accuracy graphs, showcasing the model's learning process and convergence over time.

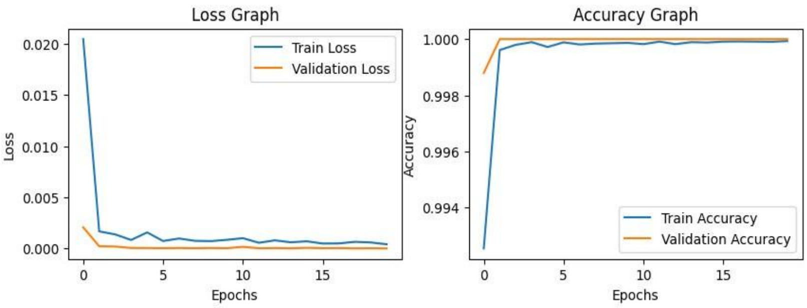


Figure 11. Loss accuracy graph for ICU healthcare dataset.

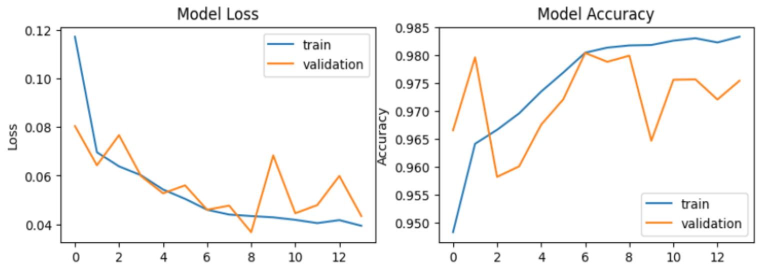


Figure 12. Loss accuracy graph for NF-TON-IoT dataset.

Table 4. Results comparison of state-of-the-art IDS techniques with proposed IDS model on IoMT devices.

| **References** | **Dataset** | **Technique** | **Precision (%)** | **Recall (%)** | **F1-score (%)** | **Accuracy (%)** | **Accuracy with Noise (%)** | **CV Accuracy (%)** |
|---|---|---|---|---|---|---|---|---|
| Kumar *et al.* [14] | TON-IoT | Ensemble Learning | 90.54 | 99.98 | 95.03 | 96.35 | . . . | . . . |
| Khan *et al.* [18] | TON-IoT | Ensemble Learning | . . . | . . . | . . . | 98.56 | . . . | . . . |
| Awotunde *Et al.* [22] | NF-TON-IoT | Swarm-Neural-Network | . . . | . . . | . . . | 89.00 | . . . | . . . |
| **Proposed** | NF-TON-IoT | Hybrid Deep Learning | 99.50 | 99.17 | 99.34 | **98.94** | 99.93 | 97.30 |
| Khan *et al.* [21] | ICU Healthcare | Hybrid Deep Learning | 96.34 | 99.11 | 100.00 | 99.83 | . . . | . . . |
| Saran *et al.* [23] | ICU Healthcare | Deep Learning | … | … | … | 99.00 | … | … |
| **Proposed** | ICU Healthcare | Hybrid Deep Learning | 99.98 | 99.99 | 99.99 | **99.99** | 99.98 | 99.99 |

Five splits for K-fold cross-validation were employed to assess the robustness and generalizability of the proposed IDS model. They obtained a Per-fold accuracy score and a mean K-fold accuracy score of 99.99% for the ICU Healthcare dataset and 97.30% for the NF-TON-IoT datasets, respectively, as depicted in Table 3. This analysis ensures the model performs consistently well across different data splits, reinforcing its reliability in diverse IoMT environments. Table 4 compares the proposed IDS model with the existing techniques for various attack-detection capabilities and performance metrics on ICU Healthcare and NF-TON-IoT datasets.

In summary, the experimental results validate the effectiveness and reliability of the proposed GRU with Attention Mechanism-based Hybrid DL IDS model for detecting intrusions in IoMT networks. The model's ability to handle complex healthcare data, robust performance metrics and comparative analysis against existing techniques underscores its potential as a critical security measure in modern healthcare infrastructures.

## 7. Conclusion and Future Scope

Securing Internet of Medical Things (IoMT) networks against known, unknown, and zero-day intrusion attacks is crucial for maintaining patient-data privacy and operational integrity in healthcare environments. In response to these challenges, we have developed and evaluated a sophisticated IDS model for IoMT applications, integrating PCA for dimensionality reduction, SMOTE for handling imbalanced datasets and a GRU with Attention Mechanism for sequence modeling. As shown in Table 3, at Section 6, our suggested IDS model performs admirably across a variety of evaluation criteria, including Precision. This research work produced a suitable model for testing new datasets by incorporating changes in the model to Recall, F1-score, Accuracy, Accuracy with noise and K-Fold Cross-validation Accuracy. We showed the model's effectiveness in detecting various types of intrusions while lowering false positives and false negatives, with 99.99% accuracy for the ICU Healthcare dataset and 98.94% accuracy for the NF-TON-IoT dataset. The evaluation of the IDS model also incorporated the analysis of Accuracy with noise and K-Fold Cross-validation Accuracy. By introducing noise into the dataset, we assessed the robustness of the model, ensuring that it maintains high accuracy even under adverse conditions, as 99.98% for the ICU Healthcare dataset and 99.93% for the NF-TON-IoT dataset. The K-Fold Cross-validation provided a comprehensive evaluation by training and validating the model across multiple data splits, resulting in an average accuracy of 99.99% for the ICU Healthcare dataset and 97.30% for the NF-TON-IoT dataset, demonstrating the model's consistent performance. We can accommodate several amendments for further enhancement and exploration to include new attack approaches to achieve robust applicability in real-life IoT and IoMT scenarios. We can test a light version of the proposed IDS model with fewer GRU units and attention layers to decrease the demands on the resources used while retaining the high detection rate of real-time intrusion detection, especially in resource-constrained IoMT environments.

## Acknowledgements

Thanks are due to Dr. Nishtha Kesswani, for her invaluable guidance, unwavering support and insightful feedback throughout this research. Her expertise and encouragement have been instrumental in completing this work.

## References

[1] F. Laricchia, "Topic: Wearables," Statista, [Online], Available: https://www.statista.com/topics/1556/wearable-tec, Accessed: Jun. 18, 2024.

[2] R. Dwivedi, D. Mehrotra and S. Chandra, "Potential of Internet of Medical Things (IoMT) Applications in Building a Smart Healthcare System: A Systematic Review," Journal of Oral Biology and Craniofacial Research, vol. 12, no. 2, pp. 302-318, 2022.

[3] S. S. Ambarkar and N. Shekokar, "Toward Smart and Secure IoT Based Healthcare System," Proc. of Internet of Things, Smart Computing and Technology: A Roadmap Ahead, Studies in Systems, Decision and Control, vol. 266, pp. 283-303, Springer, 2020.

[4] A. Tabassum, A. Erbad, A. Mohamed and M. Guizani, "Privacy-preserving Distributed IDS Using Incremental Learning for IoT Health Systems," IEEE Access, vol. 9, pp. 14271-14283, 2021.

[5] R. U. Rasool, H. F. Ahmad, W. Rafique, A. Qayyum and J. Qadir, "Security and Privacy of Internet of Medical Things: A Contemporary Review in the Age of Surveillance, Botnets and Adversarial ML," Journal of Network and Computer Applications, vol. 201, p. 103332, 2022.

[6] M. L. Hernandez-Jaimes et al., "Artificial Intelligence for IoMT Security: A Review of Intrusion Detection Systems, Attacks, Datasets and Cloud-Fog-Edge Architectures," Internet of Things, vol. 23, p. 100887, 2023.

[7] M. Wazid, A. K. Das, J. J. Rodrigues, S. Shetty and Y. Park, "IoMT Malware Detection Approaches: Analysis and Research Challenges," IEEE Access, vol. 7, pp. 182459-182476, 2019.

[8] A. Si-Ahmed et al., "Survey of Machine Learning Based Intrusion Detection Methods for Internet of Medical Things," Applied Soft Computing, vol. 140, p. 110227, 2023.

[9] Y. Rbah et al., "Machine Learning and Deep Learning Methods for Intrusion Detection Systems in IoMT: A Survey," Proc. of the 2022 2$^{nd}$ IEEE Int. Conf. on Innovative Research in Applied Science, Engineering and Technology (IRASET), pp. 1-9, Meknes, Morocco, 2022.

[10] K. Gupta et al., "A Tree Classifier Based Network Intrusion Detection Model for Internet of Medical Things," Computers and Electrical Engineering, vol. 102, p. 108158, 2022.

[11] J. Nayak, S. K. Meher, A. Souri, B. Naik and S. Vimal, "Extreme Learning Machine and Bayesian Optimization-driven Intelligent Framework for IoMT Cyber-attack Detection," The Journal of Supercomputing, vol. 78, no. 13, pp. 14866-14891, 2022.

[12] S. Liaqat et al., "SDN Orchestration to Combat Evolving Cyber Threats in Internet of Medical Things (IoMT)," Computer Communications, vol. 160, pp. 697-705, 2020.

[13] I. A. Khan et al., "XSRU-IoMT: Explainable Simple Recurrent Units for Threat Detection in Internet of Medical Things networks," Future Generation Computer Systems, vol. 127, pp. 181-193, 2022.

[14] P. Kumar et al., "An Ensemble Learning and Fog-cloud Architecture-driven Cyber-attack Detection Framework for IoMT Networks," Computer Communications, vol. 166, pp. 110- 124, 2021.

[15] S. Nandy et al., "An Intrusion Detection Mechanism for Secured IoMT Framework Based on Swarm-neural Network," IEEE J. of Biomedical and Health Informatics, vol. 26, no. 5, pp. 1969-1976, 2021.

[16] P. Singh et al., "Dew-cloud-based hierarchical federated learning for intrusion detection in IoMT," IEEE Journal of Biomedical and Health Informatics, vol. 27, no. 2, pp. 722-731, 2022.

[17] S. A. Wagan et al., "A Fuzzy-based Duo-secure Multi-modal Framework for IoMT Anomaly Detection," J. of King Saud Uni.-Computer and Information Sciences, vol. 35, no. 1, pp. 131-144, 2023.

[18] F. Khan et al., "A Secure Ensemble Learning-based Fog-cloud Approach for Cyberattack Detection in IoMT," IEEE Transactions on Industrial Informatics, vol. 19, no. 10, pp. 10125 – 10132, 2023.

[19] U. Zukaib et al., "Meta-IDS: Meta-learning Based Smart Intrusion Detection System for Internet of Medical Things (IoMT) Network," IEEE Internet of Things J., vol. 11, no. 13, pp. 23080 – 23095, 2024.

[20] S. P. RM et al., "An Effective Feature Engineering for DNN Using Hybrid PCA-GWO for Intrusion Detection in IoMT Architecture," Computer Communications, vol. 160, pp. 139-149, 2020.

[21] S. Khan and A. Akhunzada, "A hybrid DL-driven Intelligent SDN-enabled Malware Detection Framework for Internet of Medical Things (IoMT)," Computer Comm., vol. 170, pp. 209-216, 2021.

[22] J. B. Awotunde et al., "A Deep Learning-based Intrusion Detection Technique for a Secured IoMT System," Proc. of the Int. Conf. on Informatics and Intelligent Applications (ICIIA 2021), Part of the Book Series: Communications in Computer and Information Science, vol. 1547, pp. 50-62, Nov. 2021.

[23] N. Saran, N. Kesswani and R. Saharan, "Intrusion Detection System Using Deep Learning Techniques for Internet of Medical Things (IoMT)," Proc. of the International Conference on Deep Learning, Artificial Intelligence and Robotics (ICDLAIR 2023), Part of the Book Series: Lecture Notes in Networks and Systems, vol. 1001, pp. 752-763, Springer, Aug. 2024.

[24] Y. K. Saheed and M. O. Arowolo, "Efficient Cyber Attack Detection on the Internet of Medical Things-smart Environment Based on Deep Recurrent Neural Network and Machine Learning Algorithms," IEEE Access, vol. 9, pp. 161546-161554, 2021.

[25] R. Chaganti et al., "A Particle Swarm Optimization and Deep Learning Approach for Intrusion Detection System in Internet of Medical Things," Sustainability, vol. 14, no. 19, p. 12828, 2022.

[26] J. A. Alzubi, O. A. Alzubi, I. Qiqieh and A. Singh, "A Blended Deep Learning Intrusion Detection Framework for Consumable Edge-centric IoMT Industry," IEEE Transactions on Consumer Electronics, vol. 70, no. 1, pp. 2049 – 2057, 2024.

[27] F. Laghrissi, S. Douzi, K. Douzi and B. Hssina, "IDS-attention: An Efficient Algorithm for Intrusion Detection Systems Using Attention Mechanism," Journal of Big Data, vol. 8, no. 1, p. 149, 2021.

[28] M. V. Assis et al., "A GRU Deep Learning System against Attacks in Software Defined Networks," Journal of Network and Computer Applications, vol. 177, p. 102942, 2021.

[29] X. Miao, S. Li, Y. Zhu and Z. An, "A Novel Real-time Fault Diagnosis Method for Planetary Gearbox Using Transferable Hidden Layer," IEEE Sensors Journal, vol. 20, no. 15, pp. 8403-8412, 2020.

[30] F. Hussain et al., "IoT Healthcare Security Dataset," IEEE Dataport, [Online], Available: https://ieee-dataport.org/keywords/healthcare-security-dataset, 2021.

[31] F. Hussain et al., "A Framework for Malicious Traffic Detection in IoT Healthcare Environment," Sensors, vol. 21, no. 9, p. 3025, 2021.

## ملخص البحث:

لقد أدّى انتشار أجهزة إنترنت الأشياء في الرّعاية الصّحّية، وبخاصّة إنترنت الأشياء الطّبّية، إلى إحداث ثورةٍ في أنظمة رعاية المرضى والمراقبة الصّحّية. إلّا أنّ تكامل هذه الأجهزة الطّبّية المتّصلة بعضها ببعض قاد إلى إدّخال تحدّياتٍ غير مسبوقة تتعلّق بأمن البيانات، الأمر الذي يؤكّد الحاجة إلى أنظمةٍ متينةٍ لاكتشاف التّسلُّل لحماية بيانات المرضى والبنية التحتية للرعاية الصّحّية. ولحماية أجهزة إنترنت الأشياء الطّبّية من العديد من الهجمات، قام الباحثون بتطوير أنظمةٍ عديدةٍ لاكتشاف التّسلُّل لاختراق تلك الأجهزة والحصول على بيانات حسّاسة على نحوٍ غير مشروع. ومع ذلك، فإنّ تطوير نظامٍ فعّالٍ ومتينٍ لاكتشاف التّسلُّل يظلّ تحدّياً بحدّ ذاته.

نقترح في هذه الورقة نظاماً لاكتشاف التّسلُّل لإنترنت الأشياء الطّبّية باستخدام وحدة التّكرار المبوّبة مع تقنية التّعلُّم العميق الهجينة القائمة على آلية الانتباه. وقد أجريت تجارب عملية على النّظام الهجين المقترح، وبرهن على أداءٍ ممتازٍ في حماية بيانات أجهزة إنترنت الأشياء الطّبّية من الهجمات السّبرانية المختلفة. وجرى تقييم النّظام المقترح على بعض مجموعات البيانات في مجال الرّعاية الصّحّية، وذلك باستخدام عددٍ من مؤشّرات الأداء، بما فيها الدّقّة، إلى جانب مقارنة النّظام المقترح بعددً من الأنظمة المشابهة الواردة في دراسات سابقة أخرى، حيث أثبت النّظام المقترح تفوّقاً ملحوظاً على غيره من الأنظمة بتحقيقه مؤشّرات أداءٍ عالية، مما يؤكّد متانته وفعاليته، وبالتّالي نجاعته في حماية أجهزة إنترنت الأشياء الطّبّية من الهجمات والاختراقات.

Jordanian Journal of Computers and Information Technology (JJCIT), Vol. 11, No. 02, June 2025.

# HAML-IRL: OVERCOMING THE IMBALANCED RECORD-LINKAGE PROBLEM USING HYBRID ACTIVE MACHINE LEARNING

Mourad Jabrane, Mouad Jbel, Imad Hafidi and Yassir Rochd

*(Received: 14-Sep.-2024, Revised: 14-Nov.-2024, Accepted: 24-Nov.-2024)*

## ABSTRACT

*Traditional active machine-learning (AML) methods employed in Record Linkage (RL) or Entity Resolution (ER) tasks often struggle with model stability, slow convergence and handling imbalanced data. Our study introduces a novel hybrid Active Machine Learning approach to address RL, overcoming the challenges of limited labeled data and imbalanced classes. By combining and balancing informativeness, which selects record pairs to reduce model uncertainty and representativeness, it is ensured that the chosen pairs reflect the overall dataset patterns. Our hybrid approach, called Hybrid Active Machine Learning for Imbalanced Record Linkage (HAML-IRL), demonstrates significant advancements. HAML-IRL achieves an average 12% improvement in F1-scores across eleven real- world datasets, including structured, textual and dirty data, when compared to state-of-the-art AML methods. Our approach also requires up to 60% - 85% fewer labeled samples depending on the datasets, accelerates model convergence and offers superior stability across iterations, making it a robust and efficient solution for real-world record-linkage tasks.*

## KEYWORDS

*Record linkage, Entity resolution, Active machine learning, Hybrid query.*

## 1. INTRODUCTION

In the rapidly evolving field of digital data management, Record Linkage (RL)—also known as Duplicate Detection or Entity Resolution—has become increasingly vital for ensuring data integrity across a multitude of industries. As organizations continue to collect and utilize vast amounts of data from diverse sources, the need to accurately link records that refer to the same entity is paramount. This process of RL is critical for maintaining accurate and consistent data representations, which are foundational to effective data management, analytics and informed decision- making processes across various domains, such as healthcare, finance, e-commerce and government services [1]. At its core, RL involves the identification and merging of records from one or more datasets that correspond to the same real-world entity, despite potential variations in how the data is represented. This task, while conceptually straightforward, is often fraught with challenges due to issues, such as data-entry errors, incomplete records and the lack of unique identifiers across datasets. These challenges are further exacerbated in environments that rely heavily on machine learning-based RL methods, where the performance of the RL system is highly dependent on the availability and quality of labeled data [2]. The need for extensive labeled datasets to train machine-learning models poses significant obstacles, particularly in scenarios where labeled data is scarce or prohibitively expensive to obtain. This reliance on large volumes of labeled data often results in a bottleneck, slowing down the deployment and scalability of RL systems. In response to these challenges, the field has witnessed the emergence of Active Machine Learning (AML) as a promising approach to mitigate the data-dependency problem. AML is designed to enhance learning efficiency by actively selecting the most informative data points for labeling, thereby reducing the total amount of labeled data required to achieve high performance. This approach is particularly beneficial in situations where labeled data is sparse, expensive or time-consuming to acquire.

AML employs two primary strategies to optimize the learning process: informativeness and representativeness. Informativeness focuses on selecting data points that are expected to most significantly reduce the model's uncertainty, thus accelerating the learning process by focusing on the most challenging cases. Representativeness, on the other hand, ensures that the selected data points are

M. Jabrane (Corresponding Author), M. Jbel, I. Hafidi and Y. Rochd are with LIPIM Laboratory, University Sultan Moulay Slimane, Beni-Mellal 23000, Morocco. Emails: jabrane.mourad@usms.ac.ma, mouad.jbel@usms.ac.ma, i.hafidi@usms.ma and y.rochd@usms.ma

reflective of the broader dataset, helping create a training set that is more generalizable and robust. However, traditional AML approaches often prioritize one of these strategies at the expense of the other, leading to observable deficits in performance. This trade-off can result in models that are either highly specialized but prone to overfitting or general but lacking in the ability to resolve complex or uncertain cases effectively. Moreover, these traditional AML methodologies frequently struggle with issues related to model instability and slow convergence, particularly in scenarios characterized by imbalanced data. In many RL tasks, the negative class (representing non-matching pairs) vastly outnumbers the positive class (representing matching pairs), which introduces a significant bias into the dataset. As noted by Christen [3], this imbalance can severely skew the learning process, leading to models that are biased towards predicting non-matches, thus resulting in sub-optimal performance outcomes. This challenge is further compounded by the iterative nature of AML, where each round of learning and querying may amplify the inherent biases present in the data.

To address these challenges, we propose a novel Hybrid Active Machine-learning framework called HAML-IRL (Hybrid Active Machine-learning for Imbalanced Record Linkage), specifically crafted to tackle the dual challenges of limited labeled data and class imbalance in record-linkage (RL) tasks. HAML-IRL integrates a structured query strategy that systematically balances informativeness (exploitation) and representativeness (exploration). In particular, it employs a two-phase query-selection process: first, prioritizing data points that reduce model uncertainty by focusing on regions close to the decision boundary and second, ensuring that the selected samples are representative of the overall data distribution by leveraging clustering-based techniques. This dual-phase approach minimizes the risk of overfitting to minority or majority classes, a common issue in imbalanced datasets, while maximizing the coverage of potential data patterns in the training space. Through an iterative learning process, HAML-IRL dynamically adapts its focus based on model performance at each stage, allowing the query strategy to evolve as the model becomes more accurate. Our approach leverages the strengths of both strategies while mitigating their respective weaknesses through an iterative learning process. The key contributions of this work are summarized as follows:

- We introduce HAML-IRL, a novel Hybrid Active Machine-learning framework for record linkage, which integrates both informativeness and representativeness in its querying strategy. This ensures that the most informative and representative record pairs are selected, improving both the convergence speed and stability of the model.
- We provide a theoretical foundation for the HAML-IRL framework, detailing the algorithm, its scoring mechanism and its iterative training process, which is robust against imbalanced datasets and cold-start scenarios.
- We present an extensive experimental evaluation on eleven real-world datasets, including structured, textual and dirty datasets. Our results demonstrate that HAML-IRL achieves up to a 12% improvement in F1-score over state-of-the-art AML methods and performs competitively with fully supervised models.
- We validate the performance of HAML-IRL using statistical tests, including the Friedman and Nemenyi tests, to show that our method significantly outperforms other active learning strategies in handling imbalanced data.
- We show that HAML-IRL reduces the labeling burden, requiring up to between 60% and 85% fewer labeled samples compared to traditional AML approaches, making it more efficient in real-world scenarios where labeling costs are high.

The paper is structured as follows: Section 2 reviews related work on active machine learning and class-imbalance issues. Section 3 outlines the theoretical foundations of our approach, detailing the HAML-IRL algorithm, its complexities and workflow. Section 4 covers the experimental evaluation, including setup, datasets and performance criteria. In Section 5, we present and analyze the results, comparing HAML-IRL with state-of-the-art methods and validating findings using the Friedman test. Section 6 concludes with key insights, future-research directions and broader implications.

## 2. RELATED WORK

Despite the advancements of machine learning, the deployment of supervised learning models is often hindered by the scarcity of labeled data. Addressing this challenge, transfer learning has emerged as a powerful technique, enabling the adaptation of pre-trained models to new tasks with minimal labeled data [4]. Concurrently, active learning (AL) has proven to be an effective strategy for selectively querying the most informative samples for labeling, thereby enhancing model efficacy while minimizing the need for extensive data labeling [5]-[7]. In the specific domain of record linkage (RL) with supervised learning, the dependency on large, labeled datasets for training is a critical challenge. The process of manually annotating record pairs is both costly and time-consuming, presenting a significant barrier to the widespread adoption of RL models. Active Learning (AL) has been proposed as a solution to this challenge, offering a means to reduce the labeling burden by selectively identifying the most informative data points for annotation. This approach not only reduces the manual effort required, but also enhances the overall efficiency and accuracy of the RL process. Several studies have focused on the application of AL techniques to RL challenges, each contributing to the growing body of knowledge in this field. Primpeli et al. [8] introduced an unsupervised bootstrapping method that uses a minimal set of labeled data to iteratively identify and annotate informative record pairs. This method has shown promise in reducing the initial labeling effort while maintaining high accuracy. In parallel, research into uncertainty-based strategies for large-scale RL [9] has highlighted the potential of these approaches in identifying record pairs that present the most significant challenges to predictive models. These strategies have been particularly effective in scenarios where the labeled data is scarce and the models must make informed decisions under uncertainty. Interactive deduplication frameworks, as explored by Sarawagi [10] and adaptive, interactive training data-selection mechanisms, as developed by Christen [3], have further expanded the applications of AL in RL. These frameworks allow for real-time interaction between the model and the human annotator, facilitating more efficient and accurate data-labeling processes. Additionally, initiatives, such as Active Atlas [11], which employs a decision-tree ensemble and the work by Meduri et al. [12] advocating for the use of random forests, have diversified the methodological approaches to RL, providing researchers and practitioners with a broader range of tools to address the complexities of record linkage.

Recent advancements in the field have introduced innovative methods that push the boundaries of traditional RL techniques. ZeroER [13] presented a novel approach to RL that operates without the need for any labeled instances, significantly reducing the dependency on labeled data. DIAL [14], a deep Active Machine-learning (AML) strategy, represents a significant leap forward in matching disparate record representations. This method focuses on optimizing both recall during the initial clustering phase and precision in the subsequent matching task, achieving this through the unified learning of embeddings. However, all current approaches focus primarily on informativeness, often neglecting the representativeness of the queried samples. This oversight can lead to models that, while being trained on informative examples, may lack a comprehensive understanding of the data landscape, resulting in sub- optimal performance in real-world applications. Our proposed work seeks to bridge this gap by introducing a hybrid approach that integrates both informativeness and representativeness into the querying strategy. This methodology is designed to improve the efficiency and precision of RL tasks by providing the model with a holistic view of the data landscape. By challenging the model’s predictive boundaries and ensuring that the selected samples are not only informative, but also representative of the broader data distribution, our approach facilitates a more expedited and nuanced learning trajectory. This, in turn, reduces the reliance on extensive labeling efforts while enhancing the model’s ability to generalize to new, unseen data, ultimately advancing the state-of-the-art in record linkage.

## 3. Theoretical Foundations

This section outlines the core of HAML-IRL framework, which is an active learning algorithm that integrates both representativity (exploration) and informativity (exploitation) in its query strategy. This dual approach is vital for addressing the complexities of the record-linkage problem, where it is crucial not only to identify and label challenging record pairs (exploitation), but also to ensure that the model learns from a diverse set of examples (exploration) to generalize well across different scenarios and handle imbalanced data.

### 3.1 Algorithm Description

The algorithm operates as follows:

Algorithm 1 offers a structured approach to balance two critical aspects of active learning:

representativity (exploration) and informativity (exploitation), by using the following balancing mechanism:

- Informativity (exploitation): The model computes uncertainty scores for each record pair in the unlabeled dataset LD. The uncertainty score quantifies how unsure the model is about the prediction of a particular record pair. Common methods to compute this include:
    - Entropy serves as a measure of the collective uncertainty spanning all potential class predictions for a record pair x, ascertained by the class probability distribution. Elevated entropy values signify heightened informativeness due to increased uncertainty. The entropy-based informativeness is formalized as:

$$I_{Entropy}(x) = -\sum_i P(y_i|x) log_2 \, P(y_i|x) \tag{1}$$

**Algorithm 1.** HAML-IRL algorithm

```
Initialize:
Set $UD$ as the full unlabeled dataset of records
Set $LD$ as the initially labeled dataset
Train initial model $M$ on $LD$
while budget for labeling is not exhausted do
    Calculate uncertainty scores for all record pairs in $LD$ using model $M$
    Calculate representativity scores for all record pairs in $LD$
    Combine scores using a balance parameter $\alpha$:
    for each record pair $x$ in $LD$ do
        $Score(x)$ = $\alpha\times$ Uncertainty($M$, $x$)+(1 $\alpha$) Representativity($D,x$)
    end for
    Select record pair x∗ with the highest Score(x)
    Query label for x∗ and add (x∗, label) to LD
    Remove $x^*$ from $LD$
    Retrain model $M$ on updated $LD$
end while
return the trained model $M$
Optional: Return the expanded labeled dataset $LD$
```

The least confident method prioritizes record pairs with minimal confidence in their most probable class prediction, operationalized as:

$$I_{Least\_Confident}(x) = 1 - P(y_1|x) \tag{2}$$

for a record pair x. Here, $P(y_1|x)$ represents the likelihood of the most probable class, rendering scores closer to 1 indicative of higher uncertainty. This metric, varying between 0 and 1, quantifies the informativeness based on classification confidence.

These scores directly guide the exploitation aspect by prioritizing record pairs that, if labeled, are expected to provide the most information gain for the model. This directly targets improving the model's performance on similar or challenging cases.

- Representativity (exploration): Each record pair's representativity score assesses how well it represents the underlying distribution of the dataset. Record pairs that are more central or typical of the dataset's clusters will receive higher scores, as illustrated in Fig.1.One of most used methods is:
    - Density estimation: $R_{Density}(x)$ measures a record pair's alignment with the dataset's overall characteristics, computed by averaging its similarity to all pairs in *Ul*.

$$R_{Density}(x) = \frac{1}{|Ul|} \sum_{x' \in Ul} sim(x, x') \tag{3}$$

A greater $R_{Density}(x)$ value signifies a record pair's increased representativeness of the dataset's broad features, thus informing the choice of pairs that embody the data's diversity. The similarity function sim($x$,$x'$) employs measures such as Euclidean distance, Jaccard [15], Levenshtein [16] and Jaro-Winkler [17] for evaluation.

In our approach, we apply the Euclidean distance measure in conjunction with a weighted mean subtractive clustering approach [18] as indicated in Eq.4. Using this average distance measure relative to neighboring data points, each data point can be ranked by density. Referring again to Figure 1, Equation (4) enables us to identify points located in the denser (darker red) regions of the plot. This method is robust and adaptable to datasets with multiple columns, as it scales effectively across dimensions.

$$sim(x,x') = e^{-\alpha\|x-x'\|^2}, \alpha = \frac{4}{r^2} \tag{4}$$

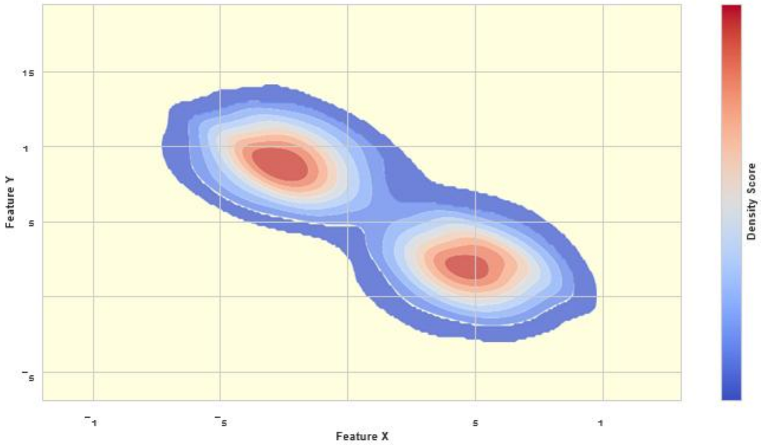


Figure 1. A 2D density plot of data distribution.

The density score at iteration k of the active learning process is calculated for each data point x based on the weighted mean subtractive clustering approach. Here, the Euclidean distance between x and other data points $x' \in Ul$ within a radius r is used to assess density.

To avoid repeatedly labeling points within the same dense areas, the density ranking is recalculated each time new labels are added, facilitating further exploration of the data space. Once a data point has been labeled, the rank of other points in its dense neighborhood is reduced in future iterations. This is achieved by adjusting the density score for points within the radius of each labeled point, as shown in Equation (5).

$$sim_{k+1}(x) = sim_k(x) - sim_{k(x_y)}e^{-\beta\|x-x_y\|^2}, \beta = \frac{4}{r_y^2}, x_y \in LD, x \neq x_y \tag{5}$$

To update the density score at iteration k + 1 of the active learning process, we adjust it based on the labels $LD$ from the previous iteration $k$ for each data point $x$ within a radius $r_y$ from each labeled point $x_y$.

This scoring promotes exploration by ensuring that the model receives training examples from across the data distribution, which helps prevent the model from being biased toward the characteristics of a few unrepresentative examples.

After updating the density rank, we retrain the model and proceed to the next iteration of the active learning loop. In this iteration, the revised rank allows us to explore newly identified dense regions within the feature space, where we present fresh samples to the Oracle to acquire labels, as illustrated in Figure 2.

- Balancing Exploration and Exploitation Score Combination: The algorithm uses a balance parameter $\alpha$, which is a weighting factor between 0 and 1, to combine the informativity $I$ and representativity $R$ scores. The formula is as follows:

$$\text{Score}(x) = \alpha \times \text{I}(M,x) + (1-\alpha) \times \text{R}(D,x) \tag{6}$$

The Score allows for a flexible balance between focusing on informative points (exploitation) and ensuring a diverse set of examples (exploration). Adjusting α: An $\alpha$ closer to 1 would prioritize record pairs that the model finds most uncertain, enhancing exploitation. An $\alpha$ closer to 0 would emphasize representativity, bolstering exploration.

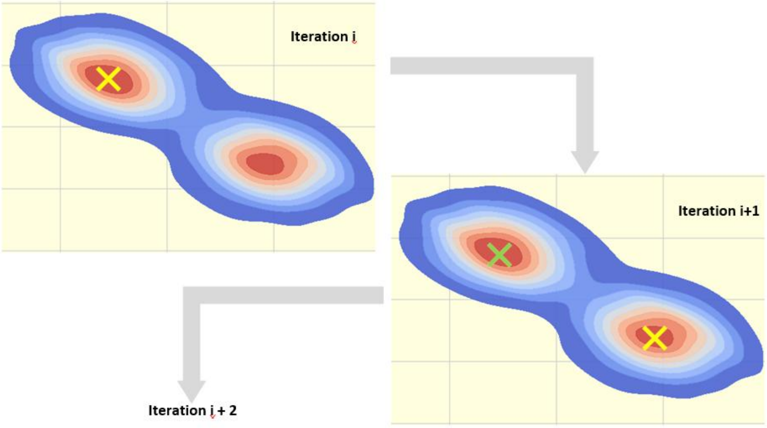


Figure 2. Active ML process.

- Iterative Learning Model Updates: After each selection, the queried label for the chosen record pair x is added to the labeled set L and the model M is retrained. This iterative refinement ensures that the model progressively improves, incorporating insights gained from both new and challenging examples and well- representative record pairs.

## 3.2 Time Complexity

At each iteration, the model M is retrained on the updated labeled dataset L. The time complexity of training depends on the type of model used. For instance, linear models may train in $O(n.d)$ where $n$ is the number of samples and d is the number of features. The computing of uncertainty for each record pair typically involves making a prediction with the model and then calculating a metric (e.g., entropy, least confident). If the model prediction takes $O(d)$ per sample and computing the metric takes constant time, the overall complexity for this step is $O(|UD|.d)$, where $|UD|$ is the size of the dataset. Additionally, representativity calculation could involve distance computations from each record pair to cluster centroids or other points. If $k$ is the number of clusters and d the number of dimensions and assuming basic Euclidean distance is used, the complexity is $O(|UD|.k.d)$. Finally, Score Combination and Selection: Combining scores and selecting the maximum can be carried out in $O(|UD|)$ after calculating the individual scores.

Overall, the time complexity per iteration can be approximated as $O(|UD|.d.\max(k, 1))$+ Time to train M. Since this is done for multiple iterations, the total complexity depends on the number of iterations, which can vary based on convergence criteria or the labeling budget.

## 3.3 HAML-RL Workflow

The workflow diagram provided in Fig. 3 outlines the process of HAML-IRL (Hybrid Active Machine-learning for Imbalanced Record Linkage), detailing the steps involved from pre-processing to the deployment of the model. Here’s a step-by-step explanation of each stage in the workflow.

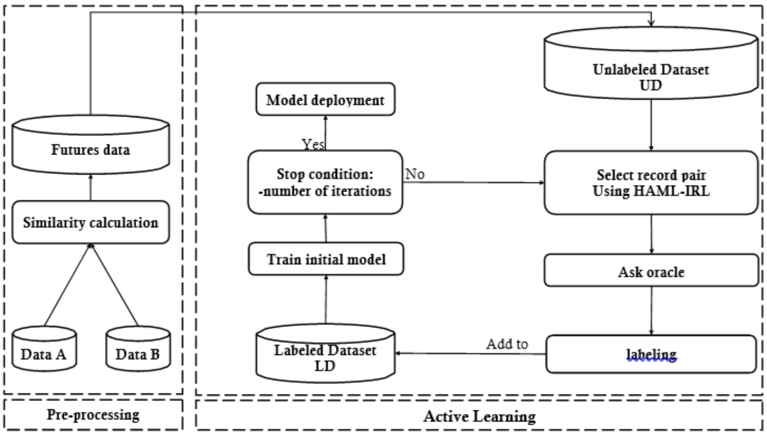


Figure 3. Workflow diagram.

### 3.3.1 Pre-processing Phase

The process begins with two datasets, referred to as Data A and Data B, which contain records that need to be matched or linked. To ensure computational efficiency, we utilize datasets from the literature where blocking has already been applied to ensure a fair comparison with other methods. After blocking, the next step involves calculating the similarity between the filtered records in Data A and Data B. This step is crucial, as it helps identify potential matches between records from the two datasets. The similarity calculation may involve various algorithms or metrics designed to measure how closely two records resemble each other based on specific features or attributes. The results of the similarity calculations are then stored in what is referred to as "Futures Data." This dataset contains pairs of records along with their computed similarity scores, which will be used in the active learning phase.

### 3.3.2 Active Learning Phase

1) **Unlabeled Dataset (UD):** The active learning phase begins with an unlabeled dataset (UD). This dataset contains the pairs of records generated during the similarity calculation, but the pairs are not yet labeled as matches or non-matches.
2) **Selecting Record Pairs Using HAML-IRL:** The core of the HAML-IRL process involves selecting record pairs from the unlabeled dataset. The selection process is guided by the HAML-IRL strategy, which is designed to prioritize pairs that will be most informative for the learning process, especially in the context of imbalanced data.
3) **Asking Oracle:** Once a pair of records is selected, the next step is to label the pair. This is done by querying oracle, which could be a human expert or a pre-existing labeled dataset, to determine whether the selected pair is a match or not. Oracle provides the true label for the record pair.
4) **Labeling:** After querying oracle, the selected pair is labeled accordingly and added to the labeled dataset (LD). This labeled data will be used to train the model.
5) **Labeled Dataset (LD):** The labeled dataset (LD) is continuously updated with new labeled pairs. As more pairs are labeled, the dataset grows, providing more training data for the model.
6) **Training the Initial Model:** Using the labeled dataset, an initial model is trained. This model is a preliminary version that will be iteratively improved as more data is labeled and added to the dataset.
7) **Stop Condition:** Number of Iterations: The process includes a stop condition based on the number of iterations. The model continues to select, label and train on new data until a pre-defined number of iterations are reached.
8) **Model Deployment:** Once the stop condition is met, the model is considered trained and ready for deployment. The final model can then be used to perform record-linkage tasks on new, unseen data.

## 4. Experimental Evaluation

This section evaluates the HAML-IRL algorithm detailed in Section 3, testing its effectiveness across diverse datasets (structured, textual, dirty). Utilizing established libraries and various datasets, we examine the algorithm performance, identifying strengths and improvement areas. These findings contribute to the discussion on AML in RL, highlighting algorithm applicability across data types.

### 4.1 Datasets

In this sub-section, we detail the ER-Magellan and EM-Primpeli datasets [8], [19] selected for evaluating HAML-IRL algorithm, ensuring a comprehensive assessment. These datasets span diverse domains, covering three specific areas of RL. Dataset specifics are provided in Table 1.

### 4.2 Performance Measurement

In RL, especially in scenarios with class imbalances, the F1−score is utilized as the metric for evaluating performance. Hand and Christen [20] characterized the F1−score as the harmonic mean of precision and recall.

The F1−score ranges from 0 to 1, with higher values denoting greater effectiveness, where the following rule represents the formula of F1–score.

$$F1 - score = \frac{2 \cdot TP}{2 \cdot TP + FP + FN}$$

Table 1. Datasets employed in RL. Within this context, $|D_i|$ indicates the total number of products in each dataset, NA represents the number of attributes each product has. Additionally, *Nl*, *Nlp* and *Nln* refer to the total of labeled pairs, matching pairs and non-matching pairs, respectively, whether in training or testing datasets. CR denotes the class ratio.

| | **Structured data-set** | | | | | **Textual data-sets** | | **Dirty data-sets** | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $D_1$ | Amazon | BeerAdvo | Fodors | iTunes | Walmart | Abt | Amazon | iTunes | Walmart | wdc | wdc |
| $D_2$ | Google | RateBeer | Zagat | Amazon | Amazon | Buy | Google item | Amazon | Amazon | phones | hdphone |
| $\lvert D_1 \rvert$ | 1363 | 4345 | 533 | 6907 | 2554 | 1081 | 1114 | 6907 | 2554 | 51 | 51 |
| $\lvert D_2 \rvert$ | 3226 | 3000 | 331 | 55923 | 22074 | 1092 | 1291 | 55923 | 22074 | 448 | 444 |
| $N_A$ | 3 | 4 | 6 | 8 | 5 | 3 | 4 | 8 | 5 | 18 | 14 |
| $Nl_{train}$ | 6874 | 268 | 567 | 321 | 6144 | 5743 | 6755 | 321 | 6144 | 1762 | 1163 |
| $Nl_{test}$ | 2293 | 91 | 189 | 109 | 2049 | 1916 | 1687 | 109 | 2049 | 440 | 290 |
| $Nlp_{train}$ | 699 | 40 | 66 | 78 | 576 | 616 | 1041 | 78 | 576 | 206 | 180 |
| $Nlp_{test}$ | 234 | 14 | 22 | 27 | 193 | 206 | 259 | 27 | 193 | 51 | 45 |
| $Nln_{train}$ | 6175 | 228 | 501 | 243 | 5568 | 5127 | 5714 | 243 | 5568 | 1556 | 983 |
| $Nln_{test}$ | 2059 | 77 | 167 | 82 | 1856 | 1710 | 1428 | 82 | 1856 | 389 | 245 |
| CR | 10.2% | 15.0% | 11.6% | 24.4% | 9.3% | 10.7% | 15.3% | 24.4% | 9.3% | 11.6% | 15.4% |

In this formula, True Positive (TP) is the number of record pairs correctly recognized as matching, False Positive (FP) is the number of record pairs wrongly recognized as matching and False Negative (FN) is the number of record pairs wrongly recognized as not matching.

## 4.3 Feature-similarity Vector-construction for RL

In our study, we address the RL challenge between two datasets, source and target, with the aligned schemata. We construct feature vectors for each entity pair by calculating similarity scores for the individual attributes. These similarity scores are computed using an array of metrics tailored to the data type: Levenshtein and Jaccard for strings; absolute difference for numeric attributes; and day, month and year differences for date attributes. In the case of string attributes exceeding an average length of six tokens, we incorporate cosine-similarity computations using the TF-IDF weighting. All calculated scores are normalized to the [0, 1] range and any missing values are assigned a score of -1 to ensure their inclusion without compromising the integrity of the dataset.

Table 2. Feature-similarity vector-construction example.

| source record S | |
|---|---|
| name | kiki dimoula |
| birthday | 05.06.1931 |

>

| target record T | |
|---|---|
| name | kiki dimula |
| birthday | 1931-06 |

| record pair id | S-T |
|---|---|
| label | true |
| cosine_tfidf | 0.73 |
| name_levenshtein | 0.91 |
| name_jaccard | 0.33 |
| name_relaxed_jaccard | 1.00 |
| name_overlap | 0.00 |
| name_containment | 0.50 |
| birth_day_sim | -1.00 |
| birth_month_sim | 1.00 |
| birth_year_sim | 1.00 |

## 5. EXPERIMENTAL RESULTS

To comprehensively assess the efficacy of our hybrid model under various conditions, we performed a detailed series of experiments using the HAML-IRL algorithm in combination with traditional methods [8], [13], [21]-[26] applied to structured, textual and unclean datasets. Figures 4, 5 and 6 depict the convergence and stability of these strategies, while Tables 3, 4 and 5 showcase their respective performances. Our experimental protocol included five independent trials without bootstrap sampling. The number of iterations was determined by the dataset sizes. Within the HAML IRL framework, we envisioned a scenario in which an unlabeled dataset LD contained all potential record pairs, beginning from an initially empty labeled set. Each iteration in this context corresponds to one manual labeling action. At every iteration, a Random Forest classifier is updated utilizing pairs from the labeled set.

The HAML-IRL benchmarking outcome on structured datasets, as depicted in Table 3, offers compelling insights into the efficacy of AML model. Significantly, the HAML-IRL algorithm showcases a competitive advantage against state-of-the-art (SOA) AML and supervised-learning F1-scores. This underscores the strategy's potential in finely tuning the balance between exploration (representativeness) and exploitation (informativeness) for enhanced data-pairing tasks. In small datasets (i.e., BeerAdvo RateBeer), the HAML-IRL strategy has proven its ability to exceed the highest SOA AML F1-score. Furthermore, in analyzing the Fodors-Zagat dataset, the HAML-IRL achieves the maximum value of F1, comparable to those observed in SOA AML and supervised ML methodologies. In the context of large datasets like Amazon-Google, iTunes Amazon and Walmart-Amazon, the HAML-IRL algorithm demonstrates also an exceptional performance, surpassing benchmarks set by current AML strategies and nearing the effectiveness of supervised ML models. This indicates that a clearly defined transition from exploration to exploitation, governed by a pre-determined labeling budget, calibrates the training task, particularly when the dataset's complexity or features are well understood beforehand. Also, this affirms HAML-IRL's robust capability in navigating through the diverse challenges presented by structured datasets, leveraging its phased approach to maximize model accuracy and learning efficiency.

Table 3. Comparative analysis on structured datasets.

| Database | Strategy | F1 | AML-F1 | Supervised-F1 |
|---|---|---|---|---|
| Amazon-Google | Representativity | 0.434 | 0.480 [13] | 0.561 [25] |
| | Informativity | 0.375 | | |
| | HAML-IRL | 0.510 | | |
| BeerAdvo-RateBeer | Representativity | 0.000 | 0.359 [25] | 0.875 [25] |
| | Informativity | 0.738 | | |
| | HAML-IRL | 0.779 | | |
| Fodors-Zagat | Representativity | 0.978 | 1.0 [13] | 1.0 [21]-[22] |
| | Informativity | 0.975 | | |
| | HAML-IRL | 1.0 | | |
| iTunes-Amazon | Representativity | 0.743 | 0.498 [25] | 0.923 [25] |
| | Informativity | 0.882 | | |
| | HAML-IRL | 0.882 | | |
| Walmart-Amazon | Representativity | 0.564 | 0.644 [25] | 0.678 [25] |
| | Informativity | 0.550 | | |
| | HAML-IRL | 0.649 | | |

The data presented in Table 4, which evaluates HAML-IRL against various AML query strategies on textual datasets, provides valuable insights into the performance of different approaches in text RL tasks. The comparison of these strategies with top-performing supervised and semi-supervised F1-scores illuminates subtle differences in their effectiveness, highlighting the critical role of strategy choice in fine-tuning AML models for text data. For the abt-buy dataset, the HAML-IRL algorithm demonstrates enhancements over purely density-based and uncertainty- based methods, as evidenced by its F1-score. This suggests that a well-structured balance between exploration and exploitation phases may be more advantageous in datasets characterized by dense and complex textual information.

Nonetheless, the HAML-IRL algorithm does not reach the SOA AML F1-score, pointing to opportunities for further improvements in managing the intricacies of textual datasets. In the case of the Amazon-Google dataset, the HAML-IRL performance exceeds leading AML F1-scores, underscoring its capacity to discern textual nuances and variations, particularly in datasets with wide-ranging textual differences. Moreover, the results from HAML-IRL approach the efficacy of established supervised-learning methods in textual RL tasks.

Table 4. Comparative analysis on textual datasets.

| Database | Strategy | F1 | AML-F1 | Supervised-F1 |
|---|---|---|---|---|
| Abt-Buy | Representativity | 0.309 | 0.674 [8] | 0.818 [8]<br>(0.628 [27]) |
| | Informativity | 0.560 | | |
| | HAML-IRL | 0.679 | | |
| Amazon-Google | Representativity | 0.637 | 0.480 [13] | 0.699 [8]<br>(0.693 [23]) |
| | Informativity | 0.468 | | |
| | HAML-IRL | 0.676 | | |

The performance analysis of the HAML-IRL algorithm on datasets with numerous errors, as shown in Table 5, highlights both challenges and possibilities when using active machine-learning (AML) techniques on problematic data. This data often contains errors, inconsistencies and gaps. When comparing this algorithm to other leading methods in terms of F1-scores, we gain detailed understanding of how effective these techniques are when data quality is poor. In small dirty datasets (i.e., "wdc phones" and "wdc headphones"), the HAML-IRL algorithm performs very well, matching or even exceeding the SOA F1-scores for AML. This performance suggests that adaptive strategies that balance data exploration and the use of existing knowledge can adeptly handle the complications of flawed data. The HAML-IRL algorithm’s success in achieving high F1-scores demonstrates that a methodical approach, starting with broad data exploration followed by targeted use of known data, can effectively reveal important insights in datasets filled with noise. In large dirty datasets like ’iTunes-Amazon’ and ’Walmart-Amazon’, the HAML-IRL algorithm also demonstrates exceptional performance, nearing the effectiveness of supervised machine-learning models, though not surpassing the benchmarks set by current AML strategies. These datasets, characterized by extensive errors, inconsistencies and missing values, present a significant challenge for any entity-resolution algorithm. The HAML-IRL algorithm’s near-benchmark performance highlights its robustness and adaptability in handling such complex and flawed data environments. The HAML-IRL algorithm’s ability to maintain high F1-scores in these large and error-prone datasets underscores the potential of hybrid active machine-learning techniques. By leveraging a methodical approach that combines broad exploratory data analysis with the strategic application of existing knowledge, the algorithm is able to navigate the intricacies of dirty data effectively. This approach allows for a nuanced understanding of the data’s structure and patterns, which in turn facilitates more accurate entity matching and resolution.

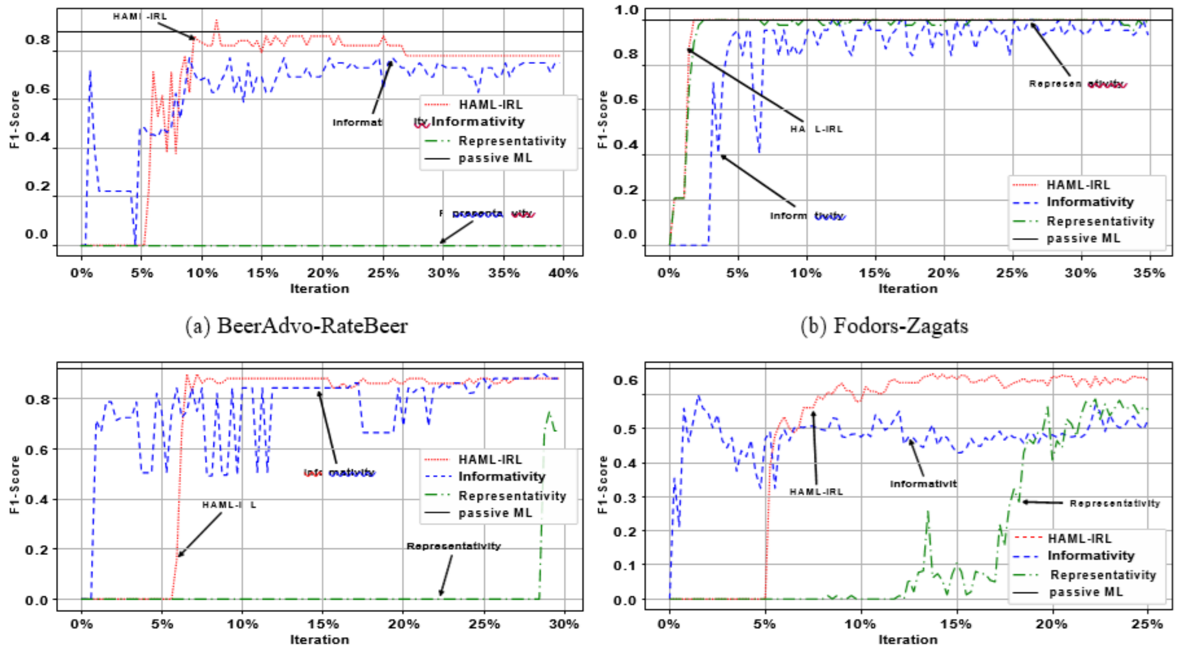
Table 5. Comparative analysis on dirty datasets.

| **Database** | **Strategy** | **F1** | **AML-F1** | **Supervised-F1** |
|---|---|---|---|---|
| iTunes-Amazon | Representativity | 0.300 | 0.638 [8] | 0.640 [25] |
| | Informativity | 0.442 | | |
| | HAML-IRL | 0.511 | | |
| Walmart-Amazon | Representativity | 0.0 | 0.513 [8] | 0.452 [25] |
| | Informativity | 0.232 | | |
| | HAML-IRL | 0.399 | | |
| WDC-phones | Representativity | 0.723 | 0.544 [8] | 0.851 [8]<br>(0.849 [24]) |
| | Informativity | 0.527 | | |
| | HAML-IRL | 0.825 | | |
| WDC-headphones | Representativity | 0.899 | 0.738 [8] | 0.966 [8]<br>(0.940 [24]) |
| | Informativity | 0.487 | | |
| | HAML-IRL | 0.945 | | |

For further understanding, Figures 4, 6 and 5 provide comprehensive comparisons of the HAML-IRL algorithm’s performance, particularly in addressing the initial cold-start problem, where no labeled data is available. These figures meticulously illustrate the algorithm’s efficacy across diverse dataset types, including structured, textual and dirty datasets. The hybrid framework is evaluated against both traditional active machine-learning (AML) approaches, such as Density and Uncertainty queries and the latest advancements in supervised methods. The figures chronicle the F1-scores at each iteration within our hybrid framework, using the F1-score as the primary metric for assessing performance throughout the analyses. In the initial stages of AML, ranging from 1% to 10% of the iterations and varying by dataset, our framework significantly outperforms traditional AML methods in developing superior predictive models across all dataset types. Also, HAML-IRL consistently produces higher-quality prediction models compared to the two standard Active ML techniques for all datasets.

Additionally, the stability of the HAML-IRL F1-scores increases after 10% of iterations and these scores begin to converge towards the performance levels observed in supervised machine learning. After 15% of the iterations, our method exhibits a remarkable enhancement in the stability of F1-scores, which start to closely approximate those from supervised techniques. Consequently, within an AML framework constrained by labeling budgets, our approach demonstrates exceptional performance by consistently yielding satisfactory results, even if the process is halted at any given iteration.

Drawing upon the empirical evidence provided by the preceding figures, it can be conclusively stated that the HAML-IRL algorithm outperforms traditional Active ML methodologies across all iterations within an Active ML context. Particularly in scenarios characterized by cold-start conditions, traditional strategies exhibited slower convergence rates and demonstrated unstable performance metrics. Thus, in an Active ML environment with budget constraints, especially when considering human annotations, our HAML-IRL solution surpasses other methods by reliably achieving satisfactory performance, even when the process is paused at any iteration.



(a) BeerAdvo-RateBeer

(b) Fodors-Zagats

(c) iTunes-Amazon

(d) Walmart-Amazon

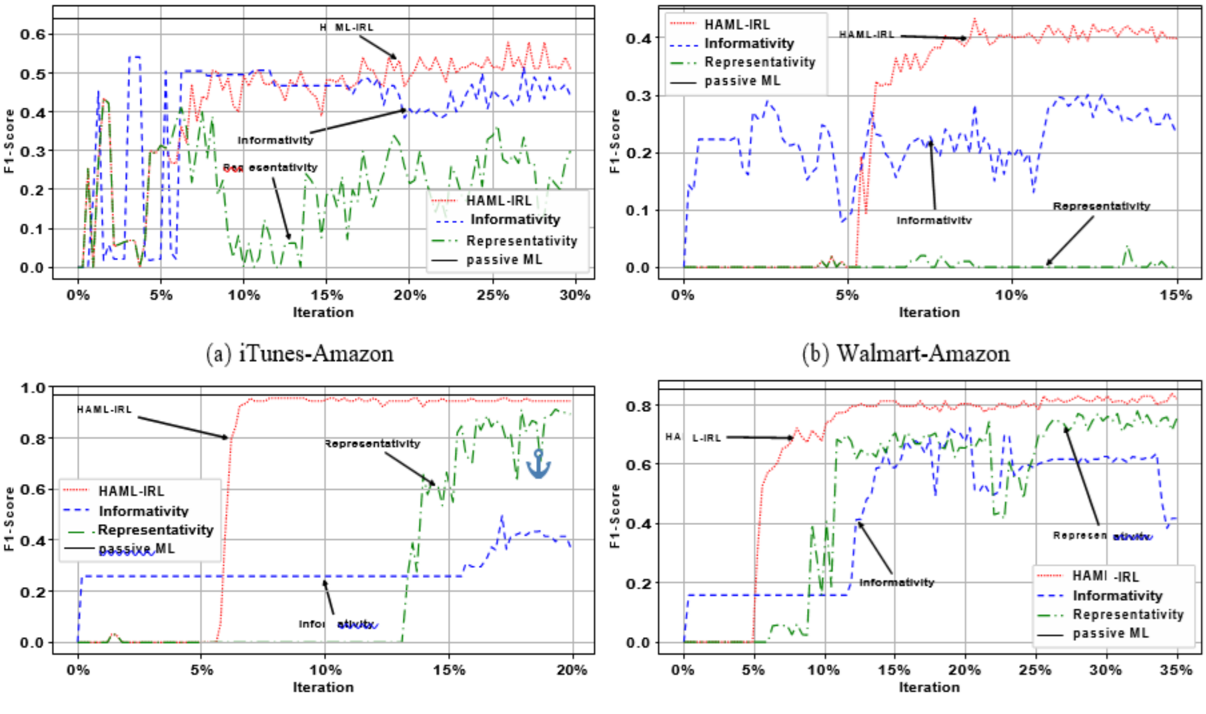
Figure 4. F1-score per AML iteration - structured datasets.

Table 6 presents a comparative analysis of the F1-scores achieved by different active machine learning strategies, including our method, HAML-IRL, across various structured, dirty and textual record-linkage datasets. This analysis provides critical insights into the effectiveness of each method in addressing the imbalanced record-linkage problem. Starting with the structured datasets, we observe that in the iTunes-Amazon dataset, both Uncertainty and HAML-IRL achieve an equal F1-score of 0.882, demonstrating their effectiveness in this context. Other methods, such as Density with a score of 0.743 and Zero-ER at 0.498, show relatively lower performance. Methods like UB-Otsus (0.646) and UB-Valley (0.689) also lag behind, indicating their limitations in handling this particular dataset.

Table 6. F1-score results across structured, dirty and textual datasets.

| **Dataset** | **Uncertainty** | **Density** | **Zero-ER** [13] | **UB-Elbow** [8] | **UB-Static** [8] | **UB-Otsus** [8] | **UB-Valley** [8] | **HAML-IRL** |
|---|---|---|---|---|---|---|---|---|
| | | | | Structured datasets | | | | |
| iTunes-Amazon | 0.882 | 0.743 | 0.498 | 0.678 | 0.655 | 0.646 | 0.689 | 0.882 |
| Walmart-Amazon | 0.550 | 0.564 | 0.644 | 0.501 | 0.393 | 0.313 | 0.424 | 0.649 |
| BeerAdvo-rateBeer | 0.738 | 0.000 | 0.359 | 0.000 | 0.481 | 0.675 | 0.675 | 0.779 |
| Amazon-Google | 0.375 | 0.434 | 0.480 | 0.325 | 0.348 | 0.278 | 0.283 | 0.510 |
| Fodors-Zagat | 0.975 | 0.978 | 1.00 | 0.964 | 0.483 | 0.578 | 0.737 | 1.00 |
| | | | | Dirty datasets | | | | |
| WDC-Phones | 0.527 | 0.723 | 0.000 | 0.523 | 0.544 | 0.438 | 0.438 | 0.825 |
| WDC-Headphones | 0.487 | 0.899 | 0.000 | 0.734 | 0.539 | 0.682 | 0.738 | 0.945 |
| iTunes-Amazon | 0.442 | 0.300 | 0.104 | 0.473 | 0.638 | 0.619 | 0.632 | 0.511 |
| Walmart-Amazon | 0.232 | 0.000 | 0.2 | 0.513 | 0.495 | 0.339 | 0.426 | 0.399 |
| | | | | Textual datasets | | | | |
| Abt-Buy | 0.560 | 0.309 | 0.52 | 0.674 | 0.660 | 0.562 | 0.630 | 0.679 |
| Amazon-Google | 0.468 | 0.637 | 0.472 | 0.588 | 0.441 | 0.600 | 0.602 | 0.676 |

In the Walmart-Amazon dataset, HAML-IRL outperforms all other methods with an F1-score of 0.649. This is particularly noteworthy as Zero-ER, a close competitor, scores 0.644. However, other methods like UB-Static and UB-Otsus perform poorly, with scores of 0.393 and 0.313, respectively, highlighting the challenges these methods face in this scenario. For the BeerAdvo-rateBeer dataset, HAML-IRL demonstrates superior performance with an F1-score of 0.779. Interestingly, Density and UB-Elbow fail completely, scoring 0, which underscores the challenges these methods face in this particular dataset. Uncertainty performs moderately well with a score of 0.738, but it still falls short of HAML IRL. In the Amazon-Google dataset, HAML IRL again leads with an F1-score of 0.510, surpassing all other methods. Zero-ER achieves 0.480 and Density comes close with 0.434, but the other methods, particularly UB-Otsus and UB-Valley, score much lower F1-scores around the 0.280 mark, indicating their inefficacy in this scenario. The Fodors-Zagat dataset presents a unique case where both Zero-ER and HAML-IRL achieve perfect scores of 1.00, showcasing their exceptional ability to match records correctly in this dataset. Uncertainty and Density also perform well with scores close to 1.00, while the remaining methods, particularly UB-Static (0.483) and UB-Otsus (0.578), fall significantly behind.



(a) iTunes-Amazon

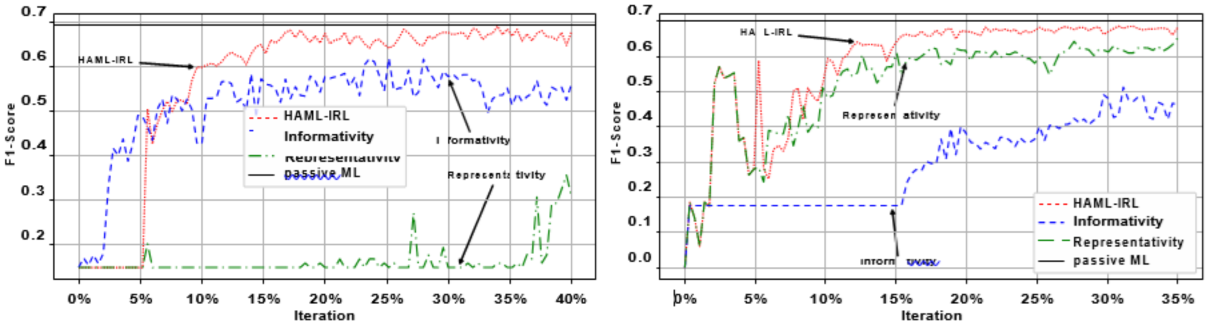
(b) Walmart-Amazon

(c) wdc-headphones

(d) wdc-phones

Figure 5. F1-score per AML iteration-dirty datasets.

Moving on to the dirty datasets, HAML-IRL continues to demonstrate its strength. In the WDC-Phones dataset, it significantly outperforms all other methods with an F1-score of 0.825. Density follows with 0.723, but Zero-ER fails completely, scoring 0. The moderate performance of UB-Otsus and UB-Valley (both at 0.438) further emphasizes the superiority of HAML-IRL in handling dirty datasets. The WDC-Headphones dataset shows a similar trend, with HAML-IRL leading with an impressive score of 0.945. Density performs well with 0.899, while Zero- ER again fails, scoring 0. The other methods, UB-Valley and UB-Otsus, perform moderately, with scores in the 0.682-0.738 range, but still, none come close to HAML-IRL's performance. In the iTunes-Amazon (Dirty) dataset, HAML-IRL achieves an F1-score of 0.511, outperforming most methods except UB-Static (0.638) and UB-Otsus (0.619). Uncertainty scores 0.442, indicating moderate effectiveness, but the overall lower scores reflect the challenges posed by this dataset. For the Walmart-Amazon (Dirty) dataset, the performance of all methods, including HAML IRL (0.399), is relatively low, indicating the dataset's complexity. Zero-ER performs slightly better with a score of 0.513, but overall, the low scores across the board suggest that this dataset is particularly challenging for record linkage.



(a) Abt-buy (b) Amazon-Google-Product

Figure 6. F1-score per AML iteration-textual datasets.

In textual datasets, HAML-IRL continues to perform strongly. In the Abt-Buy dataset, it achieves the highest F1-score of 0.679, slightly outperforming UB-Elbow (0.674) and Zero-ER (0.52). In other methods, such as Context, particularly in scenarios characterized by cold-start conditions, traditional strategies exhibited slower convergence rates and demonstrated unstable performance metrics. Thus, in an Active ML environment with budget constraints, especially when considering human annotations, our HAML-IRL solution surpasses other methods by reliably achieving satisfactory performance, even when the process is paused at any iteration. Density with 0.309 shows less effectiveness, underscoring HAML-IRL's superiority in this category.

Finally, in the Amazon-Google (textual) dataset, HAML-IRL maintains its lead with an F1-score of 0.676. Density follows with 0.637 and UB-Valley scores 0.602, while Zero-ER and UB-Static achieve moderate scores around 0.472-0.588, reflecting a closer competition in this dataset type.

Overall, the results clearly demonstrate that HAML-IRL consistently outperforms the state-of-the-art active machine learning methods across structured, dirty and textual datasets. Its ability to achieve high F1-scores, especially in challenging datasets, underscores its robustness and effectiveness in addressing the imbalanced record-linkage problem. While other methods show varying degrees of success, HAML-IRL's consistent performance across diverse datasets reaffirms its potential as a superior solution for this complex problem.

The histogram provided in Fig. 7 illustrates the mean F1-scores of various active learning models across structured, dirty and textual datasets. The F1-score, as a key metric, combines precision and recall, offering a balanced measure of a model's performance, particularly in scenarios where the class distribution is imbalanced. The uncertainty model shows a diverse range of performance across the different types of datasets. For structured datasets, it achieves a relatively high mean F1-score, indicating that the model is effective in these types of datasets, which are typically cleaner and more well-defined. However, the performance drops for dirty datasets, suggesting that the model struggles with noise and inconsistencies often present in such data. The performance in textual datasets is moderate, reflecting the model's average ability to handle the complexities of text-based record linkage.

In contrast, the Density model performs well across all dataset types, particularly in structured and textual datasets. The high mean F1-score in structured datasets shows that this model can effectively utilize the dense regions of data to make accurate predictions. Although its performance in dirty datasets is slightly lower, it remains significant, indicating that the model can manage some level of noise and variability. Its strong performance in textual datasets further highlights its adaptability to different data types.

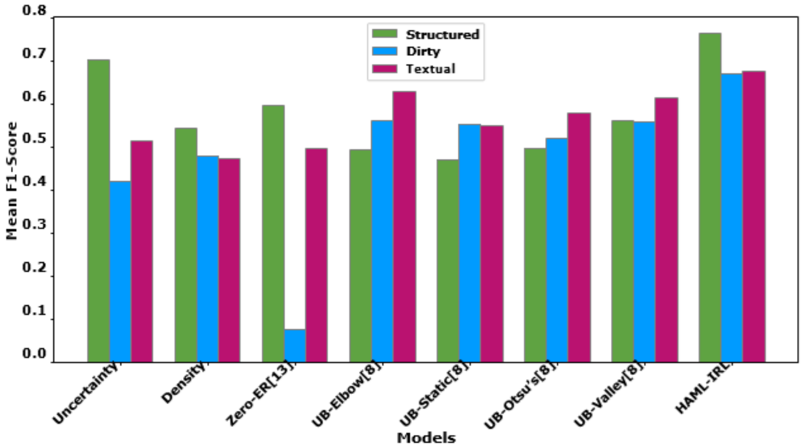


Figure 7. The mean F1-score over structured, dirty and textual datasets.

The Zero-ER model, however, exhibits poor performance, especially in dirty datasets, where its mean F1-score is nearly negligible. This suggests that Zero-ER is not well-suited to handle noise and inconsistencies, leading to poor precision and recall in these challenging scenarios. Even in structured datasets, where its performance is better, it remains subpar compared to other models, indicating limited applicability in well-defined data environments. Furthermore, the model does not fare well in textual datasets, reinforcing its limitations in handling more complex and unstructured data.

The UB-Elbow model, on the other hand, shows a balanced performance across all dataset types. Its mean F1- score in structured datasets is decent, reflecting its ability to handle clear and organized data effectively. For dirty datasets, the performance is slightly better, suggesting some robustness to noise and inconsistencies. The model also performs adequately in textual datasets, indicating a certain level of versatility across different data types.

Similarly, the UB-Static model shows strong performance in dirty datasets, achieving one of the higher mean F1-scores among the models. This indicates that UB-Static is particularly well-suited for dealing with noisy and inconsistent data, where other models might struggle. However, its performance in structured and textual datasets is moderate, suggesting that while it excels in handling variability, it may not be as effective in more structured or language-based data scenarios.

Meanwhile, the UB-Otsus model displays a relatively balanced performance across all dataset types, though it is not the top performer in any particular category. The mean F1-scores indicate that it can handle a variety of data types moderately well, but it does not particularly excel in any of them. This suggests that UB-Otsus might be a good all-rounder for general applications, but may not be the best choice for datasets with specific challenges.

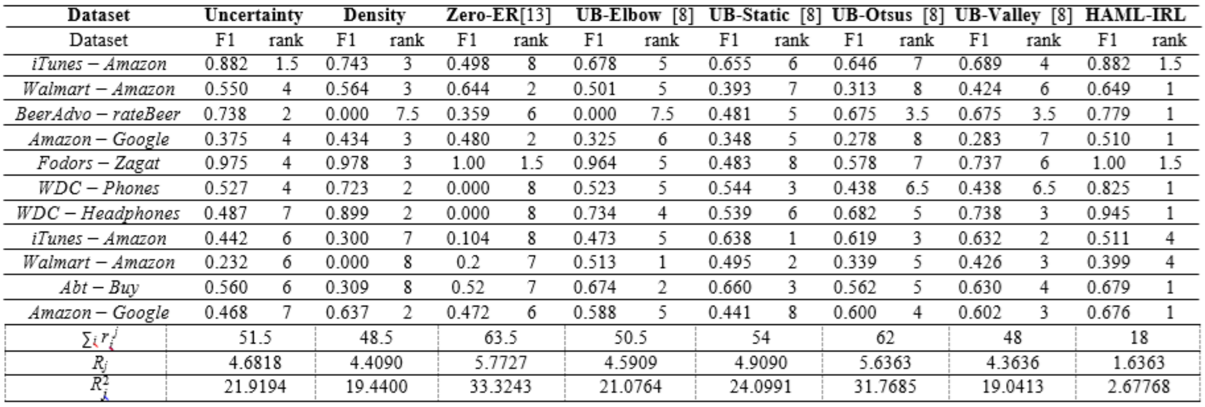
The UB-Valley model shows strong performance in both dirty and textual datasets, with relatively high mean F1- scores. This suggests that UB-Valley is effective at managing both noise and complexities of text-based record linkage. Although its performance in structured datasets is also good, it is slightly lower than in the other two categories, indicating broad applicability across different types of data.

Finally, the HAML-IRL model consistently performs the best across all dataset types, achieving the highest mean F1-scores in structured, dirty and textual datasets. This consistent top performance underscores HAML-IRL's robustness and adaptability, making it the most effective model for handling a wide range of record-linkage scenarios. The high scores across different data types demonstrate the model's ability to balance precision and recall effectively, even in challenging datasets, like dirty and textual datasets.

Overall, the histogram provides a clear comparison of the mean F1-scores for different models across structured, dirty and textual datasets. HAML-IRL emerges as the leading model, consistently achieving the highest mean F1- scores across all dataset types. This indicates its superior ability to handle both well-defined and complex, noisy data. While other models, such as UB-Valley and Density, also perform well in specific contexts, they do not match the overall effectiveness of HAML-IRL. Models like Zero-ER, which perform poorly in more challenging datasets, highlight the importance of selecting the right model based on the specific characteristics of the data being used.

## 5.1 Friedman Test

Table 7. Ranking of HAML-IRL.



| **Dataset** | **Uncertainty** | | **Density** | | **Zero-ER**[13] | | **UB-Elbow** [8] | | **UB-Static** [8] | | **UB-Otsus** [8] | | **UB-Valley** [8] | | **HAML-IRL** | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Dataset | F1 | rank | F1 | rank | F1 | rank | F1 | rank | F1 | rank | F1 | rank | F1 | rank | F1 | rank |
| *iTunes – Amazon* | 0.882 | 1.5 | 0.743 | 3 | 0.498 | 8 | 0.678 | 5 | 0.655 | 6 | 0.646 | 7 | 0.689 | 4 | 0.882 | 1.5 |
| *Walmart – Amazon* | 0.550 | 4 | 0.564 | 3 | 0.644 | 2 | 0.501 | 5 | 0.393 | 7 | 0.313 | 8 | 0.424 | 6 | 0.649 | 1 |
| *BeerAdvo – rateBeer* | 0.738 | 2 | 0.000 | 7.5 | 0.359 | 6 | 0.000 | 7.5 | 0.481 | 5 | 0.675 | 3.5 | 0.675 | 3.5 | 0.779 | 1 |
| *Amazon – Google* | 0.375 | 4 | 0.434 | 3 | 0.480 | 2 | 0.325 | 6 | 0.348 | 5 | 0.278 | 8 | 0.283 | 7 | 0.510 | 1 |
| *Fodors – Zagat* | 0.975 | 4 | 0.978 | 3 | 1.00 | 1.5 | 0.964 | 5 | 0.483 | 8 | 0.578 | 7 | 0.737 | 6 | 1.00 | 1.5 |
| *WDC – Phones* | 0.527 | 4 | 0.723 | 2 | 0.000 | 8 | 0.523 | 5 | 0.544 | 3 | 0.438 | 6.5 | 0.438 | 6.5 | 0.825 | 1 |
| *WDC – Headphones* | 0.487 | 7 | 0.899 | 2 | 0.000 | 8 | 0.734 | 4 | 0.539 | 6 | 0.682 | 5 | 0.738 | 3 | 0.945 | 1 |
| *iTunes – Amazon* | 0.442 | 6 | 0.300 | 7 | 0.104 | 8 | 0.473 | 5 | 0.638 | 1 | 0.619 | 3 | 0.632 | 2 | 0.511 | 4 |
| *Walmart – Amazon* | 0.232 | 6 | 0.000 | 8 | 0.2 | 7 | 0.513 | 1 | 0.495 | 2 | 0.339 | 5 | 0.426 | 3 | 0.399 | 4 |
| *Abt – Buy* | 0.560 | 6 | 0.309 | 8 | 0.52 | 7 | 0.674 | 2 | 0.660 | 3 | 0.562 | 5 | 0.630 | 4 | 0.679 | 1 |
| *Amazon – Google* | 0.468 | 7 | 0.637 | 2 | 0.472 | 6 | 0.588 | 5 | 0.441 | 8 | 0.600 | 4 | 0.602 | 3 | 0.676 | 1 |
| $\sum_i r_i^j$ | 51.5 | | 48.5 | | 63.5 | | 50.5 | | 54 | | 62 | | 48 | | 18 | |
| $R_j$ | 4.6818 | | 4.4090 | | 5.7727 | | 4.5909 | | 4.9090 | | 5.6363 | | 4.3636 | | 1.6363 | |
| $R_j^2$ | 21.9194 | | 19.4400 | | 33.3243 | | 21.0764 | | 24.0991 | | 31.7685 | | 19.0413 | | 2.67768 | |

In the subsequent analysis, we employ the Friedman test to evaluate the efficacy of the proposed approach on the *ER-Magellan* and *EM-Primpeli* datasets [8], [19].

The Friedman test, as initially proposed by Friedman [28], is a *non-parametric* statistical test devised to rank algorithms individually for each dataset. The algorithm demonstrating the best performance is assigned a rank of 1, the next best is assigned a rank of 2 and so on. In instances where there are ties, average ranks are allotted.

Let the rank of the *j*th algorithm among k algorithms on the *i*th of N datasets be denoted as $r_i^j$. The Friedman test compares the mean ranks of these algorithms, expressed as $R_j = \frac{1}{N}\sum_i r_i^j$.

Under the null hypothesis, which posits that all algorithms are equivalent and thus their ranks $R_j$ should be similar, the Friedman statistic is calculated as follows:

$$\chi_F^2 = \frac{12N}{k(k+1)}\left[\sum_j R_j^2 - \frac{k(k+1)^2}{4}\right].$$

When both $N$ and $k$ are sufficiently large (typically, $N > 10$ and $k > 5$), this statistic follows a *chi−square* distribution with $k-1$ degrees of freedom.

Iman and Davenport [29] observed that the Friedman $\chi_F^2$ statistic is overly conservative. They proposed an enhanced statistic:

$$F_F = \frac{(N-1)\chi_F^2}{N(k-1)-\chi_F^2}$$

This improved statistic follows an *F-distribution* with k-1 and ($k$-1)($N$-1) degrees of freedom. Critical values for this distribution can be found in statistical reference literature.

If the null hypothesis is rejected, a *post-hoc* analysis is undertaken. The Nemenyi test [30] is utilized when all classifiers are compared against each other. The performance difference between two classifiers is deemed significant if the difference between the highest and the lowest average ranks exceeds the critical difference:

$$CD = q_\alpha\sqrt{\frac{k(k+1)}{6N}}.$$

Here, the critical value of qα is derived from the Studentized range statistic divided by $\sqrt{2}$. The hypotheses for the test are stated as follows:

- $H_0$: There are no significant differences among the eight methods.
- *H1:* There are significant differences among the eight methods.

### 5.1.1 Application

The histogram provided in Fig. 8 presents the mean ranking results of various active machine-learning models using the Friedman test. The models included in the comparison are HAML-IRL, UB-Valley, Density, UB-Elbow, Uncertainty, UB-Static, UB-Otsus and Zero-ER. The mean rank values across these models are indicative of their relative performance in handling the imbalanced record-linkage problem, with lower ranks suggesting better performance. The model HAML-IRL clearly outperforms the other models, as indicated by its superior ranking.

This model achieves the lowest mean rank, signifying that it consistently performs better across the datasets included in the analysis. The results underscore HAML-IRL's robustness and adaptability, making it the most effective model for overcoming the challenges posed by imbalanced record linkage.

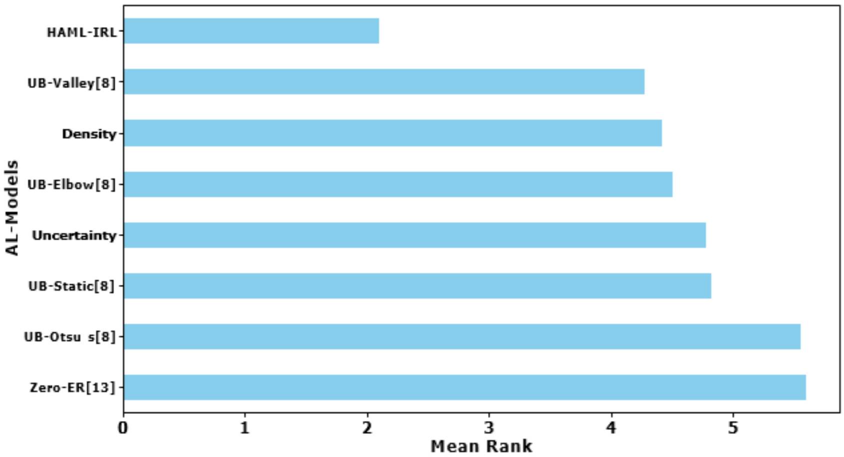


Figure 8. The mean rank over all datasets.

Following HAML-IRL, the UB-Valley and Density models are next in the ranking. Both of these models have relatively close mean ranks, suggesting that they are competitive in their performance. However, they still lag behind HAML-IRL, which suggests that while they may be effective, they do not match the comprehensive capabilities of HAML-IRL in dealing with the complexities of the datasets.

UB-Elbow and Uncertainty models are ranked in the middle range. Their mean ranks indicate moderate effectiveness, where they perform reasonably well, but are not among the top contenders. The positioning of these models suggests that they may be more suitable for specific types of datasets, but lack the broad applicability and robustness demonstrated by HAML-IRL.

On the lower end of the ranking spectrum, we find UB-Static, UB-Otsus and Zero-ER models. These models have the highest mean ranks, indicating that they are the least effective in handling the imbalanced record-linkage problem. Their poor performance in this analysis suggests that they may not be well-suited for tasks that require high accuracy in imbalanced scenarios. Zero-ER, in particular, appears to be the weakest model, as indicated by its position at the bottom of the ranking.

Overall, the Friedman test's ranking results, as depicted in the histogram, provide a clear indication of the relative effectiveness of the models under comparison. HAML-IRL stands out as the most effective model, consistently achieving the best rankings across the datasets. This outcome reflects its superior design and capability in addressing the imbalanced record-linkage problem. The other models, while showing varying degrees of effectiveness, do not reach the performance level of HAML-IRL, making it the preferred choice in this domain. In this study, we will compare $k$ = 8 methods across $N$ = 11 datasets. The methods are ranked based on their F1-scores for each dataset. Table 7 presents the results of the rankings of the proposed approach.

From the average ranking results of active learning methods across all datasets, presented in Table 7, we obtain the results of the Friedman test, including the *Chi-square* statistic ($\chi^2$), p-value, *F-distribution* ($\mathcal{F}_F$), critical difference *CD* and confidence interval (CI), as presented in Table 8. After carefully analyzing the table at a confidence level of 0.05 with degrees of freedom (7, 70), we observe that $F_{\alpha=0.05}(7, 70) = 2.143$. Since the calculated *F-value* surpasses $F_\alpha$ and the *p-value* is less than 0.05, we reject the null hypothesis $H_0$. The Friedman test, accompanied by its enhanced statistic, indicates significant differences among the eight active learning methods applied to 11 datasets, which encompass structured, dirty and textual data. Notably, our proposed approach outperforms the other algorithms when ranked by F1-score.

Table 8. Friedman-test results.

| **Approach** | $\chi^2$ | **P-value** | $\mathcal{F}_\mathcal{F}$ | **CD** | **CI** |
|---|---|---|---|---|---|
| HAML-IRL | 20.8030 | 0.003 | 3.7018 | 3.165 | [0.514, 0.914] |

After the Nemenyi test we found that the critical value $q_\alpha = 3.031$ and the corresponding *CD* = 3.03. Since the difference between the best -and the worst- performing algorithm is already greater than that, we can conclude that the *post-hoc* test is powerful enough to detect any significant differences between the algorithms. The results from *post-hoc* tests are effectively conveyed through a clear graphical representation. We utilize Autorank [31] to generate a plot that visually represents the statistical analysis for a Critical Difference. Fig. 9 displays the results derived from the data in Table 7. The top line of the diagram illustrates the axis where the average ranks of methods are plotted. This axis is oriented so that the lowest (best) ranks are on the right side, indicating that methods positioned further to the right are considered superior.

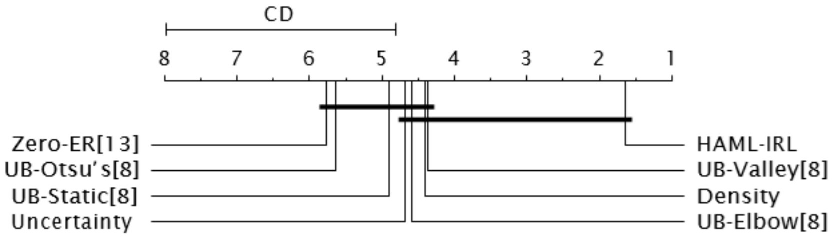


Figure 9. Comparison of all methods against each other using the Nemenyi test.

## 6. Conclusion

This research introduces a novel hybrid active machine-learning framework to address the challenge of scarce labeled data in record linkage. By balancing representativity and informativity, the framework first ensures broad data coverage, then focuses on refining the model with the most informative samples. The experiments on various datasets show that our framework outperforms traditional active learning methods and often rivals fully supervised models, especially in cold-start scenarios. The results demonstrate the framework's effectiveness in producing high- quality models with limited labeled data, offering a strategic solution for optimizing learning in record linkage.

## References

[1] Y. Aassem, I. Hafidi and N. Aboutabit, "Exploring the Power of Computation Technologies for Entity Matching," Proc. of Emerging Trends in ICT for Sustainable Development, Part of the book series: Advances in Science, Technology & Innovation, pp. 317–327, Springer, 2021.

[2] L. Alami, Y. Aassem and I. Hafidi, "KF-Swoosh: An Efficient Spark-based Entity Resolution Algorithm for Big Data," Journal of Physics, Conference Series: Proc. of the Int. Conf. on Mathematics & Data Science (ICMDS), vol. 1743, p. 012005, Khouribga, Morocco, Jan. 2021.

[3] P. Christen, D. Vatsalan and Q. Wang, "Efficient Entity Resolution with Adaptive and Interactive Training Data Selection," Proc. of the 2015 IEEE Int. Conf. on Data Mining, Atlantic City, USA, 2015.

[4] B. Zhang, D. Yang, Y. Liu and Y. Zhang, "Graph Contrastive Learning with Knowledge Transfer for Recommendation," Engineering Letters, vol. 32, no. 3, pp. 477–487, 2024.

[5] M. Jabrane, I. Hafidi and Y. Rochd, "An Improved Active Machine Learning Query Strategy for Entity Matching Problem," Proc. of the Int. Conf. of Machine Learning and Computer Science Applications,

Part of the Book Series: Lecture Notes in Networks and Systems, vol. 656 pp. 317–327, 2023.

[6] J. Mourad, T. Hiba, R. Yassir and H. Imad, "ERABQS: Entity Resolution Based on Active Machine Learning and Balancing Query Strategy," Journal of Intelligent Information Systems, vol. 62, pp. 1347-1373, Mar. 2024.

[7] M. Jabrane, H. Tabbaa, A. Hadri and I. Hafidi, "Enhancing Entity Resolution with a Hybrid Active Machine Learning Framework: Strategies for Optimal Learning in Sparse Datasets," Information Systems, vol. 125, p. 102410, Nov. 2024.

[8] A. Primpeli, C. Bizer and M. Keuper, "Unsupervised Bootstrapping of Active Learning for Entity Resolution," Proc. of European Semantic Web Conference, The Semantic Web, Part of the Book Series: Lecture Notes in Computer Science, vol. 12123, pp. 215–231, Springer, 2020.

[9] K. Qian, L. Popa and P. Sen, "Active Learning for Large-scale Entity Resolution," Proc. of the 2017 ACM on Conf. on Information and Knowledge Management, pp. 1379-1388, DOI: 10.1145/3132847.313294, 2017.

[10] S. Sarawagi and A. Bhamidipaty, "Interactive Deduplication Using Active Learning," Proc. of the 8th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining (KDD '02), pp. 269 – 278, DOI: 10.1145/775047.7750, 2002.

[11] S. Tejada, C. A. Knoblock and S. Minton, "Learning Domain-independent String Transformation Weights for High Accuracy Object Identification," Proc. of the 8th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining (KDD '02), pp. 350-359, DOI: 10.1145/775047.775099, 2002.

[12] V. V. Meduri, L. Popa, P. Sen and M. Sarwat, "A Comprehensive Benchmark Framework for Active Learning Methods in Entity Matching," Proc. of the 2020 ACM SIGMOD Int. Conf. on Management of Data, pp. 1133 – 1147, DOI: 10.1145/3318464.3380597, 2020.

[13] R. Wu, S. Chaba, S. Sawlani, X. Chu and S. Thirumuruganathan, "ZeroER: Entity Resolution Using Zero Labeled Examples," Proc. of the 2020 ACM SIGMOD Int. Conf. on Management of Data, pp. 1149 – 1164, DOI: 10.1145/3318464.3389743, 2020.

[14] A. Jain, S. Sarawagi and P. Sen, "Deep Indexed Active Learning for Matching Heterogeneous Entity Representations," Proc. of the VLDB Endowment, vol. 15, no. 1, pp. 31–45, 2021.

[15] R. Dharavath and A. K. Singh, "Entity Resolution-based Jaccard Similarity Coefficient for Heterogeneous Distributed Databases," Proc. of the 2nd Int. Conf. on Computer and Communication Technologies, Advances in Intelligent Systems and Computing, vol. 379, pp. 497–507, Sept. 2015.

[16] V. I. Levenshtein, "Binary Codes Capable of Correcting Deletions, Insertions and Reversals," Soviet Physics-Doklady, vol. 10, pp. 707–710, 1965.

[17] M. A. Jaro, "Advances in Record-linkage Methodology As Applied to Matching the 1985 Census of Tampa, Florida," Journal of the American Statistical Association, vol. 84, no. 406, pp. 414–420, 1989.

[18] J. Chen, Z. Qin and J. Jia, "A Weighted Mean Subtractive Clustering Algorithm," Information Technology Journal, vol. 7, no. 2, pp. 356–360, 2008.

[19] S. Das, A. Doan, P. S. G. C., C. Gokhale, P. Konda, Y. Govind and D. Paulsen, "The Magellan Data Repository," [Online], Available: https://sites.google.com/site/anhaidgroup/projects/data.

[20] D. Hand and P. Christen, "Using the F-measure for Evaluating Record Linkage Algorithms," Statistics and Computing, vol. 28, no. 3, pp. 539–547, 2017.

[21] Y. Li, J. Li, Y. Suhara, A. Doan and W.-C. Tan, "Effective Entity Matching with Transformers," The VLDB Journal, vol. 32, pp. 1215-1235, 2023.

[22] S. Li and H. Wu, "Transformer-based Denoising Adversarial Variational Entity Resolution," Journal of Intelligent Information Systems, vol. 61, pp. 631-650, 2023.

[23] S. Mudgal et al., "Deep Learning for Entity Matching: A Design Space Exploration," Proc. of the 2018 Int. Conf. on Management of Data (SIGMOD '18), pp. 19-34, DOI: 10.1145/3183713.3196926, 2018.

[24] P. Petrovski and C. Bizer, "Learning Expressive Linkage Rules from Sparse Data," Semantic Web, vol. 11, no. 3, pp. 549–567, 2020.

[25] G. Papadakis, N. Kirielle, P. Christen and T. Palpanas, "A Critical Re-evaluation of Benchmark Datasets for (Deep) Learning-based Matching Algorithms," ArXiv: 2307.01231, 2023.

[26] R. Chen, Y. Shen and D. Zhang, "GNEM: A Generic One-to-Set Neural Entity Matching Framework," Proc. of the Web Conf. 2021, DOI: 10.1145/3442381.3450119 Ljubljana, Slovenia, 2021.

[27] D. Chen, Y. Lin, W. Li, P. Li, J. Zhou and X. Sun, "Measuring and Relieving the Over-smoothing Problem for Graph Neural Networks from the Topological View," Proc. of the AAAI Conf. on Artificial Intelligence, vol. 34, no. 4, pp. 3438–3445, 2020.

[28] M. Friedman, "The Use of Ranks to Avoid the Assumption of Normality Implicit in the Analysis of Variance," Journal of the American Statistical Association, vol. 32, no. 200, pp. 675–701, 1937.

[29] R. L. Iman and J. M. Davenport, "Approximations of the Critical Region of the fbietkan Statistic," Communications in Statistics - Theory and Methods, vol. 9, no. 6, pp. 571–595, 1980.

[30] P. B. Nemenyi, Distribution-free Multiple Comparisons, PhD Thesis, Princeton University, 1963.

[31] S. Herbold, "Autorank: A Python Package for Automated Ranking of Classifiers," Journal of Open Source Software, vol. 5, p. 2173, Apr. 2020.

## ملخص البحث:

تُكافح الطّرق التّقليدية للتّعلُّم الآلي النّشط الّتي تُوظّف في ربْط السّجلّات من أجْل التّغلُّب على مشكلاتٍ تشمل البيانات غير المتوازنة.

نُقدّم في هذه الورقة البحثية نظاماً مُبتكراً للتّعلُّم الآلي النّشط الهجين؛ بهدف التّغلُّب على مشكلة ربْط السّجلّات غير المتوازن. ويتمّ ذلك عن طريق اختيار أجزاء معينة من السّجلّات لتقليل اللّايقين في النظام، مع مراعاة أن تعْكس الأجزاء المختارة من السّجلّات كامل الأنماط المتضمَّنة في مجموعة البيانات. وقد حقّق نظامنا الهجين المقترح تطوراتٍ ملحوظةٍ من حيث مؤشّرات الأداء مقارنةً بمثيلاته الواردة في أدبيات الموضوع عند تطبيقه على مجموعات بياناتٍ من العالم الحقيقي.

ويحتاج نظامنا المقترح إلى ما يتراوح بين 60% و 85% أقلّ من غيره من العينات الموسومة، اعتماداً على مجموعة البيانات الّتي يُطبّق عليها. وهذا يجعل منه نظاماً متيناً و فعّالاً لحلّ مشكلة ربْط السّجلّات غير المتوازن.

# LAIOV-5G: Lightweight Authentication Scheme for IoV Based on 5G Technology in Smart-city Environment

Murtadha A. Alazzawi[1], Saad Ali Alfadhli[1], Ahmed Al-Shammari[2], Zaid Ameen Abduljabbar[3] and Vincent Omollo Nyangaresi[4]

*(Received: 12-Oct.-2024, Revised: 3-Dec.-2024, Accepted: 6-Dec.-2024)*

## Abstract

*The Fifth Generation (5G) networks have enabled the development of smart cities, in which massive amounts of data are collected, stored and disseminated. The ultimate objective of these smart cities is to cut costs and improve security performance. In this environment, Internet of Vehicles (IoV) helps connect vehicles, pedestrians, control rooms and some roadside infrastructure. Owing to the insecure nature of the communication channel utilized in IoV to exchange information, it is important to develop practical techniques to preserve data confidentiality and privacy. To this end, numerous security solutions have been proposed over the recent past. Unfortunately, most of these authentication techniques have security flaws, which endangers the transmitted data, while some of them are highly inefficient. To address these gaps, we present a Lightweight Authentication Scheme for the Internet of Vehicles (IoV) based on 5G technology (LAIOV-5G).The security analysis carried out demonstrates that LAIOV-5G mitigates numerous potential attacks that threaten the IoV communication in a smart-city environment. In addition, the performance analysis of LAIOV-5G verifies its effectiveness and efficiency.*

## Keywords

*Authentication, 5G, Security, VANETs, Smart City.*

## 1. Introduction

The Internet of Things (IoT) encompasses modern wireless technologies or applications that sense, process, manage and control large volumes of data used for service or application-level enhancements [1]. These advancements are not just theoretical, but they have a direct impact on our daily lives. For instance, the smart-city applications, such as smart homes, IoV, Intelligent Transportation System (ITS) and smart industrial manufacturing, have facilitated scalable and efficient information exchanges that meet various domain requirements [2]-[3]. As explained in [4], a real-time IoV computing environment has been facilitated by the exponential growth of today's automotive technologies, combining numerous approaches, like IoV, VANETs and cloud. This helps address a variety of challenges that may arise on roadways due to congestions and other traffic-related concerns [4]. This practical application of IoT in addressing real-world problems underscores its relevance and importance.

The increasing integration of IoT into smart cities has revealed new possibilities for enhancing efficiency and productivity in various areas, such as intelligent transportation systems, critical infrastructure management and industrial automation [5]-[7]. Among all these technologies, IoT has emerged as a crucial enabler for services, such as real-time traffic control, accident-avoidance mechanisms and vehicle-to-infrastructure (V2I) communication. However, these developments pose significant security hurdles, such as protecting confidential information, ensuring communication integrity and thwarting unauthorized access. This investigation addresses these hurdles by proposing a simplified authentication scheme specifically designed for IoV networks based on 5G technology. By leveraging fast data-transfer speeds, reduced latency and improved reliability of 5G technology, this scheme offers a robust and effective answer for secure and seamless connectivity in smart urban environments [8]-[9].

1. M. A. Alazzawi and S. A. Alfadhli are with Department of Computer Techniques Engineering, Imam Alkadhim University College (IKU), 10001, Baghdad, Iraq. Emails: murtadhaali@alkadhum-col.edu.iq and Saadali@alkadhum-col.edu.iq
2. A. Al-Shammari is with Department of Computer Science, College of Computer Science and Information Technology, University of Al-Qadisiyah, Al Diwaniyah, 58002, Iraq. Email: ahmed.alshammari@qu.edu.iq
3. Z. A. Abduljabbar is with Department of Computer Science, College of Education for Pure Sciences, University of Basrah, Basrah, Iraq and with Huazhong University of Science and Technology, Shenzhen Institute, Shenzhen, China. Email: zaid.ameen@uobasrah.edu.iq
4. V. O. Nyangaresi is with Department of Computer Science and Software Engineering, Jaramogi Oginga Odinga Uni. of Science & Technology, Bondo, Kenya and with Department of Applied Electronics, Saveetha School of Eng., SIMATS, Chennai, Tami lnadu, India. Email: vnyangaresi@jooust.ac.ke

Recently, the rise in vehicle production has made IoV the longest-lasting technical trend in the world today [10]. With IoV, a self-organized network may be formed and messages can be broadcast to the moving vehicles. It offers several advantages, exampled by integrated warning systems which alert drivers about accidents. Afterwards, drivers may make decisions quickly depending on the information provided. Still, the accuracy and safety of self-driving cars could be increased by sharing more complex information among them [11]. However, if there is no substantial security and privacy protection in place, adversaries can quickly access sensitive and private information belonging to car users [12]. Apart from privacy issues, data authenticity and integrity are other important security topics in IoV. For instance, malicious IoV entities can forward false information to human drivers or self-deriving cars, which can result in wrong judgments and decisions. Ultimately, this can lead to tragic events, such as serious road accidents that result in loss of lives. It is also possible for malicious entities to infiltrate IoV networks in order to carry out terrorist attacks. Moreover, falsified information may lure customers to dangerous zones or rival parking lots where evils, such as kidnapping, can be executed. This potential misuse of IoV underscores the need for robust security measures. As discussed in [13]-[15], significant investments in wireless-communication technologies has led to the development of 5G networks. In these networks, mobile data rates can be increased 1000 folds, resulting in transmission rates of up to 10 Gbps. As such, 5G networks have increased speeds compared to their predecessors, such as the Fourth-Generation (4G) networks. Moreover, 5G networks have reduced latencies and increased efficiency, which improves the battery life of their network elements. This helps in creating a conducive environment for the deployment of many battery-powered devices in the IoT [16].

Motivated by the inefficiency and security vulnerabilities of most existing authentication schemes, we propose a lightweight authentication technique for 5G-based IoV networks in a smart-city environment. The proposed LAIoV-5G scheme solves the security challenges by introducing a lightweight authentication scheme specifically designed for 5G-based IoV. By leveraging high data transfer rates of 5G, reducing latency and improving reliability, the LAIoV-5G scheme provides a robust solution for secure and efficient communications in smart-city environments. The proposed LAIoV-5G scheme aims to improve security, privacy and resilience against potential attacks, through reliable authentication across full assessments. Specifically, the major contributions of our work are as follows:

- The authentication method is developed based on a lightweight and secure cryptographic primitive; namely, ECC, hash function and timestamp to make the source-authentication process secure and efficient. In fact, a two-factor authentication mechanism is presented that is lightweight, efficient, dependable and secure for IoV applications in a smart-city environment.
- We have designed LAIoV-5G scheme to be extremely lightweight, ensuring its high performance in the IoV system. The improved security performance of our proposed LAIoV-5G scheme is crucial for the IoV in which communications take place over insecure communication media.
- We have conducted a comprehensive evaluation of the resistance of our proposed LAIoV-5G scheme to various security intrusions. The results indicate that LAIoV-5G scheme has robust security features.

The rest of this work is structured as follows: Section 2 describes some of related works in this domain while Section 3 presents a background of lightweight authentication schemes, which is followed by the proposed LAIoV-5G scheme in Section 4. Section 5 presents the security analysis. Section 6 discusses the performance analysis. The paper is finally concluded in Section 7.

## 2. RELATED WORK

This section explores the IoV studies based on 5G technology. IoV, compared to conventional wireless networks, presents a host of technical and security obstacles [17]. For instance, issues such as privacy, key distribution, bootstrap, mobility, incentives and poor error tolerance are yet to be addressed. Therefore, both industry and academia have developed several methods to protect privacy and ensure the authenticity of vehicle users in response to these challenges. For instance, Public Key Infrastructure (PKI) has been developed to facilitate key distribution and mutual authentication across IoV users [18]-[24]. In 2005, authentication schemes have been presented in [18] and [19]. In these two protocols, vehicle location and public-key signatures are utilized to prevent attackers waiting on the side of the road from pretending to be an authorized vehicle user on a highway. However, the deployed PKI makes these schemes inefficient, especially in dense IoV networks. In addition, large storage is required for storage of these public-key signatures. To address some of these concerns, hash chain-based authentication mechanisms are developed in [20]-[22]. However, user anonymity is not provided in these schemes and hence, attackers can obtain sensitive driver information, such as registration plates and driver identities. To address this concern, anonymous authentication techniques have been suggested in [23] and [24]. In these schemes, unique pseudo-identities are deployed to conceal true identities and hence mitigate privacy leakage. Here, only the trusted authority (TA) can recover the true identities from these pseudo-identifications.

When it comes to high density of vehicle populations, the task of gathering and storing traffic-related data becomes complex. To tackle this issue, several strategies have been suggested for integrating cloud computing into automotive networks. Basically, the cloud allows vehicles to share resources, like storage, computation and bandwidth. As seen in [25]-[27], these strategies comprise of center, vehicular and roadside clouds. These three clouds have diverse considerations. For instance, the authors in [25] have incorporated autonomous vehicular clouds to utilize unused resources. On the other hand, the platform as a service cloud platform has been incorporated for interactive, mobile and functional clients in [26]. However, the IoV clouds in [27] have been classified as being hybrid vehicular clouds (HVCs), vehicular clouds (VCs) or vehicle-utilizing clouds (VuCs). The unique nature of these solutions emanate from the fact that vehicles can act as cloud service providers (for VCs), customers (for VuCs) and both customers and cloud service providers (for HVCs).

Recent research works in [28]–[33] have proposed authentication techniques to address vehicle networks' privacy and security aspects. In addition, identity-based methods [28]–[34] have been developed to leverage on Bilinear Pair (BP)-related cryptographic procedures for message signing and signature validation. However, BP procedures are computationally extensive. In addition, signature signing and validation require heavy computations and message exchanges. To address these issues, an Elliptic Curve Cryptography (ECC) and identity-based approach is developed in [35]. Although this technique solves the high-computation problems in BP procedures, it has some performance challenges. For instance, as the number of participating nodes increases, the time consumption of ECC procedures also increases, highlighting the urgency of finding a solution. Similarly, several authentication systems based on ECC have been presented in [35]–[42] to address vehicular communication's privacy and security requirements. However, they face the same challenges as the ones in [35].

The most recent schemes utilize vehicle networks supported by 5G technologies [42]–[46] to eliminate the need for Roadside Units (RSUs). In essence, these schemes utilize a vehicle network provided by 5G technology, bypassing the involvement of RSUs in the authentication process. To establish a 5G-enabled vehicle network for RSUs, it is crucial to meticulously analyze and address several key concerns. The 5G wireless network, renowned for its efficiency, enables immediate and low-latency transmission of data, a vital feature for the Vehicle to Everything (V2X) protocol. Vehicles can seamlessly connect with RSUs and other vehicles using 5G modems, sensors and on-board units (OBUs). Relay stations play a pivotal role as intermediaries, facilitating communication between cars and the network backbone, a feature that enhances the network's capabilities. The core network, equipped with resources, efficiently manages data traffic, performs processing tasks and conducts analytics. These resources can be strategically located, either centrally or at the network's periphery. The access network, comprising 5G base stations, ensures comprehensive coverage to RSUs and cars.

Instead, LAIOV-5G leverages the transceiver circuit and algorithmic innovation to circumvent these limitations. In the field of large-scale IoT networks, LAIOV-5G provides lightweight, scalable and efficient authentication mechanisms by taking full advantage of emerging 5G network capabilities, such as ultra-low latency, high data-transfer rates and increased reliability. It is a fully digital scheme with limited computational cost for the authentication process, enhancing higher security features while reducing computational cost compared to existing schemes. This enables fast and secure authentication in real time, especially in dynamic situations, such as intelligent transportation systems (ITSs) and critical infrastructure management. Moreover, LAIOV-5G is specifically built to address the unique problems of smart-city settings, where millions of devices and vehicles must communicate securely and efficiently. The adoption of 5G technology enables the system to handle massive amounts of data and promotes seamless vehicle-to-infrastructure (V2I) communication, which is critical for applications, such as self-driving cars and intelligent traffic management. This makes LAIOV-5G not only more efficient, but also more versatile, as it can meet the security requirements of future IoV systems in smart cities.

## 3. BACKGROUND

In this part, we describe the network structure as well as the security goals of our LAIoV-5G scheme. Table 1 gives a brief description of all the notations used in our LAIoV-5G scheme.

Table 1. Symbols of the proposed work.

| Notations | Definition |
|---|---|
| $TA$ | Trusted Authority |
| $V_i$ | Vehicle |
| $SK_i$ | a shared session key |
| $ID_v$ | Identity of vehicle |
| $PW_i$ | Password |
| $r_v$ | Random number |
| $\oplus$ | Exclusive OR operation |
| $SC_i$ | Smart card |
| $K_s$, $K_p$ | Public and private keys |
| $\|$ | String concatenation |
| $h_i()$ | Cryptography hash function |

### 3.1 Network Structure

This sub-section explains the three network components that make up the network structure of our proposed LAIoV-5G scheme. This includes the vehicles, 5G base station (5G-BS) and the trusted authority, TA. The components shown in Figure 1 are briefly described in the following steps [47].

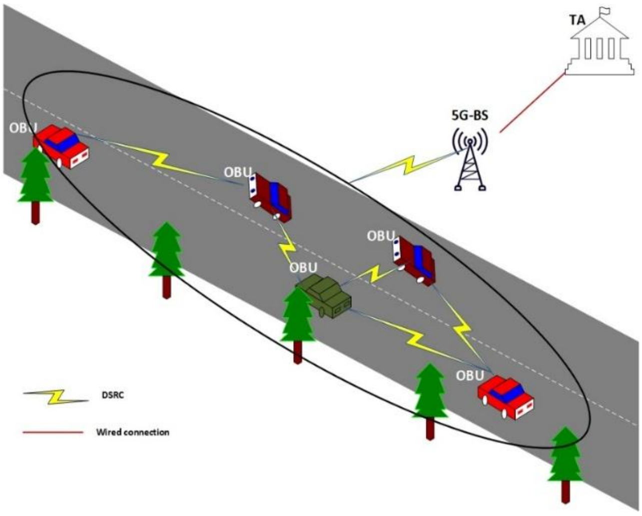


Figure 1. Network structure.

**TA**: This is a powerful computer system, which is a key player in 5G-enabled vehicular networks. It has a large storage capacity to store data. It also issue private keys for very matching vehicles as well as generating system parameters. To uphold network reliability, prevent single points of failure as well as network bottleneck, a number of redundant TAs are deployed in the IoV network.

**5G-BS:** This wireless-communication device is positioned at road intersections and other high-traffic areas. The 5G-BS transceiver has breakneck transmission speeds and wide-area coverage. To prevent attacks, this **5G-BS** is properly safeguarded, for instance, by the use of layered security architecture. It basically acts as an intermediary between the network nodes (vehicles) and the trusted authority, TA. Due to the nature of the processing that it carries out, this **5G-BS** is equipped with large storage, which is necessary during its verification procedures.

**Vehicle:** To facilitate the exchange of traffic-related data in IoV, each vehicle is equipped with an On-Board Unit (OBU). In an effort to prevent unauthorized access, modifications and other attacks, each OBU incorporates a Tamper-Proof Device (TPD). This helps safeguard essential data received from TA and other network elements.

### 3.2 Threat Model

In this sub-section, we model the attacker to have a range of capabilities that can be used in the process of trying to compromise the proposed scheme. Here, the adversary poses the following risks:

- Can fully take charge of the wireless-communication channels. Afterwards, attackers can intercept, capture, modify, erase and insert bogus messages into the communication channel.
- Can steal a user’s smart card or access a user’s password. Thereafter, these security tokens can be utilized to commit numerous cases of system compromise.
- Using techniques, such as power analysis, attackers in possession of a user’s smart card can retrieve the sensitive security values stored in it.
- It is possible for attackers to determine the identities of every server and all users.

### 3.3 Security Goals

To counter the capabilities of the attacker advocated above and ensure robust security for IoV communication using 5G technology, our proposed LAIoV-5G scheme must fulfill the following requirements.

1. ***Mutual Authentication and Integrity:*** These are not just crucial elements, but also the backbone of our proposed LAIoV-5 G scheme. They are the pillars on which our communication security stands, ensuring that only approved entities engage in the interaction process and that the transmitted or stored data remains unaltered and unchanged.
2. ***Unlinkability:*** Adversaries should be incapable of associating any session or messages to any particular network element.
3. ***TA Impersonation Attack:*** This is not just a type of cybercrime, but also a serious threat to our LAIoV-5 G scheme. In this attack, an attacker pretends to be a trusted authority, potentially causing significant damage to our system. Therefore, adversaries should be unable to launch this attack against our LAIoV-5 G scheme.
4. ***Social Engineering Attacks*****:** Here, the attacker pretends to be a familiar person to the target, such as a known user or a trusted entity, in order to gain trust and exploit access privileges.
5. ***Maintaining Privacy for Users*****:** Maintaining user anonymity involves keeping a user's identity concealed or undisclosed to safeguard his/her privacy through encryption methods.
6. ***Replay Attack*****:** A legitimate transmission is required in our LAIoV-5 G scheme. Therefore, previously transmitted messages should not be sent again to a target system to trigger unauthorized actions or data breaches.
7. ***Smart-card Threats*****:** These are dangerous attacks in which a physical smart card containing sensitive data or cryptographic keys is used to obtain unauthorized access to systems or resources.
8. **S*tolen Verifier and Privileged Insider Attacks*****:** This type of attack involves an insider with privileged access to a system that steals a verifier device, such as a token or hardware-security module (HSM). These stolen verifiers can then bypass authentication mechanisms and gain unauthorized access.

### 3.4 Hash Functions

In this sub-section, the one-way hashing function h (.) takes $o$ (string of arbitrary length) as the input.

Thereafter, it produces an output of fixed length, referred to as the hash code. Therefore, hash code = $h(o)$ and any small alteration in the value of the input string can have profound effects on this hash code. According to [43], the hash h (.) has the characteristics below:

- For a given input string, it is simple to find hash code $= h(o)$.

- Given the hash code $h(o)$, its is mathematically difficult to determine $o$.
- For any two inputs of $o_1$ and $o_2$, it is cumbersome to find $h(o_1) = h(o_2)$. This hash function with this property is said to be collision resistant.

## 4. THE LAIOV-5G SCHEME

Our proposed scheme consists of four main phases, including initialization, registration, login and password change, each of which plays a critical role in securing IoV communications. The initialization phase is the foundation, where the TA generates the cryptographic parameters required for the scheme. Using ECC, the TA generates and shares common and public parameters, such as curve points and hash functions, with all participating entities. These parameters allow for lightweight and secure cryptographic computations while maintaining efficient resource utilization. In the second phase, each vehicle is securely registered with the TA. Upon successful completion, the TA assigns a unique vehicle ID and securely embeds the registration details on a smart card provided to the vehicle. This phase is crucial in ensuring that only authorized and verified vehicles are granted access to the IoV network, effectively mitigating the risk of unauthorized entities infiltrating the system.

The login phase is responsible for establishing secure communication channels. The vehicle initiates a session; it sends an encrypted request containing its identity and a timestamp to the TA. The TA verifies the request, ensuring the vehicle's legitimacy. Mutual authentication is then performed between the vehicle and the TA, after which a session key is generated. This session key is generated using lightweight cryptographic exchange, ensuring that all subsequent communications remain confidential and tamper-resistant. Finally, the password-update phase allows the vehicle to securely change its credentials. To do this, the vehicle must confirm its current credentials with the TA. Once verified, the TA simplifies the secure update of both the password and the secret key, ensuring that the process is protected from unauthorized changes.

These four phases work together to form a comprehensive security framework for IoV environments. The interactions and computations between entities are illustrated in Figures 2 and 3 of the manuscript, providing a clear overview of the protocol's operation. This structured approach balances strong security with lightweight requirements for IoV systems, making them efficient and practical for deployment in real-world scenarios. Specific descriptions of f these stages are detailed in the following sub-sections.

### 4.1 Initialization Phase

This phase is responsible for creating and distributing system parameters *via* TA as the following steps:

- Choosing two prime numbers $p$ and $q$.
- Generating random numbers $a$ and $\in F_p$.
- Choosing an elliptic curve $EC$, such that $4a^3 + 27b^2 \neq 0$
- Select the private key $K_s$, where $K_s \in [1, a * b]$.
- Selecting $G$ as a base point on the $EC$.
- Calculating the public key $K_p = GK_s$.
- Determining the cryptography hash function $h(.)$.
- At the end, trusted authority TA publishes parameters $\{q, K_p, G, h(.)\}$.

### 4.2 Registration Phase

Every vehicle that aspires to be part of the IoV network plays a crucial role and must first register. If a vehicle $V_i$ decides to register with the $TA$, the following steps should be followed.

- A user of $V_i$ chooses the identity$ID_v$, Password$PW_i$ and an arbitrary number$r_v \in Z_p^* \ and$ sends $\{ID_v, h(ID_v \parallel PW_i \parallel r_v) \oplus r_v\}$ as request for registration to the $TA$, *via* a highly secure channel, ensuring the safety of the data.
- On receiving the message $\{ID_v, h(ID_v \parallel PW_i \parallel r_v) \oplus r_v\}$, the $TA$ computes $= h(ID_v \parallel K_s) \oplus h(ID_v \parallel PW_i \parallel r_v) \oplus r_v$. Thereafter, it is sent back to them *via* a secure communication medium.
- After getting $A$, the $V_i$ computes the following:

- $B = A \oplus r_v$
- $B = h(ID_v \parallel K_s) \oplus h(ID_v \parallel PW_i \parallel r_v)$
- $C = h(ID_v \parallel PW_i \parallel r_v)$
- Then, the values $\{B, C, r_v, h(\,)\}$ (which include the vehicle's unique identifier and registration details) are uploaded on the smart card $SC_i$ for future verification.

### 4.3 Log-in Phase

The goal of this phase is to have the user of vehicle $V_i$ sign-in into a system with the given $SC_i$ credentials. Thereafter, a secure communication channel is created with a $TA$ server by following the steps outlined below:

**Step 1**. The user $V_i$ inserts the $SC_i$ and inputs his/her credentials $ID_v$, $PW_i$, the $OBU$, then computes $C^* = h(ID_v \parallel PW_i \parallel r_v)$ and confirms it against stored data on the $SC_i$. The session will be terminated if the values $C$ and $C^*$ do not match. Otherwise, $V_i$ will start a secure communication with $TA$ by generating an arbitrary number $a \in Z_p^*$ and achieving the following equations:

- $X = aP$
- $Y = ID_v \oplus (aK_p)$
- $\sigma = h(ID_v \parallel X \parallel h(ID_v \parallel PW_i \parallel r_v) \parallel T_1)$

Then, it sends the encrypted message $\{X, Y, \sigma, B, T_1\}$ to the $TA$.

**Step 2**. On receiving the message $\{X, Y, \sigma, B, T_1\}$, $TA$ achieves the following equations:

- $ID_v = Y \oplus (K_s X)$
- $U_{TA} = B \oplus h(ID_v \parallel K_s)$
- $\sigma^* = h(ID_v \parallel X \parallel U_{TA} \parallel T_1)$.
- It hecks $? = \sigma^*$; the session will be terminated if the check is not verified. Otherwise, the TA will compute the session secret key $\mathrm{SK_i}$ as follows:

$$SK_i = h((K_s X) \parallel ID_v \parallel h(ID_v \parallel K_s))$$
$$Auth_{TA} = h(SK_i \parallel (K_s X), T_2)$$

- Finally, $TA$ sends back the $\{Auth_{TA}, T_2\}$ to $V_i$.

**Step 3**. On receiving the message $\{Auth_{TA}, T_2\}$, the $V_i$ achieves the following equations:

- $U_v = A \oplus h(ID_v \parallel PW_i \parallel r_v)$
- $SK_i = h\big((aK_p) \parallel ID_v \parallel U_v\big)$
- $Auth_{TA}^* = h(SK_i \parallel \big(aK_p\big), T_2)$.
- It checks $Auth_{TA} ?= Auth_{TA}^*$; the session will be terminated if the check is not verified. Otherwise, the $V_i$ will send $\{Auth_v, T_3\}$ to the $TA$ as a response message confirming that the vehicle received the session key correctly, where $Auth_v = h(ID_v \parallel \big(aK_p\big) \parallel U_v \parallel SK_i \parallel T_3)$.

**Step 4**. On receiving the message $\{Auth_v, T_3\}$, the $TA$ computes $Auth_v^* = h(ID_v \parallel (K_s X) \parallel U_{TA} \parallel SK_i \parallel T_3)$ and checks $Auth_v ?= Auth_v^*$. If the check is not verified, the log-in process will be terminated. If not, both $TA$ and $V_i$ consent on using $SK_i$ as a shared session key.

### 4.4 Password-change Phase

The procedures carried out in this sub-section are crucial, since they give the user of vehicle $V_i$ the ability to update his/her password at their discretion. Both $TA$ and $V_i$ parties are involved in the following steps:

**Step 1**. The user of $V_i$ logs in to the vehicle, as explained in the previous phase.

**Step 2**. The user of $V_i$ enters a new password $PW_{i-new}$.

**Step 3**. The smart card $SC_i$, a key player in this process, selects a new arbitrary number $r_{v-new}$ and performs the following equations:

- $B_{new} = B \oplus h(ID_v \parallel PW_i \parallel r_v) \oplus h(ID_v \parallel PW_{i-new} \parallel r_{v-new})$
- $C_{new} = h(ID_v \parallel PW_{i-new} \parallel r_{v-new})$

**Step 4**. The $SC_i$ stores both $B_{new}$ and $C_{new}$ instead of $B$ and $C$ respectively.

**Step 5**. The $V_i$ send both new values of $B_{new}$ and $C_{new}$ to the $TA$ after encrypting them by the session key $SK_i$, ensuring the highest level of security.

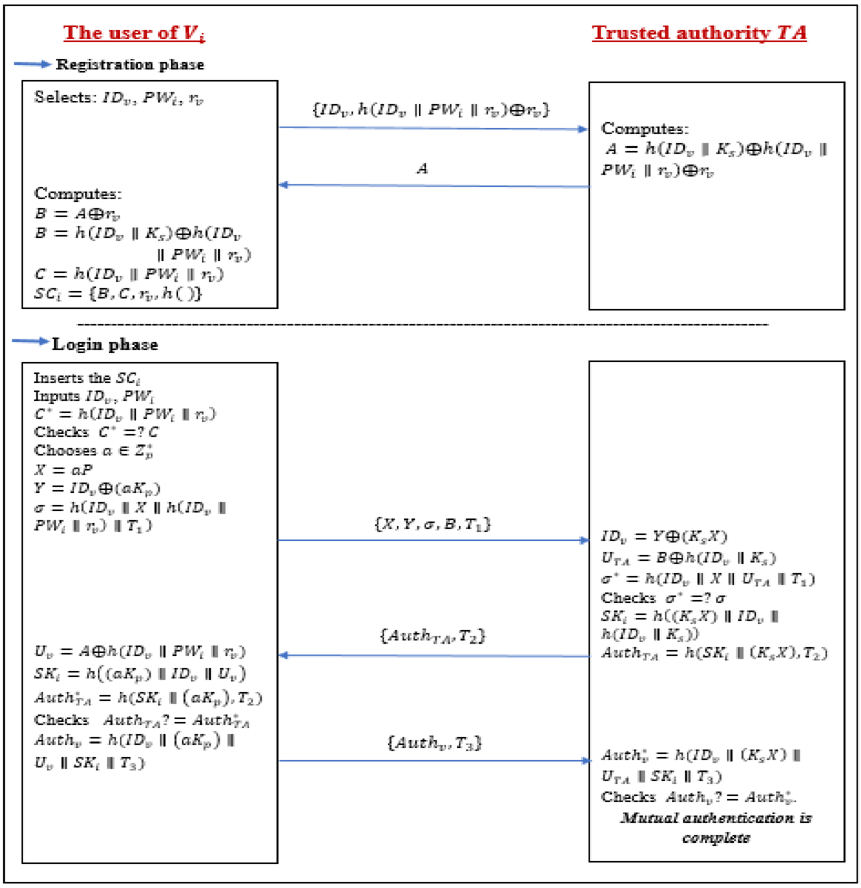

The user of $V_i$
Trusted authority $TA$
Registration phase
Selects: $ID_v$, $PW_i$, $r_v$
$\{ID_v, h(ID_v \parallel PW_i \parallel r_v)\oplus r_v\}$
Computes:
$A = h(ID_v \parallel K_s)\oplus h(ID_v \parallel PW_i \parallel r_v)\oplus r_v$
$A$
Computes:
$B = A\oplus r_v$
$B = h(ID_v \parallel K_s)\oplus h(ID_v \parallel PW_i \parallel r_v)$
$C = h(ID_v \parallel PW_i \parallel r_v)$
$SC_i = \{B, C, r_v, h(\,)\}$
Login phase
Inserts the $SC_i$
Inputs $ID_v$, $PW_i$
$C^* = h(ID_v \parallel PW_i \parallel r_v)$
Checks $C^* =? C$
Chooses $a \in Z_p^*$
$X = aP$
$Y = ID_v\oplus(aK_p)$
$\sigma = h(ID_v \parallel X \parallel h(ID_v \parallel PW_i \parallel r_v) \parallel T_1)$
$\{X, Y, \sigma, B, T_1\}$
$ID_v = Y\oplus(K_s X)$
$U_{TA} = B\oplus h(ID_v \parallel K_s)$
$\sigma^* = h(ID_v \parallel X \parallel U_{TA} \parallel T_1)$
Checks $\sigma^* =? \sigma$
$SK_i = h((K_s X) \parallel ID_v \parallel h(ID_v \parallel K_s))$
$Auth_{TA} = h(SK_i \parallel (K_s X), T_2)$
$\{Auth_{TA}, T_2\}$
$U_v = A\oplus h(ID_v \parallel PW_i \parallel r_v)$
$SK_i = h((aK_p) \parallel ID_v \parallel U_v)$
$Auth_{TA}^* = h(SK_i \parallel (aK_p), T_2)$
Checks $Auth_{TA}? = Auth_{TA}^*$
$Auth_v = h(ID_v \parallel (aK_p) \parallel U_v \parallel SK_i \parallel T_3)$
$\{Auth_v, T_3\}$
$Auth_v^* = h(ID_v \parallel (K_s X) \parallel U_{TA} \parallel SK_i \parallel T_3)$
Checks $Auth_v? = Auth_v^*$.
Mutual authentication is complete

Figure 2. Registration and log-in phases.

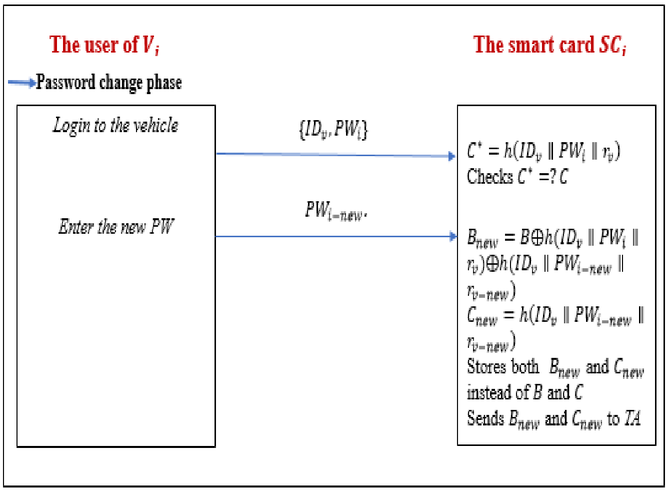

The user of $V_i$
The smart card $SC_i$
Password change phase
Login to the vehicle
$\{ID_v, PW_i\}$
$C^* = h(ID_v \parallel PW_i \parallel r_v)$
Checks $C^* =? C$
Enter the new PW
$PW_{i-new}$.
$B_{new} = B\oplus h(ID_v \parallel PW_i \parallel r_v)\oplus h(ID_v \parallel PW_{i-new} \parallel r_{v-new})$
$C_{new} = h(ID_v \parallel PW_{i-new} \parallel r_{v-new})$
Stores both $B_{new}$ and $C_{new}$ instead of $B$ and $C$
Sends $B_{new}$ and $C_{new}$ to $TA$

Figure 3. Password-change phase.

## 5. Security Analysis

The essence of this section is to present security analysis of our LAIoV-5G scheme. This analysis confirm the proposed LAIoV-5G scheme's robustness and highlights its resistance to various attacks. We further demonstrate that the LAIoV-5G scheme's security is unaffected by various potential circumstances. As shown in Table 2, our LAIoV-5G scheme meets key security requirements, as compared to several related schemes. This should reassure you of its effectiveness.

Table 2. Security comparison.

| **Security requirements** | **Wu, T. Y. et al. [49]** | **Karim, S. et al. [50]** | **Salami, Y, et al. [51]** | **Xie et al. [52]** | **LAIoV-5G** |
|---|---|---|---|---|---|
| Mutual authentication and integrity | Yes | Yes | Yes | Yes | Yes |
| Unlinkability | Yes | No | Yes | No | Yes |
| $TA$ impersonation attack | No | Yes | Yes | Yes | Yes |
| User of $V_i$ impersonation attack | Yes | Yes | Yes | Yes | Yes |
| User anonymity | Yes | Yes | No | Yes | Yes |
| Replay attack | No | Yes | Yes | No | Yes |
| Stolen smart card threat | Yes | No | No | Yes | Yes |
| Stolen verifier and privileged insider threats | Yes | No | Yes | Yes | Yes |

### 1. *Mutual Authentication and Integrity*

The authentication and integrity of our LAIoV-5G scheme is provided as follows:

First message $\{X, Y, \sigma, B, T_1\}$: The $TA$ authenticates the received message $\{X, Y, \sigma, B, T_1\}$. Accordingly, it computes the $ID_v$ and $U_{TA}$ by the deployment of private key $K_s$, then it checks $\sigma ?= \sigma^*$.

Second message $\{Auth_{TA}, T_2\}$: The vehicle user $V_i$ authenticates the received message $\{Auth_{TA}, T_2\}$ according to the equation $Auth_{TA} ?= Auth_{TA}^*$. Only a genuine $TA$ can compute $Auth_{TA}$ since it owns the system's secret key $K_s$. In the same way, at the $V_i$ part, $Auth_{TA}^*$ contains the session key $SK_i$, which includes $ID_v$ concatenated with $U_v$. Moreover, $U_v$ is computed by using the $ID_v$ and $PW_i$. Thus, only a genuine $V_i$ can compute $Auth_{TA}^*$.

Third message $\{Auth_v, T_3\}$: The $TA$ authenticates the received message $\{Auth_v, T_3\}$ according to the equation $Auth_v ?= Auth_v^*$. As $Auth_v^* = h(ID_v \parallel (K_s X) \parallel U_{TA} \parallel SK_i \parallel T_3)$ includes the system's secret key $K_s$ and one-way hash function $h()$ is used, it is impossible for the attacker to compute it.

Hence, the proposed LAIoV-5G scheme offers mutual verification and integrity protection.

### 2. *Unlinkability*

The design of the messages sent in our LAIoV-5 G scheme, such as $\{X, Y, \sigma, B, T_1\}$, is a testament to its technical complexity. It has no static value according to an arbitrary number $a \in Z_p^*$., ensuring that all messages for the exact vehicle are different. This level of complexity makes it impossible for attackers to establish whether any two beacons are being generated by the same vehicle. Hence, the proposed LAIoV-5G scheme offers the unlinkability, a feat of technical ingenuity.

### 3. *$TA$ Impersonation Attack*

In our LAIoV-5G scheme, to pretend to be a legitimate $TA$, an adversary must be in possession of the system's private key $K_s$so as to facilitate the computation of $U_{TA} = A \oplus h(ID_v \parallel K_s)$. Additionally, the session key $SK_i = h((K_s X) \parallel ID_v \parallel h(ID_v \parallel K_s))$ will calculate if having the $K_s$. Likewise, the $TA$'s signature $Auth_{TA} = h(SK_i \parallel (K_s X), T_2)$ contains both $K_s$ and $SK_i$. Thus, only the genuine $TA$ can compute all these security parameters. For this reason, our LAIoV-5G scheme can resist the $TA$ masquerade threats.

### 4. *Vehicle $V_i$ User Impersonation Attack*

In our LAIoV-5G scheme, let's assume that an attacker captures the log-in message $\{X, Y, \sigma, A, T_1\}$, he/she cannot modify this message due to changing the $Y$ for each session. Furthermore, $\sigma = h(ID_v \parallel X \parallel h(ID_v \parallel PW_i \parallel r_v) \parallel T_1)$ contains $ID_v$, $PW_i$ and hash function. Hence, our LAIoV-5G scheme can mitigate vehicle $V_i$ user impersonations.

### 5. *Anonymous Communication*

In our LAIoV-5G scheme, the user of $V_i$ sends a message $\{X, Y, \sigma, A, T_1\}$ through the open-access environment that $ID_v$ is not in the plain text, during the log-in phase. If any challenger intercepts the message, whose role is to test the user's authenticity, he/she cannot obtain the $ID_v$, because in $Y = ID_v \oplus (aK_p)$, the arbitrary nonce $a$ is exposed to a multiplication operation with the public key $K_p$.

Besides, $XOR$ is applied between $ID_v$ and the $aK_p$. Additionally, in $\sigma = h(ID_v \parallel X \parallel h(ID_v \parallel PW_i \parallel r_v) \parallel T_1)$, $ID_v$ is concatenated with $X$, $C^*$and then encrypted with hashing function $h()$. Hence, the proposed LAIoV-5G scheme offers anonymous communication.

### 6. *Message Replay Attacks*

For the proposed LAIoV-5G scheme, timestamp $T_i$ is applied to all sending messages $\{X, Y, \sigma, A, T_1\}$; $\{Auth_{TA}, T_2\}$; $\{Auth_v, T_3\}$,, the receiver avoids the replay attack by refusing the message if the timestamp expires. Hence, our LAIoV-5G scheme can prevent replay attacks.

### 7. *Stolen Smart Card Attack*

Our LAIoV-5G scheme is built with a strong focus on security. The smart card securely stores data $B = h(ID_v \parallel K_s) \oplus h(ID_v \parallel PW_i \parallel r_v)$ and $C = h(ID_v \parallel PW_i \parallel r_v)$., making it impossible for an attacker to obtain any parameter used to guess the $ID_v$ and $PW_i$ or the secret data. Even if the attacker manages to get the user's information $SC$, he/she cannot utilize the stored data for his/her own benefit. This robust security design of our LAIoV-5G scheme effectively prevents smart card-loss attacks.

### 8. *Privileged Insider and Stolen Verifier Threats*

In our LAIoV-5G scheme, we do not preserve any database and $TA$ authenticates the message received from the $V_i$ using the private key$K_s$. Also, the $ID_v$ and $PW_i$are not sent to the $TA$ in plaintext. So, our LAIoV-5G scheme can resist the privileged-insider and stolen-verifier threats.

## 6. PERFORMANCE EVALUATION

The security features supported by the proposed LAIoV-5 G scheme with those offered by its peers [49]–[52] are presented in Table 2. It is clear that our scheme mitigates numerous threats, including privileged insider, user impersonation, stolen verifiers, server impersonation and stolen smart-card threats. The added benefit of user anonymity further enhances the appeal of the suggested protocol. Based on the information shown in Table 2, it is clear that the related protocols contain a few security issues, whereas our LAIoV-5 G scheme is fully secure against such threats.

In this section, an examination of the effectiveness of our scheme, including computational and communication costs, is presented. We demonstrate the performance of our scheme by comparing it with the schemes of Wu, T. Y. et al. [49], Karim, S. et al. [50], Salami, Y. et al. [51] and Xie et al. [52]. Our evaluation of the computational complexities of our LAIoV-5G scheme and its peers yielded impressive results. We adopted the time of cryptographic operations as managed by Xie et al. [52] which are executed on a 64-bit laptop with Windows 10 Pro environment installed and 16 GB of RAM, running on an Intel i5 6300 GHz CPU. Table 3 shows the time taken to run different cryptographic operations.

Table 3. Execution time.

| Operation | Notation | Time cost (ms) |
|---|---|---|
| Hash function | $T_h$ | 0.019 |
| Multiplication of point on ECC | $T_m$ | 2.610 |
| Symmetric encryption/decryption | $T_{enc-dec}$ | 0.511 |

In the scheme of Wu, T. Y. et al. [49], the following operations are executed: (12 scalar multiplications) and (22 secure hash functions). Thus, the total computation time is $22T_h$+ $12T_m = 31.738$. In the scheme of Karim, S. et al. [50], the following operations are executed: (6 scalar multiplications) and (10 secure hash functions). Thus, the total computation time is $10T_h$+ $6T_m = 15.85$. In the scheme of Salami, Y, et al. [51], the following operations are executed: (8 scalar multiplications) and (30 secure hash functions). Thus, the total computation time is $30T_h$+ $8T_m = 21.45$. In the scheme of Xie et al. [52], the following operations are executed: (6 scalar multiplications) and (18 secure hash functions) and (1 Symmetric Encryption/Decryption). Thus, the total computation time is $18T_h$ + $6T_m$ + $1T_{enc-dec} = 16.513$ . On the other hand, our LAIoV-5G scheme needs only (3 scalar multiplications) and (13 secure hash functions). Thus, the total computation time of our LAIoV-5G scheme is $13T_h$+ $3T_m = 8.077$. Table 4 gives the comparative analysis of the obtained computation complexities.

Table 4. Computation-cost comparison.

| **Schemes** | $T_h$ | $T_m$ | $T_{enc-dec}$ | **Total** | **Computation cost (ms)** |
|---|---|---|---|---|---|
| Wu, T. Y. et al. [49] | 22 | 12 | 0 | $22T_h$+ $12T_m$ | 31.738 |
| Karim, S. et al. [50] | 10 | 6 | 0 | $10T_h$+ $6T_m$ | 15.85 |
| Salami, Y, et al. [51] | 30 | 8 | 0 | $30T_h$+ $8T_m$ | 21.45 |
| Xie et al. [52] | 18 | 6 | 1 | $18T_h$+ $6T_m$+$1T_{enc-dec}$ | 16.513 |
| LAIoV-5G | 13 | 3 | 0 | $13T_h$+ $3T_m$ | 8.077 |

In terms of communication cost, Table 5 shows the sizes of different cryptographic operations, while Table 6 provides a comparative analysis of the communication complexity of our scheme *versus* its counterparts. In Karim, S. et al. [50], four messages are transmitted; namely (Mssg1 = RIDVn, CertifVn, AVn, DsignVn, TS1), (Mssg2 = RIDRSU, CertifRSU, BRSU, SKey-VerRSU−V, TS2) and (Mssg3 = ACKVn−RSU, TS3), which include (3 ECC points), (2 physical identities), (4 hash function outputs) and (3 timestamps). Thus, a total of 2400 bits are transmitted. In the same way, the communication cost is calculated for Wu, T. Y. et al. [49], Salami, Y, et al. [51], Xie et al. [52] and our LAIoV-5G schemes.

Table 5. Cryptographic-operation output sizes.

| **Operations** | **Cost (bits)** |
|---|---|
| Elliptic Curve Point | 256 bits |
| Actual identity | 256 bits |
| One-way hash function | 256 bits |
| Timestamps | 32 bits |
| Arbitrary nonce | 256 bits |
| Symmetric encryption/decryption | AES-128 bits |

Table 6. Communication-cost comparison

| **Schemes** | **No. of messages** | **Communication cost (bit)** |
|---|---|---|
| Wu, T. Y. et al. [49] | 5 | 3744 |
| Karim, S. et al. [50] | 3 | 2400 |
| Salami, Y, et al. [51] | 5 | 3520 |
| Xie et al. [52] | 4 | 2976 |
| LAIoV-5G | 3 | 1632 |

As shown in Table 4 and Table 6, the computation time of our LAIoV-5G scheme is 8.077 ms, which is 74.6%, 49%, 62.3% and 51% lower than those of Wu, T. Y. et al. [49], Karim, S. et al. [50], Salami, Y. et al. [51] and Xie et al. [52], respectively. The communication cost of our LAIoV-5G scheme is 1632 bits, which is 56.4%, 32%, 53.6% and 45.1% lower than those of Wu, T. Y. et al. [49], Karim, S. et al. [50], Salami, Y. et al. [51] and Xie et al. [52], respectively.

Table 7. Improvement of our LAIoV-5G scheme over other schemes.

| **Schemes** | **Computation improvement** | **Communication improvement** |
|---|---|---|
| Wu, T. Y. et al. [49] | 74.6% | 56.4% |
| Karim, S. et al. [50] | 49% | 32% |
| Salami, Y, et al. [51] | 62.3% | 53.6% |
| Xie et al. [52] | 51% | 45.1% |

Table 7 shows the improvement of our LAIoV-5G scheme compared with other schemes in terms of computation and communication costs. The results unequivocally demonstrate the superiority of computational and communication efficiency of our scheme over other related schemes. Moreover, our scheme achieves a robust security posture at lower-bandwidth requirements, further solidifying its effectiveness and impressiveness.

## 7. CONCLUSION

This paper presents a highly effective LAIOV-5G protocol to secure message exchanges in IoV enabled smart cities. The proposed scheme enables a unique authentication method and demonstrates cost-effectiveness in terms of computation and communication complexities. The comparative evaluation results show that it incurs the lowest costs when contrasted against its peer authentication protocols. Specifically, security evaluations show that LAIOV-5G protocol withstands significant known security attacks. Some of these attacks include stolen smart card, privileged insider, impersonation and message-replay attacks. Hence, the suggested methodology has been demonstrated to be effective, dependable and secure. In future work, we plan to conduct a detailed evaluation of the performance of the proposed scheme in large-scale smart-vehicle networks and address the challenges related to real-world applications, which were beyond the scope of this study.

## REFERENCES

[1] S. Chen, H. Xu, D. Liu, B. Hu and H. Wang, "A Vision of IoT: Applications, Challenges and Opportunities with China Perspective," IEEE Internet of Things J., vol. 1, no. 4, pp. 349-359, 2014.

[2] C.-M. Chen, Z. Li, A. K. Das, S. A. Chaudhry and P. Lorenz, "Provably Secure Authentication Scheme for Fog Computing-enabled Intelligent Social Internet of Vehicles," IEEE Transactions on Vehicular Technology, vol. 73, no. 9, pp. 13600-13610, DOI: 10.1109/TVT.2024.3382971, Sept. 2024.

[3] S. Mumtaz, A. Bo, A. Al-Dulaimi and K.-F. Tsang, "Guest Editorial 5G and Beyond Mobile Technologies and Applications for Industrial IoT (IIoT)," IEEE Transactions on Industrial Informatics, vol. 14, no. 6, pp. 25882591, Jun. 2018.

[4] S. Garg et al., "MobQoS: Mobility-aware and QoS-driven SDN Framework for Autonomous Vehicles," IEEE Wireless Communications, vol. 26, no. 4, pp. 1220, Aug. 2019.

[5] M. A. Al Sibahee et al., "Blockchain-based Authentication Schemes in Smart Environments: A Systematic Literature Review," IEEE Internet of Things J., vol. 11, no. 21, pp. 34774-34796, 2024.

[6] V.O. Nyangaresi et al., "Energy Efficient Dynamic Symmetric Key Based Protocol for Secure Traffic Exchanges in Smart Homes," Applied Sciences, vol. 12, no. 24, p. 12688, 2022.

[7] V.O. Nyangaresi et al., "Smart City Energy Efficient Data Privacy Preservation Protocol Based on Biometrics and Fuzzy Commitment Scheme," Scientific Reports, vol. 14, Article no. 16223, 2024.

[8] V. O. Nyangaresi et al., "Authentication and Key Agreement Protocol for Secure Traffic Signaling in 5G Networks," Proc. of the 2021 IEEE 2nd Int. Conf. on Signal, Control and Communication (SCC), pp. 188-193, DOI: 10.1109/SCC53769.2021.9768338, Tunis, Tunisia, 2021.

[9] V. O. Nyangaresi et al., "Towards Security and Privacy Preservation in 5G Networks," Proc. of the 2021 29th Telecommuni. Forum (TELFOR), pp. 1-4, DOI: 10.1109/TELFOR52709.2021.9653385, Belgrade, Serbia, 2021.

[10] G. Rathee et al., "Trusted Orchestration for Smart Decision-making in Internet of Vehicles," IEEE Access, vol. 8, pp. 157427-157436, 2020.

[11] C.-M. Chen, Q. Miao, S. Kumari, M. K. Khan and J. J. P. C. Rodrigues, "A Privacy-preserving Authentication Protocol for Electric Vehicle Battery Swapping Based on Intelligent Blockchain," IEEE Internet of Things J., vol. 11, no. 10, pp. 17538-17551, 15 May15, 2024.

[12] C.-M. Chen et al., "A Secure Authentication Protocol for Internet of Vehicles," IEEE Access, vol. 7, pp. 12047-12057, 2019.

[13] J. G. Andrews et al., "What Will 5G Be?" IEEE Journal on Selected Areas in Communications, vol. 32, pp. 1065–1082, 2014.

[14] X. Huang, R. Yu, J. Kang, Y. He and Y. Zhang, "Exploring Mobile Edge Computing for 5G-enabled Software Defined Vehicular Networks," IEEE Wireless Communications, vol. 24, pp. 55–63, 2017.

[15] S. A. A. Shah, E. Ahmed, M. Imran and S. Zeadally, "5G for Vehicular Communications," IEEE Communications Magazine, vol. 56, pp. 111–117, 2018.

[16] M. A. Al-Shareeda et al., "Password-guessing Attack-aware Authentication Scheme Based on Chinese Remainder Theorem for 5G-enabled Vehicular Networks," Applied Sciences, vol. 12, no. 3, p. 383, 2022.

[17] B. Parno and A. Perrig, "Challenges in Securing Vehicular Networks," Proc. of Workshop Hot Topics Network (HotNets-IV), pp. 1_6, Annapolis, MD, USA, 2005.

[18] S. Capkun and J.-P. Hubaux, "Secure Positioning of Wireless Devices with Application to Sensor Networks," Proc. of the IEEE 24th Annual Joint Conference of the IEEE Computer and Communications Societies, vol. 3, pp. 1917-1928, Miami, USA, Mar. 2005.

[19] L. Lazos, R. Poovendran and S. Apkun, "ROPE: Robust Position Estimation in Wireless Sensor Networks," Proc. of the IEEE 4th Int. Symposium on Information Processing in Sensor Networks (IPSN 2005), Boise, USA, p. 43, 2005.

[20] A. Studer, F. Bai, B. Bellur and A. Perrig, "Flexible, Extensible and Efficient VANET Authentication,"

Journal of Communications and Networks, vol. 11, no. 6, pp. 574-588, Dec. 2009.

[21] X. Lin et al., "TSVC: Timed Efficient and Secure Vehicular Communications with Privacy Preserving," IEEE Transactions on Wireless Communications, vol. 7, no. 12, pp. 4987-4998, Dec. 2008.

[22] B. Ying, D. Makrakis and H. T. Mouftah, "Privacy Preserving Broadcast Message Authentication Protocol for VANETs," J. of Network and Computer Applications, vol. 36, no. 5, pp. 1352-1364, 2013.

[23] C. Zhang, R. Lu, X. Lin, P.-H. Ho and X. Shen, "An Efficient Identity-based Batch Verification Scheme for Vehicular Sensor Networks," Proc. of the 27th IEE Conf. on Computer Communications (IEEE INFOCOM 2008), pp. 246-250, Phoenix, USA, Apr. 2008.

[24] C. Zhang, P.-H. Ho and J. Tapolcai, "On Batch Verification with Group Testing for Vehicular Communications," Wireless Networks, vol. 17, no. 8, p. 1851, 2011.

[25] M. Eltoweissy, S. Olariu and M. Younis, "Towards Autonomous Vehicular Clouds," Proc. of the Int. Conf. on *Ad Hoc* Networks, Part of the Book Series: Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Eng., vol. 49 pp. 1-16, Berlin, Germany, 2010.

[26] D. Bernstein, N. Vidovic and S. Modi, "A Cloud PAAS for High Scale, Function and Velocity Mobile Applications - With Reference Application As the Fully Connected Car," Proc. of the 2010 IEEE 5th Int. Conf. on Systems and Networks Communications (ICSNC), pp. 117-123, Nice, France, Aug. 2010.

[27] R. Hussain, J. Son, H. Eun, S. Kim and H. Oh, "Rethinking Vehicular Communications: Merging VANET with Cloud Computing," Proc. of the IEEE 4th Int. Conf. on Cloud Computing Technology and Science Proceedings (CloudCom), pp. 606-609, Taipei, Taiwan, Dec. 2012.

[28] H. Zhong, S. Han, J. Cui, J. Zhang and Y. Xu, "Privacy-preserving Authentication Scheme with Full Aggregation," Information Sciences, vol. 476, pp. 211–221, 2019.

[29] M. Azees, P. Vijayakumar and L. J. Deboarh, "EAAP: Efficient Anonymous Authentication with Conditional Privacy-preserving Scheme for Vehicular *Ad Hoc* Networks," IEEE Transactions on Intelligent Transportation Systems, vol. 18, pp. 2467–2476, 2017.

[30] L. Zhang et al., "Distributed Aggregate Privacy-preserving Authentication in VANETs," IEEE Transactions on Intelligent Transportation Systems, vol. 18, pp. 516–526, 2016.

[31] M. Bayat et al., "A New and Efficient Authentication Scheme for Vehicular *Ad Hoc* Networks," Journal of Intelligent Transportation Systems, vol. 24, pp. 171–183, 2020.

[32] M. Bayat, M. Pournaghi, M. Rahimi and M. Barmshoory, "NERA: A New and Efficient RSU-based Authentication Scheme for VANETs," Wireless Networks, vol. 26, pp. 3083–3098, 2020.

[33] S. M. Pournaghi et al., "NECPPA: A Novel and Efficient Conditional Privacy-preserving Authentication Scheme for VANET," Computer Networks, vol. 134, pp. 78–92, 2018.

[34] M. A. Al-Shareeda et al., "SE-CPPA: A Secure and Efficient Conditional Privacy-preserving Authentication Scheme in Vehicular *Ad Hoc* Networks," Sensors, vol. 21, p. 8206, 2021.

[35] D. He, S. Zeadally, B. Xu and X. Huang, "An Efficient Identity-based Conditional Privacy-preserving Authentication Scheme for Vehicular *Ad Hoc* Networks," IEEE Transactions on Information Forensics and Security, vol. 10, pp. 2681–2691, 2015.

[36] M. R. Asaar, M. Salmasizadeh, W. Susilo and A. Majidi, "A Secure and Efficient Authentication Technique for Vehicular Ad-hoc Networks," IEEE Transactions on Vehicular Technology, vol. 67, no. 6, pp. 5409–5423, 2018.

[37] M. A. Al-Shareeda, M. Anbar, S. Manickam and I. H. Hasbullah, "Towards Identity-based Conditional Privacy-preserving Authentication Scheme for Vehicular Ad Hoc Networks," IEEE Access, vol. 9, pp. 113226–113238, 2021.

[38] M. A. Al-Shareeda, M. Anbar, M. A. Alazzawi, S. Manickam and A. S. Al-Hiti, "LSWBVM: A Lightweight Security without Using Batch Verification Method Scheme for a Vehicle Ad Hoc Network," IEEE Access, vol. 8, pp. 170507-170518, DOI: 10.1109/ACCESS.2020.3024587, 2020.

[39] J. S. Alshudukhi, B. A. Mohammed and Z. G. Al-Mekhlafi, "Conditional Privacy-preserving Authentication Scheme without Using Point Multiplication Operations Based on Elliptic Curve Cryptography (ECC)," IEEE Access, vol. 8, pp. 222032–222040, 2020.

[40] M. Alazzawi, H. Lu, A. Yassin and K. Chen, "Efficient Conditional Anonymity with Message Integrity and Authentication in a Vehicular Ad hoc Network," IEEE Access, vol. 7, pp. 71424–71435, 2019.

[41] M. A. Alazzawi et al., "ID-PPA: Robust Identity-based Privacy-preserving Authentication Scheme for a Vehicular Ad-Hoc Network," Proc. of Advances in Cyber Security (ACeS 2020), Part of Book Series: Communications in Computer and Information Science, vol. 1347, DOI: 10.1007/978-981-33-6835-4_6, Springer, Singapore, 2021.

[42] J. S. Alshudukhi, Z. G. Al-Mekhlafi and B. A. Mohammed, "A Lightweight Authentication with Privacy-preserving Scheme for Vehicular Ad Hoc Networks Based on Elliptic Curve Cryptography," IEEE Access, vol. 9, pp. 15633–15642, 2021.

[43] J. Cui, J. Chen, H. Zhong, J. Zhang and L. Liu, "Reliable and Efficient Content Sharing for 5G-enabled Vehicular Networks," IEEE Trans. on Intelligent Transportation Syst., vol. 23, no. 2, pp. 1–13, 2020.

[44] J. Cui, X. Zhang, H. Zhong, Z. Ying and L. Liu, "RSMA: Reputation System-based Lightweight Message Authentication Framework and Protocol for 5G-enabled Vehicular Networks," IEEE Internet of Things

J., vol. 6, no. 4, pp. 6417–6428, 2019.

[45] J. Zhang, H. Zhong, J. Cui, M. Tian, Y. Xu and L. Liu, "Edge Computing-based Privacy-preserving Authentication Framework and Protocol for 5G-enabled Vehicular Networks," IEEE Transactions on Vehicular Technology, vol. 69, no. 7, pp. 7940–7954, 2020.

[46] M. A. Al-Shareeda et al., "CM-CPPA: Chaotic Map-based Conditional Privacy-preserving Authentication Scheme in 5G-enabled Vehicular Networks," Sensors, vol. 22, no. 13, p. 5026, 2022.

[47] M. A. Al-Shareeda et al., "Efficient Conditional Privacy Preservation with Mutual Authentication in Vehicular Ad Hoc Networks," IEEE Access, vol. 8, pp. 144957–144968, 2020.

[48] M. A. Alazzawi, H. Lu, A. A.Yassin and K. Chen, "Robust Conditional Privacy-preserving Authentication based on Pseudonym Root with Cuckoo Filter in Vehicular Ad Hoc Networks," KSII Trans. on Internet and Information Systems, vol. 13, no. 12, DOI: 10.3837/tiis.2019.12.018, 2019.

[49] T.-Y. Wu, Z. Lee, L. Yang and C.-M. Chen, "A Provably Secure Authentication and Key Exchange Protocol in Vehicular Ad Hoc Networks," Security and Communication Networks, vol. 2021, no. 1, p. 9944460, 2021.

[50] S. M. Karim et al., "BSDCE-IoV: Blockchain-based Secure Data Collection and Exchange Scheme for IoV in 5G Environment," IEEE Access, vol. 11, pp. 36158-36175, 2023.

[51] Y. Salami, V. Khajehvand and E. Zeinali, "SAIFC: A Secure Authentication Scheme for IOV Based on Fog-cloud Federation," Security and Communication Networks, vol. 2023, no. 1, p. 9143563, DOI: 10.1155/2023/9143563, 2023.

[52] Q. Xie and J. Huang, "Improvement of a Conditional Privacy-preserving and Desynchronization-Resistant Authentication Protocol for IoV," Applied Sciences, vol. 14, no. 6, p. 2451, 2024.

## ملخص البحث:

لقد مكّنت شبكات الجيل الخامس من تطوير مُدنٍ ذكية يتمّ فيها جمْع كمّيّاتٍ هائلةٍ من البيانات وتخزينها ونشرها. ويتمثّل الهدف النّهائي لتلك المدن الذّكية في تقليل التّكلفة ورفع مستوى أمان الأداء. وفي هذه البيئة، تساعد إنترنت المركبات في ربْط المركبات والمشاة وغُرف التّحكُّم وبعض البِنى التّحتية للطرق. ونظراً للطّبيعة غير الآمنة لقنوات الاتّصال في إنترنت المركبات لتبادل المعلومات، فإنّ من المهمّ تطوير تقنياتٍ عمليةٍ للحفاظ على سرّية المعلومات وعلى الخصوصية.

وقد تمّ اقتراح العديد من الحُلول المرتبطة بالأمان في الماضي القريب. ولسوء الحظّ، فإنّ غالبية تقنيات المصادقة لها عيوب فيما يتعلّق بالأمان، الأمر الّذي يهدّد البيانات المنقولة، كما أنّ بعضها يتّصف بقدرٍ عالٍ من عدم الفاعلية.

ولِجَسْر تلك الفجوات، نقدّم في هذه الورقة البحثية مخطّط مصادقةٍ "خفيف الوزْن" لإنترنت المركبات، مبنياً على تقنية الجيل الخامس. وقد جرى تحليل المخطّط المقترح من حيث الأمان، وبيّنت نتائج التّحليل أنّ المخطّط المقترح تمكَّن من كبْح العديد من الهجمات الّتي تهدّد اتّصالات إنترنت المركبات في بيئة المدينة الذّكية. ومن ناحيةٍ أخرى، أثبت مخطّط المصادقة المقترح مستوىً عالياً من الفاعلية في تجارب تحليل الأداء.

# TOWARDS OPTIMIZING THE DOWNLINK TRANSMIT POWER IN UAV-INTEGRATED IRS WIRELESS SYSTEMS

Ali Reda, Tamer Mekkawy and Ashraf Mahran

*(Received: 24-Sep.-2024, Revised: 28-Nov.-2024, Accepted: 15-Dec.-2024)*

## *ABSTRACT*

*Air-to-ground interference poses a critical challenge in integrating unmanned aerial vehicles (UAVs) into cellular networks. In downlink scenarios, UAVs can withstand significant interference from co-channel base stations (BSs) due to the guaranteed line-of-sight (LoS) link with ground users. Our research focuses on power optimization in BSs and applying green-energy principles to pave the way for more sustainable and energy-efficient BSs within UAV-integrated wireless systems. To this end, this paper investigates a downlink UAV communication scenario in which the intelligent reflecting surface (IRS) is mounted on the UAV to practically nullify the interference originating from the co-channel BSs. We formulate the IRS beamforming matrix to reduce transmit power by optimizing passive beamforming for IRS elements, incorporating adjustments to phase shifts and amplitude coefficients while considering the positioning of the UAV. The proposed optimization problem is non-convex and thus a successive convex-approximation (SCA) method is adopted to convert all constraints into a quadratic approximation. Simulation results demonstrate that the proposed SCA algorithm provides an efficient transmit-power minimization approach with low computational complexity for large IRS, since it achieves close-to-optimal performance and significantly outperforms conventional systems without IRS. In interference scenarios and with different numbers of IRS meta-atoms, the proposed algorithm achieves a power reduction of approximately 8 and 13 dBm, while maintaining the same required signal-to-interference-plus-noise ratio.*

## *KEYWORDS*

*Intelligent reflecting surfaces (IRS), Successive convex approximation (SCA), Cone programming, Unmanned aerial vehicles (UAVs).*

## 1. INTRODUCTION

Unmanned aerial vehicles (UAVs) have become a critical enabler for delivering wireless *ad-hoc* connectivity, especially in disaster areas and harsh environments [1]. However, because line-of-sight (LoS) transmission is dominant in aeronautical communication, UAVs are frequently affected by signal interference from terrestrial networks. On the other hand, intelligent reflecting surface (IRS) has emerged recently as a practical way to enhance the wireless-propagation environment by improving communication and coverage performance [2]. An IRS is made up of several passive meta-atoms that are coordinated by a microcontroller in order to create a directional reflection of the incident signal towards a desirable direction. The reflected signals are added either constructively or destructively to the received signal to strengthen or weaken the signal-to-interference-plus-noise ratio (SINR) by carefully tweaking the phase shifts. For systems beyond the fifth generation (5G), IRS-assisted wireless networks have demonstrated considerable spectral and power efficiency [3]. IRS-assisted UAV enables intelligent reflection over the air. Compared to terrestrial IRS set-ups, IRS-supported UAV systems are more effective at establishing robust LoS connections with ground BSs due to the UAV's elevated position, consequently reducing signal blockage [4].

The next-generation wireless networks are poised to transform wireless communication, facilitating the realization of advanced applications, like holographic telepresence, autonomous vehicles and pervasive sensing. However, future networks will require higher data rates, reduced latency and highly reliable, secure communications to support these innovations. This demand will require the deployment of additional BSs and network components, resulting in heightened energy consumption [5]. The cooperative-RSMA system with IRS assistance has proven to reduce the system's energy consumption [6]. Consequently, it is imperative to explore potential energy-saving opportunities within the next generation. One crucial area of focus involves optimizing the phase-shift configurations of multiple IRS meta-atoms to achieve the desired signal amplification or interference

A. Reda, T. Mekkawy and A. Mahran are with the Avionics Department, Military Technical College, Cairo 11331, Egypt. Emails: ali.aboelyazeed@ieee.org, mekkawy@ieee.org and a.mahran@ieee.org

suppression in IRS-assisted systems. However, it is computationally demanding to solve the joint optimization of the passive beamforming at the IRS and the active beamforming at the BS directly due to its non-convex nature. Some algorithms based on alternating optimization (AO) [7], semi-definite relaxation (SDR) [8] and manifold optimization [9] maximize the performance of passive IRS beamforming. Consequently, the aforementioned approaches necessitate many iterations or significant computing complexity to reach convergence for large IRS elements. IRS applications have garnered significant attention for enhancing the quality of both uplink and downlink transmissions. A joint transmit beamforming of the access point (AP) and the passive reflection of the IRS was frequently discussed in technical literature. More specifically, in [10], the authors maximized the secrecy rate of the communication link with a single eavesdropper in a multi-input-single-output (MISO) system. In [11], the authors utilized the transmit beamforming and the reflecting coefficients to maximize the weighted sum secrecy rate. Additionally, the authors in [12] studied the secure-communication capabilities of the IRS-assisted multi-user in a MISO interference channel. We use the best active transmit beamforming strategy along with passive phase shifts for the IRSs in this work. These are found using SDR and successive convex approximation (SCA) methods. The passive-beamforming capabilities of IRSs were used to attain an optimal performance balance between terrestrial aerial users and UAVs in [13].

The authors in [14] investigated the security and spectrum-efficiency aspects of secondary users in Cognitive Radio Networks (CRNs) with the assistance of the IRS. Specifically, they studied the implementation of the IRS in an underlay CRN and co-designed the transmit and reflect beamforming vectors at the IRS. To address the challenge of maximizing secrecy capacity in communication systems, they proposed an iterative AO algorithm. In [15], the authors introduced the IRS and utilized an AO approach to maximize the secrecy rate in Multiple-Input Multiple-Output (MIMO) wiretap channels. Jiang et al. investigated the use of IRS in MIMO cognitive radio systems in [16]. To solve the sum-rate maximization problem, they introduced the weighted minimal mean square error (WMMSE) approach and an AO-based method. The authors in [17] addressed the problem of optimizing the beampattern for an eavesdropping target in an IRS-aided integrated sensing and communication (ISAC) system. They employed an SCA algorithm to jointly optimize the transmit beamforming of the AP and the phase-shift matrix of the IRS, aiming to maximize the beampattern gain.

## 1.1 Related Work

The selection of a phase shifter poses a challenge in IRS-aided communication systems [18]. Previous works on IRS generally assume that each reflecting unit functions as a continuous phase shifter, enabling the phase-shift matrix to adjust for reflective beamforming [19]-[20]. The authors in [8] introduced the joint active and passive-beamforming problem, utilizing the SDR method to minimize transmit power at the BS. However, implementing this approach in practice is challenging due to hardware limitations. Conversely, in [21], the authors investigated IRS-aided wireless networks assuming that only a discrete number of phase shifts are deployed at each reflecting unit. Furthermore, the utilization of non-orthogonal multiple access (NOMA) in conventional multi-IRS downlink systems is investigated in [22], where the authors presented a successive phase-rotation approach to determine the phase shifts sequentially.

In recent years, several papers have focused on IRS-assisted wireless communication [23]. A significant challenge in this system is the joint optimization of phase shifts at the IRS and beamforming vectors at the BS. In [24], a study was conducted on a massive multiple-input multiple-output communication system, where the problem of maximizing the minimum signal-to-interference-plus-noise ratio was addressed. This was achieved by jointly optimizing the signal power, transmit precoding vector and the effective phase shifts at the large intelligent surface. Additionally, optimizing both the active beamforming at the BS and the passive beamforming at the IRS can minimize the BS transmit power to meet mobile-user rate requirements [8]. Users can flexibly modify the channel conditions at the IRS, introducing an additional degree of freedom (DoF) to enhance system performance. In [25], a power-minimization framework was explored within an IRS-enabled NOMA network. The optimization of transmit beamforming and IRS phase shifts was addressed and solved using an alternating optimization approach. The authors in [26] addressed the problem of minimizing power consumption in a downlink-communication scenario assisted by the IRS. This optimization aimed to satisfy minimum SINR thresholds. This paper takes a joint approach, considering both power control and IRS reflection coefficients in the optimization process. On the other hand, the UAV-assisted IRS can significantly improve communication-system reliability and overcome obstacles by adjusting both its position and phase shifts. For instance, in [27], the authors explored the concept of a UAV-carried IRS, aiming to minimize the average total-power consumption of the system through joint optimization of the UAV's trajectory/velocity and the IRS's phase shifts. In [28], the authors investigated a joint active and passive-beamforming approach using a generalized Bender's decomposition-based algorithm. This approach utilizes codebook-based passive beamforming across multiple IRS in a multi-user MISO system. The primary objective is to minimize the transmit power while optimizing the system performance. Moreover, the authors in [29] utilized the SCA and SDR techniques to address the non-convex transmit beamforming-optimization problem. Their objective was to maximize the energy efficiency of a multi-user MISO system aided by IRS.

Unlike terrestrial wireless channels, which often suffer from significant path loss and multi-path effects, the elevated position of UAVs typically provides more dominant LoS channels, leading to improved communication performance. However, these strong LoS links also make UAVs more vulnerable to attacks from terrestrial nodes, such as eavesdropping and jamming [30]. IRS-assisted UAVs offer several advantages, including the ability to dynamically adjust the IRS's position through UAV maneuvering. This introduces a new degree of freedom for IRS optimization. Moreover, mounting the IRS on the UAV enables full- angle reflection and increases the number of LoS links [31]. Numerous studies have focused on optimizing transmit power in IRS-assisted UAV systems [32].

## 1.2 Motivation and Contributions

Compared to traditional, fixed-insert IRS-aided transmission, the benefits of combining the IRS with a UAV are clear. But, there is a lack of research; therefore, further study is required. Our suggested technique has lower per-iteration complexity compared to [8], which is mostly caused by the system's reliance on SDR to solve the optimization problem. The use of AO in [7] makes the problem easier to understand, but the solution that comes out might not be very good because of how the design factors are interdependent. Furthermore, convergence of a penalty-based strategy necessitates a large number of iterations. In this study, we look at how IRS-aided UAVs deal with the problem of balancing transmit power and quality-of-service (QoS) needs in downlink communication. This paper introduces a powerful optimization approach that ensures solution convergence by combining SCA and second- order cone-programming (SOCP) techniques. By breaking the original problem into manageable convex sub-problems with closed-form solutions, the suggested SCA method effectively handles the non-convexity of the original problem. This is achieved by utilizing linear approximations and convex lower limits. Monte Carlo simulations show that the low-complexity SCA achieves efficient performance, far better than benchmark schemes, evaluating the usefulness of this method. Finally, the following are the technical contributions of our paper:

- SCA is adopted to counterbalance the non-convex characteristic of the optimization problem, aiming to minimize the transmit power by optimizing the reflection coefficients of the IRS meta-atoms. This optimization guarantees the fulfillment of both QoS and minimum power for a specific SINR value. By formulating the problem as a SOCP, all optimization variables are updated simultaneously in each iteration. This approach guarantees the algorithm's provable convergence.
- We report an interference-mitigation scheme that utilizes a single UAV equipped with an IRS, eliminating the need for deploying multiple IRSs near each ground BS [13]. This approach effectively addresses interference issues between UAV and ground users during the UAV's downlink communications. The IRS-aided UAV employs passive beamforming to mitigate the interference originating from the co-channel BSs.
- The problem is formulated as a SOCP and the convergence is shown by updating the variables simultaneously. Moreover, the algorithm successfully reduces the power consumption by around 13 dBm, while maintaining the required SINR level.

The paper is structured as follows: Section 2 presents the system model and formulates the optimization problem. In Section 3, we decompose the power optimization and passive beamforming using the SCA algorithm. Section 4 summarizes the convergence and complexity analysis. Section 5 shows the numerical results obtained. Finally, the conclusions are summarized in Section 6.

### Notations

Regular and bold small letters stand for scalars and vectors, while bold capital letters are used for matrices. The magnitude of a scalar $x$ is represented by $|x|$, while the Euclidean norm of a vector $\boldsymbol{x}$ is denoted as $\| \mathbf{x} \|$. The transpose of the matrix $\mathbf{X}$ is represented by $\mathbf{X}^{\mathrm{T}}$, the conjugate transpose by $\mathbf{X}^{\mathrm{H}}$ and the rank is $\mathrm{rank}(\mathbf{X})$. The notation $\mathrm{diag}(\mathbf{X})$ refers to a diagonal matrix. The notation $\mathbb{C}^{\boldsymbol{m}\times\boldsymbol{n}}$ represents the set of $m \times n$ complex matrices. Complex numbers' real and imaginary parts are presented as $\Re(\cdot)$ and $\Im(\cdot)$, respectively. Finally, the letter $\boldsymbol{j}$ represents the imaginary unit $\sqrt{-1}$.

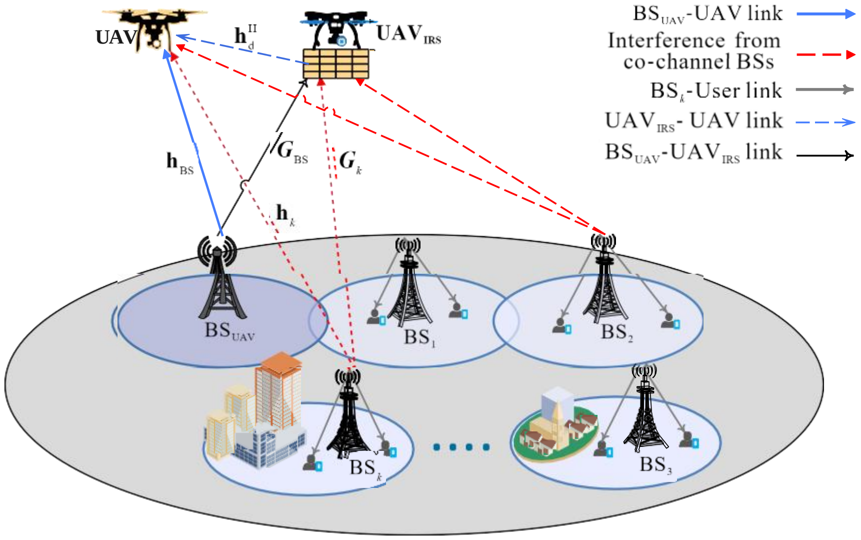

UAV
$\mathbf{h}_d^H$
UAV$_{\text{IRS}}$
BS$_{\text{UAV}}$-UAV link
Interference from co-channel BSs
BS$_k$-User link
UAV$_{\text{IRS}}$- UAV link
BS$_{\text{UAV}}$-UAV$_{\text{IRS}}$ link
$\mathbf{h}_{\text{BS}}$
$\boldsymbol{G}_{\text{BS}}$
$\boldsymbol{G}_k$
$\mathbf{h}_k$
BS$_{\text{UAV}}$
BS$_1$
BS$_2$
BS$_k$
BS$_3$

Figure 1. Downlink interference in a cellular-UAV scenario.

## 2. System Model and Problem Definition

This section introduces the system model under consideration and analyzes the functionality of the IRS modules in minimizing the required transmit power.

### 2.1 System Model

We investigate a downlink cellular-connected UAV scenario, assuming all BSs have $N_r$ antennas and a UAV is connected directly with BS$_{\text{UAV}}$. This scenario includes $K$ additional co-channel BSs serving ground UEs, introducing potential interference for the UAV. In contrast to conventional terrestrial inter-cell interference (ICI), cellular-connected UAV results in an interference problem that includes aerial-ground interference; *i.e.*, interference resulting from co-channel terrestrial BSs and inter-UAV interference; *i.e.*, interference resulting among co-channel UAVs. An aided integrating UAV with IRS, UAV$_{\text{IRS}}$, is deployed to minimize the interference that is caused in the $k^{th}$ BS, BS$_k$, $\forall k \in \mathcal{S} = \{1,2,\ldots,K\}$. Here, we assume that only aerial-ground interference exists, as shown in Fig. 1. Let UAV$_{\text{IRS}}$ have fixed altitude and elevation angle and the IRS has $N$ meta-atoms.

The meta-atoms in UAV$_{\text{IRS}}$ aim to enhance the power efficiency of the communication link between the BS$_{\text{UAV}}$ and the UAV. We assume that the amplitude, $\alpha_i \in [0,1]$ and the phase, $\theta_i \in [0,2\pi)$, $\forall i \in \{1,2,\ldots,N\}$, can be independently adjusted for each meta-atom. While the phase and amplitude responses of the meta-atom are physically interconnected, there exist design approaches that effectively mitigate this inter-dependency [33]. Consequently, we denote the reflection diagonal matrix associated with the IRS as:

$$\mathbf{\Theta} = \mathrm{diag}\left[\alpha_1 e^{j\theta_1} \quad \alpha_2 e^{j\theta_2} \ldots. \quad \alpha_N e^{j\theta_N}\right], \in \mathbb{C}^{N\times N} \tag{1}$$

The BSs are positioned at the Cartesian coordinates $(x_{\text{BS},k}, y_{\text{BS},k}, h_{\text{BS},k})$, while the coordinates of the UAV-IRS are $(x_{\text{U}}, y_{\text{U}}, h_{\text{U}})$, which are assigned during the time interval. In this scheme, it is assumed that the UAV and the UAV$_{\text{IRS}}$ operate as a swarm system, with the UAV maintained at a fixed altitude,

$h_{\text{UAV}}$ and at a close distance with a defined flight period $T$ and maximum speed $V_{\max}$.

Given that the UAV operates at a sufficient altitude, we assume that the link between the UAV and the BS is only LoS, as described in [34]. Additionally, we assume that a single antenna is installed on the UAV. Let $\text{h}_{\text{BS}}$ and $\mathbf{h}_k \in \mathbb{C}^{1\times N_r}$ be the direct channel from $\text{BS}_{\text{UAV}}$ to UAV and the interfering channel from $\text{BS}_k$ to UAV, respectively, while the channel from IRS to the UAV is $\mathbf{h}_{\text{d}}^{\text{H}} \in \mathbb{C}^{1\times N}$. $\mathbf{G}_{\text{BS}}$ and $\mathbf{G}_k \in \mathbb{C}^{N\times N_r}$ represent the channels from $\text{BS}_{\text{UAV}}$ and $\text{BS}_k$ to $\text{UAV}_{\text{IRS}}$, respectively. Moreover, $\boldsymbol{p}_k$ and $\boldsymbol{p}_{\text{BS}} \in \mathbb{C}^{N_r\times 1}$ denote the transmit-power vector of $\text{BS}_k$ and $\text{BS}_{\text{UAV}}$, respectively. Therefore, the received signal at the UAV is represented as:

$$\begin{aligned} y_u \quad &= \left(\mathbf{h}_{\text{BS}} + \mathbf{h}_{\text{d}}^{\text{H}}\mathbf{\Theta}\mathbf{G}_{\text{BS}}\right)\boldsymbol{p}_{\text{BS}}x_{\text{BS}} \\ &+ \textstyle\sum_{k\in\mathcal{S}} \left(\mathbf{h}_k + \mathbf{h}_{\text{d}}^{\text{H}}\mathbf{\Theta}\mathbf{G}_k\right)\boldsymbol{p}_k x_k + n_u \end{aligned} \tag{2}$$

where the symbols transmitted by $\text{BS}_{\text{UAV}}$ and $\text{BS}_k$ are respectively denoted as $x_{\text{BS}}$, $x_k$ and $n_u \sim \mathcal{CN}(0, \sigma_u)$ is the additive white Gaussian noise (AWGN) with zero-mean and variance $\sigma_u^2$ at the UAV. Therefore, the SINR at the target UAV is given by:

$$\beta = \frac{|\boldsymbol{\ell}_{\text{BS}}\boldsymbol{p}_{\text{BS}}|^2}{\sum_{k\in\mathcal{S}} |\boldsymbol{\ell}_k\boldsymbol{p}_k|^2 + \sigma_u^2} \tag{3}$$

where $\boldsymbol{\ell}_{\text{BS}} = \mathbf{h}_{\text{BS}} + \mathbf{h}_{\text{d}}^{\text{H}}\mathbf{\Theta}\mathbf{G}_{\text{BS}}$ and $\boldsymbol{\ell}_k = \mathbf{h}_k + \mathbf{h}_{\text{d}}^{\text{H}}\mathbf{\Theta}\mathbf{G}_k$. Obviously, the interference from the co-channels is directly affected by the configuration of IRS meta-atoms. Therefore, choosing $\alpha_i$ and $\theta_i$ for each IRS element has a vital role in maximizing $\beta$. IRS elements can occasionally function as active components for channel estimation; however, this paper focuses only on passive elements and provides perfect Channel State Information (CSI) which is acquired between the UAV and each BS [13]. This assumption is reasonable due to the strong LoS links between the elevated UAV and ground terminals.

## 2.2 Problem Definition

With the overwhelming rise of data traffic, data rates, reduced latency and the need for highly reliable and secure communications, network energy consumption poses a significant economic challenge and a major hurdle for next-generation wireless networks. This emphasizes the need for energy-efficient solutions, paving the way for green cellular networks. The objective of this paper is to minimize the required transmit power from the $\text{BS}_{\text{UAV}}$ by simultaneously optimizing IRS coefficients, $\mathbf{\Theta}$. The optimization problem that minimizes the transmit power while ensuring the QoS can be set as follows:

$$\min_{\mathbf{\Theta}} \quad ||\boldsymbol{p}_{\text{BS}}||^2 \tag{4a}$$

$$s.t. \quad \beta \geq \gamma \tag{4b}$$

$$\alpha_i \in [0,1], \forall i \in \mathcal{N} \tag{4c}$$

$$\theta_i \in [0,2\pi), \forall i \in \mathcal{N} \tag{4d}$$

$$\sqrt{(x_{\text{U}}^{t+1} - x_{\text{U}}^t)^2 + (y_{\text{U}}^{t+1} - y_{\text{U}}^t)^2} \leq V_{\max} \tag{4e}$$

$$x_{\min} \leq x_{\text{U}}^t \leq x_{\max} \tag{4f}$$

$$y_{\min} \leq y_{\text{U}}^t \leq y_{\max} \tag{4g}$$

where $\{x_{\min}, x_{\max}\}$ and $\{y_{\min}, y_{\max}\}$ denote the boundary of the coordinate of the UAV (or $\text{UAV}_{\text{IRS}}$) $x_{\text{U}}$ and $x_{\text{U}}$, respectively. (4b) represents the SINR threshold, $\gamma$, that ensures the acceptable QoS of UAV. In (4e), it is ensured that the UAV does not exceed its maximum speed. The range of tasks that the UAV can perform is constrained to those specified in (4f) and (4g). We consider that the BS has access to perfect CSI for all links, as explained in [8], to fully assess the capabilities of the IRS. Clearly, the problem in (4) is non-convex due to the left-hand-side of (4b), which is not jointly concave with respect to $p_{\text{BS}}$ and $\mathbf{\Theta}$.

The challenge in solving (4) arises due to the interdependence between the design variables $\boldsymbol{p}_{\text{BS}}$ and $\mathbf{\Theta}$ in (4a)-(4b). To address this, Hua et al. [7] introduced a dual loop AO method based on penalty approaches. This algorithm updates the auxiliary variables by solving a quadratically constrained quadratic program. While the AO approach with a penalty technique achieves variable decoupling, it suffers from limited solution efficiency. Furthermore, the complexity per iteration of the penalty-based method was $\mathcal{O}(N^3)$. On the other hand, SDR with Gaussian randomization [8] can solve (4). However, the utilization of SDR becomes increasingly complex as the number of meta-atoms grows. Furthermore, obtaining a rank-1 feasible solution necessitates a considerable number of randomization steps, thereby substantially contributing to the overall complexity. Finally, since the coupling between $\boldsymbol{\Theta}$ and $\boldsymbol{p}_{\mathrm{BS}}$ is high, AO-based methods are typically inefficient, as they can not guarantee a theoretically stationary point.

## 3. Optimal Transmit Power Using SCA

In this section, we present a solution to non-convexity of (4b). We address the non-convexity in (4) by using an SCA to produce a high-performance solution. Our approach relies on a set of (in)equalities for arbitrary complex-valued vectors $a$ and $b$, as described in [35]:

$$||\boldsymbol{a}||^2 \geq 2\Re\{\boldsymbol{b}^{\mathrm{H}}\boldsymbol{a}\} - ||\boldsymbol{b}||^2, \tag{5a}$$

$$\Re\{\boldsymbol{a}^{\mathrm{H}}\boldsymbol{b}\} = \frac{1}{4}(||\boldsymbol{a}+\boldsymbol{b}||^2 - ||\boldsymbol{a}-\boldsymbol{b}||^2), \tag{5b}$$

$$\Im\{\boldsymbol{a}^{\mathrm{H}}\boldsymbol{b}\} = \frac{1}{4}(||\boldsymbol{a}-j\boldsymbol{b}||^2 - ||\boldsymbol{a}+j\boldsymbol{b}||^2) \tag{5c}$$

First, to handle (4b), we employ the method of SCA, where both $\boldsymbol{\Theta}$ and $p_{\mathrm{BS}}$ are optimized in each iteration. Initially, we use (3) to transform (4b) into an equivalent form given by:

$$\frac{|\boldsymbol{\ell}_{\mathrm{BS}}\boldsymbol{p}_{\mathrm{BS}}|^2}{\gamma} \geq \textstyle\sum_{k\in\mathcal{S}} |\boldsymbol{\ell}_k\boldsymbol{p}_k|^2 + \sigma_u^2 \tag{6}$$

We observe that the term in (4b) is non-convex. However, since we aim to maximize the right-hand side of (6), we can extract a concave lower bound for it using the following approach:

$$\begin{aligned}
|\boldsymbol{\ell}_{\mathrm{BS}}\boldsymbol{p}_{\mathrm{BS}}|^2 \quad & \overset{(a)}{\geq} \; 2\Re\{(\mathrm{u}^{(m)})^{\mathrm{H}}\boldsymbol{\ell}_{\mathrm{BS}}\boldsymbol{p}_{\mathrm{BS}}\} - |\mathrm{u}^{(m)}|^2 \\
& \overset{(b)}{\geq} \frac{1}{2}\{||\mathrm{u}^{(m)}\boldsymbol{\ell}\,^{\mathrm{H}}_{\mathrm{BS}} + \boldsymbol{p}_{\mathrm{BS}}||^2 \\
& \quad -||\mathrm{u}^{(m)}\boldsymbol{\ell}\,^{\mathrm{H}}_{\mathrm{BS}} - \boldsymbol{p}_{\mathrm{BS}}||^2\} - |\mathrm{u}^{(m)}|^2 \\
& \overset{(c)}{\geq} \; [\,\Re\{(\mathbf{v}^{(m)})^{\mathrm{H}}[\mathrm{u}^{(m)}\boldsymbol{\ell}\,^{\mathrm{H}}_{\mathrm{BS}} + \boldsymbol{p}_{\mathrm{BS}}]\,\} - \frac{1}{2}||\boldsymbol{v}^{(m)}||^2 \\
& \quad -\frac{1}{2}||\mathrm{u}^{(m)}\boldsymbol{\ell}\,^{\mathrm{H}}_{\mathrm{BS}} - \boldsymbol{p}_{\mathrm{BS}}||^2 - |\mathrm{u}^{(m)}|^2] \\
& \quad = f(\boldsymbol{p}_{\mathrm{BS}};\boldsymbol{\Theta};{\boldsymbol{p}_{\mathrm{BS}}}^{(m)};\boldsymbol{\Theta}^{(m)}),
\end{aligned} \tag{7}$$

where $\mathrm{u}^{(m)} = \boldsymbol{\ell}^{(m)}_{\mathrm{BS}}{\boldsymbol{p}_{\mathrm{BS}}}^{(m)}$, $\boldsymbol{v}^{(m)} = \mathrm{u}^{(m)}(\boldsymbol{\ell}^{(m)}_{\mathrm{BS}})^{\mathrm{H}} + {\boldsymbol{p}_{\mathrm{BS}}}^{(m)}$ and $\boldsymbol{\ell}\,^{(m)}_{\mathrm{BS}} = \mathbf{h}_{\mathrm{BS}} + \mathbf{h}_d^{\mathrm{H}}\boldsymbol{\Theta}^{(m)}\mathbf{h}_{\mathrm{BS}}$. Here, $\boldsymbol{\Theta}^{(m)}$ and $\boldsymbol{\ell}^{(m)}_{\mathrm{BS}}$ denote the values of $\boldsymbol{\Theta}$ and $\boldsymbol{\ell}\,_{\mathrm{BS}}$ in the $m^{th}$ iteration of the SCA model, respectively. Additionally, $(a)$ and $(c)$ in (7) result from (5a), while $(b)$ results from (5b). It can be proven that $f(\boldsymbol{p}_{\mathrm{BS}};\boldsymbol{\Theta};{\boldsymbol{p}_{\mathrm{BS}}}^{(m)};\boldsymbol{\Theta}^{(m)})$ is jointly concave w.r.t. $\boldsymbol{\Theta}$ and $\boldsymbol{p}_{\mathrm{BS}}$.

Next, to address (4b), we utilize the concept of SCA, where we optimize both $\boldsymbol{\Theta}$ and $\boldsymbol{p}_{\mathrm{BS}}$ in each iteration. Initially, we transform (4b) into an equivalent form as:

$$\frac{|\boldsymbol{\ell}_{\mathrm{BS}}\boldsymbol{p}_{\mathrm{BS}}|^2}{\gamma} \geq {\sigma_u}^2 + \textstyle\sum_{k\in\mathcal{S}} ({\varrho_k}^2 + {\bar{\varrho}_k}^2), \tag{8a}$$

$$\varrho_k \geq |\Re\{\boldsymbol{\ell}_k\boldsymbol{p}_k\}|, \forall k \in \mathcal{S}, \tag{8b}$$

$$\bar{\varrho}_k \geq |\Im\{\boldsymbol{\ell}_k\boldsymbol{p}_k\}|, \forall k \in \mathcal{S} \tag{8c}$$

where the new slack variables $\varrho_k$ and $\bar{\varrho}_k$ are used. It is simple to observe that if (2.2b) is feasible, then (8) is also feasible and *vice versa*. Since the right-hand side (RHS) of (8a) is convex, we need to find a concave lower bound for the term $|\boldsymbol{\ell}_{\mathrm{BS}}\boldsymbol{p}_{\mathrm{BS}}|^2$ in (8a). Let ${\boldsymbol{p}_{\mathrm{BS}}}^{(m)}$ and $\boldsymbol{\Theta}^{(m)}$ represent the value of $\boldsymbol{\Theta}$ and $\boldsymbol{p}_{\mathrm{BS}}$ in the $m^{th}$ iteration of the SCA process, respectively. Similarly to (7), this can be represented as:

$$\begin{aligned}
\frac{|\boldsymbol{\ell}_{\mathrm{BS}}\boldsymbol{p}_{\mathrm{BS}}|^2}{\gamma} \quad & \geq \frac{1}{\gamma}[\,\Re\{(\mathbf{v}^{(m)})^{\mathrm{H}}[\,\mathrm{u}^{(m)}\boldsymbol{\ell}^{\mathrm{H}}_{\mathrm{BS}} - \boldsymbol{p}_{\mathrm{BS}}]\,\} \\
-\frac{1}{2}||\boldsymbol{v}^{(m)}||^2 \quad & -\frac{1}{2}||\mathrm{u}^{(m)}\boldsymbol{\ell}^{\mathrm{H}}_{\mathrm{BS}} - \boldsymbol{p}_{\mathrm{BS}}||^2 - |\mathrm{u}^{(m)}|^2] \\
& = f(\boldsymbol{p}_{\mathrm{BS}};\boldsymbol{\Theta};{\boldsymbol{p}_{\mathrm{BS}}}^{(m)};\boldsymbol{\Theta}^{(m)}).
\end{aligned} \tag{9}$$

By utilizing the property that $x \geq |y|$ if and only if $x \geq y$ and $x \geq -y$ and incorporating (5b), we can express (8b) in an equivalent form as:

$$\varrho_k \geq \Re\{\boldsymbol{\ell}_k \boldsymbol{p}_k\} = \frac{1}{4}(||(\boldsymbol{\ell}_k + \boldsymbol{p}_k||^2 - ||(\boldsymbol{\ell}_k - \boldsymbol{p}_k||^2), \tag{10a}$$

$$\varrho_k \geq -\Re\{\boldsymbol{\ell}_k \boldsymbol{p}_k\} = \frac{1}{4}(||(\boldsymbol{\ell}_k - \boldsymbol{p}_k||^2 - ||(\boldsymbol{\ell}_k + \boldsymbol{p}_k||^2) \tag{10b}$$

Due to the presence of the negative quadratic term on the right-hand side of (10a), it becomes non-convex. Therefore, we can transform it into convex by applying the inequality in (5a) as:

$$\begin{aligned}\varrho_k \quad &\geq \frac{1}{4}[\left\|\boldsymbol{\ell}_k^{\mathrm{H}} + \boldsymbol{p}_k\right\|^2 - 2\Re\{(\boldsymbol{\ell}_k^{(m)} - (\boldsymbol{p}_k{}^{(m)})^{\mathrm{H}}(\boldsymbol{\ell}_k^{\mathrm{H}} - \boldsymbol{p}_k)\} \\ &+||(\boldsymbol{\ell}_k{}^{(m)})^{\mathrm{H}} - \boldsymbol{p}_k{}^{(m)}||^2] = \rho_k(\boldsymbol{p}_k; \boldsymbol{\Theta}; \boldsymbol{p}_k{}^{(m)}; \boldsymbol{\Theta}^{(m)})\end{aligned} \tag{11}$$

Applying a similar argument,(10b) results in:

$$\begin{aligned}\varrho_k \quad &\geq \frac{1}{4}[\left\|\boldsymbol{\ell}_k^{\mathrm{H}} + \boldsymbol{p}_k\right\|^2 - 2\Re\{(\boldsymbol{\ell}_k^{(m)} + \left(\boldsymbol{p}_k{}^{(m)}\right)^{\mathrm{H}}(\boldsymbol{\ell}_k^{\mathrm{H}} + \boldsymbol{p}_k)\} \\ &+||(\boldsymbol{\ell}_k{}^{(m)})^{\mathrm{H}} + \boldsymbol{p}_k{}^{(m)}||^2] = \bar{\rho}_k(\boldsymbol{p}_k; \boldsymbol{\Theta}; \boldsymbol{p}_{\boldsymbol{k}}{}^{(m)}; \boldsymbol{\Theta}^{(m)})\end{aligned} \tag{12}$$

Similarly, in (11) and (12), (8c) results in the following inequalities:

$$\begin{aligned}\bar{\varrho}_k \quad &\geq \frac{1}{4}[||\boldsymbol{\ell}_k^{\mathrm{H}} - j\boldsymbol{p}_{\boldsymbol{k}}||^2 - 2\Re\{(\boldsymbol{\ell}_k^{(m)} - j(\boldsymbol{p}_k{}^{(m)})^{\mathrm{H}}(\boldsymbol{\ell}_k^{\mathrm{H}} + j\boldsymbol{p}_{\boldsymbol{k}})\} \\ &+||(\boldsymbol{\ell}_k{}^{(m)})^{\mathrm{H}} + j\boldsymbol{p}_k{}^{(m)}||^2] = \omega_k(\boldsymbol{p}_k; \boldsymbol{\Theta}; \boldsymbol{p}_k{}^{(m)}; \boldsymbol{\Theta}^{(m)})\end{aligned} \tag{13}$$

$$\begin{aligned}\bar{\varrho}_k \quad &\geq \frac{1}{4}[||\boldsymbol{\ell}_k^{\mathrm{H}} + j\boldsymbol{p}_k||^2 - 2\Re\{(\boldsymbol{\ell}_k^{(m)} + j(\boldsymbol{p}_k{}^{(m)})^{\mathrm{H}}(\boldsymbol{\ell}_k^{\mathrm{H}} - j\boldsymbol{p}_k)\} \\ &+||(\boldsymbol{\ell}_k{}^{(m)})^{\mathrm{H}} - j\boldsymbol{p}_k{}^{(m)}||^2] = \overline{\omega}_k(\boldsymbol{p}_k; \boldsymbol{\Theta}; \boldsymbol{p}_k^{(m)}; \boldsymbol{\Theta}^{(m)})\end{aligned} \tag{14}$$

Subsequently, the only remaining issue is the non-convexity of (4c) and (4d). To tackle this issue, we begin by transforming the equality constraint in (4c) and (4d) into a convex inequality constraint. We introduce a regularization term into the objective function to guarantee the satisfaction of the inequality constraint as an equality at convergence. Additionally, To tackle the non-convexity of the resulting objective function, we employ a first-order approximation of the regularization term centered around $\boldsymbol{\Theta}^{(m)}$. As a result, we can restate the non-convex problem in (4) to approximate the convex sub-problem of the SCA model to an equivalent form as:

$$\min_{\boldsymbol{\Theta},\tau,\bar{\tau}} \quad ||p_{\mathrm{BS}}||^2 - \delta[2\Re\{(\boldsymbol{\Theta}^{(m)})\boldsymbol{\Theta}\} - \| \boldsymbol{\Theta}^{(m)} \|^2], \tag{15a}$$

$$s.t. \quad f(\boldsymbol{p}_{\mathrm{BS}}; \boldsymbol{\Theta}; \boldsymbol{p}_{\mathrm{BS}}{}^{(m)}; \boldsymbol{\Theta}^{(m)}) \geq \sigma_u{}^2 + \textstyle\sum_{k\in\mathcal{S}} (\varrho_k{}^2 + \bar{\varrho}_k{}^2), \quad \forall k \in \mathcal{S} \tag{15b}$$

$$(11) - (14), \quad \forall k \in \mathcal{S} \tag{15c}$$

$$\alpha_i \in [0,1], \quad \forall i \in \mathcal{N} \tag{15d}$$

$$\theta_i \in [0,2\pi), \quad \forall i \in \mathcal{N} \tag{15e}$$

where the regularization parameter is defined as $\delta > 0$, $m$ is the iteration number, $\tau = \{\varrho_k\}_{k\in\mathcal{S}}$ and $\bar{\tau} = \{\bar{\varrho}_k\}_{k\in\mathcal{S}}$. Note that the non-convexity of the regularization parameter $\delta$ in (15a) leads to the non-convexity of (15). However, to address this issue, we utilize (5a) to convexify (15a). It is evident that all the constraints specified in (15) can be expressed using quadratic cones. Consequently, (15) qualifies as an SOCP problem that can be efficiently solved using the MOSEK solver [36]. When the objective function gradually falls, it eventually reaches a specific threshold $\epsilon$ and the optimal $\boldsymbol{\Theta}^{\star}$ is achieved based on the eigenvalues' decomposition.

## 4. Algorithm Analysis

The proposed algorithm is summarized in Algorithm 1, where the optimization of $\boldsymbol{\Theta}$ aims to minimize the BS transmit power, which is achieved by tuning $\epsilon$ using the MOSEK solver. In this section, the feasibility of interference elimination is discussed for LoS channels, then convergence behavior and computational complexity are presented.

### 4.1 Feasibility of Interference Elimination in LoS Channels

We investigate a special case of pure LoS channels between BS-UAV, UAV$_{\mathrm{IRS}}$-UAV and BS-UAV$_{\mathrm{IRS}}$ to understand the impact of the number of reflecting elements on interference-cancellation feasibility.

This scenario is significant in practical applications owing to the elevated altitude of the UAV and UAV$_{\mathrm{IRS}}$. Consequently, we can determine the channel gains for the direct and cascaded paths related to B$S_k$.

**Algorithm 1:** SCA-based algorithm for solving (15)

```
Input: Maximum iteration number $M$, the initial value ($\boldsymbol{p}_{\mathrm{BS}}^{(0)}$, $\mathbf{\Theta}^{(0)}$, $\delta > 0$)
Output: The optimal value {$\boldsymbol{p}_{\mathrm{BS}}^{*}$, $\mathbf{\Theta}^{*}$}
Initialize: $m = 0$;
repeat
    The optimal solution of (15) based-MOSEK tool is $\boldsymbol{p}_{\mathrm{BS}}^{(0)}, \mathbf{\Theta}^{(0)}$.
    Update $\boldsymbol{p}_{\mathrm{BS}}^{(m+1)} \leftarrow \boldsymbol{p}_{\mathrm{BS}}^{*}$,        $\mathbf{\Theta}^{(m+1)} \leftarrow \mathbf{\Theta}^{\star}$
    $m \leftarrow m+1$
until The objective value of (15) converges;
```

$$||\mathbf{h}_k||_1 = \frac{\sqrt{\beta_0}}{d_{\mathrm{BU},k}}, k \in \mathcal{S} \tag{16a}$$

$$\left\|\mathbf{h}_{\mathrm{d}}^{\mathrm{H}}\mathbf{\Theta}\mathbf{G}_{\boldsymbol{k}}\right\|_1 = \frac{N\beta_0}{d_{\mathrm{BI},k}d_{\mathrm{IU}}}, k \in \mathcal{S} \tag{16b}$$

Here, $\beta_0$ represents the LoS path loss at a reference distance of 1 meter. $d_{\mathrm{BU},k}$, $d_{\mathrm{IU}}$ and $d_{\mathrm{BI},k}$ denote the distances between the $\mathrm{BS}_k$-UAV, $\mathrm{UAV}_{\mathrm{IRS}}$-UAV and $\mathrm{BS}_k$-$\mathrm{UAV}_{\mathrm{IRS}}$, respectively. Substituting (16a) and (16b) into the complete interference-nulling condition, $||\mathbf{h}_{\mathrm{d}}^{\mathrm{H}}\mathbf{\Theta}\mathbf{G}_k||_1 \geq ||\mathbf{h}_k||_1$, we obtain:

$$\frac{N\beta_0}{d_{\mathrm{BI},k}d_{\mathrm{IU}}} \geq \frac{\sqrt{\beta_0}}{d_{\mathrm{BU},k}}, k \in \mathcal{S} \tag{17}$$

In this scenario, the elevated altitude of both the UAV and $\mathrm{UAV}_{\mathrm{IRS}}$ and their close proximity result in approximately equal distances, indicated by $d_{\mathrm{BU},k} \approx d_{\mathrm{BI},k}$. Consequently, we can simplify the expression for the minimum number of reflecting elements required to achieve interference nulling at the UAV as $N_{min} = \left\lceil \frac{d_{\mathrm{IU}}}{\sqrt{\beta_0}} \right\rceil$. This suggests that a larger number of reflecting elements facilitates the achievement of interference nulling, as demonstrated through simulations in Section 5.

## 4.2 Convergence Behavior of Algorithm

Without loss of generality and for a given $\delta$, let $f(\boldsymbol{p}_{\mathrm{BS}}{}^{(m)}; \mathbf{\Theta}^{(m)})$ be the total transmit power and passive beamforming IRS coefficients' matrix at the $m^{th}$ iteration. Thus, we observe:

$$f(\boldsymbol{p}_{\mathrm{BS}}{}^{(m)}; \mathbf{\Theta}^{(m)}) \geq f(\boldsymbol{p}_{\mathrm{BS}}{}^{(m+1)}; \mathbf{\Theta}^{(m)}) \tag{18}$$

where (18) holds since (15) represents a convex function and the transmit power $\boldsymbol{p}_{\mathrm{BS}}{}^{(m+1)}$ is optimized and updated at $(m+1)^{\mathrm{th}}$ iteration based on the passive beamforming $\mathbf{\Theta}^{(m)}$. Interestingly though, $\boldsymbol{p}_{\mathrm{BS}}{}^{(m+1)}$ remains unchanged during the $(m+1)^{\mathrm{th}}$ iteration, even with the updated passive beamforming $\mathbf{\Theta}^{(m+1)}$. Therefore, (18) holds. Considering the constraint (15c), the objective sequence is $f(\boldsymbol{p}_{\mathrm{BS}}; \mathbf{\Theta}) \geq -\delta M$, indicating that the objective function $f(\boldsymbol{p}_{\mathrm{BS}}{}^{(m)}; \mathbf{\Theta}^{(m)})$ converges. Also, taking into account the limited system resources and the minimum SINR, we assume that the objective function in (15) must have a lower bound, which is a finite value. Consequently, we can prove the convergence of Algorithm 1 towards a feasible solution.

## 4.3 Computational-complexity Analysis

Let's suppose that we have only one UAV. It can be easily demonstrated that there are $2(K+N+1)+1$ optimization variables in total in (15), considering that they are real-valued. There are $N+2$ total second-order conic restrictions. As a result, using the justifications in [37], the suggested SOCP-based method's overall per-iteration complexity is given by:

$$\mathcal{O}[2(4+N)^{0.5}(1+K+N)(4+16K+8N+20K^2+8KL+4N^2)] \tag{19}$$

In realistic scenarios, the IRS meta-atoms are anticipated to be significantly greater than the number of interfering BSs. Consequently, the complexity of Algorithm 1-based SCA can be accurately approximated as $\mathcal{O}(N^{3.5})$. Although the use of AO in [7] simplifies the optimization problem, the complex interdependence among the design variables may prevent it from producing a solution of high quality. Furthermore, as demonstrated in Section 5, employing a penalty-based algorithm necessitates a substantial number of iterations to meet convergence, resulting in prolonged problem-solving time. On the other hand, the complexity of the SDR-based algorithm discussed in [8] is dominated by solving the semi-definite program (SDP) in (8), resulting in a per-iteration complexity of approximately $\mathcal{O}(N^7)$ [37]. This is significantly higher than the complexity of our proposed algorithm, especially for large numbers of IRS meta-atoms. As we will demonstrate numerically in the next section, our algorithm offers significant computational advantages.

## 5. Numerical Results

This section documents numerical results that demonstrate the performance of Algorithm 1. For the simulation, we examine a system that incorporates an IRS to assist a UAV, where the $\mathrm{BS}_{\mathrm{UAV}}$ and $\mathrm{UAV}_{\mathrm{IRS}}$ are positioned at (0, 0, 10 m) and $(x_{\mathrm{U}}, y_{\mathrm{U}}, 50\ m)$, respectively. The remaining co-channel BSs are situated at 10 m height, the locations are randomly and uniformly generated within their respective cells, which are denoted by coordinates $(x_{\mathrm{BS},k}, y_{\mathrm{BS},k}, 10\ m)$. The $\mathrm{UAV}_{\mathrm{IRS}}$ operates at a constant altitude $h_{\mathrm{U}}$, with a flight duration $T$ set to 80 seconds and the maximum speed of the UAV is $V_{\max} = 25\ m/s$. With a total bandwidth of 10 MHz, we suppose that the system operates at its center frequency of 2.4 GHz. Additionally, we consider parameters $\delta = 0.001$ and $\sigma^2 = -174$ dBm. The UAV-$\mathrm{UAV}_{\mathrm{IRS}}$ channel is assumed to exhibit Rayleigh fading, whereas the BS-UAV channels are modeled using Rician fading with a Rician factor of K=5 and distance-dependent path-loss for the communication links, as described in [38].

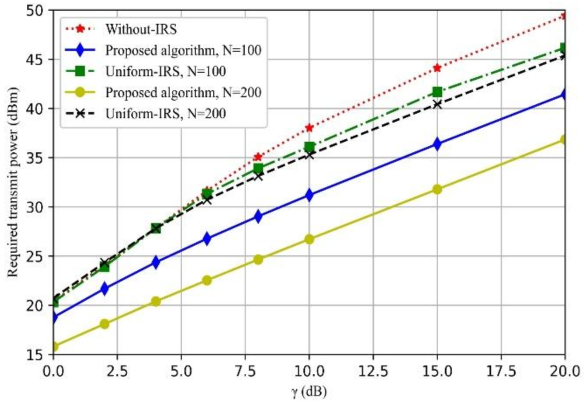


Figure 2. Total transmit power for the proposed algorithm at $K = 6$, $N_r = 4$.

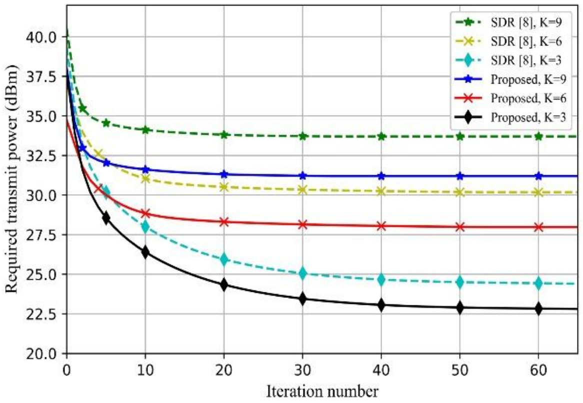


Figure 3. Performance of proposed algorithm at different K values and N=200.

Fig. 2 quantifies the relationship between the transmit power and threshold SINR with $K = 6$, where the effect of Algorithm 1 is shown with varying numbers of IRS meta-atoms. In this figure, we calculate the required transmit power without IRS (*i.e.*, $\mathbf{\Theta} = 0$) and compare the result with uniform IRS (*i.e.*, $\mathbf{\Theta} = 1$) at different values of $N$. We observe that the uniform IRS demonstrates limited effectiveness when the threshold SINR falls below 5 dBm, while uniform IRS enhances the received signal for higher $\gamma$. Its impact is not substantial enough to overcome the interference signals introduced by the BS when the SINR threshold is below 5 dBm. Consequently, the required transmit power is decreased by increasing the number of IRS meta-atoms. Moreover, Fig. 2 shows the required transmit power for Algorithm 1 and compares it with the uniform IRS at $N = 100$ and $N = 200$. We notice that the required transmit power for Algorithm 1 is less than the required power for uniform IRS. For example, when $\gamma = 10$ dB, the required transmit power for Algorithm 1 at $N = 100$ is approximately 5 dBm less than that of the uniform IRS. Similarly, when $N = 200$, the reduction is approximately 13 dBm. These results emphasize the advantage of the IRS-aided communication system, enabling the UAV to achieve reduced required transmit power while maintaining the desired QoS and SINR level. Additionally, it shows the savings in power when utilizing the proposed algorithm.

Figure 3 illustrates the performance of Algorithm 1 compared to the SDR-based approach [8] using a predefined threshold of $\gamma = 10$ and $N_r = 4$. Through comparisons involving varying numbers of interference base stations while maintaining the same number of IRS meta-atoms (*i.e.*, $K = 3$, $K = 6$ and $K = 9$ with $N = 200$), it is observed that the necessary transmit power increases with a rise in the number of interference base stations. The figure distinctly shows that Algorithm 1 surpasses the SDR-based algorithm in terms of the required transmit power. Additionally, this effect is attributed to the destructive beam effects towards the primary network, which satisfies constraint (4b) and consequently reduces interference, resulting in an improved required transmit power.

The convergence analysis of Algorithm 1 is illustrated in Figure 4. We use different initial values for $\boldsymbol{\Theta}_i$, uniform ($\boldsymbol{\Theta} = 1$), randomly distributed ($\boldsymbol{\Theta}$ with randomized $\theta$ and $\alpha$), or $\boldsymbol{\Theta} = 0$. The convergence is shown for $K = 6$, $N = 100$, $N_r = 4$ and two values for $\gamma$. We observe the rapid convergence of the proposed algorithm reaching its convergence point within the $8^{th}$ iteration for any initial point. Also, we show that the required transmit power is increased by increasing the threshold value of $\gamma$. In other words, increasing the QoS for the UAV requires more base station's transmit power.

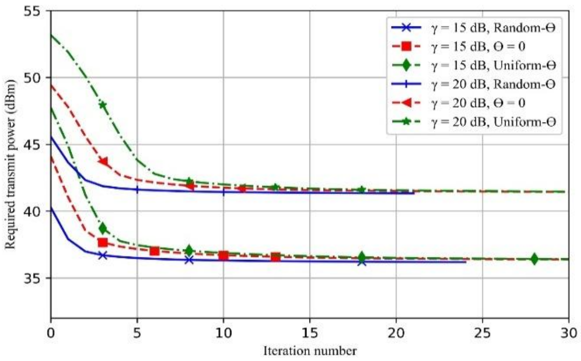


Figure 4. Convergence behavior of Algorithm 1 (N=100).

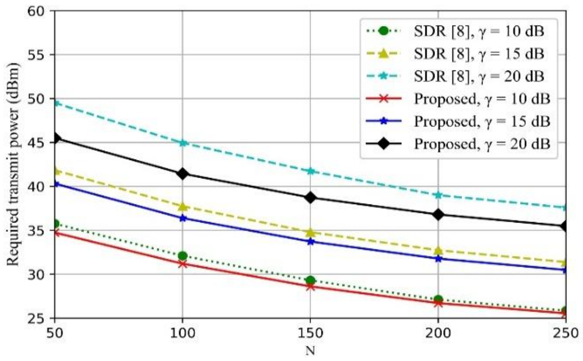


Figure 5. Total required transmit power *versus* SCA-IRS meta-atoms for K=6 and Nr=4 under different schemes.

Figure 5 illustrates the relationship between the required transmit power and the number of IRS meta-atoms for various threshold SINR values, considering the calculated optimal IRS coefficient $\boldsymbol{\Theta}$. Additionally, the performance of the SDR-based method [8] is presented. As anticipated, a decreasing trend in the necessary transmit power is observed with an increasing number of IRS meta-atoms. This trend can be attributed to the passive nature of the IRS, where a higher count of phase shifters allows for more power to be reflected from the BS. Consequently, this amplifies the SINR, resulting in a reduced required transmit power. Both the transmit power of $\mathrm{BS_{UAV}}$ and that of each co-channel $\mathrm{BS}_k$ are set to be equal, denoted as $\boldsymbol{p}_{\mathrm{BS}} = \boldsymbol{p}_k$ [13]. It is important to note that the required transmit power of the UAV to achieve a specific SINR and the received interference power exhibit a steady increase or decrease with $N$. However, as $N$ increases, the UAV transmit power decreases due to substantial interference suppression by the IRS, leading to a more prominent desired signal power. Furthermore, the figure indicates that the SDR-based methods exhibit similar performance for small values of $\gamma$. Conversely, for larger $\gamma$, Algorithm 1 outperforms the SDR-based methods.

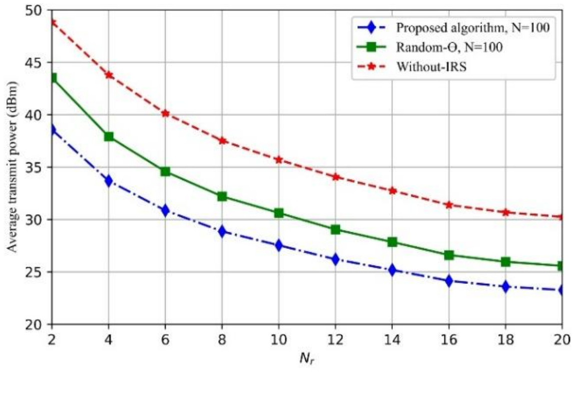


Figure 6. Transmit power of the BS *versus* number of antennas at $\mathrm{BS_{UAV}}$.

Finally, Figure 6 illustrates the anticipated decrease in average transmit power required for information transmission as the number of transmit antennas at the $BS_{\text{UAV}}$ increases for $\gamma = 10$ and $K = 6$. This effect is attributed to the availability of more DoF for spatial multiplexing with a larger antenna array, allowing for lower transmit at the BS. Furthermore, it can be observed that employing phase shifts at the IRS yields significant performance gains compared to random phase or the absence of IRS. These findings highlight the effectiveness of multiple-antenna techniques in reducing transmit power.

## 6. CONCLUSIONS

In this paper, we have presented a framework that leverages the SCA algorithm and SOCP relaxation to optimize both the phase-shift and amplitude coefficients of the IRS matrix. This approach effectively addresses the challenge of minimizing transmit power for UAV-downlink communication while maintaining QoS. Notably, the algorithm updates all optimization variables simultaneously in each iteration, ensuring efficient convergence. Furthermore, we investigate the impact of exploiting both the direct channel and the reflected signal from the IRS, enabling constructive interference and enhanced reception at the target UAV. The combined use of SOCP and SCA facilitates rapid convergence to an effective solution. Simulations demonstrate that our IRS-aided UAV communication system achieves significant power savings of approximately 13 dBm compared to conventional approaches while maintaining the desired SINR level across various interference scenarios. In our future work, we aim to address the challenge of analyzing how altitude and speed impact interference mitigation. Additionally, we will explore the challenges of dynamic channel conditions and imperfect CSI.

## REFERENCES

[1] Y. Zeng, Q. Wu and R. Zhang, "Accessing from the Sky: A Tutorial on UAV Communications for 5G and beyond," Proceedings of the IEEE, vol. 107, no. 12, pp. 2327–2375, 2019.

[2] A.-A. A. Boulogeorgos and A. Alexiou, "Coverage Analysis of Reconfigurable Intelligent Surface Assisted THz Wireless Systems," IEEE Open J. of Vehicular Technology, vol. 2, pp. 94–110, 2021.

[3] A. Ihsan, W. Chen, M. Asif, W. U. Khan, Q. Wu and J. Li, "Energy-efficient IRS-aided NOMA Beamforming for 6G Wireless Communications," IEEE Transactions on Green Communications and Networking, vol. 6, no. 4, pp. 1945–1956, 2022.

[4] Q. Wu, S. Zhang, B. Zheng, C. You and R. Zhang, "Intelligent Reflecting Surface-aided Wireless Communications: A Tutorial," IEEE Trans. on Communications, vol. 69, no. 5, pp. 3313–3351, 2021.

[5] T. Song, D. Lopez, M. Meo, N. Piovesan and D. Renga, "High Altitude Platform Stations: The New Network Energy Efficiency Enabler in the 6G Era," Proc. of the 2024 IEEE Wireless Communications and Networking Conference (WCNC), pp. 1–6, Dubai, UAE, 2024.

[6] S. Khisa, M. Elhattab, C. Assi and S. Sharafeddine, "Energy Consumption Optimization in RIS-assisted Cooperative RSMA Cellular Networks," IEEE Trans. on Communications, vol. 71, no. 7, pp. 4300–4312, 2023.

[7] M. Hua, Q. Wu, W. Chen, O. A. Dobre and A. Lee Swindlehurst, "Secure Intelligent Reflecting Surface Aided Integrated Sensing and Communication," IEEE Trans. on Wireless Communications, vol. 23, no. 1, pp. 575–591, 2023.

[8] Q. Wu and R. Zhang, "Intelligent Reflecting Surface Enhanced Wireless Network *via* Joint Active and Passive Beamforming," IEEE Trans. on Wireless Communi., vol. 18, no. 11, pp. 5394–5409, 2019.

[9] M. A. ElMossallamy, K. G. Seddik, W. Chen, L. Wang, G. Y. Li and Z. Han, "RIS Optimization on the Complex Circle Manifold for Interference Mitigation in Interference Channels," IEEE Transactions on Vehicular Technology, vol. 70, no. 6, pp. 6184–6189, 2021.

[10] M. Cui, G. Zhang and R. Zhang, "Secure Wireless Communication via Intelligent Reflecting Surface," IEEE Wireless Communications Letters, vol. 8, no. 5, pp. 1410–1414, 2019.

[11] H. Niu, Z. Chu, F. Zhou and Z. Zhu, "Simultaneous Transmission and Reflection Reconfigurable Intelligent Surface-assisted Secrecy MISO Networks," IEEE Communications Letters, vol. 25, no. 11, pp. 3498– 3502, 2021.

[12] Y. Liu, J. Yang, K. Huang, X. Sun and Y. Wang, "Secure Wireless Communications in the Multi-user MISO Interference Channel Assisted by Multiple Reconfigurable Intelligent Surfaces," Journal of Communications and Networks, vol. 24, no. 5, pp. 530–540, 2022.

[13] X. Pang, W. Mei, N. Zhao and R. Zhang, "Intelligent Reflecting Surface-assisted Interference Mitigation for Cellular-connected UAV," IEEE Wireless Communications Letters, vol. 11, no. 8, pp. 1708–1712, 2022.

[14] X. Wu, J. Ma, Z. Xing, C. Gu, X. Xue and X. Zeng, "Secure and Energy Efficient Transmission for IRS-assisted Cognitive Radio Networks," IEEE Trans. on Cognitive Communications and Networking, vol. 8, no. 1, pp. 170–185, 2022.

[15] M. Kim and D. Park, "Intelligent Reflecting Surface-aided MIMO Secrecy Rate Maximization," ICT Express, vol. 8, no. 4, pp. 518–524, 2022.

[16] W. Jiang, Y. Zhang, J. Zhao, Z. Xiong and Z. Ding, "Joint Transmit Precoding and Reflect Beamforming Design for IRS-assisted MIMO Cognitive Radio Systems," IEEE Transactions on Wireless Communications, vol. 21, no. 6, pp. 3617–3631, 2022.

[17] V. Kumar, M. Chafii, A. L. Swindlehurst, L.-N. Tran and M. F. Flanagan, "SCA-based Beamforming Optimization for IRS-enabled Secure Integrated Sensing and Communication," arXiv preprint, arXiv: 2305.03831, 2023.

[18] P. Liu, G. Jing, H. Liu, L. Yang and T. A. Tsiftsis, "Intelligent Reflecting Surface-assisted Cognitive Radio-inspired Rate-splitting Multiple Access Systems," Digital Communications and Networks, vol. 9, no. 3, pp. 655–666, 2023.

[19] J. Ye, S. Guo and M.-S. Alouini, "Joint Reflecting and Precoding Designs for SER Minimization in Reconfigurable Intelligent Surfaces-assisted MIMO Systems," IEEE Trans. Wireless Commun., vol. 19, no. 8, pp. 5561–5574, 2020.

[20] Z. Albataineh, K. F. Hayajneh, H. Shakhatreh, R. A. Athamneh and M. Anan, "Channel Estimation for Reconfigurable Intelligent Surface-assisted mmWave based on Re'nyi Entropy Function," Scientific Reports, vol. 12, no. 1, p. 22301, 2022.

[21] Q. Wu and R. Zhang, "Beamforming Optimization for Wireless Network Aided by Intelligent Reflecting Surface with Discrete Phase Shifts," IEEE Trans. on Communications, vol. 68, no. 3, pp. 1838–1851, 2020.

[22] H. Wang, Z. Shi, Y. Fu and R. Song, "Downlink Multi-IRS Aided NOMA System with Second-order Reflection," IEEE Wireless Communications Letters, vol. 12, no. 6, pp. 1022–1026, 2023.

[23] J. Wang and H. Yu, "Rf Energy Harvesting Schemes for Intelligent Reflecting Surface-aided Cognitive Radio Sensor Networks," Scientific Reports, vol. 12, no. 1, p. 22462, 2022.

[24] Q.-U.-A. Nadeem, A. Kammoun, A. Chaaban, M. Debbah and M.-S. Alouini, "Asymptotic Max-min SINR Analysis of Reconfigurable Intelligent Surface Assisted MISO Systems," IEEE Trans. on Wireless Communications, vol. 19, no. 12, pp. 7748–7764, 2020.

[25] W. Feng, J. Tang, Q. Wu, Y. Fu, X. Zhang, D. K. C. So and K.-K. Wong, "Resource Allocation for Power Minimization in RIS-assisted Multi-UAV Networks with NOMA," IEEE Trans. on Communications, vol. 71, no. 11, pp. 6662–6676, 2023.

[26] H. Wang, C. Liu, Z. Shi, Y. Fu and R. Song, "On Power Minimization for IRS-aided Downlink NOMA Systems," IEEE Wireless on Communications Letters, vol. 9, no. 11, pp. 1808–1811, 2020.

[27] Y. Cai, Z. Wei, S. Hu, D. W. K. Ng and J. Yuan, "Resource Allocation for Power-efficient IRS-assisted UAV Communications," Proc. of the 2020 IEEE Int. Conf. on Communications Workshops (ICC Workshops), pp. 1–7, Dublin, Ireland, 2020.

[28] H. Huang, Y. Zhang, H. Zhang, Z. Zhao, C. Zhang and Z. Han, "Multi-IRS-aided Millimeter-wave Multi-user MISO Systems for Power Minimization Using Generalized Benders Decomposition," IEEE Trans. on Wireless Communications, vol. 22, no. 11, pp. 7873–7886, 2023.

[29] B. Zheng and R. Zhang, "Intelligent Reflecting Surface-enhanced OFDM: Channel Estimation and Reflection Optimization," IEEE Wireless on Communications Letters, vol. 9, no. 4, pp. 518–522, 2020.

[30] Z. Ji, W. Yang, X. Guan, X. Zhao, G. Li and Q. Wu, "Trajectory and Transmit Power Optimization for IRS-assisted UAV Communication under Malicious Jamming," IEEE Trans. on Vehicular Technology, vol. 71, no. 10, pp. 11262–11266, 2022.

[31] C. You, Z. Kang, Y. Zeng and R. Zhang, "Enabling Smart Reflection in Integrated Air-Ground Wireless Network: IRS Meets UAV," IEEE Wireless Communications, vol. 28, no. 6, pp. 138–144, 2021.

[32] A. Khalili, E. M. Monfared, S. Zargari, M. R. Javan, N. M. Yamchi and E. A. Jorswieck, "Resource Management for Transmit Power Minimization in UAV-assisted RIS Hetnets Supported by Dual Connectivity," IEEE Trans. on Wireless Communications, vol. 21, no. 3, pp. 1806–1822, 2022.

[33] K. Ntontin et al., "Autonomous Reconfigurable Intelligent Surfaces through Wireless Energy

Harvesting," 2022 IEEE 95th Vehicular Technology Conference: (VTC2022-Spring), pp. 1–6, Helsinki, Finland, 2022.

[34] X. Shao and R. Zhang, "Target-mounted Intelligent Reflecting Surface for Secure Wireless Sensing," IEEE Trans. on Wireless Communications, vol. 23, no. 8, pp. 9745–9758, 2024.

[35] V. Kumar, R. Zhang, M. D. Renzo and L.-N. Tran, "A Novel SCA-based Method for Beamforming Optimization in IRS/RIS-assisted MU-MISO Downlink," IEEE Wireless Communications Letters, vol. 12, no. 2, pp. 297–301, 2023.

[36] M. Aps, "Mosek Fusion API for Python 10.0.46," [Online], Available: https://docs.mosek.com/10.0/pythonfusion.pdf.

[37] A. Ben-Tal and A. Nemirovski, Lectures on Modern Convex Optimization: Analysis, Algorithms and Engineering Applications, SIAM Publications Library, 2001.

[38] N. S. Perović, L.-N. Tran, M. Di Renzo and M. F. Flanagan, "On the Maximum Achievable Sum-rate of the RIS-aided MIMO Broadcast Channel," IEEE Trans. on Signal Processing, vol. 70, pp. 6316–6331, 2022.

## ملخص البحث:

إنّ التّداخل بين الجوّ والأرض يشكّل تحدّياً في تحقيق التّكامل بين المركبات الجوّية غير المأهولة وشبكات الاتّصالات. ويمكن للمركبات الجوّية غير المأهولة أن تقاوم قدْراً كبيراً من التّداخل من المحطّات الأرضية إذا تمّ ضمان خطّ نظرٍ مباشر يربط بينها وبين المستخدمين على الأرض. لذا تهدف هذه الدّراسة إلى التّركيز على تحسين القُدرة في المحطّات الأرضية وتطبيق مبادىء الطّاقة الخضراء لتمهيد الطّريق نحو محطّات أرضية أكثر استدامةً وفعالية من حيث الطّاقة في أنظمة الاتّصال اللاسلكية المتكاملة مع المركبات الجوّية المأهولة. كما تبحث هذه الدّراسة في تحسين قُدرة الإرسال في رابط التنزيل في أنظمة السّطح العاكس الذّكي اللاسلكية المتكاملة مع الطّائرات بدون طيّار؛ إذ يتمّ تركيب السّطح العاكس الذّكي على الطّائرة بدون طيّار من أجل إزالة التّداخل الصّادر من المحطّات الأرضية التّابعة للشّبكة ذاتها.

وتبين نتائج المحاكاة أنّ النّظام المقترح في هذه الدّراسة هو نظام ذو فاعلية كبيرة من حيث تحسين قدرة الأرسال، إلى جانب قدْرٍ منخفض من تعقيد الحسابات، وأداءٍ قريب من المثالية يتفوّق إلى حدٍّ كبيرٍ على الأنظمة التّقليدية الّتي تخْلو من أنظمة السّطح العاكس الذّكي. وتُحقّق الخوارزمية المستخدمة خفْضاً في القدرة يتراوح بين 8 و 13 ديسيبل (dBm)، مع الحفاظ على نسبة الإشارة إلى التّداخل والضّجيج المطلوبة.

Jordanian Journal of Computers and Information Technology (JJCIT), Vol. 11, No. 02, June 2025.

# SECURING WIRELESS COMMUNICATIONS WITH ENERGY HARVESTING AND MULTI-ANTENNA DIVERSITY

Nguyen Quang Sang[1], Tran Cong Hung[2], Tran Trung Duy[1], Minh Tran[3]
and Byung Seo Kim[4]

*(Received: 29-Nov.-2024, Revised: 19-Jan.-2025, Accepted: 10-Feb.-2025)*

## *ABSTRACT*

*This paper presents a secure wireless communication system that integrates Physical Layer Security (PLS) with Energy Harvesting (EH) to enhance both data confidentiality and network sustainability. The proposed system uniquely employs Maximal Ratio Combining (MRC) and Selection Combining (SC) techniques at the multi-antenna destination node D, which is a novel approach in EH-driven PLS systems. The system model features a source node S, powered by energy harvested from spatially distributed power stations, a multi-antenna destination node D and an eavesdropper node E within the communication range. A time-switching protocol allows the source node S to alternate between energy harvesting and secure data transmission. To improve signal quality and security, the destination node D employs Maximal Ratio Combining (MRC) and Selection Combining (SC) techniques to mitigate fading and eavesdropping risks. Analytical expressions for the Signal-to-Noise Ratios (SNRs) at the destination and eavesdropper are derived, along with the Probability Density Function (PDF) and Cumulative Distribution Function (CDF) of these SNRs under block Rayleigh fading. We also provide an exact formulation for Secrecy Outage Probability (SOP), quantifying the likelihood of information leakage under different system configurations. The model is validated through Monte Carlo simulations, confirming the accuracy of the theoretical analysis. Simulation results highlight the impact of key parameters—energy harvesting efficiency $\eta$, time- switching parameter $\alpha$, number of antennas $M$ , number of beacon nodes $N$ and the power of beacon nodes—on Secrecy Outage Probability (SOP), offering valuable insights for optimizing secure and energy- efficient communication in wireless networks. An asymptotic analysis is also provided to characterize system performance at high SNR.*

## *KEYWORDS*

*Physical layer security, Energy harvesting, Selection combining, Maximal ratio combining.*

## 1. INTRODUCTION

In wireless communications, ensuring data security is a critical concern due to the inherent vulnerability of wireless channels to eavesdropping and interference [1]. Traditional security measures often rely on cryptographic techniques at higher layers; however, these can be resource-intensive and may not be fully effective in dynamic or low-power environments. Physical Layer Security (PLS), first conceptualized in the mid-20th century and developed further in the 2000s, leverages the physical properties of the wireless channel to protect data from interception. PLS focuses on optimizing the signal-to-noise ratio and channel conditions in favor of legitimate users while limiting the information available to potential eavesdroppers. Presently, research in PLS involves integrating advanced techniques, such as beamforming, artificial noise generation and cooperative relay strategies, to enhance security while minimizing energy costs [2].

In cooperative communication systems, Decode-and-Forward (DF) and Amplify-and-Forward (AF) are two widely adopted relaying protocols that enhance the robustness and coverage of wireless networks. Studies have shown that improving the capacity in various relay models, such as the half-duplex relay channel, can further optimize these protocols by addressing specific phase-transmission

1. N. Q. Sang and T. T. Duy are with Posts and Telecommunications Institute of Technology, Ho Chi Minh City, Vietnam. Emails: sangnq@ptit.edu.vn and duytt@ptit.edu.vn
2. T. C. Hung is with the School of Computer Science & Engineering, The SaiGon International University, Ho Chi Minh City, Vietnam. Email: tranconghung@siu.edu.vn
3. M. Tran (Corresponding Author) is with the Advanced Intelligent Technology Research Group, Faculty of Electrical and Electronics Engineering, Ton Duc Thang University, Ho Chi Minh City, Vietnam. Email: tranhoangquangminh@tdtu.edu.vn
4. B.-S. Kim is with the Department of Software and Communications Engineering, Hongik University, Sejong, South Korea. Email: jsnbs@hongik.ac.kr

challenges [3]. In DF, the relay node decodes the received signal, processes it and then retransmits it, effectively mitigating noise but adding processing delay. This approach is particularly advantageous in scenarios where data integrity is critical [4]-[7]. On the other hand, AF relays amplify the received signal, including any noise, before forwarding it, resulting in a simpler implementation, but potentially amplifying noise as well [8]-[9]. The choice between DF and AF often depends on specific network requirements, such as the desired balance between complexity, latency and reliability [10]. Recent studies in secure cooperative communications have demonstrated the impact of DF and AF on system security and efficiency, especially in energy-constrained and eavesdropping-prone environments [11]-[15]. For example, various secure cooperative transmission protocols have been developed for two-way energy-constrained relaying networks, which improve secrecy outage and throughput performance even in the presence of multiple eavesdroppers through strategic relay and jammer selection. Notably, protocols for secure two- way communication in energy-constrained relaying networks demonstrate improved secrecy outage and throughput performance by implementing cooperative relay strategies, including relay and jammer selection to mitigate eavesdropping [16]. Combining binary jamming at relay nodes with network coding at source nodes has demonstrated improvements in outage performance by limiting eavesdroppers' ability to decode the transmitted messages in two-way relaying networks [17]. The work in [18] introduces a relay-assisted model combined with friendly interference collaboration, achieving improved secrecy performance in multi-destination transmissions.

Energy harvesting (EH) is a transformative approach to prolonging the lifespan of wireless devices by collecting energy from the environment, including sources like solar, wind and even radio-frequency (RF) signals from nearby devices or dedicated beacon nodes. In wireless systems, EH allows nodes to operate autonomously, reducing the dependency on traditional power sources. Two prevalent EH techniques are Time Switching (TS) and Power Splitting (PS) [19]-[23]. Time switching separates data and energy reception into distinct time slots, allowing devices to focus on either energy harvesting or data transmission at any moment. Power splitting, on the other hand, enables simultaneous data and energy reception by dividing the incoming signal into two paths; one for energy harvesting and the other for information processing. Hybrid protocols, such as the Hybrid Time Switching and Power Splitting-based Relaying (HTPR) protocol, have been shown to further optimize the throughput in cooperative SWIPT networks by leveraging the benefits of both approaches and using techniques like Maximum Ratio Combining (MRC) at the destination [24]. Both methods are widely researched and continue to be optimized for maximum efficiency and practical deployment in real-world wireless systems. The research demonstrates that energy harvesting with power splitting in cooperative networks can significantly enhance performance, even under complex channel conditions like Nakagamim/Rayleigh fading [25]. Recent research highlights that optimizing for user performance and handling hardware impairments in ambient backscatter systems can significantly improve system reliability and efficiency [26].

Integrating PLS and EH is highly significant in wireless communications, as it addresses both security and energy sustainability [27]-[28]. Studies on decode-and-forward full-duplex networks using power-splitting and self-energy recycling techniques underscore the balance between system security and reliability, even with eavesdroppers present [29]. By incorporating EH, nodes can continually replenish their energy, supporting the implementation of PLS without straining power resources. The integration of simultaneous wireless information and power transfer (SWIPT) in amplify-and-forward (AF) IoT networks provides a significant trade-off between security and reliability, highlighting the advantages of employing friendly jammers alongside power-splitting relaying strategies to mitigate eavesdropping risks [30]. A study on the physical layer security in SWIPT-based decode-and-forward relay networks shows that employing dynamic power splitting significantly enhances outage and secrecy performance in the presence of eavesdroppers [31]. Additionally, PLS with RF energy harvesting in SWIPT cooperative networks enhances information-transmission security and prolongs network lifetime, as discussed in recent studies [32]. The interplay between EH and friendly jammers has been shown to substantially improve both reliability and security in wireless-powered networks, especially in hostile eavesdropping environments, as demonstrated by research on cooperative jamming techniques [33]. Moreover, recent studies also highlight security and reliability enhancements in satellite-terrestrial networks, where a satellite transmits confidential information *via* multiple relay nodes, incorporating friendly jammers to improve secure transmission amidst imperfect channel conditions [34]. To enhance system outage performance in energy harvesting-based two-way relaying protocols, relay-selection methods were proposed, demonstrating significant reliability improvements in data transmission over fading channels [35]. This combination is especially beneficial in systems where nodes operate remotely or autonomously, such as in sensor networks or IoT applications [36]. The performance analysis of time-switching energy harvesting in half-duplex sensor networks under hardware impairments reveals critical insights into outage probability and throughput, emphasizing the viability of energy-harvesting strategies in Rician fading environments [37]. The authors in [38] evaluate the secure performance of multi-hop relay networks by employing joint relay and jammer-selection strategies under imperfect channel conditions, enhancing the system's resistance to multiple eavesdroppers. In the context of cognitive radio networks, cooperative multi-hop transmission protocols can enhance secrecy performance, especially in the presence of hardware impairments, as demonstrated by analyzing the effective signal-to-interference-plus-noise ratio and deriving expressions for end-to-end secrecy outage probability [39]. Utilizing EH alongside PLS allows for sustainable, secure communication channels capable of resisting eavesdropping attempts while ensuring long-term operational viability, even in energy-limited settings. The authors in [40] highlight the impact of power beacon-assisted energy harvesting on device-to-device communication networks, particularly under the influence of co-channel interference and eavesdropping threats, offering closed-form expressions for outage and secrecy outage probabilities. The authors in [41] analyze the security and reliability of power splitting-based relaying schemes in IoT networks, revealing the advantages of dynamically adjusting power-splitting ratios to enhance system performance.

For further enhancing system performance, especially in secure and energy-harvesting-based wireless systems, the use of multiple antennas at the receiving node provides substantial advantages. Multiple antennas increase spatial diversity, which improves both reliability and security in data transmission. Techniques like Selection Combining (SC) and Maximal Ratio Combining (MRC) are commonly employed [42]-[43]. SC chooses the antenna with the highest received signal-to-noise ratio (SNR), simplifying the hardware requirements while maintaining reasonable performance gains. MRC, meanwhile, combines signals from all antennas in proportion to their SNR, resulting in maximal signal enhancement. Both techniques enhance the robustness of the communication link, with MRC often offering superior performance in environments with high interference or noise.

In this paper, we develop and analyze a wireless communication model where a source node, powered by energy harvested from nearby beacon nodes, transmits data securely to a destination node with multiple antennas. An eavesdropper node attempts to intercept the transmission, but the system's security is ensured through PLS techniques. The destination node employs SC and MRC to maximize the signal quality. We derive the security outage probability to assess the system's performance and validate our analytical results through Monte Carlo simulations in Matlab, highlighting the model's effectiveness in secure, sustainable wireless communications.

The list of important contributions of this paper is shown as follows:

1) We develop a secure wireless communication model, integrating PLS and EH, where the destination node uses multiple antennas and selection combining (SC) or maximal ratio combining (MRC) techniques to enhance signal quality.
2) We derive SOP for the system and provide detailed mathematical formulations, which are validated through Monte Carlo simulations.
3) To validate the analytical results, we conduct extensive numerical simulations, evaluating the system's performance under the effects of various parameters, including the power and number of beacon nodes, the number of antennas at the destination and the time-switching factor. Additionally, we investigate the asymptotic behavior to analyze the system's performance under high SNR conditions.

To better highlight the novelty of our work and how it differs from existing studies, we present a comparison with relevant papers in Table 1. This comparison emphasizes the unique aspects of our proposed approach, particularly in the integration of PLS with EH and the use of advanced combining techniques, like SC and MRC.

Table 1. Comparison between our work and previous papers in terms of novelty.

| **Ref. / Prop.** | **PLS** | **EH** | **SC and MRC** |
|---|---|---|---|
| [2] | √ | X | X |
| [24] | X | √ | X |
| [25] | X | √ | X |
| [29] | √ | √ | X |
| [30] | √ | √ | X |
| [33] | √ | √ | X |
| [41] | √ | √ | X |
| [42] | X | √ | √ |
| [43] | √ | X | √ |
| Our study | √ | √ | √ |

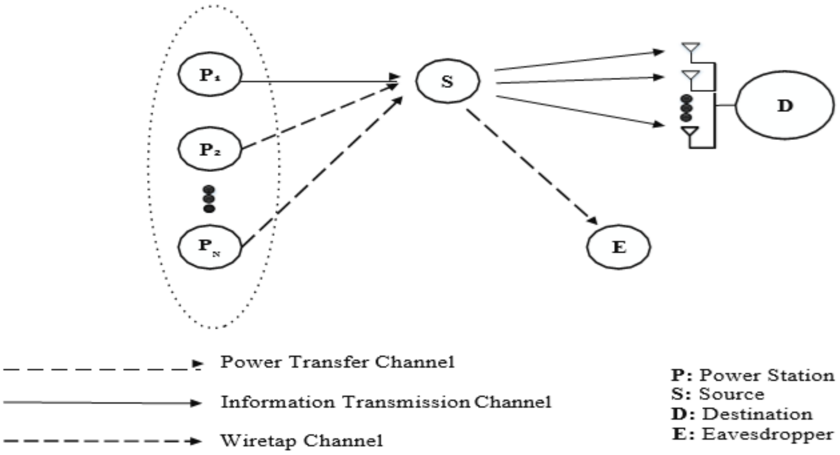


Figure 1. System model of secure wireless communication with PLS and EH, including power stations {$P_n$}, source *S*, destination *D* and eavesdropper *E*.

*Organization:* The remainder of this paper is organized as follows. Section 2 details the system model. In Section 3, we present the performance analysis and Section 4 follows with simulation results to evaluate system performance. Finally, Section 5 concludes the paper, summarizing key insights and potential avenues for future research.

## 2. SYSTEM MODEL

In this study, we consider a secure wireless communication model that integrates both PLS and EH to enhance data confidentiality and system sustainability. The system comprises four primary components: a set of power stations, denoted as $\{P_n | n = 1, \dots, N\}$, a source node *S*, a destination node *D* and an eavesdropper node *E*. The nodes $P_n$, *S* and *E* are each equipped with a single antenna, while the destination node *D* is equipped with *M* antennas.

Table 2. Time allocation for the proposed secure communication scheme.

| **Phase** | **Duration** | **Description** |
|---|---|---|
| Energy Harvesting | $\alpha T$ | *S* harvests energy from *N* beacon nodes $P_n$ for $n = 1,\dots,N$. |
| Information Transmission | $(1-\alpha)T$ | *S* transmits data to *D* using SC or MRC, while *E* attempts to intercept the data during the same time. |

### 2.1 Energy Harvesting from Power Stations

The source node *S* is powered by energy harvested from multiple power stations $P_n$, $n = 1, \dots, N$, spatially distributed around *S*. Each power station transmits energy over a dedicated Power Transfer Channel, modeled as a block Rayleigh fading channel. The harvested power at *S*, denoted by $P_S$, is given by:

$$P_S = \frac{E_h}{(1-\alpha)T} = \frac{\eta\alpha T P_{\mathrm{P}} \sum_{n=1}^{N}|hP_nS|^2}{(1-\alpha)T} = \varkappa P_{\mathrm{P}} \sum_{n=1}^{N}|hP_nS|^2 \tag{1}$$

where:

- $\gamma_{P_n}S = |h_{P_n}S\,|^2$ is the *channel gain* between the power station $P_n$ and the source node *S*,
- $E_h$ represents the total energy harvested at the source node *S*,
- $\alpha$ is the fraction of time dedicated to energy harvesting,
- $T$ is the total time duration of one transmission block,
- *$\eta$ is the energy-conversion efficiency of the harvesting process,*
- $P_{\mathrm{P}}$ *is the transmit power of each power station,*
- $\kappa = \frac{\eta\alpha}{(1-\alpha)}$ *is a constant that consolidates several parameters for simplicity.*

This harvested power enables *S* to operate autonomously, sustaining secure communication without reliance on conventional power sources. A time-switching (TS) strategy is employed at *S*, alternating between energy harvesting and information processing.

### 2.2 Secure Information Transmission to the Destination Node and Eavesdropping Threat from a Wiretap Channel

The source node *S* transmits confidential information to the destination node *D* over the primary *Information Transmission Channel*, modeled as a block Rayleigh fading channel. This channel is subject to fading and potential eavesdropping, with an eavesdropper node *E* positioned within the vicinity of *S*, posing a significant security threat by intercepting the transmitted signal over a *Wiretap Channel*, also modeled as a block Rayleigh fading channel.

To counteract these vulnerabilities, the destination *D* is equipped with M antennas, denoted as $D_m$ for $m$ = 1, . . . , *M* and employs two diversity-combining techniques: Selection Combining (SC) and Maximal Ratio Combining (MRC). SC enhances energy efficiency by selecting the antenna with the highest signal- to-noise ratio (SNR), while MRC linearly combines signals from all antennas in proportion to their SNRs, maximizing the received signal strength. This multi-antenna setup at *D* significantly improves the system's security and resilience against fading, interference and eavesdropping.

In this phase, the received signals at the destination *D* and at the eavesdropper *E* are expressed as follows:

$$y_D^{\zeta} = \sqrt{P_S}h_{SD}^{\xi}x_S + n_D^{\zeta}$$
$$y_E = \sqrt{P_S}h_{SE}x_S + n_E \tag{2}$$

where $n_D^{\zeta}$ and $n_E$ are zero-mean Additive White Gaussian Noise (AWGN) terms with variance $N_0$, $\zeta \in \{\mathrm{SC}, \mathrm{MRC}\}$ indicates the diversity-combining technique employed at *D* and *E* $\{\bullet\}$ denotes the expectation operator.

In this phase, the received signals at the destination *D* and at the eavesdropper *E* are expressed as follows:

$$y_D^{\zeta} = \sqrt{P_S}h_{SD}^{\zeta}x_S + n_D^{\zeta}$$
$$y_E = \sqrt{P_S}h_{SE}x_S + n_E \tag{3}$$

where:

- $\zeta \in \{\mathrm{SC}, \mathrm{MRC}\}$ represents the diversity-combining technique employed at the destination node *D*. Specifically, $\zeta$ can take the value "SC" for Selection Combining (SC) or "MRC" for Maximal Ratio Combining (MRC).

- $x_S$ represents the transmitted signal from the source node *S*. Specifically, it is the data signal that is transmitted to both the destination node *D* and the eavesdropper node *E*. The signal $x_S$ is assumed to have a unit power, i.e., $\mathbb{E}\{x_S^2\}$= 1, where $\mathbb{E}\{\cdot\}$ denotes the expectation operator.

- $n_D^{\zeta}$ *and* $n_E$ are zero-mean Additive White Gaussian Noise (AWGN) terms with variance $N_0$, present at the destination node *D* and the eavesdropper node *E*, respectively.

The SNRs at the destination *D* and the eavesdropper *E*, which determine the ability to successfully decode the transmitted signal *x*S, are given by:

$$\gamma_D^{\zeta} = \frac{P_S\gamma_{SD}^{\zeta}}{N_0},$$
$$\gamma_E = \frac{P_S\gamma_{SE}}{N_0} \quad (4)$$

where $\gamma_{SD}^{\zeta}$ and $\gamma_{SE}$ represent the effective channel gains from *S* to *D* and from *S* to *E*, respectively. This configuration allows the system to dynamically optimize its security by leveraging the SC or MRC technique at *D* to either maximize energy efficiency or signal strength, effectively countering the interception attempts by *E* and ensuring robust, secure communication.

By substituting (1) into (3), we have:

$$\gamma_D^{\zeta} = \varkappa\Psi\gamma_{SD}^{\zeta}\gamma P_N S$$
$$\gamma_E = \varkappa\Psi\gamma_{SE}\,\gamma P_N S \quad (5)$$

where $\Psi = \frac{P_{\mathrm{p}}}{N_o}$ represents the ratio of transmit power from the power station to the noise power at the receiver, indicating the effectiveness of energy harvesting. The term $\gamma P_N S = \sum_{n=1}^{N}|hP_nS|^2$ signifies the cumulative channel gain from all power stations to the source node S, reflecting the overall channel quality experienced by *S*.

Considering all channels characterized by block Rayleigh fading, we can express the cumulative distribution function (CDF) and probability density function (PDF) for the squared amplitudes of the channel gains as follows:

$$F_{\gamma_{SE}}(x) = 1 - \exp(-\lambda_{SE}x) \quad (6)$$

$$f_{\gamma_{SE}}(x) = \frac{\partial F_{\gamma_{SE}}(x)}{\partial x} = \lambda_{SE}\exp(-\lambda_{SE}x) \quad (7)$$

Here, $\lambda_{SE}$ represents the mean of the exponential random variable $\gamma SE$. In this context, it is important to note that similar definitions apply to other channel gains, including $\gamma_{SD}$ and $\gamma P_n S$, reflecting the overall channel conditions across the network.

To incorporate path loss into our model, we define the parameters as:

$$\lambda_{SE} = (d_{SE})^{\beta} \quad (8)$$

where $d_{SE}$ denotes the link distance between nodes *S* and *E* and $\beta$ is the path loss exponent.

## 3. Performance Analysis

### 3.1 Derivation of CDF for $\gamma_{SD}^{\zeta}$ and $\gamma P_N S$

In this sub-section, we undertake the derivation of the Cumulative Distribution Functions (CDFs) for the random variables $\gamma_{SD}^{\zeta}$ and $\gamma P_N S$, as delineated in Equation (4). The determination of these CDFs is critical for evaluating the performance of the system, particularly in terms of reliability and security. We will provide a comprehensive mathematical derivation of these CDFs to facilitate a deeper analysis of system performance.

#### 3.1.1 MRC Case

In the MRC scenario, we calculate the PDF and CDF of $\gamma_{\mathrm{SD}}^{\mathrm{MRC}}$ as well as the PDF for $\gamma P_N S$. The PDF of $\gamma_{\mathrm{SD}}^{\mathrm{MRC}} = \sum_{m=1}^{M}\left|h_{SD_m}\right|^2$ can be expressed as follows [44]:

$$f_{\gamma_{\mathrm{SD}}^{\mathrm{MRC}}} = \frac{(\lambda_{SD})^M}{(M-1)!}x^{M-1}\exp(-\lambda_{SD}x) \quad (9)$$

where $\lambda_{SD} = \lambda_{SDm}$, $\forall m \in (1, 2, ..., M)$ represents the mean of the random variable (RV) $\gamma_{\mathrm{SD}}^{\mathrm{MRC}}$.

Next, based on this PDF, the CDF of $\gamma_{\mathrm{SD}}^{\mathrm{MRC}}$ can be derived as:

$$F_{\gamma_{\mathrm{SD}}^{\mathrm{MRC}}}(x) = \int_0^x f_{\gamma_{\mathrm{SD}}^{\mathrm{MRC}}}(t)dt = \frac{1}{\Gamma(M)} \times \gamma(M, \lambda_{SD}x) \quad (10)$$

where Γ(●) and γ (*a*, *b*) denote the Gamma function and the lower incomplete Gamma function, respectively.

For the PDF of $\gamma P_N S = \sum_{n=1}^{N} |hP_n S|^2$, we can express it as:

$$f_{\gamma P_N S}(x) = \frac{(\lambda_{PS})^N}{(N-1)!} x^{N-1} \exp(-\lambda_{PS} x) \tag{11}$$

where $\lambda_{PS} = \lambda_{PnS}$, ∀n ∈ (1, 2, ..., *N*) is the mean of the RV $\gamma \mathrm{P_N} S$.

### 3.1.2 SC Case

In this sub-section, we focus on deriving the PDF and CDF of $\gamma_{\mathrm{SD}}^{\mathrm{SC}}$ = max $(|h_{SD_1}|^2, |h_{SD_2}|^2, \ldots, |h_{SD_M}|^2)$, which can be derived as follows [8]:

$$F_{\gamma_{\mathrm{SD}}^{\mathrm{SC}}}(x) = (1 - \exp(-\lambda_{SD} x))^M = 1 + \sum_{m=1}^{M} (-1)^m \binom{M}{m} \exp(-m\lambda_{SD} x) \tag{12}$$

where $\lambda_{SD} = \lambda_{SDm}$, ∀m ∈ (1, 2, ..., M ) denotes the mean of the RV $\gamma_{\mathrm{SD}}^{\mathrm{SC}}$.

## 3.2 Secrecy Outage Probability (SOP) Analysis

In the domain of physical layer security, considerable attention has been devoted to the capacity to transmit confidential messages at a positive rate—termed the secrecy rate—between a source and a legitimate destination, while ensuring that an eavesdropper remains uninformed. The successful transmission hinges on the condition that the source-destination channel exhibits superior performance compared to the source-eavesdropper channel. Notably, the secrecy rate improves as the disparity in channel strengths increases, allowing for more secure communications.

The secrecy rate is mathematically expressed as [44]:

$$C_{\mathrm{sec}} = \max(C_{\mathrm{D}} - C_{\mathrm{E}}, 0), \tag{13}$$

where CD = (1 − α) log2 (1 + $\gamma_D^{\zeta}$) is the achievable rate at the destination and CE = (1 − α) log2 (1+ $\gamma_{\mathrm{E}}$) is the rate at the eavesdropper. Here, α represents the fraction of time allocated for secure transmission, while $\gamma_D^{\zeta}$ and $\gamma_{\mathrm{E}}$ denote the signal-to-noise ratios (SNRs) at the destination and eavesdropper, respectively.

Secrecy outage occurs when the secrecy capacity drops below a specified target secrecy rate, an event that poses significant challenges for secure communication. The Secrecy Outage Probability (SOP) is defined as:

$$SOP = \Pr(C_{sec} < C_{th}) = \Pr\left(\frac{1+\gamma_D^{\zeta}}{1+\gamma_E} < \gamma_{th}\right) \tag{14}$$

where $C_{th}$ is the threshold secrecy rate and $\gamma_{th} = 2^{\frac{C_{th}}{1-\alpha}}$ defines the critical boundary for secure transmission. This formulation underscores the relationship between channel conditions and the achievable secrecy rate, thus informing strategies for optimizing secure communication performance under varying operational scenarios.

## 3.3 Exact Analytical Expression for Secrecy Outage Probability (SOP)

### 3.3.1 SOP for MRC Case

Substituting (4) into (13), we can assert:

$$\begin{aligned} \mathrm{SOP^{MRC}} &= \Pr\left(\frac{1+\kappa\Psi\gamma_{SD}^{MRC}\gamma P_N S}{1+\kappa\Psi\gamma_{SE}\gamma P_N S} < \gamma_{th}\right) = \Pr\left(\kappa\Psi\gamma_{SD}^{MRC}\gamma P_N S < \gamma_{th}\kappa\Psi\gamma_{SE}\gamma P_N S + \tilde{\gamma}_{th}\right) \\ &= \int_0^{+\infty} \underbrace{\Pr\left(\kappa\Psi\gamma_{SD}^{MRC} x < \gamma_{th}\kappa\Psi\gamma_{SE} x + \tilde{\gamma}_{th}\right)}_{\Upsilon} . f_{\gamma P_N S}(x)dx \end{aligned} \tag{15}$$

where $\tilde{\gamma}_{th} = \gamma_{th} - 1$.

From (14), ϒ can be computed as follows:

$$\Upsilon = \Pr\left(\kappa\Psi\gamma_{SD}^{MRC}\gamma P_N S < \gamma_{th}\kappa\Psi\gamma_{SE}\gamma P_N S + \tilde{\gamma}_{th}\right) = 1 - \Pr\left(\gamma_{th}\kappa\Psi\gamma_{SE} x \le \kappa\Psi\gamma_{SD}^{MRC} x - \tilde{\gamma}_{th}\right)$$

$$= 1 - \int_0^{\frac{\tilde{\gamma}_{th}}{\kappa\Psi x}} F_{\gamma_{SE}}\left(\frac{y}{\gamma_{th}} - \frac{\tilde{\gamma}_{th}}{\gamma_{th}\kappa\Psi x}\right) f_{\gamma_{SD}^{MRC}}(y)dy$$

$$= 1 - \int_0^{\frac{\tilde{\gamma}_{th}}{\kappa\Psi \mathrm{x}}} \left\{1 - \exp\left(-\lambda_{SE}\left[\frac{y}{\gamma_{th}} - \frac{\tilde{\gamma}_{th}}{\gamma_{th}\kappa\Psi x}\right]\right)\right\} \frac{\lambda_{SD}^M}{(M-1)!} y^{M-1}\exp(-\lambda_{SD}y)dy$$

$$= 1 - \frac{\lambda_{SD}^M}{(M-1)!}\int_0^{\frac{\tilde{\gamma}_{th}}{\kappa\Psi x}} y^{M-1} exp\,(-\lambda_{SD}y)dy + \frac{\lambda_{SD}^M}{(M-1)!}\exp\left(\frac{\lambda_{SE}\tilde{\gamma}_{th}}{\gamma_{th}\kappa\Psi x}\right)\int_0^{\frac{\tilde{\gamma}_{th}}{\kappa\Psi x}} y^{M-1}\exp(-y[\frac{\lambda_{SE}}{\gamma_{th}} + \lambda_{SD}])dy \quad (16)$$

Using Equation (3.381.1) from [45], we derive:

$$\Upsilon = 1 - \frac{\gamma\left(M, \frac{\lambda_{SD}\tilde{\gamma}_{th}}{\kappa\Psi x}\right)}{\Gamma(M)} + \left(\frac{\lambda_{SE}}{\gamma_{th}\lambda_{SD}} + 1\right)^{-M} \times \frac{\exp\left(\frac{\lambda_{SE}\tilde{\gamma}_{th}}{\gamma_{th}\kappa\Psi x}\right)}{\Gamma(M)} \times \gamma(M, \frac{\tilde{\gamma}_{th}}{\kappa\Psi x}\left[\frac{\lambda_{SE}}{\gamma_{th}} + \lambda_{SD}\right]) \quad (17)$$

Finally, substituting (10) and (16) into (14) allows us to express:

$$\mathrm{SOP}^{\mathrm{MRC}} = \frac{\lambda_{PS}^N}{(N-1)!}\int_0^{+\infty} \Upsilon.\, x^{N-1}\exp(-\lambda_{PS}x)\,dx. \quad (18)$$

### 3.3.2 SOP for SC Case

Following a similar approach as in Equation (14), we derive SOP for the SC scenario, denoted as SOP$^{\mathrm{SC}}$:

$$\mathrm{SOP}^{\mathrm{SC}} = \int_0^{+\infty} \underbrace{\Pr(\,\kappa\Psi\gamma_{SD}^{SC}x < \gamma_{th}\kappa\Psi\gamma_{SE}x + \tilde{\gamma}_{th})}_{\Xi}.f_{\gamma P_N S}(x)dx \quad (19)$$

In this formulation, Ξ is expressed in Equation (18) as:

$$\Xi = \int_0^{+\infty} F_{\gamma_{SD}^{SC}}(\gamma_{th}y + \frac{\tilde{\gamma}_{th}}{\kappa\Psi x}) \times f_{\gamma_{SE}}(y)dy \quad (20)$$

By combining Equations (6) and (11), we can further expand Equation (19) as:

$$\Xi = 1 + \sum_{m=1}^M (-1)^m\lambda_{SE}\binom{M}{m}\exp(-\frac{m\lambda_{SD}\tilde{\gamma}_{th}}{\kappa\Psi x})\int_0^{+\infty}\exp(-y[m\lambda_{SD}\gamma_{th} + \lambda_{SE}])\,dy =$$
$$1 + \sum_{m=1}^M (\frac{(-1)^m\lambda_{SE}}{m\lambda_{SD}\gamma_{th}+\lambda_{SE}})\binom{M}{m}\exp(-\frac{m\lambda_{SD}\tilde{\gamma}_{th}}{\kappa\Psi x}) \quad (21)$$

where $\binom{M}{m} = \frac{M!}{m!(M-m)!}$ as the combination of *M* items taken m at a time, *M*! is the factorial of *M* and *m*! is the factorial of *m*.

Inserting Equation (20) into (18), the SOP for SC, SOP$^{\mathrm{SC}}$, can be computed as:

$$\mathrm{SOP}^{\mathrm{SC}} = 1 + \sum_{m=1}^M (\frac{(-1)^m\lambda_{SE}}{m\lambda_{SD}\gamma_{th}+\lambda_{SE}})\frac{(\lambda_{PS})^N}{(N-1)!}\binom{M}{m} \times \int_0^{+\infty} x^{N-1}\exp(-\frac{m\lambda_{SD}\tilde{\gamma}_{th}}{\kappa\Psi x} - \lambda_{PS}x)dx \quad (22)$$

Utilizing the integral identity (3.471.9) in [45], we obtain the final expression:

$$\mathrm{SOP}^{\mathrm{SC}} = 1 + 2\sum_{m=1}^M \left(\frac{(-1)^m\lambda_{SE}}{m\lambda_{SD}\gamma_{th}+\lambda_{SE}}\right)\frac{1}{(N-1)!}\binom{M}{m} \times \left(-\frac{m\lambda_{SD}\lambda_{PS}\tilde{\gamma}_{th}}{\kappa\Psi}\right)^{\frac{N}{2}} K_N(2\sqrt{\frac{m\lambda_{SD}\tilde{\gamma}_{th}}{\kappa\Psi\lambda_{PS}}}) \quad (23)$$

where $K_v(\cdot)$ represents the modified Bessel function of the second kind and *v* order.

## 4. PERFORMANCE EVALUATION THROUGH SIMULATION

In the context of modern wireless communication systems, performance evaluation through simulations is essential for validating theoretical models and ensuring practical applicability. This section presents a comprehensive analysis of the performance of the proposed system through simulations, focusing on various parameters, including Signal-to-Noise Ratio (SNR), energy-harvesting efficiency and the impact of different combining techniques, such as MRC and SC. Monte Carlo simulations, implemented using MATLAB, were employed to generate the results, providing an accurate representation of the system's behavior under various scenarios. The simulation parameters used for generating Figures 2 to 5 are detailed in Table 3, ensuring reproducibility and transparency of the presented results. By varying these parameters, we gain valuable insights into the Secrecy Outage Probability (SOP) and how it is influenced by the interplay of these factors. The results obtained from the simulations provide a deeper understanding of the trade-offs involved in enhancing security and reliability in wireless communication systems. Following the simulations, we discuss the implications of the observed results, as illustrated in the figures, to highlight the effectiveness of our approach in mitigating eavesdropping risks.

Figure 2 presents SOP as a function of the Signal-to-Noise Ratio denoted by $\Psi$ across different combining techniques: MRC and SC. The results indicate that MRC consistently outperforms SC, achieving lower SOP values across the entire range of $\Psi$. Notably, as $\Psi$ increases, SOP decreases for both techniques, with MRC demonstrating a more significant reduction. For instance, at $\Psi = 5$ dB, MRC yields an SOP of approximately 0.0884, compared to SC's 0.2512. Additionally, the asymptotic behavior of the SOP reveals that MRC stabilizes at 0.0871, while SC converges to 0.2468 as SNR approaches infinity. The close alignment between simulation and analytical results underscores the reliability of the mathematical analysis. These findings highlight the enhanced security and reliability of MRC in mitigating eavesdropping risks in secure wireless communication systems.

Table 3. Simulation parameter settings for performance analysis in Figures 2–5.

| **Para.** | **Description** | **Figure 2** | **Figure 3** | **Figure 4** | **Figure 5** |
|---|---|---|---|---|---|
| $\lambda_{\mathrm{SD}}$ | Average channel gain for Source-Destination | 1 | 1 | 1 | 1 |
| $\lambda_{\mathrm{PS}}$ | Average channel gain for Power Station | 1 | 1 | 1 | 1 |
| $\lambda_{\mathrm{SE}}$ | Average channel gain for Source-Eavesdropper | 1 | 1 | 1 | 1 |
| $C_{th}$ | Secrecy capacity threshold | 0.1 | 0.1 | [0.1,0.2] | 0.1 |
| $\Psi$ (dB) | Signal-to-Noise Ratio in dB | [−20 : 30] | 5 | 5 | 5 |
| $M$ | Number of diversity branches in MRC/SC | 4 | [1 : 7] | 4 | 4 |
| $N$ | Number of relay nodes | 4 | [2, 4] | 2 | 2 |
| $\eta$ | Energy-harvesting efficiency | 0.6 | 0.6 | [0 : 1] | 0.6 |
| $\alpha$ | Power-splitting ratio | 0.6 | 0.6 | 0.6 | [0.1 : 0.9] |
| loop | Number of Monte Carlo simulation iterations | $10^5$ | $10^5$ | $10^5$ | $10^5$ |

Figure 3 illustrates SOP for MRC and SC techniques, considering two distinct scenarios: $N = 2$ and $N = 4$ power beacons. The results highlight the significant impact of the number of antennas at the destination ($M$) on SOP performance, with higher values of $M$ leading to improved secrecy performance across both combining techniques. The simulation results (denoted by markers) are in close agreement with the analytical models (solid lines), validating the accuracy of the derived expressions. Notably, the SOP decreases as the number of antennas increases, demonstrating the effectiveness of antenna diversity in enhancing security. Additionally, the comparison between MRC and SC shows that MRC consistently outperforms SC in terms of secrecy outage, especially when the number of antennas is large. These findings underscore the importance of antenna selection in optimizing secure communication performance in practical wireless networks, particularly in energy-constrained environments.

Figure 4 illustrates SOP as a function of the energy-harvesting factor η for both MRC and SC schemes, considering two different threshold capacities $C_{th} = 0.1$ and $C_{th} = 0.2$. As $\eta$ increases, a significant reduction in SOP is observed, which indicates an improvement in the system's security performance due to more efficient energy harvesting. This behavior is attributed to the fact that higher $\eta$ values provide more available energy for secure communication, thereby lowering the probability of secrecy outage. However, after reaching a certain threshold of $\eta$, the SOP curve begins to level off, signifying that further increases in energy harvesting yield marginal benefits. This phenomenon can be explained by the fact that once the energy harvested exceeds the minimal requirement for reliable transmission, additional energy does not substantially affect the SOP, leading to a saturation effect.

In terms of combining techniques, the MRC approach consistently outperforms SC, as shown by its lower SOP values across all scenarios. This is expected, given that MRC utilizes all available signal paths to maximize the received signal strength, leading to a more reliable secure transmission compared to SC, which only selects the best available path. Furthermore, the results demonstrate that a higher threshold capacity $C_{th}$ results in an increased SOP, implying that as the required transmission rate (or secrecy rate) becomes more stringent, the system becomes more vulnerable to secrecy outages. This trade-off underscores the importance of balancing the energy-harvesting capabilities with the required secrecy performance in practical wireless communication systems. These findings provide valuable insights into optimizing energy-harvesting techniques and combining strategies for secure and efficient communication.

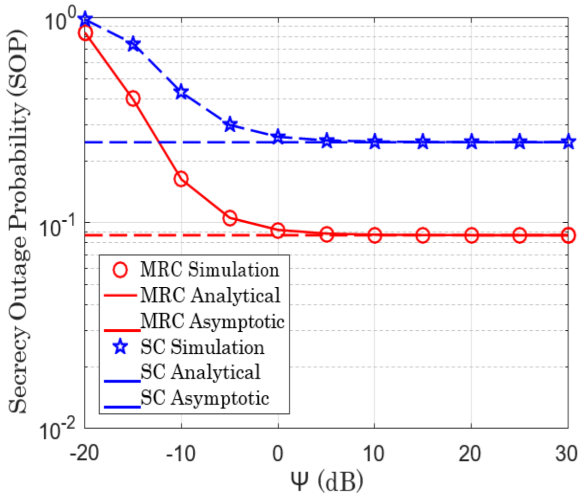


Figure 2. Secrecy outage probability as a function of the signal-to-noise ratio denoted by Ψ for MRC and SC.

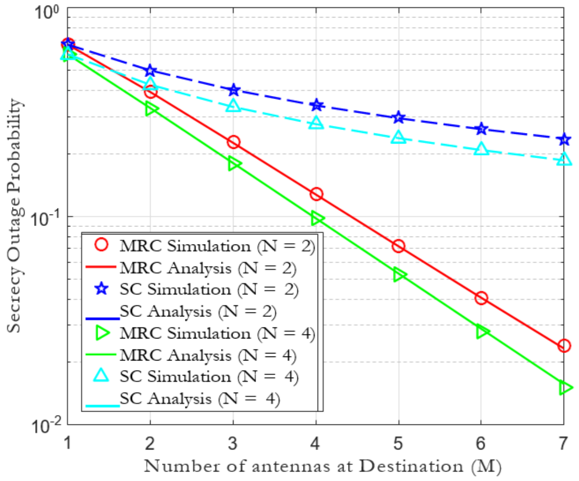


Figure 3. Secrecy outage probability (SOP) *vs.* number of antennas at the destination (*M*) for different scenarios of MRC and SC.

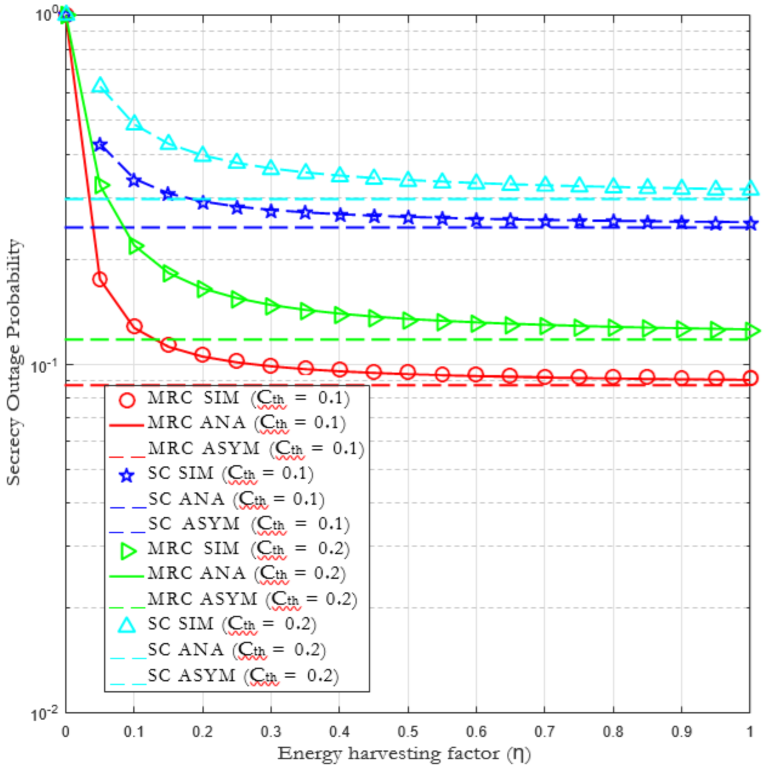


Figure 4. Impact of $\eta$ and $C_{th}$ on secrecy outage probability (SOP) for various MRC and SC configurations.

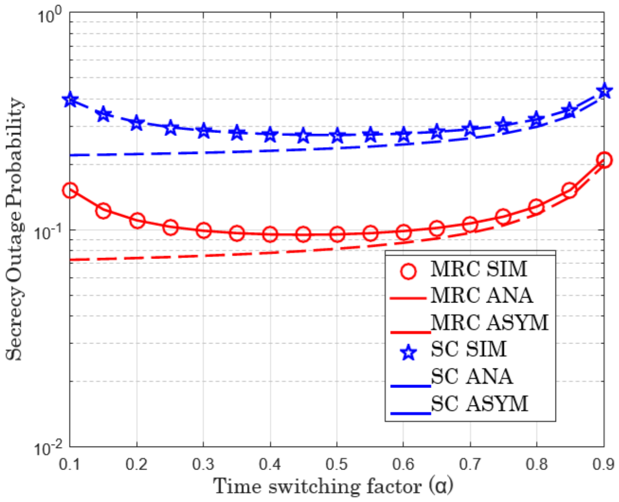


Figure 5. Effect of time switching factor ($\alpha$) on secrecy outage probability for MRC and SC.

Figure 5 illustrates the impact of the time switching factor ($\alpha$) on secrecy outage probability (SOP) for both MRC and SC schemes. As observed, increasing α from 0.1 results in a decrease in SOP, indicating improved system performance. This is due to the increased time available for energy harvesting at the source node (S), allowing more energy to be used for transmitting information to the destination node (D). However, when $\alpha$ exceeds a threshold of approximately 0.7, SOP begins to rise. This can be attributed to the fact that as $\alpha$ increases, more time is devoted to energy harvesting, leaving less time for signal transmission, which reduces the achievable rate at the destination node (D) and thus increases the probability of secrecy outage. This demonstrates the trade-off between energy harvesting and communication efficiency in energy-constrained systems.

## 5. CONCLUSION

In this paper, we analyzed the performance of a secure wireless communication system that integrates Physical Layer Security (PLS) and Energy Harvesting (EH) under various system configurations. We considered a cooperative communication model where a source node transmits data to a destination node, equipped with multiple antennas, while harvesting energy from beacon nodes in the presence of an eavesdropper. The analytical expressions for the Secrecy Outage Probability (SOP) were derived, incorporating key parameters such as the Signal-to-Noise Ratio (SNR), energy-harvesting efficiency ($\eta$), the number of interference nodes ($M$), the number of beacon nodes ($N$) and the time-switching factor ($\alpha$).

Monte Carlo simulations were employed to assess the impact of these parameters on SOP. The results indicate that increasing $\eta$ and $\Psi$ enhances SOP performance by improving both the system's energy-harvesting efficiency and the quality of the received signals. The time-switching factor $\alpha$ plays a crucial role in balancing energy harvesting and data transmission: higher values of $\alpha$ prioritize energy harvesting, which may reduce the time available for data transmission, leading to increased SOP when α exceeds a certain threshold. Furthermore, an increase in the number of antennas at the destination node and the number of beacon nodes $N$ contributes to a reduction in SOP, thereby improving both signal diversity and energy availability. In contrast, a higher number of interference nodes $M$ tends to increase SOP, emphasizing the trade-offs in secure communication system design.

The theoretical results derived in this work were validated through simulations, demonstrating the accuracy and robustness of the proposed analytical models. These findings underscore the potential of combining EH and PLS to enhance both security and efficiency in wireless communication networks. Future work could explore adaptive time-switching strategies, multi-relay configurations and alternative energy-allocation methods to optimize system performance and security in dynamic environments. Adaptive time-switching techniques could be implemented to dynamically adjust the time allocation between energy harvesting and transmission, based on real-time environmental conditions, improving energy efficiency and communication reliability. Multi-relay configurations, leveraging energy-harvesting relays, could increase system reliability and coverage, especially in challenging environments with limited direct links. Additionally, the techniques presented in [46] and

[47], which apply convolutional neural networks (CNNs) for surface-defect detection, could potentially enhance the current system by introducing advanced machine-learning models to improve decision-making processes and system efficiency in wireless communication security. Emerging technologies, like 6G networks and machine learning offer significant potential to complement and further enhance the proposed system, especially in complex, real-world scenarios with unpredictable conditions.

## REFERENCES

[1] D. Grenar, J. Frolka, K. Slavicek, O. Dostal and M. Kyselak, "Network Physical Layer Attack in the Very High Capacity Networks," Advances in Electrical and Electronic Eng., vol. 21, no. 1, pp. 37-47, Mar. 2023.

[2] W. Guo, C. Song, X. Xia, F. Hu, H. Zhao, S. Shao and Y. Tang, "Analysis of Cooperative Jamming Cancellation with Imperfect Time Synchronization in Physical Layer Security," IEEE Wireless Communications Letters, vol. 10, no. 2, pp. 335-338, Feb. 2021.

[3] Z. Al-qudah and K. A. Darabkh, "A Simple Encoding Scheme to Achieve the Capacity of Half-duplex Relay Channel," Advances in Electrical and Electronic Eng., vol. 20, no. 1, pp. 33-42, Mar. 2022.

[4] T. N. Nguyen, P. T. Tran, T. H. Q. Minh, M. Voznak and L. Sevcik, "Two-way Half Duplex Decode and Forward Relaying Network with Hardware Impairment over Rician Fading Channel: System Performance Analysis," ELEKTRONIKA IR ELEKTROTECHNIKA, vol. 24, no. 2, pp. 74-78, 2018.

[5] T. N. Nguyen, M. Tran, T.-L. Nguyen and M. Voznak, "Adaptive Relaying Protocol for Decode and Forward Full-duplex System over Rician Fading Channel: System Performance Analysis," China Communications, vol. 16, no. 3, pp. 92-102, Mar. 2019.

[6] P. T. Tin, N. T. Luan, T. N. Nguyen, M. Tran and T. T. Duy, "Throughput Enhancement for Multi-hop Decode-and-Forward Protocol Using Interference Cancellation with Hardware Imperfection," Alexandria Engineering Journal, vol. 61, no. 8, pp. 5837-5849, Aug. 2022.

[7] T. N. Nguyen, L.-T. Tu, D.-H. Tran, V.-D. Phan, M. Voznak and S. Chatzinotas, "Outage Performance of Satellite Terrestrial Full-duplex Relaying Networks with Co-channel Interference," IEEE Wireless Communications Letters, vol. 11, no. 7, pp. 1478-1482, Jul. 2022.

[8] T. N. Nguyen, T. T. Duy, P. T. Tran, M. Voznak, X. Li and H. V. Poor, "Partial and Full Relay Selection Algorithms for AF Multi-relay Full-duplex Networks With Self-energy Recycling in Non-identically Distributed Fading Channels," IEEE Transactions on Vehicular Technology, vol. 71, no. 6, pp. 6173-6188, Mar. 2022.

[9] Y. Lee, "End-to-end Error-rate Based Incremental Relaying for AF Cooperative Communications," IEEE Communications Letters, vol. 17, no. 9, pp. 1806-1809, Sep. 2013.

[10] W. Su and X. Liu, "On Optimum Selection Relaying Protocols in Cooperative Wireless Networks," IEEE Transactions on Communications, vol. 58, no. 1, pp. 52-57, Jan. 2010.

[11] R. Saini, D. Mishra and S. De, "OFDMA-based DF Secure Cooperative Communication with Untrusted Users," IEEE Communications Letters, vol. 20, no. 4, pp. 716-719, Apr. 2016.

[12] J. Mo, M. Tao and Y. Liu, "Relay Placement for Physical Layer Security: A Secure Connection Perspective," IEEE Communications Letters, vol. 16, no. 6, pp. 878-881, Jun. 2012.

[13] R. Bassily and S. Ulukus, "Secure Communication in Multiple Relay Networks through Decode-and-Forward Strategies," Journal of Communications and Network, vol. 14, no. 4, pp. 352-363, Aug. 2012.

[14] H.-M.Wang, M. Luo, X.-G. Xia and Q. Yin, "Joint Cooperative Beamforming and Jamming to Secure AF Relay Systems with Individual Power Constraint and No Eavesdropper's CSI," IEEE Signal Processing Letters, vol. 20, no. 1, pp. 39-42, Jan. 2013.

[15] C. Jeong, I.-M. Kim and D. I. Kim, "Joint Secure Beamforming Design at the Source and the Relay for an Amplify-and-Forward MIMO Untrusted Relay System," IEEE Transactions on Signal Processing, vol. 60, no. 1, pp. 310-325, Jan. 2012.

[16] S. Q. Nguyen and H. Y. Kong, "Improving Secrecy Outage and Throughput Performance in Two-way Energy-Constrained Relaying Networks under Physical Layer Security," Wireless Personal Communications, vol. 96, pp. 6425-6457, May 2017.

[17] S. Q. Nguyen and H. Y. Kong, "Combining Binary Jamming and Network Coding to Improve Outage Performance in Two-way Relaying Networks under Physical Layer Security," Wireless Personal Communications, vol. 85, pp. 2431-2446, July 2015.

[18] L. Liang, X. Li, H. Huang, Z. Yin, N. Zhang and D. Zhang, "Securing Multi-destination Transmissions with Relay and Friendly Interference Collaboration," IEEE Internet of Things Journal, vol. 11, no. 10, pp. 18782-18795, May 2024.

[19] X. Zhou, R. Zhang and C. K. Ho, "Wireless Information and Power Transfer: A Dynamic Power Splitting Approach," IEEE Transactions on Communications, vol. 61, no. 9, pp. 3991-4003, Sep. 2013.

[20] R. Jiang, K. Xiong, P. Fan, Y. Zhang and Z. Zhong, "Power Minimization in SWIPT Networks with

Coexisting Power-Splitting and Time-switching Users under Non-linear EH Model," IEEE Internet of Things Journal, vol. 6, no. 5, pp. 8853-8869, DOI: 10.1109/JIOT.2019.2923977, Oct. 2019.

[21] X. Zhou, R. Zhang and C. K. Ho, "Wireless Information and Power Transfer in Multi-user OFDM Systems," IEEE Transactions on Wireless Communications, vol. 13, no. 4, pp. 2282-2294, Apr. 2014.

[22] F. K. Ojo and M. F. M. Salleh, "Throughput Analysis of a Hybridized Power-Time Splitting Based Relaying Protocol for Wireless Information and Power Transfer in Cooperative Networks," IEEE Access, vol. 6, pp. 24137-24147, DOI: 10.1109/ACCESS.2018.2828121, Apr. 2018.

[23] R. Tao, A. Salem and K. A. Hamdi, "Adaptive Relaying Protocol for Wireless Power Transfer and Information Processing," IEEE Communications Letters, vol. 20, no. 10, pp. 2027-2030, Oct. 2016.

[24] P. S. Lakshmi and M. G. Jibukumar, "A Hybrid Protocol for SWIPT in Cooperative Networks," Advances in Electrical and Electronic Eng., vol. 19, No. 1, pp. 28-41, Mar. 2021.

[25] T. N. Nguyen, M. Tran, T.-L. Nguyen, D.-H. Ha and M. Voznak, "Performance Analysis of a User Selection Protocol in Cooperative Networks with Power Splitting Protocol-based Energy Harvesting over Nakagami-m/Rayleigh Channels," Electronics, vol. 8, no. 4, 2019.

[26] M.-S. V. Nguyen and H.-P. Dang, "Exploiting Performance of Ambient Backscatter Systems in Presence of Hardware Impairment," Advances in Electrical and Electronic Eng., vol. 19, no. 4, pp. 314-321, 2021.

[27] B. C. Nguyen et al., "Cooperative Communications for Improving the Performance of Bidirectional Full-duplex System with Multiple Reconfigurable Intelligent Surfaces", IEEE Access, vol. 9, pp. 134733 - 134742, Nov. 2021.

[28] K. Lee, J.-P. Hong, H.-H. Choi and T. Q. S. Quek, "Wireless-powered Two-way Relaying Protocols for Optimizing Physical Layer Security," IEEE Transactions on Information Forensics and Security, vol. 14, no. 1, pp. 162-174, Jan. 2019.

[29] B. V. Minh et al., "Self-energy Recycling in DF Full-duplex Relay Network: Security-Reliability Analysis," Advances in Electrical and Electronic Eng., vol. 22, no. 2, pp. 85-95, 2024.

[30] T. N. Nguyen, D.-H. Tran, T. V. Chien, V.-D. Phan, M. Voznak, adn P. T. Tin, "Security–Reliability Trade-off Analysis for SWIPT- and AF-based IoT Networks with Friendly Jammers," IEEE Internet of Things Journal, vol. 9, no. 21, pp. 21662-21675, June 2022.

[31] V-. D. Pham et al., "A Study of Physical Layer Security in SWIPT-based Decode-and-Forward Relay Networks with Dynamic Power Splitting," Sensors, vol. 21, no. 7, Aug. 2021.

[32] T. N. Nguyen et al., "Physical Layer Security in AF-based Cooperative SWIPT Sensor Networks," IEEE Sensors Journal, vol. 23, no. 1, pp. 689-705, Jan. 2023.

[33] V. D. Phan, T. L. Nguyen, T. T. Phu and V. V. Nguyen, "Reliability-Security in Wireless-powered Cooperative Network with Friendly Jammer," Advances in Electrical and Electronic Engineering Journal, vol. 20, no. 4, pp. 584-591, Jan. 2022.

[34] W. Zeng, J. Zhang, D. W. K. Ng, B. Ai and Z. Zhong, "Two-way Hybrid Terrestrial-satellite Relaying Systems: Performance Analysis and Relay Selection," IEEE Transactions on Vehicular Technology, vol. 68, no. 7, pp. 7011-7023, Jul. 2019.

[35] T. N. Nguyen et al., "Performance Enhancement for Energy Harvesting Based Two-way Relay Protocols in Wireless *Ad-Hoc* Networks with Partial and Full Relay Selection Methods," *Ad Hoc* Networks, vol. 81, pp. 178-187, Mar. 2019.

[36] P. T. Tin, T. N. Nguyen, M. Tran, T. T. Trang and L. Sevcik, "Exploiting Direct Link in Two-way Half-duplex Sensor Network over Block Rayleigh Fading Channel: Upper Bound Ergodic Capacity and Exact SER Analysis," Sensors, vol. 20, no. 4, Feb. 2019.

[37] T. N. Nguyen, T. H. Q. Minh, P. T. Tran and M. Voznak, "Energy Harvesting over Rician Fading Channel: A Performance Analysis for Half-duplex Bidirectional Sensor Networks under Hardware Impairments," Sensors, vol. 81, no. 6, 2018.

[38] C. T. Dung, T. M. Hoang, N. N. Thang, M. Tran and T. T. Phuong, "Secrecy Performance of Multi-user Multi-hop Cluster-based Network with Joint Relay and Jammer Selection under Imperfect Channel State Information," Performance Evaluation, vol. 147, p. 102193, 2021.

[39] P. T. Tin, D. T. Hung, T. N. Nguyen, T. T. Duy and M. Voznak, "Secrecy Performance Enhancement for Underlay Cognitive Radio Networks Employing Cooperative Multi-hop Transmission with and without Presence of Hardware Impairments," Entropy, vol. 21, no. 2, 2019.

[40] B. V. Minh, T. H. Q. Minh, V. D. Phan and H. T. Nguyen, "D2D Communication Network with the Assistance of Power Beacon under the Impact of Co-channel Interferences and Eavesdropper: Performance Analysis," Advances in Electrical and Electronic Eng., vol. 21, no. 4, pp. 351-359, 2023.

[41] M. Tran et al., "Security and Reliability Analysis of the Power Splitting-based Relaying in Wireless Sensors Network," Sensors, vol. 24, no. 4, Feb. 2024.

[42] T. M. Hoang, X. N. Tran, B. C. Nguyen and L. T. Dung, "On the Performance of MIMO Full-Duplex Relaying System with SWIPT under Outdated CSI," IEEE Transactions on Vehicular Technology, vol. 69, no. 12, pp. 15580-15593, Dec. 2020.

[43] L. Liang, X. Li, H. Huang, Z. Yin, N. Zhang and D. Zhang, "Securing Multi-destination Transmissions

with Relay and Friendly Interference Collaboration," IEEE Internet of Things Journal, vol. 11, no. 10, pp. 18782-18795, Mar. 2024.

[44] H. D.-Hung, T. N. Nguyen, M. Tran, X. Li, P. T. Tran and M. Voznak, "Security Analysis of a Two-way Half-duplex Wireless Relaying Network Using Partial Relay Selection and Hybrid TPSR Energy Harvesting at Relay Nodes," IEEE Access, vol. 8, pp. 187165-187181, Oct. 2020.

[45] I. S. Gradshteyn and I. M. Ryzhik, Table of Integrals, Series and Products, 7th Edn, ISBN: 978-0-12-373637-6, Editor: A. Jeffrey, Burlington, MA: Academic Press, 2007.

[46] D. Zhang, X. Hao, D. Wang, C. Qin, B. Zhao, L. Liang and W. Liu, "An Efficient Lightweight Convolutional Neural Network for Industrial Surface Defect Detection," Artificial Intelligence Review, vol. 56, pp. 10651-10677, March 2023.

[47] Z. Dehua, H. Xinyuan, L. Linlin, L. Wei and Q. Chunbin, "A Novel Deep Convolutional Neural Network Algorithm for Surface Defect Detection," Journal of Computational Design and Engineering, vol. 9, no. 5, pp. 1616-1632, 2022.

## ملخص البحث:

تقدِّم هذه الورقة نظام اتّصالاتٍ لاسلكيةٍ آمناً يجمع بين أمان الطبقّات الفيزيائية (PLS) وحَصَاد الطّاقة (EH) لتحسين كلٍّ من سِرّية البيانات واستدامة الشّبكة. ويوظّف النّظام المقترح تجميع النّسب القصوى (MRC) وتجميع الانتقاء (SC) لعُقدة الهدف (D) متعدّدة الهوائيات، باعتبارها طريقة مبتكرة لأنظمة أمانِ الطّبقات الفيزيائية المشغَّلة بحصاد الطّاقة.

ويرتكز النّظام في نموذجهِ على عُقدة مَصْدَرٍ يتمّ تشغيلها بطاقةٍ يجري حصادُها من محطّاتِ قُدْرَةٍ موزَّعة توزيعاً حَيّزيّاً، وعُقدة هدفٍ متعدَّدة الهوائيات، وعُقدة اختراقٍ، ضمن مدى الاتّصال. ويعمل بروتوكول تبديل زمني على جعل عُقدة المصدر تتناوب بين حصاد الطّاقة والنّقل الآمن للبيانات. ولتحسين جودة الإشارة وأمانها، توظِّف عقدة الهدف تقنيات تجميع النّسب القصوى وتجميع الانتقاء للتّخفيف من مخاطر الاضمحلال والاختراق.

من ناحيةٍ أخرى، نقدّم تشكيلاً دقيقاً لاحتمالية خروج السّرّية للوصْف الكمّيّ لاحتمالية تسرُّب المعلومات تحت تشكيلاتٍ مختلفة للنّظام. وقد جرى التّحقُّق من النّموذج من خلال محاكاة مونتي كارلو، لإثبات دقّة التّحليلات النّظرية. وتُلقي نتائج المحاكاةِ الضّوءَ على المتغيّرات الأساسية (فعالية حَصاد الطّاقة، و التّبديل الزّمني، و عدد الهوائيات، و عدد العُقد العاملة كمناراتٍ وقُدْرتها) على احتمالية خروج السّرّية (SOP)، مقدِّمين بذلك تحليلاً عميقاً لأجْل تحسين أداء أنظمة اتّصالٍ لاسلكية آمنة وذات فعالية من حيث الطّاقة. كذلك قدّمْنا تحليلاً لوصفِ أداء النّظام عند نسب إشارةٍ إلى ضجيج عالية.

Jordanian Journal of Computers and Information Technology (JJCIT), Vol. 11, No. 02, June 2025.

# STUDY OF RECENT IMAGE RESTORATION TECHNIQUES: A COMPREHENSIVE SURVEY

Nikita Singhal, Anup Kadam, Pravesh Kumar, Hritik Singh, Aaryan Thakur and Pranay

*(Received: 24-Dec.-2024, Revised: 22-Feb.-2025, Accepted: 26-Mar.-2025)*

## ABSTRACT

*The rapid advancements in digital imaging technologies, including image restoration (IR), have created a growing demand for effective image-restoration techniques. Various kinds of degradation, including noise, blur and low resolution, should be handled with these techniques. Restoration is important in many applications, including medical imaging, surveillance, photography and remote sensing, where image quality will be critical to the correctness of analysis and decision. This article provides an all-inclusive review of state-of-the-art (SOTA) methods in image restoration, covering traditional methods as well as modern techniques like deep learning (DL) and transformer-based models. Traditional image-restoration techniques include deblurring, denoising and super-resolution based on mathematical models and handcrafted algorithms. These methods were indeed effective for certain types of noise or blur, but generalized poorly to various real-world scenarios. Recent advances in machine learning (ML), especially DL using convolutional neural networks (CNNs), have made data-driven approaches that learn directly from large datasets much more effective. Recently, transformer-based models, such as Vision Transformers and Swin Transformers, have shown the ability to capture global dependencies in images, leading to superior performance on complex restoration tasks. It is also to mention the challenge of generalization across the type of degradation, say mixed noise or blur, and across different datasets. The proposed survey indicates the limitations of existing approaches, including computational cost and generalization challenges and offers insights into possible directions for future research. Considering these challenges and achievements, this article attempts to provide helpful guidance on methods for future research on restoring images.*

## KEYWORDS

*Image restoration, Deep learning, Transformer-based architectures, Noise reduction, Cross-domain models.*

## 1. INTRODUCTION

Image restoration, which aims to preserve high-quality images from deteriorating or damaged ones, has gained increased attention in modern multimedia-driven society due to the growing usage of digital photos. Degradations such as noise and blur can significantly affect image quality, affecting everything from everyday photography to medical imaging. Because it preserves details and improves visual clarity—two things that are often required for jobs involving image analysis, image restoration is therefore an important field of research in computer vision and image processing. The challenge of addressing various forms of degeneration has led academics to explore innovative methods for accurate and efficient image restoration.

Traditional approaches to image deblurring, denoising and super-resolution focused on specially designed algorithms that introduced regularization and filtering techniques to attempt to make use of mathematical models in recovering lost information. Such approaches proved to be very effective for some classes of noise or blur, but fared rather poorly at generalizing to other classes of degradation and sometimes produced sub-optimal results when applied directly to real-world problems. Thus, the entire domain has undergone major changes with recent developments in the fields of ML and DL, where a model becomes capable of learning from data rather than from rules.

The recent resurgence of interest in image restoration is due to architectures that have been designed primarily within the context of natural-language processing, particularly those built on the Transformer model. Here, among heroes, Swin Transformers and Vision Transformers, or ViTs, have proven to capture long-range dependencies and accurately model global interactions within images and return much detail lost in more traditional approaches for restoring images. So, mainly, wide pre-training on megascale data provides those Transformer-based models with a rather strong sense of

N. Singhal, A Kadam, P. Kumar, H. Singh,A. Thakur and Pranay are with Department of Computer Engineering, Army Institute of Technology, Pune, India. Emails: [ngupta, akadam, praveshkumar_21195, hritiksingh_21186, aaryanthakur_21161, pranay_21211] @aitpune.edu.in

both global and local features. So, this kind of model proves to be extremely efficient in many restoration tasks, from video- and image-compression enhancement to the repair of damaged medical images. For this reason, Swin Transformers and ViTs are now the main representatives of modern image restoration. It is here that success lies-largest capacity for recovering the finest detail and significantly raising the quality of degraded images.

In recent years, multiple ML techniques are implemented to solve complicated tasks in image restoration and related problems. Those methods consist mainly of traditional machine learning, deep learning-based methods and more advanced models that include Transformers and GAN-based approaches. Each technique presents advantages and limitations and offers a specific solution for challenges. Table 1 summarizes these diverse methodologies, highlighting some important studies related to each approach. This general summary serves as a foundation for the development of techniques in machine learning and offers some insight into just how each approach uniquely contributes to image restoration. Notably, traditional methods continue to dominate baseline comparisons, but DL, Transformer and diffusion-based models are beginning to take the field, because they pose SOTA performance on complex restoration tasks.

Table 1. Summary of machine-learning approaches and related studies.

| Machine-learning Approach | Related Studies |
|---|---|
| Traditional Machine-learning Approaches | [1],[2],[3],[4] |
| DL-based Approaches | [5],[6],[7],[8],[9],[10],[11],[12],[13],[14] |
| Transformer-based Models | [15],[16],[17],[18],[19] |
| Multitask and Meta-learning Approaches | [20],[21] |
| GAN-based Approaches | [22],[23] |
| Diffusion-based Models | [11],[24],[25],[26],[27],[28],[29] |
| Hybrid Models | [9],[30] |
| Domain-specific Approaches | [31],[32],[33],[34],[35],[36],[37],[38] |

Despite the remarkable progress made so far, the research in image restoration remains a burdensome task with several difficulties. Those include a significant reduction in the computational cost of restoring methods for real-time applications, handling multiple degradations together and boosting the generality of models across different domains. Lack of high-quality annotated datasets for specific domains, such as medical images and setting up cross-domain restoration models are vital today. This work considers the techniques developed for image restoration, focusing on deep-learning strategies, traditional methods and more recent transformer-based models. At the same time, we pass through the main datasets that are generally utilized alongside performance indicators and challenges that characterize the state of image restoration research today and point out possible lines for further research.

## 1.1 Comprehensive Comparison between Existing Survey Papers

In the field of image restoration, numerous survey articles have been published, each providing unique insights into various algorithms, methodologies and applications. However, these surveys differ in focus, evaluation criteria and comprehensiveness. Table 2 provides a comparative summary of prominent survey articles, which outline their respective strengths and limitations. This comparison enables a clearer understanding of the existing literature, helping to identify common approaches, as well as gaps in coverage that may benefit from further research. By examining the merits and demerits, this review aims to position our study in the context of existing work and to highlight areas where our approach may offer additional insights.

Table 2. Comparison of paper with existing surveys.

| Review Paper | Objective | Merits | Demerits |
|---|---|---|---|
| [39], 2021 | To explore the application of DL methods in SAR image restoration. | Offers a detailed analysis of SAR-specific restoration challenges with deep learning. | Limited to SAR images, not generalizable to other image modalities. |

Continuation of Table 2

| Review Paper | Objective | Merits | Demerits |
|---|---|---|---|
| [40], 2021 | To review GAN-based methods for image reconstruction in medical imaging. | Explores the successful use of GANs in improving medical-imaging quality and accuracy. | Primarily focuses on medical imaging, limiting its applicability to other fields. |
| [41], 2022 | To explore DL and smart technologies for image super-resolution. | Provides a critical analysis of recent advancements in super-resolution techniques. | Focuses primarily on super-resolution, lacks coverage of other restoration techniques. |
| [42], 2022 | To review DL approaches for demoiring screen-shot images. | Focuses on a niche issue in image restoration, providing specialized solutions. | Limited to demoiring applications, lacks broader applicability to other restoration tasks. |
| [43], 2022 | To review various image-restoration methods for different image types. | Provides a broad review of restoration methods across diverse image types and applications. | Lacks depth in any specific domain due to its broad scope. |
| [44], 2023 | Surveys diffusion models for image restoration and enhancement,analyzing their advantages, challenges and recent improvements. | Provides a structured taxonomy of diffusion models and their applications in denoising, super- resolution and deblurring. | Diffusion models often require high computational resources, which is not thoroughly discussed in terms of practical deployment. |
| [45], 2023 | To review various IR methods designed to handle salt and pepper noise. | Provides an extensive survey of both linear and non-linear filtering techniques to restore ground- truth images. | Primarily focuses on salt and pepper noise, limiting its generalizability to other types of image degradation. |
| [46], 2023 | To compare GAN-based approaches for image deblurring. | Offers a comparative analysis of multiple GAN- based methods for deblurring. | Limited to GAN-based approaches, excluding other potential techniques. |
| [47], 2023 | To review DL-based techniques for image restoration in real-world settings. | Provides a comprehensive analysis of different DL techniques for image | Does not focus on specific restoration domains, making it broad in scope. |
| [48], 2023 | To review underwater image-restoration techniques. | Addresses specific challenges of underwater imaging and offers solutions for image | Limited to underwater optical imaging, lacks generalization to other image types. |
| [49], 2023 | To review quality assessment algorithms for realistic blurred images. | Provides insights into quality assessment for blurred images using a comprehensive database. | Limited to quality assessment, lacking exploration of actual image restoration techniques. |
| [50], 2025 | To analyze and compare machine learning-based techniques for improving image quality. | Offers a broad comparison of machine-learning models, highlighting their strengths and applications in image enhancement. | Lacks in-depth discussion on DL-based approaches, focusing more on traditional ML techniques. |
| [51], 2024 | To develop and analyze a GAN-based image-restoration algorithm for engineering applications. | Highlights the potential of GANs in reconstructing and restoring images with high precision in engineering contexts. | Focuses on engineering applications, with limited exploration of general-purpose image-restoration techniques. |
| [52], 2024 | To analyze deep-learning techniques applied to image denoising. | Provides an in-depth review of SOTA methods in denoising using deep learning. | Primarily focuses on denoising, with limited exploration of other restoration tasks. |
| [53], 2024 | To report on the NTIRE 2024 challenge focused on bracketing image restoration. | Highlights the latest advancements from the NTIRE challenge with benchmark results. | Results are constrained to the challenge datasets, limiting real-world applicability. |
| **Proposed Survey** | To offer an overview of image restoration methods with a focus on recent advancements. | Provides acomprehensive analysis of image restoration techniques,including GAN, hybrid, DL and transformer-based methods, …etc. Additionally, it presents key metrics such as inference time, PSNR and SSIM, enabling detailed evaluation of each method’s efficacy. | |

### 1.2 Performance Comparison of Image-restoration and denoising Techniques

Table 3 presents a comparative evaluation of the effectiveness of various denoising and IR methods. DnCNN easily handles both known and unknown noise levels while achieving good PSNR in a range of denoising applications. As noise-reduction settings are changed, the efficacy of Wiener filtering improves, offering a balanced approach to both noise reduction and feature retention. An extremely useful technique for minimizing noise and preserving significant image edges is total variation regularization. QTP loss improves perceptual quality by addressing problems, such as inadequate augmentation and misleading color. The Three-stage CNN exhibits remarkable performance in color-image restoration, especially in denoising and demosaicking. Finally, VCRNet is a strong candidate for real-world settings, since it effectively resolves no-reference image-quality assessment (NR-IQA) tasks, retrieving images even in the absence of reference data. Depending on the particular needs of image-restoration activities, these approaches provide a variety of possibilities.

Table 3. Performance comparison of image-restoration and denoising techniques.

| **Model/Technique** | **Accuracy/Performance** |
|---|---|
| Deep Neural Networks (DnCNNs) | High PSNR results across various tasks; specific improvements noted in denoising. |
| Wiener Filtering | Performance improves with better noise reduction; often provides a good balance. |
| Total Variation Regularization | Effective in reducing noise while preserving image edges. |
| Quality-Task-Perception (QTP) Loss | Enhanced perceptual quality for images during restoration. |
| Three-stage CNN | Effectively restores color images with high performance in various tasks. |
| Visual Compensation Network (VCRNet) | Efficiently handles NR-IQA tasks, showing significant promise in restoring images. |

## 2. LITERATURE SURVEY

This section explores the various algorithms used in Image Restoration. Figure 1 depicts a taxonomy of image restoration that divides the methods into several categories of model-based approaches. The systematic classification above indicates an overview of methods proposed to address various problems associated with image restoration, from diffusion-based models, GANs, transformer-based models, deep-learning techniques, hybrid approaches, multi-task/meta-learning approaches, to conventional machine learning-based models. Each category has several methods proposed to address a specific type of vision impairment.

### 2.1 Traditional Machine Learning-based Approaches

Traditional machine-learning techniques have played a very crucial role in image restoration. They were applied to various restoration tasks under conditions, like diffraction effects and limited visibility. The techniques use mathematical models and feature-driven approaches to solve the problem of image degradation, where the model's understanding of local image properties and image production is essential. Although these approaches were significant advances in particular domains, they were also inherently limited as techniques based on rigid priors and handcrafted features. These models were much less flexible than the subsequently developed DL-based techniques, because they frequently required intense fine-tuning to work generically over many applications. However, they formed a strong basis for adaptive previous use and image formation that went into the formulation of modern image-restoration frameworks. A comparison of various traditional image-restoration methods, highlighting their key characteristics and performance, is presented in Table 4.

Generalized Image Formation Model (GIFM) is a framework in computer vision and image processing that describes the process of capturing and reconstructing images. It generalizes the traditional image-formation models, thus enabling a broader range of applications and accommodating various imaging modalities. Liang et al. [1] objective was to rebuild images shot in low-visibility conditions using the GIFM. Using a machine learning-based approach, this tactic integrated domain information relevant to creating images in challenging conditions, such as fog and dim illumination. The recommended method worked well on datasets with poor visibility, successfully enhancing image clarity. However, its applicability to a greater range of vision problems was restricted by its inability to adapt to significant variations in lighting conditions.

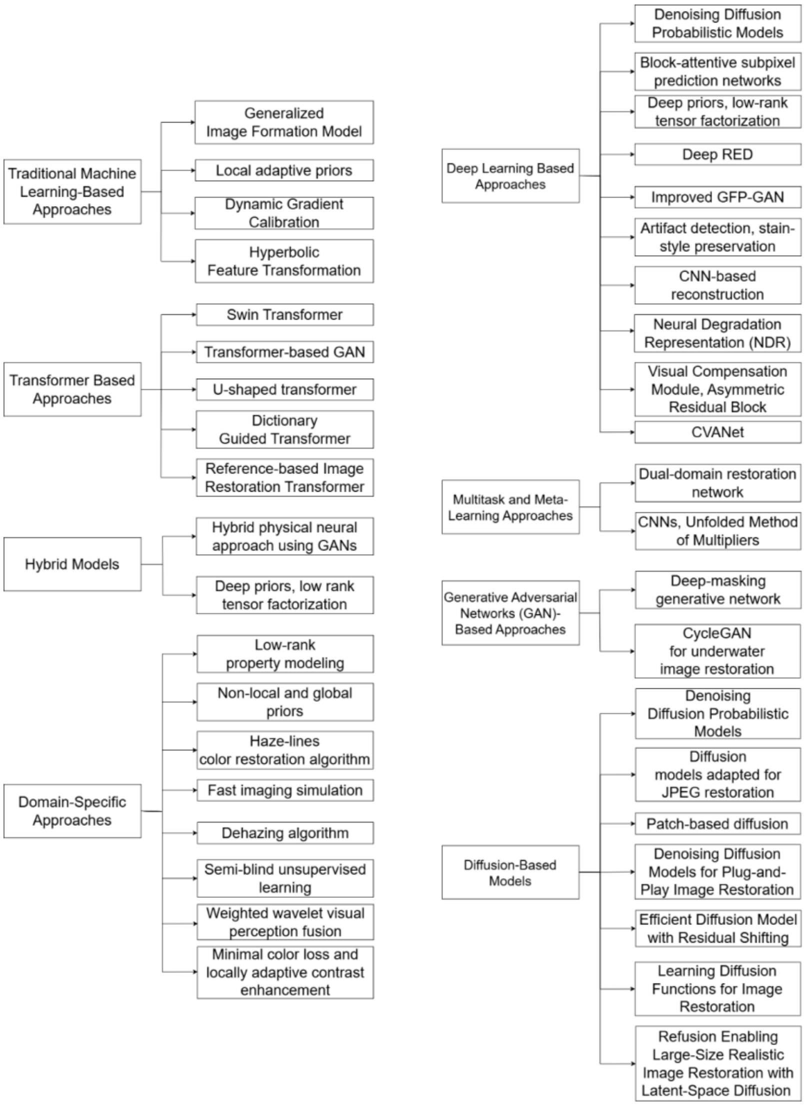

Traditional Machine Learning-Based Approaches
Generalized Image Formation Model
Local adaptive priors
Dynamic Gradient Calibration
Hyperbolic Feature Transformation
Transformer Based Approaches
Swin Transformer
Transformer-based GAN
U-shaped transformer
Dictionary Guided Transformer
Reference-based Image Restoration Transformer
Hybrid Models
Hybrid physical neural approach using GANs
Deep priors, low rank tensor factorization
Domain-Specific Approaches
Low-rank property modeling
Non-local and global priors
Haze-lines color restoration algorithm
Fast imaging simulation
Dehazing algorithm
Semi-blind unsupervised learning
Weighted wavelet visual perception fusion
Minimal color loss and locally adaptive contrast enhancement
Deep Learning Based Approaches
Denoising Diffusion Probabilistic Models
Block-attentive subpixel prediction networks
Deep priors, low-rank tensor factorization
Deep RED
Improved GFP-GAN
Artifact detection, stain-style preservation
CNN-based reconstruction
Neural Degradation Representation (NDR)
Visual Compensation Module, Asymmetric Residual Block
CVANet
Multitask and Meta-Learning Approaches
Dual-domain restoration network
CNNs, Unfolded Method of Multipliers
Generative Adversarial Networks (GAN)-Based Approaches
Deep-masking generative network
CycleGAN for underwater image restoration
Diffusion-Based Models
Denoising Diffusion Probabilistic Models
Diffusion models adapted for JPEG restoration
Patch-based diffusion
Denoising Diffusion Models for Plug-and-Play Image Restoration
Efficient Diffusion Model with Residual Shifting
Learning Diffusion Functions for Image Restoration
Refusion Enabling Large-Size Realistic Image Restoration with Latent-Space Diffusion

Figure 1. Taxonomy diagram for image restoration.

Local Adaptive Prior-based Image Restoration (LAPIR) is a technique in image processing that deals with the process of recovering or enhancing images by using prior knowledge about the content of the image, particularly in areas degraded or noisy. Jiang et al. [2] introduced a LAPIR technique specifically tailored for space diffraction imaging systems. Using priors that adapt to the specifics of the diffraction process, the technique effectively reduced noise and enhanced the restoration of features in diffraction-distorted images. This technique, which focused on the distinctive properties of space diffraction, significantly improved image quality and was particularly helpful in fields where effects of diffraction are frequent, such as astronomy and remote sensing. The algorithm fared better than traditional methods in a number of instances when it came to recovering structural information. Its use in fixing other image issues, such as motion blur or general low-light photography, was constrained by its difficulty in generalizing beyond its original application due to its uniqueness to spatial diffraction.

Yang et al. [3] formalized a generic Image Restoration framework for Visual Recognition (IRVR) designed to facilitate holistic semantic recovery across various high-level tasks in image restoration. To improve generalization, they maximized semantic recovery during the training of IR models and used image regression as an additional regularization term. For compatibility with any potentially unseen recognition models, they adjusted the gradient of the primary objective with the regularization gradient. The IRVR was recognition-agnostic and integrated as a plug-and-play module into existing IR techniques without adding computational cost at inference time. Through extensive experiments, Yang et al. [3] demonstrated IRVR's effectiveness and its strong generalization across different downstream high-level tasks. This precise recovery of intrinsic semantic details proved critical for advanced machine analysis, ensuring integrity and authenticity in multi-media content.

Recent breakthroughs in the restoration of aged photos have improved greatly by generative networks, while the restoration quality still remains heavily affected by the latent space properties, which captures the necessary semantic information essential for successful recovery. To resolve this problem, Chen et al. [4] developed a new generative network that uses hyperbolic embeddings to regenerate old photos affected by multiple degradations. For further improving hierarchical representational capability, the intermediate hyperbolic features were processed with channel mixing and group convolutions. Furthermore, an attention-based aggregation mechanism in hyperbolic space was employed; this enabled the latent vectors to capture important semantic factors that contribute to higher-quality restoration. A diversity loss function was also defined for steering each latent vector toward the disentangling of semantically different aspects. Extensive experiments showed that this method outperforms existing restoration techniques with visually pleasing results even for complex degradations.

Table 4. Comparison of traditional image-restoration methods.

| Study | Methods/ Algorithms Used | Dataset Used | Accuracy/ Performance | Limitations |
|---|---|---|---|---|
| [1], 2022 | Generalized Image Formation Model | Poor-visibility datasets | PSNR: 17.198,<br>SSIM: 0.565,<br>CIEDE: 14.402 | Struggles with extreme lighting variations |
| [2], 2023 | Local Adaptive Priors | Space-diffraction datasets | SSIM: 0.8503,<br>VIF: 0.6425,<br>SNR: 22.9858 | Limited generalization to other imaging systems |
| [3], 2024 | Dynamic Gradient Calibration, Intrinsic Semantic Consistency Constraint, Ground-truth Augmentation Strategy | CUB DATASET | PSNR: 29.94,<br>SSIM: 0.8892 | Limited testing on real-time applications, requires more exploration for model robustness in dynamic conditions |
| [4], 2024 | Hyperbolic Feature Transformation, Group-wise Feature Aggregation | TJU-OPR, FFHQ [54] | PSNR: 23.64,<br>SSIM: 0.8206,<br>LPIPS: 0.25,<br>FID: 13.175 | Sensitive to latent space selection, which affects stability in complex images; computational complexity due to hyperbolic transformations, challenging for large-scale or real- time applications |

## 2.2 Deep Learning-based Approaches

Deep-learning methods have revolutionized photo restoration by utilizing the capacity of neural networks to automatically recognize complex relationships and patterns in images. DL models can extract hierarchical representations from unprocessed image data, allowing for more complex and efficient restoration solutions than traditional machine-learning models that depend on manually created features. Deep learning is particularly well-suited to image-restoration applications, because CNNs efficiently maintain spatial information throughout feature-extraction layers. Table 5 presents a comparison of DL-based IR methods, outlining their key features and effectiveness. These techniques have greatly improved image restoration by removing the need for manually constructed features and allowing models to learn directly from data. Despite challenges with computational resources and data requirements, deep learning-based methods continue to advance, adopting innovations that increase their accuracy and adaptability across a variety of image-restoration applications. The generalized deep-learning architecture, depicted in Figure 2, highlights the key components and workflow of a typical neural-network model.

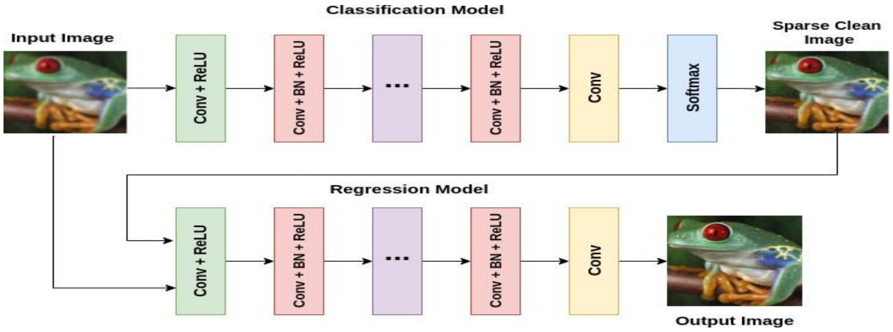


Figure 2. Overview of a generalized deep-learning framework.

Block-attentive subpixel-prediction networks (BASPNs) represent a neural-network architecture used for high-resolution image generation tasks and also in terms of performance, a high level of improvement for the case of sub-pixel image prediction. This network architecture is applied in the related applications, such as: image super-resolution, enhancement, or video-frame prediction. In place of full-resolution images, this paper introduced a novel family of networks known as Subpixel Prediction Networks (SPNs), which predict reshaped and spatially down-sampled block-wise tensors. This novel method significantly increased network speed by reducing the impact of spatial downsampling on restoration performance. Kim et al. [5] included a unique Subpixel Block Attention module which reduced discontinuities between blocks by recalibration of block-wise characteristics in order to further improve performance. Experimental results showed that these networks successfully matched computational efficiency and restoration quality in three important image restoration tasks: image augmentation, color-image denoising and image-compression artifact removal. This study demonstrated how SPNs can improve image-restoration procedures' speed and effectiveness.

Deep Residual Encoder-Decoder (RED) is a neural-network architecture for IR tasks, particularly deblurring, inpainting and denoising. It combines the deep-learning approach with the residual-learning technique to boost performance in recovering images from distorted or low-quality inputs. The Deep Unfolding Network (DUN) developed an effective framework for image restoration by combining a regularization module with a data-fitting module. In classic DUN models, which often used a DCNN for regularization, data fitting was done prior to regularization at each stage. The regularization module was positioned before the data-fitting module in the enhanced DUN that the authors of this study deployed. The Regularization by Denoising (RED) method served as the foundation for this regularization model, which included a recently developed DCNN. For the data-fitting part, Kong et al. [6] employed a closed- form method based on the Faster Fourier Transform (FFT). Among the many advantages of the proposed DRED-DUN model were its capacity to integrate the interpretability of RED with the adaptability of discovered image-adaptive regularization; its full end-to-end trainability, which allowed for cooperative regularization-network optimization with extra parameters; and its superior performance compared to both model-based and learning-based methods, as evidenced by higher PSNR values and better visual quality. Notably, this approach performed better than cut CNN-based Reconstruction, which refers to the use of CNNs for the task of reconstructing images or signals from incomplete or degraded data. Perdios et al. [7] allowed full-view frame capture at rates more than 1 kHz, ultrafast ultrasound (US) which has greatly improved biomedical imaging and paved the way for novel methods, such as shear-wave elastography. However, diffraction artifacts from sidelobes and grating lobes provide difficulties. Frame rates are decreased by the need for several acquisitions for sufficient image quality in traditional methods. A two-step image-reconstruction technique based on CNNs was developed for real-time imaging in order to address this issue. This method uses a residual CNN trained to eliminate diffraction artifacts to perform a high-quality restoration after beginning with a poor-quality estimation from a back-projection-based operation. The mean signed logarithmic absolute error was established as the training loss function to address the high dynamic range of radio frequency US images. Tests using a linear transducer array showed that this technique could achieve a dynamic range of more than 60 dB and rebuild images taken from single plane-wave acquisitions with quality on par with the best artificial aperture imaging.

Based on the free-energy principle, no-reference image-quality assessment techniques have attracted much attention and applied GANs recently. As a result, they achieve more accuracy in quality prediction than the former methods. However, most of the GAN-based methods can barely recover very poor-quality images, resulting in the broken relationship of distorted images and restored images between their quality reconstruction. To solve this problem, Pan et al. [8] proposed a VCRNet based on the non-adversarial model for better compensation of heavily distorted images. The innovations in this model would be a visual compensation module, an optimized asymmetric residual block and a mixed loss function based on error maps. All these further enhance the restoration capacity of the visual restoration network (VRN) by better handling the visual restorations. VCRNet further enhances the ability to accurately estimate the qualities of severely degraded images with multi-level restoration features coming from the VRN. Performing SOTA in all seven widely used IQA databases demonstrates the effectiveness of the proposed VCRNet for image-quality assessment.

These are concepts that have been used for various applications in machine learning, image processing and computer vision, including representation learning, denoising and image reconstruction. Deep priors are neural networks that serve as implicit priors in generative modeling. Low-rank tensor factorization is a mathematical method to decompose an array (tensor) that exists in multi-dimensional form into the sum of lower-dimensional tensors. This is very important to handle high-dimensional data by retaining significant structures and patterns. Zhang et al. [9] looked at the difficulties related to processing mixed noise pollution in hyperspectral images (HSIs). Although a number of approaches have been put out to address this problem, they typically fall into one of the following categories: model-driven or data-driven. Model-driven approaches were frequently criticized for being sensitive to changes in parameters and having high processing costs due to iterative optimization. However, data-driven techniques often performed poorly due to overfitting. This study suggested a unique approach to HSI restoration that blends low- rank tensor factorization (DP-LRTF) with deep denoising priors to get beyond these restrictions. Tucker tensor factorization was used to enforce global spectral low-rank requirements and two deep denoising priors were used to improve the spectral orthogonal basis and spatial reduced factor. This combined strategy effectively used the low-rank structure of HSIs and the powerful feature-extraction capabilities of deep learning. Experimental evaluations demonstrated that DP-LRTF significantly exceeded both model- driven and data-driven methods in terms of blended noise reduction and execution efficiency in a range of simulated and real-world scenarios.

Table 5. Comparison of deep learning image restoration methods.

| Study | Methods/ Algorithms Used | Dataset Used | Accuracy/ Performance | Limitations |
|---|---|---|---|---|
| [5], 2021 | Block-attentive sub-pixel prediction networks | Div2k dataset [56] | PSNR: 33.89,<br>SSIM: 0.934,<br>LPIPS: 0.0990 | Risk of overfitting |
| [6], 2021 | Deep RED (Regularization by Denoising) | Multiple benchmark datasets | PSNR: 35.98 | Limited to certain types of degradations |
| [7], 2021 | CNN-based reconstruction | Ultrasound imaging datasets | PSNR:14.23,<br>SSIM: 0.31 | Limited to specific ultrasound configurations |
| [8], 2022 | Visual Compensation Module, Asymmetric Residual Block, Error Map-based Mixed Loss Function | Seven representative IQA databases | SROCC: 0.973,<br>PLCC: 0.974 | May face challenges in cases of extreme degradation |
| [9], 2023 | Deep priors, low-rank tensor factorization | Hyperspectral datasets | PSNR: 32.943,<br>SSIM: 0.9704 | Computational complexity |
| [10], 2023 | Artifact detection, stain-style preservation | Histology datasets | PSNR: 26.37,<br>SSIM: 0.9359,<br>SRE: 56.31 | Limited generalizability |
| [14], 2024 | A DL-based super-resolution method that utilizes feature, channel and pixel attention mechanisms to enhance image details. | DIV2K [57], Set 5, Set 14, BSD100, Urban100 | PSNR: 38.19,<br>SSIM: 0.9613 | Computationally more expensive and the robustness of the complex network remains a challenge |
| [11], 2024 | Denoising Diffusion Probabilistic Models | CelebA-HQ [55] and FFHQ [54] | PSNR: 33.2055,<br>SSIM: 0.8662,<br>LPIPS: 0.0966 | Slow Convergence |
| [12], 2024 | Improved GFP-GAN | Miner face dataset | PSNR:26.1061,<br>SSIM:0.7236,<br>LPIPS: 0.3827,<br>FID: 46.51 | Specific to miner face images |
| [13], 2024 | Neural Degradation Representation (NDR), Degradation Query and Injection Modules, Bidirectional Optimization Strategy | BSD68, UR-BAN100 | PSNR: 26.02,<br>SSIM: 0.8657 | Potential complexity in handling highly heterogeneous degradations |

Artifact detection and Stain-style preservation are two closely associated ideas found in the broad category of image processing and computer vision, often utilized specifically within medical imaging and digital art, as well as within the area of image restoration. Artifact detection consists of detecting distortions or errors within an image not existing in the scene under photo. Ke et al. [10] addressed these artifacts through manual quality control, where the level of automation in image analysis is significantly reduced. By detecting and fixing artifacts, a systematic pre-processing technique was proposed to bridge this gap and lessen its impact on subsequent AI diagnostic tasks. At first, the AR-Classifier artefact-detection network distinguished between normal tissues and common artifacts, such as out-of-focus regions, spots, marking dye, tattoo pigment and tissue folds. It also categorized artifact fixes based on how restorable they were. Then, in an effort to preserve tissue architecture and stain styles, the AR-CycleGAN artifact restoration network performed de-artifact processing. A standard for performance evaluation was built using both publicly available datasets of breast and colorectal cancer and clinically gathered whole slide images. The functional structures of the suggested method were rigorously evaluated across multiple metrics in a variety of tasks, including artifact restoration and classification, as well as downstream diagnostic tasks, like tumor classification and cell segmentation.

DCNNs demonstrated remarkable capabilities in feature extraction and detail reconstruction for single-image super-resolution (SISR). However, previous DCNN-based approaches often failed to fully leverage the complementary strengths among feature maps, channels and pixels, which limited their ability to capture rich image details. To address these challenges, Zhang et al. [14] introduced a Cascaded Visual Attention Network (CVANet). This network was designed to mimic the human visual-attention mechanism to enhance detail reconstruction. The proposed approach incorporated three key modules: a Feature Attention Module (FAM) for feature-level attention learning, a Channel Attention Module (CAM) to strengthen feature maps through channel-level attention and a Pixel Attention Module (PAM) that adaptively selected representative features from previous layers to generate a high-resolution output. By effectively exploring feature-representation capabilities and human visual-perception properties, CVANet significantly improved image resolution. Experimental evaluations on four benchmark datasets demonstrated that CVANet outperformed SOTA methods in terms of subjective visual perception, PSNR and SSIM.

Denoising Diffusion Probabilistic Models (DDPMs) is a class of generative models that have gained acceptance for their ability to provide high-quality images and successfully perform various tasks in the field of computer vision, especially in image generation and in painting as well as denoising. Pang et al. [11] examined a facial image-restoration technique that made use of a pre-trained unconditional DDPM model in order to offer more flexible restoration procedures. The overall quality of the restored photos was found to suffer from low iterations throughout the resampling process. The study suggested an optimization technique for the inversion process that combined continuous sampling and sample scheduling in order to lessen this problem and improve image quality. The suggested strategy outperformed current techniques in facial image restoration, according to extensive testing utilizing the CelebA-HQ [55] and FFHQ datasets [54]. In terms of LPIPS and PSNR measures, the outcomes showed an excellent performance. Additionally, face-recognition accuracy improved by 15.7% when photos were restored using random masks and by a significant 26% when images were restored using central masks.

Improved GFP-GAN is an advanced model of the original GFP-GAN model, designed with the intention of high-quality facial image generation and restoration. GFP-GAN pays special attention to generating more realistic human faces while also preserving details and improving quality. A New Blind Restoration Approach for Miner Face Images Utilizes an Enhanced GFP-GAN Model. The challenges presented by miner face images, which are crucial for information exchange and for the digital transformation and astute management of mining firms, were addressed. To solve the issues of complex degradation variables including noise, blurring and low resolution, Zhang et al. [12] proposed a blind-restoration model built on an improved GFP-GAN. This concept attempted to achieve a balance between integrity and authenticity throughout the repair process. The authors successfully removed the complex degeneration from the miner face photos by first integrating a UNet++ network, using the pre- trained StyleGAN2 network as a source of previous knowledge. To improve the use of previous features from the pre-training network, they also added a channel-attention technique to the channel-split spatial feature-transform layer. This method allowed the miner face photographs to more accurately and authentically portray their end result. According to experimental data, the suggested approach significantly outperformed competing model methods in terms of reconstructing miner face photos.

At present, the conventional techniques used in the restoration of images are adequate for only one type of degradation in the image. However, in real-time applications, the nature of degradation varies and is mostly unknown. This mismatch may lead to a considerable drop in the performance of the model under consideration. With the motivation to overcome the mentioned problem, Yao et al. [13] came up with the all-in-one image-restoration network for managing multiple degradation types inside a single framework. The core of this approach is a neural degradation representation (NDR), which captures the unique characteristics of different degradation types. The NDR acts like a neural dictionary, which can adaptively decompose various degradations into fundamental components and allows the network to generalize across multiple degradation types. The authors introduced a degradation-query module and a degradation-injection module to utilize the NDR in order to approximate and inject the specific degradation patterns according to the learned representation, thus allowing the network to handle diverse degradations in a unified way. Moreover, it makes use of two-way optimization strategy: It actually degrades and reconstructs in the process, one after another, in order to enhance the degradation representation.

## 2.3 Transformer-based Models

Recently, transformer-based models have emerged as highly successful image-restoration techniques by leveraging the ability to detect local and global dependencies in image data. First developed for applications in natural-language processing, transformers established the self-attention mechanism that enables them to look at data in their entirety by evaluating the significance of different input elements. Since transformers can understand more complex patterns in images, as well as contextual links across the entire image, they tend to perform well in restoring images compared to traditional CNNs that basically rely on localized features. Even though the transformer-based models for image restoration are still nascent, there is definitely a lot of room to improve, because they can fit well and also capture high-order correlations in images. Such transformers are bound to be part of future SOTA image-restoration systems, not to mention the critical components of the systems, since efficiency gains and architectural advances are ongoing. Figure 3 illustrates the structure of a Transformer-based model, which processes information by focusing on different parts of the input data. This approach allowed the model to understand relationships between elements efficiently.

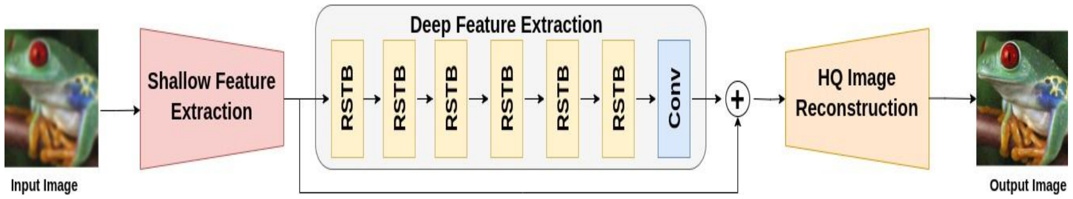


Figure 3. Overview of a generalized transformer-based model.

SwinIR includes three main parts, which are shallow feature extraction, deep-feature extraction and high-quality image reconstruction. Deep Feature Extraction combines several residual Swin Transformer blocks, having several layers of Swin Transformers combined with residual connections. Liang et al. [15] tested the model on three sample tasks: image super-resolution (including classical, lightweight and real-world scenarios), image denoising (including both grayscale and color images) and JPEG compression artifact removal. Experimental results demonstrate that SwinIR outperforms SOTA approaches in performance by 0.14 to 0.45 dB, while also achieving a decrease of up to 67% in the total number of parameters.

RFormer (Reconstruction Transformer) combines reconstruction tasks with transformer architectures for applications in image processing and computer vision. This method leverages the transformers' ability to capture long-range dependencies and learn complex data patterns in applications, such as image inpainting, noise removal and super-resolution. Deng et al. [16] presented this approach coupled with a new dataset, Real Fundus, consisting of 120 pairs of low and high-quality fundus images; this dataset focuses on addressing the difficulties of reconstructing clinical fundus images. Their contribution introduced a Transformer-based Generative Adversarial Network (GAN), which addresses real-world degradation in clinical fundus images. At the heart of this architecture is the Window-based Self-attention Block, which captures the long-range dependencies and the non-local self-similarity in an efficient manner. Furthermore, a Transformer-based discriminator was used to further improve the visual quality of reconstructed images. Experiments on the RF dataset showed that

RFormer performed substantially better than the SOTA methods.

Wang et al. [17] developed the Uformer, a Transformer-based architecture for image restoration which balances efficiency with effectiveness using Transformer blocks in forming a hierarchical encoder-decoder network. There are two novel designs, one is the locally-enhanced window Transformer block and the other one is a learnable multi-scale restoration modulator. LeWin Transformer block employs non-overlapping window-based self-attention to efficiently capture the local context without consuming considerable computation for high-resolution feature maps. Meanwhile, the multi-scale restoration modulator is, in fact, a kind of multi-scale spatial bias that refines features from the decoder's layers while enhancing detail restoration without computational overhead or significant increases in parameters. These improvements can help Uformer model essential dependencies for image restoration at different levels; namely, local and global. Thorough testing has been conducted on various tasks of restoration and as a result, Uformer has shown comparable or sometimes superior performance to SOTA methods while maintaining architectural simplicity.

The Under-Display Camera (UDC) allows users to realize an all-screen experience based on the placement of a camera below the display panel, but this setup heavily degrades the image quality due to the unique properties which affect the display, so restoration is really challenging. Although multiple solutions have been proposed toward dealing with the UDC image-restoration issue, there are yet no specific methods and databases used for restoring UDC face images, which happen to be a basic problem when taking into consideration UDC applications. In response to the same, Tan et al. [18] designed a two-stage network and named it UDC Degradation Model Network (UDC-DMNet). This simulates the color filtering effect, brightness attenuation and diffraction effects seen when using UDC imaging as it synthesizes UDC images. The authors developed dedicated UDC face training and testing datasets named FFHQ and CelebA- Test in aid of UDC face restoration by making use of UDC-DMNet in combination with good-quality face images; namely, from FFHQ [54] and CelebA-Test. They introduced a new kind of dictionary-guided transformer network known as DGFormer that comes up with facial component dictionary, with image characteristic accounting for the particular features of UDC image and can hence blindly recover a face related specifically to a UDC scenario. Experimental results show that the proposed DGFormer and UDC-DMNet have the SOTA performance in UDC image restoration.

Zhang et al. [19] introduced a multi-stage image-restoration (IR) approach for progressively restoring images with multiple degradations by transferring similar edges and textures from a reference image, referred to as the Reference-based Image Restoration Transformer (Ref-IRT). The proposed method operates in three stages. In the first stage, a cascaded U-Transformer network performs the preliminary recovery of the degraded image. This network comprises two U-Transformer architectures connected by feature-fusion layers at both encoder and decoder levels, enabling each U-Transformer to predict the residual image step-by-step, progressing from simple to complex and from coarse to fine toward complete recovery. The second and third stages aim to enhance the restoration quality by transferring textures from a reference image to the partially restored target image. To achieve accurate content and texture matching between the reference and target images, the authors propose a quality-degradation-restoration method. A texture-transfer and reconstruction network then maps these transferred features to generate the final high-quality output. By progressively refining degraded inputs, the method enhances restoration quality, particularly in cases involving severe distortions. This approach demonstrates effectiveness in handling complex degradations by incorporating contextual information from high-quality references. Experiments conducted on three benchmark datasets confirm the superior performance of Ref-IRT in comparison to other cutting-edge techniques for multi-degraded image restoration.

Table 6 provides a comparison of transformer-based models, highlighting their architectures and performance in image restoration.

## 2.4 Multi-task and Meta-learning Approaches

Recent image-restoration techniques have become popular due to multi-tasking and meta-learning techniques, because they offer solutions that can work based on shared data-related activities or quickly adapt novel restoration settings. They overcome the pitfalls with single-task models, which struggle most of the time not to generalize across other degradations and various context-specific types of images by allowing the model to learn common representations of improvements in performance in various tasks. All things considered, the multi-task and meta-learning techniques have enlarged the scope of image restoration in that they provide tools for not only improving the performance on known tasks, but also allowing models to be well equipped in coping with new and challenging degradation conditions. Further development is expected to advance resilient and flexible models in image restoration that could solve a range of dynamic problems of image deterioration. A detailed comparison of multi-task and meta-learning approaches, showcasing their key strategies and performance in image restoration, is provided in Table 7.

Table 6. Comparison of transformer-based models.

| Study | Methods/ Algorithms Used | Dataset Used | Accuracy/ Performance | Limitations |
|---|---|---|---|---|
| [15], 2021 | Swin Transformer for image restoration | Classic5 [58], LIVE1 [59], Flickr2K | PSNR: 34.52, SSIM: 0.908 | High resource consumption |
| [16], 2022 | Transformer-based GAN | Fundus clinical dataset | PSNR: 28.38, SSIM: 0.873 | Specific to fundus images |
| [17], 2022 | U-shaped transformer, Window-based Self-attention, hierarchical encoder-decoder structure, skip connections | SIDD (Smartphone Image Denoising Dataset) [60], GoPro Dataset, DIV2K Dataset [56] | PSNR: 26.28 , | High computational cost due to transformer-based architecture |
| [18], 2023 | DGFormer, UDC-DMNet | FFHQ-P/T, CelebA-Test-P/T | PSNR: 38.35, SSIM: 0.9678, LPIPS: 0.0720 | Limited to UDC-specific scenarios |
| [19], 2024 | Reference-based Image Restoration Transformer (Ref- IRT), Cascaded U-Transformer Network, Texture Transfer and Reconstruction Network | CUFED5, WR_SR, XRIR | PSNR: 28.893, SSIM: 0.905, LPIPS: 0.421 | Requires reference image for optimal restoration |

Table 7. Comparison of multi-task and Meta-learning approaches.

| Study | Methods/ Algorithms Used | Dataset Used | Accuracy/ Performance | Limitations |
|---|---|---|---|---|
| [20], 2022 | Dual-domain restoration network | CT and low-dose imaging datasets | PSNR: 42.03, SSIM: 0.966, RMSE: 20.18 | Specific to CT and low-dose images |
| [21], 2022 | CNNs, Unfolded of Multi-method pliers | Multispectral datasets are used | PSNR: 36.47, SSIM: 0.9873 | Increased complexity and computation limited |

DuDoUFNet is a specialized DL model that is set to solve the challenges of image restoration by progressively reconstructing images in a dual-domain framework. It is a dual domain under-to-fully complete progressive restoration network which was created in this work with the goal of combining low-dose computed tomography (LDCT) with metal artifact removal (MAR). Due to the increasing use of low-dose computed tomography (LDCT) to reduce radiation exposure in patients, image quality is frequently compromised by noise, particularly in cases where patients have metallic implants. This can lead to extra streak artifacts and increased noise, which can impair medical diagnoses and related applications. The main emphasis of previous studies was either full-dose CT MAR or denoising LDCT images without considering the effect of metallic implants. Reconstructions from MARLD may not be as good as they may be if conventional MAR or LDCT methods are used. Zhou et al. [20] used a two-stage progressive restoration network to effectively restore from the sinogram to the image domain while drastically lowering noise and artifacts.

Marivani et al. [21] examined MIR and fusion by framing the problem as a linked convolutional sparse coding challenge, employing the Method of Multipliers (MM) for resolution. The MM-based strategy drove the building of a CNN encoder, relying on the concepts of deep unfolding. Marivani et al. [21] suggested two multimodal models that combined the specified encoder, followed by a customized decoder that transformed the learned representations into the appropriate output. Unlike most current deep learning techniques, which often featured several encoding branches blended by concatenation or linear combination, this technique offered a more efficient and systematic approach for fusing input at various stages of the network. This method resulted in representations that permitted accurate image reconstruction. Marivani et al. [21] evaluated the models on three image-restoration domains and two image-fusion domains. Those quantitative and qualitative comparisons with other SOTA analytical and deep-learning techniques further emphasized that the proposed framework outperforms.

## 2.5 Generative Adversarial Network (GAN)-based Approaches

GAN is a strong image-restoration technology that emerges from a novel framework, which is capable of generating realistic images from the damaged input. They are composed of two neural networks, the discriminator and the generator. GANs have been trained against each other. The goal of the generator is to generate high-quality restored images. The discriminator judges the realism of the generated images by distinguishing between made and genuine outputs. The adversarial structure of GANs makes them particularly effective in restoration tasks that preserve the texture and details of the original image. The generator has to produce images that look more realistic. GAN-based approaches focus on the generation of images that should have realistic textures and structural features and hence these are useful for applications where perceptual quality is the requirement. They are very effective whenever the classical models fail, as for example, in extreme noise reduction or super-resolution at very high levels, they have the capacity to learn complex distributions. Figure 4 illustrates the structure of a GAN, which consists of two components—a generator and a discriminator. The generator creates data samples, while the discriminator evaluates their authenticity, enabling the model to generate high-quality outputs through an adversarial training process.

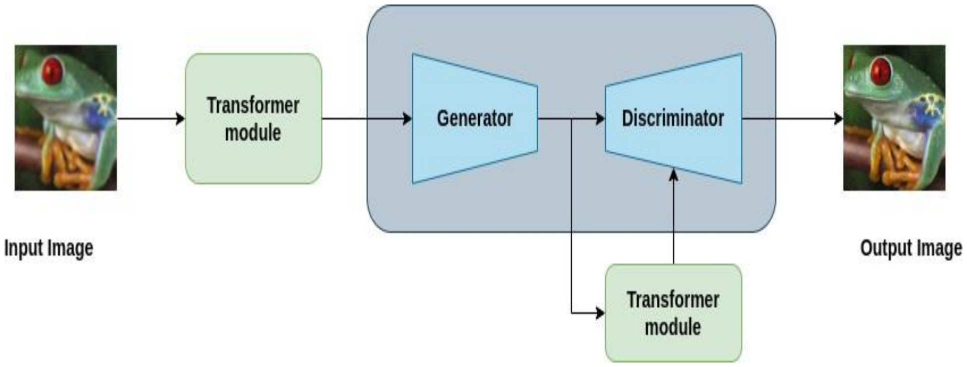


Figure 4. Generalized representation of a GAN architecture.

Deep-Masking Generative Network (DMGN) is specifically a deep-learning model set up for numerous image-generation and restoration applications, typically where selective masking of various parts of the images under consideration is involved. It is a unified technique for recovering backgrounds from images that have been superimposed. Feng et al. [22] unified framework for background restoration from overlain images that successfully handles different kinds of noise—the DMGN—was presented. The generative technique used by the DMGN is coarse-to-fine. It starts by producing a noise image and a coarse background image simultaneously. The background image is then improved in quality by using the noise image for refinement. The unique Residual Deep-Masking Cell, which enhances the extraction of pertinent information while reducing noise using a learnt gating mask that regulates information flow, lies at the heart of the DMGN. The DMGN gradually produces noisy images and high-quality background images by repeatedly applying this cell. To help with backdrop refining, a two-pronged approach is also used to take use of the created noise image as contrasted signals. Extensive tests on three challenges (image dehazing, image reflection removal and rain streak removal) showed that the DMGN consistently beats the SOTA techniques customized for each particular job.

UW-CycleGAN refers to the advanced version of the CycleGAN model which has been designed specifically with unsupervised image-to-image translation tasks in the forefront. It uses transformations involving wavelets to refine traditional CycleGAN architectures on their performance. Yan et al. [23] suggested Model-driven and cycle-consistent generative adversarial network (CycleGAN) which is inspired by the underwater image-creation model. Targeting the transmission map, scene depth, attenuation coefficient and background light directly was the goal of the model. Extensive trials proved that this technique produced recovered photographs with improved color saturation and brightness, surpassing other approaches for restoring underwater images in both quantitative and qualitative aspects. The efficiency of the CycleGAN in enhancing detection accuracy was further shown by research on underwater item detection.

Table 8 presents an overview of GAN-based approaches, emphasizing their methodologies and effectiveness in image restoration.

Table 8. Comparison of GAN-based approaches.

| Study | Methods/ Algorithms Used | Dataset Used | Accuracy/ Performance | Limitations |
|---|---|---|---|---|
| [22], 2021 | Deep-masking generative network | PLNet [61] | PSNR:23.05, SSIM:0.823 | Struggles with complex occlusions |
| [23], 2023 | CycleGAN for underwater image restoration | NYU-V2 dataset [62] | PSNR: 21.14, SSIM: 0.83 , UIQM: 2.24, CCF: 50.10 | Struggles with complex underwater conditions |

## 2.6 Diffusion-based Models

Diffusion-based models have been the latest advancement in image restoration. Image models can enhance damaged images with probabilistic-modeling simulation of physical-diffusion processes iteratively. Such models work to revert a noisy image to its original state through progressive and reversible procedures that borrow inspirations from concepts, such as image denoising and noise diffusion. Unlike the rest of the restoration algorithms that predict restored results almost instantaneously, diffusion models learn complicated noise patterns and gradually eliminate them to produce images with a very fine degree of control over the restoration process. This is novel in the context of diffusion-based models and with it, the future prospects are promising in image restoration. As the techniques for more efficient computations and faster sampling continue to advance, it can be anticipated that diffusion-based models will increasingly be integrated into flexible frameworks of high-fidelity image restoration toward wide applications requiring high quality and flexibility over many types of degradations. Figure 5 depictes the framework of a diffusion model, where data is gradually transformed into noise in a forward process and then reconstructed in the reverse process.

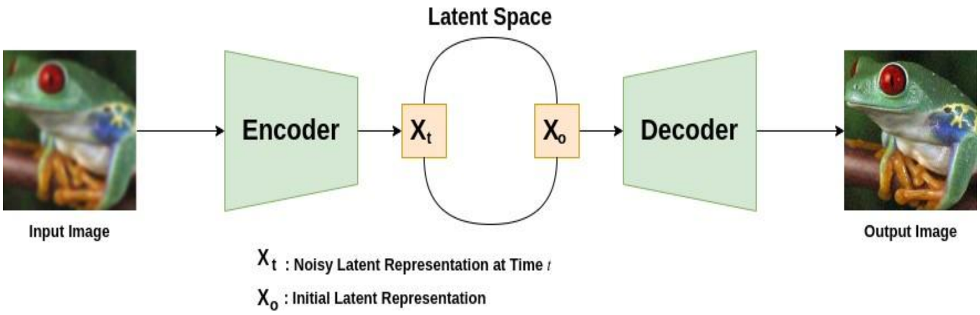

Latent Space
Encoder
$X_t$
$X_o$
Decoder
Input Image
Output Image
$X_t$ : Noisy Latent Representation at Time $t$
$X_o$ : Initial Latent Representation

Figure 5. Architecture of a diffusion-based generative model.

Plug-and-Play IR covered a well-established and flexible way to solve inverse problems by making use of pre-trained denoisers as implicit image priors. Most current methods were discriminative Gaussian denoisers. However, there was no research on diffusion models as generative denoiser priors. Although some research incorporated diffusion models into image restoration, they had either sub-optimal performance or needed an extreme number of Neural Function Evaluations (NFEs) during inference. To overcome such limitations, Zhu et al. [24] introduced DiffPIR that combined the plug-and-play scheme with the diffusion sampling. In contrast to traditional plug-and-play IR schemes relying on Gaussian denoisers, DiffPIR utilized the generation ability of diffusion models to produce better image restoration. The scheme was tested on three prominent IR tasks; i.e., super-resolution, deblurring and inpainting. Experimental performance on the FFHQ and ImageNet benchmarks confirmed that DiffPIR achieved SOTA reconstruction accuracy as well as visual quality and, at the same time, kept an inference process within 100 NFEs.

Luo et al. [27] aimed to enhance the usability of diffusion models in IR by optimizing important factors, like network architecture, noise intensity, denoising steps, training image size and optimization methods. Luo et al. [27] introduced Refusion, a U-Net-based latent diffusion model, that performed diffusion in a low-resolution latent space with high-resolution details left for decoding. In contrast to other latent diffusion models that employed VAE-GAN for compression, their model was more stable and produced very precise reconstructions without adversarial training. Such improvement enabled the model to efficiently handle various image-restoration tasks like real-world shadow removal, high-resolution non-homogeneous dehazing, stereo super-resolution and bokeh-effect conversion. Refusion was shown to handle large-scale images (e.g. 6000×4000×3 in high-resolution dehazing) without sacrificing robust performance on various restoration tasks. Notably, it achieved the best perceptual performance in the NTIRE 2023 Image Shadow Removal Challenge.

Ortega et al. [28] analyzed the role of anisotropic-diffusion models in image restoration and emphasized the role of the diffusion function, where, traditionally, this diffusion function was fixed as part of the classical approach. The idea is introduced on learning this function dynamically using either a Fields of Experts (FoE) or a U-Net, demonstrating that their approach outperformed conventional and SOTA models with some numerical experiments. Ortega et al. [28] Perona-Malik model combined with machine-learning techniques was leveraged to directly learn an optimized diffusion function from data. By combining classical approaches with data-driven methods, a balance between interpretability in a mathematical sense and improved restoration was achieved. By demonstrating generalization to a host of image-restoration tasks, this approach offered the possibility of offering a more stable and effective replacement for purely deep learning-based models, such as blind denoising.

Although diffusion-based IR techniques have shown impressive results, their poor inference speeds—which required hundreds or even thousands of sample steps—hampered their applicability. Current acceleration methods tried to expedite this process, but they frequently resulted in performance issues and very blurry restored photos. In order to overcome this restriction, Yue et al. [29] put out an effective IR diffusion model that greatly decreased the number of necessary diffusion steps without sacrificing image quality. By doing away with the requirement for post-acceleration during inference, their method prevented performance deterioration. By modifying their residuals, they specifically created a Markov chain to ease the transitions between high- and low-quality images, significantly increasing transition efficiency. Furthermore, in order to regulate the noise strength and the varying speed during the diffusion process, they created a noise schedule. According to experimental assessments, the suggested method only required four sample steps and performed better than or on par with SOTA methods in four important IR tasks: image high resolution, inpainting, blind facial restoration and deblurring.

In order to increase versatility in face-image restoration, Pang et al. [11] created a method using DDPM and made use of an unbiased DDPM model that had already been trained. Pang et al. [11] found that the quality of the recovered photos suffered when there were not as many iterations in the resampling procedure. An optimization strategy for the inversion process was put out to address this problem and produce better restoration quality by combining sample scheduling with progressive sampling. Numerous tests with the CelebA-HQ[55] and FFHQ datasets[54] showed that their approach outperformed other methods in face-image restoration. It performed outstandingly in terms of LPIPS and PSNR measurements, specifically. Additionally, the restoration method increased the accuracy of detection of faces by 15.7% for facial photos with random masks and by 26% for images with central masks.

Welker et al. [25] tackled the problem of blind JPEG restoration at high compression levels by leveraging the high-fidelity generating capabilities of diffusion models. S. Welker et al. [25] named their approach DriftRec and suggested a change to the forward stochastic differential equation in diffusion models. DriftRec successfully avoided the blurriness typical in other approaches and substantially better restored the distribution of clean images, as evidenced by a comparative study against an L2 regression baseline using the same network design and cutting-edge JPEG restoration techniques. This method's applicability to different restoration jobs is increased, because it merely needed a dataset of clean/corrupted image pairings and did not require any prior knowledge of the corruption process. DriftRec took use of the closeness of both clean and damaged image distributions, which are far closer to one another than they are to the usual Gaussian prior utilized in diffusion models, in contrast to other conditional and unconditional diffusion models. Because of this, even in the absence of additional improvements, it only required small amounts of extra noise and fewer sample steps. Despite not being trained on instances of this nature, the study demonstrated that DriftRec extended well to difficult circumstances, including unaligned double JPEG compression and blind restoration of JPEGs received from the internet.

PD-CR is patch-based diffusion with constrained refinement that uses the diffusion processes to improve images through refinement of local patches. The method introduced by Cho et al. [26] improved the noise estimates that were produced by patch-based diffusion models so that the restored image, with maintained brightness of the damaged input image, could be given. In the proposed method, patch- based diffusion models were applied to efficiently address high-resolution photos with minimal memory usage. The experimental results indicated that the proposed method was superior to existing leading-edge approaches in various image-restoration tasks, which included image denoising and raindrop removal.

Table 9 presents a comparison of diffusion-based models, outlining their methodologies, datasets, accuracy and key limitations in image restoration.

Table 9. Comparison of diffusion-based models.

| Study | Methods/ Algorithms Used | Dataset Used | Accuracy/ Performance | Limitations |
|---|---|---|---|---|
| [24], 2023 | Denoising Diffusion Models for Plug-and-Play Image Restoration | FFHQ[54], ImageNet | PSNR: 31.01, LPIPS: 0.152 | High computational cost and slow inference due to iterative denoising steps |
| [27], 2023 | Enabling Large-Size Realistic Image Restoration with Latent-Space Diffusion Models | Flickr1024 | PSNR: 21.88, SSIM: 0.6977, LPIPS: 0.121 | Limited generalization to diverse degradations and training stability challenges. |
| [28], 2024 | Learning Diffusion Functions for Image Restoration | BSD500 [63] | PSNR: 29.5, SSIM: 0.83, LPIPS :0.15 | High computational cost |
| [29], 2024 | Efficient Diffusion Model for Image Restoration by Residual Shifting | RealSR-V3, Re- alSet80 | PSNR: 25.02, SSIM: 0.6833, LPIPS: 0.2076 | Limited generalization to unseen noise types |
| [11], 2024 | Denoising Diffusion Probabilistic Models | CelebA-HQ [55] and FFHQ [54] | PSNR: 33.2055, SSIM: 0.8662, LPIPS: 0.0966 | Slow convergence |
| [25], 2024 | Diffusion models adapted for JPEG restoration | JPEG image datasets | PSNR: 25.78, SSIM: 0.73, FID: 29.7 | Limited to JPEG artifacts |
| [26], 2024 | Patch-based diffusion | SIDD [60] and Raindrop dataset [64] | PSNR: 38.21, SSIM: 0.901, LPIPS: 0.134, NIQE: 13.72 | Edge artifacts |

## 2.7 Hybrid Models

Hybrid models in image restoration exploited the advantages of multiple methods, including generative models, deep learning and traditional machine learning for added performance and adaptability. Hybrid models can be constructed by combining the global contextual understanding of transformers, localizing the feature-extraction capacities of CNNs to specific interest regions and the qualities of image-creation realism as provided by diffusion models or GANs. This synergy allows for stronger image restoration solutions that can handle a variety of degradation types and challenging restoration tasks. Future research on hybrid models is likely to focus on more effective structures that balance performance with complexity. Innovations, such as attention processing, adaptive feature extraction and knowledge distillation, may enhance the effectiveness of hybrid methods even further. Hybrid models are expected to be pivotal for optimal performance in many image-restoration tasks as the current developments progress. Table 10 compares various hybrid approaches, highlighting their combined methodologies, performance across datasets, accuracy and associated limitations in image restoration.

Hybrid Unfolding Reconstruction (HybrUR) is a deep-learning model aimed towards image-reconstruction tasks in MRI and other imaging modalities. It combines elements of traditional image-reconstruction techniques and modern deep-learning methods with the goal of improving image-restoration quality and efficiency. Yan et al. [30] provided an unsupervised framework for underwater photo restoration using unpaired underwater and airborne photos, based on data and physics. To improve image quality and perform effective colour correction, an explicit degeneration model of underwater photos was developed using well-established optical-physics concepts. The loss of underwater vision was modelled using neural networks and a generator based on the Jaffe-McGlamery degeneration theory was created. The scene depth and degeneration factors for backscattering estimate were additionally physically restricted in order to solve the vanishing-gradient problem during hybrid physical-neural model training. The experimental results demonstrated that the proposed method successfully restored high-quality unmanaged underwater photographs without supervision. On many benchmarks, their technique outperformed several cutting- edge supervised and unsupervised algorithms, indicating that it will perform well in real-world situations.

Table 10. Comparison of Hybrid approaches.

| **Study** | **Methods/ Algorithms Used** | **Dataset Used** | **Accuracy/ Performance** | **Limitations** |
|---|---|---|---|---|
| [30], 2023 | Hybrid physical- neural approach using GANs | RUIE[65] | UICM: 5.142, UCIQE: 0.495 | Struggles with edge retention |
| [9], 2023 | Deep priors, low- rank tensor factorization | Hyperspectral datasets | PSNR: 32.943, SSIM: 0.9704 | Computational complexity |

Combined Deep Priors with Low-rank Tensor Factorization for Hyperspectral Image (HSI) Restoration is a novel approach designed to improve the quality of HSIs, which often suffer from noise, distortions and incomplete data. This method integrates deep-learning techniques with low-rank tensor factorization to effectively restore and reconstruct HSIs while preserving essential details and spectral information.

The global spectral low-rank criterion was represented by Tucker-tensor factorization in the proposed technique. Two deep denoising priors were then used to optimize the spectral orthogonal basis and the spatial reduction factor. With this combined approach, Zhang et al. [9] were able to benefit from the low-rank characteristics of HSIs and the potent feature-extraction capabilities of deep learning for HSI restoration. The DP-LRTF outperformed both model-driven and data-driven approaches in terms of execution efficiency and mixed-noise removal from HSIs in a number of simulated and real-world studies.

## 2.8 Domain-specific Approaches (Underwater, Hyperspectral, Remote Sensing, Medical Imaging, …etc.)

In image restoration, domain-specific approaches focus on adapting methods and algorithms to better adapt to the specific challenges each particular application domain has, such as medical imaging, remote sensing, underwater imaging and hyperspectral imaging. These domains often display characteristic degradation types and quality of restoration requirements that necessetate highly specialized procedures that capitalize upon the features of the domains and the type of images. Table 11 summarizes domain-specific approaches, focusing on their tailored methodologies, performance on specialized datasets and key limitations in image restoration.

Chang et al [31] explored the low-rank features across spatial, spectral and non-local self-similarity modes in hyperspectral images (HSIs), showing that the internal low-rank correlations within every mode affect restoration results to different extents. Their results identified the potential of spectral, along with non-local induced low-rank features toward HSI modeling, therefore resulting in the development of an optimal low-rank tensor (OLRT) model for improved HSI recovery. This work also investigated the existence of low-rank properties in both the image and sparse error parts, such as stripe noise in HSIs. Taking advantage of low-rank tensor priors for sparse errors and HSIs, OLRT developed into OLRT-robust principal-component analysis. The earlier methods were not versatile; they were often designed for specific HSI tasks, whereas the ideal low-rank prior was highly versatile across different HSI restoration applications. Thorough assessments on different benchmarks indicated that the proposed approaches significantly outperformed the SOTA methods.

The iterative model Non-local meets Global provides an all-inclusive approach for hyperspectral image (HSI) restoration, which is based on both non-local and global information for better quality. He et al. [32] proposed that the spectral sub-spaces of each full-band patch group align with the global spectral low-rank sub-space, which covers the whole HSI. This observation led to the development of a unified model for HSI restoration that integrates both spectral and spatial elements. The approach uses non-local spatial denoising and low-rank orthogonal basis exploration to streamline computational demands. The restoration process begins with updating the latent input image by resolving a fidelity term, followed by implementing an efficient alternating minimization method with adaptive-rank selection. It learns an orthogonal basis in the low-dimensional space and decreases the image representation. Re-iteration of non-local low-rank denoising refines restoration further. The experiments conducted on both the simulated as well as the real-world dataset show that the proposed approach achieves superior performance compared to any existing SOTA HSI restoration techniques.

The low contrast and color distortion in underwater photographs brought on by wavelength-dependent light attenuation were the subjects of this investigation. Color restoration is more difficult with underwater images than with terrestrial ones due to the different attenuation across wavelengths that depends on the water body and the three-dimensional structure of the scene. Berman et al. [33] proposed the method, which considered multiple spectral profiles of different types of water and reduced the problem to single-image dehazing by computing two global parameters: the attenuation ratios of the blue-red and blue-green channels. Because the type of water was unknown, a variety of characteristics from an existing library of water types were evaluated. The color distribution was utilized to automatically identify the optimal solution. The collection includes 57 underwater photographs taken in various locations; stereo photography was used to determine the 3D structure and color charts were applied to the scenes for ground truth.

Zhang et al. [37] addressed the problems of limited visibility and color aberrations in underwater photographs brought on by light scattering and absorption that varies with wavelength. Zhang et al. [37] developed MLLE, an effective and reliable technique for enhancing underwater images, to get around these problems. Using a maximum attenuation map-guided fusion technique and a minimum color-loss concept, they first locally altered an image’s color and features. In order to adaptively improve the image contrast, the mean and variance of local image blocks were then calculated using integral and squared integral maps. Furthermore, a color-balance technique was presented to rectify color discrepancies between CIELAB color space channels a and b. The improved photos had more contrast, vibrant colors and better detail retention. Three datasets for underwater-picture enhancement were used in extensive studies, which showed that MLLE performed better than the SOTA techniques. A single CPU could handle 1024 x 1024 × 3 photos in a single second, demonstrating the method’s computational efficiency. Further tests showed that the MLLE-realized improvement greatly enhanced saliency detection, keypoint recognition and underwater-picture segmentation.

Zhang et al. [38] focused on how light scattering and absorption degraded underwater image quality, making them less useful for analysis and applications. Zhang et al. [38] developed Weighted Wavelet Visual Perception Fusion (WWPF), an underwater image-augmentation technique, to address these problems. To fix color aberrations in underwater photos, they first used a color-correction technique guided by an attenuation map. To enhance the overall contrast, they then used a maximum information entropy optimized global contrast-augmentation approach. At the same time, localized details were enhanced using a quick integration optimized local contrast-enhancement technique. A WWPF technique was presented in order to integrate the advantages of both local and global contrast-enhanced images. High-quality underwater photos were created by fusing low-frequency and high-frequency components at various scales. Comprehensive tests on three benchmark datasets showed that WWPF performed better than current SOTA techniques in both qualitative and quantitative assessments.

Li et al. [34] proposed a fast simulation approach for image acquisition with the remote sensing TDI camera, employing image resampling to simulate degraded image qualities with high accuracy. This process considered various degradation factors, enabling the creation of a rather large dataset suitable for most modern supervised learning-based approaches to image restoration. Moreover, the work presented a new network architecture, containing a row-attention block and a row-encoder block, especially tailored to tackle row-variant blur and restore degraded images efficiently. The method was tested through real-world images and simulated degraded datasets with good experimental performances. In contrast to previously blind image-restoration techniques, the technique here showed superior results without resorting to multi-spectral bands or high-frequency sensor data.

Table 11. Domain-specific approaches.

| Study | Methods/ Algorithms Used | Dataset Used | Accuracy/ Performance | Limitations |
|---|---|---|---|---|
| [31], 2020 | Low-rank property modeling | Hyperspectral datasets | PSNR: 57.02,<br>SSIM: 0.9985,<br>SAM: 0.0216 | Computationally intensive |
| [32], 2020 | Non-local and global priors | CAVE dataset | PSNR: 3.03,<br>SSIM: 0.9807 | Computational complexity |
| [33], 2020 | Haze-lines, color restoration algorithm | New quantitative underwater dataset | PCC: 0.85 | Limited to specific underwater conditions |
| [37], 2022 | Minimal Color Loss and Locally Adaptive Contrast Enhancement (adaptive enhancement of contrast and color preservation) | UCCS, UIQS, UIEB | PCQI: 1.136,<br>UIQM: 5.293,<br>CCF: 46.872 | Cannot handle the underwater images acquired in low light conditions well |
| [38], 2023 | Weighted Wavelet Visual Perception Fusion (WWPF) using wavelet transform for multi-scale frequency decomposition and contrast enhancement | UCCS, UIQS, UIEB | UCIQE: 0.617,<br>AG: 10.818,<br>CCF: 40.851 | Cannot suppress image noise well |
| [34], 2023 | Fast imaging simulation, image resampling | Remote sensing datasets | PSNR: 30.970,<br>SSIM: 0.882 | Trade-off between Speed and Accuracy |
| [35], 2023 | Dehazing algorithm | Outdoor/remote sensing datasets | PSNR: 27.08,<br>SSIM : 0.94,<br>PI: 2.24 | Limited to specific atmospheric conditions |
| [36], 2024 | Semiblind unsupervised learning for co-phase errors | Optical synthetic aperture imaging datasets | PSNR: 25.72,<br>SSIM: 0.758 | Dependence on Phase Initialization |

An efficient image-dehazing method that works with both outdoor and remote-sensing photos is presented by Li et al. [35]. The plan combined the benefits of image enhancement and repair methods. To increase transmittance and fix errors in transmittance estimations reported in previous methods, the researchers employed Gaussian-weighted image fusion. After dehazing, color distortion was also corrected using an unsharp mask technique. The approach suggested by Li et al. [35] outperformed current dehazing techniques in the effective removal of haze from images, according to experimental results on both synthetic and real-world datasets. The solution outperformed other approaches with a PSNR of 27.08 and SSIM of 0.94 when applied to the RICE dataset.

Zhong et al. [36] introduced RPIR, a semi-blind, unsupervised learning technique for image restoration in OSAI systems with co-phase faults. Based on the traditional maximum a posteriori (MAP) model, RPIR used a multi-scale neural network that required no prior training. This network gathered input blur kernel flaws for use as residual priors in the MAP model. To solve the data and earlier terms, they employed alternating minimization. RPIR reduced erroneous blur kernels in OSAI systems due to co-phase error variations. The results indicated that RPIR considerably enhanced image resolution and clarity in treating co-phase faults in OSAI systems, exceeding other unsupervised deep-learning techniques and standard deconvolution methods.

Machine-learning models are frequently tailored to meet the unique requirements of specific domains, such as underwater imaging, hyperspectral analysis, remote sensing and medical imaging. While general- purpose architectures, such as CNNs, transformers, GANs and diffusion models, are frequently used in a variety of applications, applying them directly to domain-specific tasks may not necessarily produce the best results. These tasks frequently necessitate specialized architectures, task-specific loss functions and domain-aware pre-processing approaches to improve model performance. For example, underwater-image restoration requires models capable of correcting color aberrations and scattering effects, which are specific to aquatic environments. Likewise, medical-imaging models need to consider low contrast and anatomical features, necessitating domain-specific training procedures and domain-aware regularization methods. In remote sensing and hyperspectral imaging, models must maintain spectral fidelity and handle high-dimensional data efficiently. By adding such domain-specific adaptations, machine-learning models can perform much better than their general-purpose equivalents.

## 2.9 Comprehensive Image-restoration and Denoising Datasets

In recent years, a wide variety of datasets have been developed to benchmark the performance of image-restoration and denoising algorithms. These datasets vary in content, type of degradation and complexity, providing diverse scenarios for evaluating model effectiveness. Table 12 summarizes key datasets frequently used in image-restoration and denoising research, highlighting characteristics, such as dataset size, types of degradation (e.g. noise, blur, low resolution) and typical applications. This compilation serves as a foundation for comparing algorithm performance across different degradation scenarios and understanding the suitability of specific datasets for various restoration tasks.

Table 12. Comprehensive image-restoration and denoising datasets.

| Name of Dataset | Year | Brief Description |
|---|---|---|
| BSD500 [63] | 2012 | Part of the Berkeley Segmentation Dataset, containing 500 images used for enoising and segmentation. |
| URBAN100 | 2015 | Contains 100 high-resolution images of urban scenes, featuring buildings, streets and architectural structures. It is widely used in image super-resolution and restoration tasks to evaluate model performance on complex textures and fine details. |
| GoPro [66] | 2017 | Contains paired blurred and sharp images from GoPro cameras, used for motion-deblurring research. |
| DIV2K [57] | 2017 | High-quality dataset with 1,000 images for super-resolution and general image restoration, with multiple degradation levels. |
| SIDD [60] | 2018 | A dataset consisting of more than 30,000 noisy images under different lighting conditions, along with ground-truth images. |
| Color BSD68 [67] | 2018 | Part of Berkeley Segmentation Dataset and Benchmark, it contains 68 images for measuring image-denoising algorithms' performance. |
| PIRM [68] | 2018 | Comprises 200 diverse images divided for validation and testing, used for perceptual image-restoration tasks. |
| HAC [69] | 2019 | Contains 316K pairs show casing various weather conditions for testing restoration under adverse circumstances. |
| FFHQ [70] | 2019 | Contains 70000 high-quality face dataset from NVIDIA, used for inpainting and denoising, with diverse ages, ethnicities and lighting conditions. |
| SCISR [71] | 2019 | Synthetic and camera-based image super-resolution dataset, used for super-resolution in low-quality smartphone-captured images. Contains 50000+ images. |
| Hide [72] | 2019 | It consists of 8,422 blurry images and with them their corresponding image pairs with 65,784 densely annotated FG human bounding boxes. |
| Raindrop [73] | 2020 | A dataset containing 1,119 pairs of images, where one is degraded by raindrops and the other is clean. |
| UHDS [74] | 2022 | A dataset of 29,500 rain and rain-free image pairs covering various natural rain scenarios. |
| Sentinel-2 Satellite Images [75] | 2022 | Includes 3,740 pairs of overlapping image crops with cross-band and cross-detector parallax effects for analysis. |
| TinyPerson [76] | 2022 | A dataset focusing on tiny objects with 72,651 annotated images, collected from high-resolution videos. |
| LSDIR [77] | 2023 | A large-scale dataset containing 84,991 training images, 1,000 validation images and 1,000 test images. |
| HQ-50K [78] | 2023 | Introduces 50,000 high-quality images with rich textures for image-restoration applications. |

## 3. Evaluation Metrics Used for Image Restoration

Various evaluation metrics have been utilized in the literature survey to effectively evaluate performance and compare different image restoration algorithms. They are essential to objectively quantify how good image quality is after restoration, thus allowing researchers to compare different approaches in different degradation conditions. Table 13 provides a detailed summary of widely used evaluation metrics, highlighting their specific purposes and the aspects of image quality they focus on.

## 4. Results and Discussion

The evaluation of multiple image-restoration models was conducted on a high-performance hardware configuration comprising a Tesla T4 GPU equipped with 15,360 MB of VRAM, supported by NVIDIA-SMI 535.104.05, Driver Version 535.104.05 and CUDA Version 12.2. The hardware operated under optimal conditions, maintaining a stable temperature of 54°C. The models were evaluated using the SIDD (Smartphone Image Denoising Dataset) and DND (Darmstadt Noise Dataset), a benchmark dataset widely utilized for assessing image-restoration techniques. This experimental setup facilitated precise benchmarking of inference times and PSNR values, which are pivotal metrics for assessing the performance of pre-trained models in image-restoration tasks. The results, summarized in Table 14, provide a comprehensive comparison of the evaluated models.

Table 13. Summary of evaluation metrics for image restoration.

| **Evaluation Metric** | **Description** |
|---|---|
| Peak Signal to Noise Ratio (PSNR) | Calculates the ratio between maximum signal power and noise power. |
| Structural Similarity Index Measure (SSIM) | Evaluates similarity between two images based on brightness, contrast and structure. Values closer to 1 indicate higher similarity. |
| Perception-based Contrast Quality Index (PCQI) | Measures image quality by evaluating contrast and structural fidelity, considering human visual perception. |
| Underwater Image Quality Measure (UIQM) | A composite metric used to assess underwater image quality based on colorfulness (UICM), sharpness (UISM) and contrast (UIConM). Higher UIQM values indicate better visual quality. |
| Color Contrast Factor (CCF) | Quantifies color contrast in images by analyzing pixel-intensity differences across different channels. Higher values indicate more vibrant and enhanced images. |
| Fréchet Inception Distance (FID) | It quantifies the disparity in feature distributions between real and generated images using deep-learning features, where lower FID scores signify improved visual quality. |
| Average Gradient (AG) | Evaluates image sharpness by calculating the mean gradient magnitude across the image. |
| Natural Image Quality Evaluator (NIQE) | A no-reference image-quality assessment metric that compares statistical deviations from natural image characteristics. Lower NIQE scores indicate better image quality. |
| Signal-to-Reconstruction Error (SRE) | Measures the ratio of signal strength to reconstruction error, assessing restoration accuracy. Higher SRE values indicate better restoration performance. |
| Mean Absolute Error (MAE) | Computes the average of absolute differences between original and restored images. Lower MAE indicates higher restoration accuracy. |
| Normalized Root Mean Squared Error (NRMSE) | Provides a normalized measure of deviation between restored and original images, making it suitable for comparing images with different brightness levels. |
| Feature Similarity Index Measure (FSIM) | Assesses similarity by focusing on high-frequency components, capturing perceptual differences aligned with human vision. |
| Visual Information Fidelity (VIF) | Measures the amount of visual information preserved in the restored image compared to the original, reflecting human visual perception. |
| LPIPS | Uses deep neural-network features to evaluate perceptual similarity, emphasizing visual quality as perceived by humans. |
| Diversity Index for Image Denoising | Analyzes the variability between multiple denoising outputs for the same noisy input, useful for exploring alternative restoration methods. |
| Perceptual Loss Metric | Leverages features from pretrained models to assess perceptual quality, optimizing image restoration for human-like perception rather than pixel-level accuracy. |
| No-reference Evaluation Metric | Evaluates image quality without needing the original image, using methods like NIQE and BRISQUE to assess naturalness and perceptual features important to human observers. |

Transformer models showed high restoration performance, with Restormer realizing a PSNR of 39.12 and an SSIM of 0.913 at an inference rate of 0.7581 images per second, while MIRNet realized a PSNR of 38.86 and an SSIM of 0.940 but at a much slower processing rate. SwinIR kept a balance between accuracy and efficiency with a PSNR of 36.30 and an inference rate of 1.12 images per second.

CNN-based models, including SRCNN, MPRNet and NAFNet, demonstrated mixed compromises between accuracy and speed. MPRNet had a comparable inference rate of 2.18 images per second with a PSNR of 33.86. In the same way, NAFNet had a PSNR of 32.93 along with an SSIM of 0.867, proving its stability. From models of denoising, DDNM recorded a PSNR of 24.32 and an SSIM of 0.794, while in this class, CycleISP led others with a PSNR of 39.43 and an SSIM of 0.955. Other models of denoising, including CBDNet, RidNet and DREAMNet, also performed robustly in removing noise, with RidNet recording a PSNR of 38.26 and an SSIM of 0.945.

Table 14. Comparison of various models based on inference time, PSNR and SSIM for both SIDD and DND datasets.

| Model | Inference Time | SIDD Dataset | | DND Dataset | |
|---|---|---|---|---|---|
| | | PSNR | SSIM | PSNR | SSIM |
| Restormer | 0.7581 | 39.12 | 0.913 | 36.41 | 0.926 |
| SRCNN | 0.075 | 34.80 | 0.7184 | 32.75 | 0.862 |
| MIRNet | 0.033 | 36.30 | 0.950 | 38.86 | 0.940 |
| DDNM | 0.153 | 22.07 | 0.832 | 24.32 | 0.794 |
| CWR | 0.0315 | 27.85 | 0.817 | 26.76 | 0.851 |
| SwinIR | 1.120 | 36.30 | 0.847 | 34.52 | 0.896 |
| MPRNet | 2.180 | 33.86 | 0.844 | 34.56 | 0.817 |
| NAFNET | 0.7382 | 32.93 | 0.867 | 30.09 | 0.865 |
| HINET | 0.8737 | 29.97 | 0.906 | 30.65 | 0.894 |
| GFPGAN | 1.063 | 26.01 | 0.763 | 26.42 | 0.713 |
| ESRGAN | 0.790 | 27.24 | 0.791 | 26.30 | 0.711 |
| DnCNN | 0.058 | 21.96 | 0.571 | 31.74 | 0.780 |
| CycleISP | 0.132 | 36.81 | 0.930 | 39.43 | 0.955 |
| CBDNet | 0.15 | 30.44 | 0.795 | 36.12 | 0.920 |
| RidNEt | 0.3921 | 36.01 | 0.903 | 38.26 | 0.945 |
| DREAMNET | 0.417 | 35.72 | 0.916 | 38.23 | 0.940 |

Super-resolution and enhancement frameworks, such as GFPGAN and ESRGAN, weighed perceptual quality against efficiency, with GFPGAN delivering a PSNR of 26.42 at a rate of 1.063 images per second. ESRGAN, however, had a PSNR of 27.24, making it suitable for perceptual restoration. The Contrastive Underwater Restoration (CWR) framework, developed specifically for underwater image restoration, posted a PSNR of 27.85 and an SSIM of 0.817, making it more domain-specific to restoration.

The SSID and DND dataset served as a critical benchmark, offering realistic noisy images captured from smartphones, which posed a challenging yet relevant scenario for image restoration models. These findings underline the diverse trade-offs between speed and accuracy among contemporary image restoration techniques. The summarized results provide valuable insights for guiding future advancements in the development of image restoration methodologies.

In general, the research points to significant trade-offs between restoration performance and computational cost, with transformer-based models producing high image quality at the expense of processing speed, while CNN-based methods keep accuracy and real-time feasibility in balance. Denoising methods, especially CycleISP and RidNet, exhibited robust noise reduction performance and super-resolution models such as ESRGAN improved image perceptual quality well. The results indicate that hybrid methods combining transformers, CNNs and self-supervised learning would potentially further enhance image restoration performance under various imaging conditions.

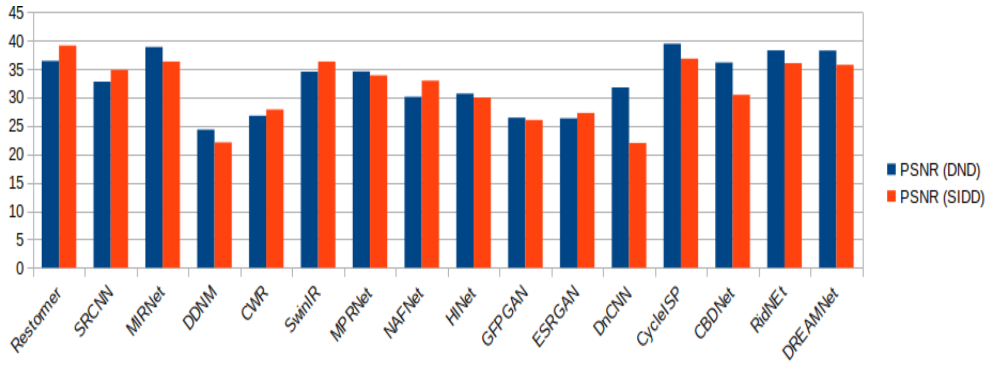


Figure 6. A comparison of PSNR values across various models on the SIDD and DND datasets.

To offer a thorough comparison, the PSNR and SSIM values of several image-restoration techniques, assessed on the SIDD and DND datasets, were shown in Figure 6 and Figure 7. These bar graphs illustrated how performance varied among several methods, emphasizing variances in restoration quality. In addition to providing insights on general-performance trends in picture restoration, the visual representation highlighted how different approaches maintained structural features and perceived quality.

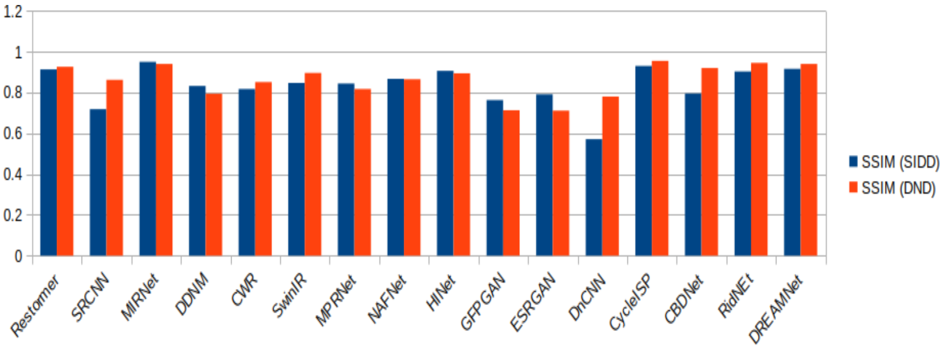


Figure 7. A comparison of SSIM values across various models on the SIDD and DND datasets.

## 5. Open Challenges for Future Research

Despite substantial advancements in image-restoration techniques, several open challenges remain that warrant further investigation. One significant area requiring attention is the treatment of multiple and complex types of degradation commonly encountered in real-world situations. These include combinations of blur, noise and compression artifacts, where existing models, often calibrated for a single type of degradation, frequently underperform in multi-faceted scenarios. Therefore, there is a critical need for methods that can adaptively address a range of real-world degradation levels, ensuring that restoration techniques are robust against diverse degradation types.

Another pressing challenge is achieving an optimal balance between model complexity and processing speed. Many advanced restoration techniques demand substantial computational resources, which limits their scalability and practicality in real-time applications, particularly on mobile devices. Future research should focus on developing lightweight, yet effective, models that facilitate image restoration across a broader spectrum of applications, particularly in resource-constrained environments where efficient processing is essential. Maintaining the fidelity of natural textures and minute details in restored images presents an ongoing challenge. Restoration techniques that excessively enhance sharpness or contrast can distort the original essence of the scene, leading to unrealistic outcomes. It is imperative that successful restoration processes not only enhance visual appeal, but also faithfully represent the scene as it was originally depicted, preserving the integrity of visual information.

Lastly, the lack of effective domain adaptation poses a significant limitation to the applicability of image-restoration models. General-purpose restoration frameworks often fail to capture features unique to specific application domains, such as medical imaging, underwater photography and remote sensing. To enhance the impact and functionality of IR techniques, the development of adaptable or domain-specific methodologies is essential, as these can effectively address the unique requirements of diverse contexts. In summary, addressing these challenges will not only improve existing image-restoration techniques, but also expand their applicability across various fields, paving the way for innovative solutions in an increasingly visual-centric world.

## 6. Conclusion

A comprehensive analysis of the existing literature reveals both the advantages and limitations of current approaches, as well as outlining prospective directions for future research. In light of recent advancements in deep learning, traditional machine-learning and innovative architectures, such as Transformers and GANs, significant progress has been made in enhancing image restoration across various applications, including mobile photography, remote sensing and medical imaging. These methodologies have consistently demonstrated improvements in the clarity, quality and utility of degraded images. Nevertheless, several challenges persist that warrant further investigation. These challenges encompass the effective management of complex mixed degradation types, achieving a balance between computational efficiency and restoration quality and ensuring adaptability across diverse imaging scenarios.

Future-research endeavors are anticipated to concentrate on the development of more flexible and efficient models capable of addressing a wide spectrum of degradation scenarios while remaining suitable for real-time applications, particularly on resource-constrained devices. Furthermore, advancing hybrid models alongside domain-specific strategies will be essential in propelling research initiatives forward. Addressing these unresolved issues and enhancing image quality and accessibility will amplify the impact of image-restoration technologies in an increasingly visual-centric society.

## REFERENCES

[1] Z. Liang et al., "GIFM: An Image Restoration Method with Generalized Image Formation Model for Poor Visible Conditions," IEEE Trans. on Geoscience and Remote Sensing, vol. 60, pp. 1–16, 2022.

[2] S. Jiang et al., "Local Adaptive Prior-based Image Restoration Method for Space Diffraction Imaging Systems," IEEE Transactions on Geoscience and Remote Sensing, vol. 61, pp. 1–10, 2023.

[3] Z. Yang, J. Huang, M. Zhou, N. Zheng and F. Zhao, "IRVR: A General Image Restoration Framework for Visual Recognition," IEEE Transactions on Multimedia, vol. 26, pp. 7012-7026, 2024.

[4] R. Chen, T. Guo, Y. Mu and L. Shen, "Learning Compact Hyperbolic Representations of Latent Space for Old Photo Restoration," IEEE Transactions on Image Processing, vol. 33, pp. 3578–3589, 2024.

[5] T. Kim, C. Shin, S. Lee and S. Lee, "Block-attentive Subpixel Prediction Networks for Computationally Efficient Image Restoration," IEEE Access, vol. 9, pp. 90881–90895, 2021.

[6] S. Kong, W. Wang, X. Feng and X. Jia, "Deep RED Unfolding Network for Image Restoration," IEEE Transactions on Image Processing, vol. 31, pp. 852–867, 2021.

[7] D. Perdios et al., "CNN-based Image Reconstruction Method for Ultrafast Ultrasound Imaging," IEEE Transactions on Ultrasonics, Ferroelectrics and Frequency Control, vol. 69, no. 4, pp. 1154–1168, 2021.

[8] Z. Pan et al., "VCRNet: Visual Compensation Restoration Network for No-reference Image Quality Assessment," IEEE Transactions on Image Processing, vol. 31, pp. 1613–1627, 2022.

[9] Q. Zhang et al., "Combined Deep Priors with Low-rank Tensor Factorization for Hyperspectral Image Restoration," IEEE Geoscience and Remote Sensing Letters, vol. 20, pp. 1–5, 2023.

[10] J. Ke et al., "Artifact Detection and Restoration in Histology Images with Stain-style and Structural Preservation," IEEE Transactions on Medical Imaging, vol. 42, no. 12, pp. 3487–3500, 2023.

[11] Y. Pang, J. Mao, L. He, H. Lin and Z. Qiang, "An Improved Face Image Restoration Method Based on Denoising Diffusion Probabilistic Models," IEEE Access, vol. 12, pp. 3581-3596, 2024.

[12] X. Zhang and J. Feng, "A Novel Blind Restoration Method for Miner Face Images Based on Improved GFP-GAN Model," IEEE Access, vol. 12, pp. 104676–104687, 2024.

[13] M. Yao, R. Xu, Y. Guan, J. Huang and Z. Xiong, "Neural Degradation Representation Learning for All-in-one Image Restoration," IEEE Transactions on Image Processing, vol. 33, pp. 5408–5423, 2024.

[14] W. Zhang, W. Zhao, J. Li, P. Zhuang, H. Sun, Y. Xu and C. Li, "Cvanet: Cascaded Visual Attention Network for Single Image Super-resolution," Neural Networks, vol. 170, pp. 622–634, 2024.

[15] J. Liang et al., "SwinIR: Image Restoration Using Swin Transformer," Proc. of the IEEE/CVF Int. Conf. on Computer Vision, pp. 1833–1844, Montreal, Canada, 2021.

[16] Z. Deng et al., "RFormer: Transformer-based Generative Adversarial Network for Real Fundus Image Restoration on a New Clinical Benchmark," IEEE Journal of Biomedical and Health Informatics, vol. 26, no. 9, pp. 4645–4655, 2022.

[17] Z. Wang et al., "Uformer: A General U-shaped Transformer for Image Restoration," Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recog., pp. 17683–17693, New Orleans, USA, 2022.

[18] J. Tan et al., "Blind Face Restoration for Under-display Camera *via* Dictionary Guided Transformer," IEEE Transactions on Circuits and Systems for Video Technology, vol. 34, no. 6, pp. 4914–4927, 2023.

[19] Y. Zhang, Q. Yang, D. M. Chandler and X. Mou, "Reference-based Multi-stage Progressive Restoration for Multi-degraded Images," IEEE Transactions on Image Processing, vol. 33, pp. 4982–4997, 2024.

[20] B. Zhou et al., "DuDoUFNet: Dual-domain Under-to-fully-complete Progressive Restoration Network for Simultaneous Metal Artifact Reduction and Low-dosect Reconstruction," IEEE Transactions on Medical Imaging, vol. 41, no. 12, pp. 3587–3599, 2022.

[21] I. Marivani et al., "Designing CNNS for Multimodal Image Restoration and Fusion *via* Unfolding the Method of Multipliers," IEEE (TCSVT) Journal, vol. 32, no. 9, pp. 5830–5845, 2022.

[22] X. Feng et al., "Deep-masking Generative Network: A Unified Framework for Background Restoration from Superimposed Images," IEEE Transactions on Image Processing, vol. 30, pp. 4867–4882, 2021.

[23] H. Yan et al., "UW-CycleGAN: Model-driven CycleGAN for Underwater Image Restoration," IEEE Transactions on Geoscience and Remote Sensing, vol. 61, DOI: 10.1109/TGRS.2023.3315772, 2023.

[24] Y. Zhu et al., "Denoising Diffusion Models for Plug-and-play Image Restoration," Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recogn., pp. 1219–1229, Vancouver, Canada, 2023.

[25] S. Welker, H. N. Chapman and T. Gerkmann, "DriftRec: Adapting Diffusion Models to Blind JPEG Restoration," IEEE Transactions on Image Processing, vol. 33, pp. 2795-2807, 2024.

[26] H. Cho, H.-K. Shin et al., "PD-CR: Patch-based Diffusion Using Constrained Refinement for Image Restoration," IEEE Signal Processing Letters, vol. 31, pp. 949–953, 2024.

[27] Z. Luo et al., "Refusion: Enabling Large-size Realistic Image Restoration with Latent-space Diffusion Models," Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition, pp. 1680–1691, Vancouver, Canada, 2023.

[28] J. V. Ortega, M. Haas and A. Effland, "Learning Diffusion Functions for Image Restoration," Proc. of the 2024 IEEE Int. Symposium on Biomedical Imaging (ISBI), Athens, Greece, 2024.

[29] Z. Yue, J. Wang and C. C. Loy, "Efficient Diffusion Model for Image Restoration by Residual Shifting," arXiv preprint, arXiv: 2403.07319, 2024.

[30] S. Yan et al., "HybrUR: A Hybrid Physical-neural Solution for Unsupervised Underwater Image Restoration," IEEE Trans. on Image Processing, vol. 32, pp. 5004–5016, 2023.

[31] Y. Chang et al., "Hyperspectral Image Restoration: Where Does the Low-rank Property Exist?" IEEE Trans. on Geoscience and Remote Sensing, vol. 59, no. 8, pp. 6869–6884, 2020.

[32] W. He et al., "Non-local Meets Global: An Iterative Paradigm for Hyperspectral Image Restoration," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 44, no. 4, pp. 2089–2107, 2020.

[33] D. Berman, D. Levy, S. Avidan and T. Treibitz, "Underwater Single Image Color Restoration Using Haze-lines and a New Quantitative Dataset," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 43, no. 8, pp. 2822–2837, 2020.

[34] M. Li et al., "Imaging Simulation and Learning-based Image Restoration for Remote Sensing Time Delay and Integration Cameras," IEEE Transactions on Geoscience and Remote Sensing, vol. 61, 2023.

[35] C. Li et al., "Efficient Dehazing Method for Outdoor and Remote Sensing Images," IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, vol. 16, pp. 4516–4528, 2023.

[36] S. Zhong et al., "RPIR: A Semi-blind Unsupervised Learning Image Restoration Method for Optical Synthetic Aperture Imaging Systems with Co-phase Errors," IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, vol. 17, pp. 15344-15358 2024.

[37] W. Zhang et al., "Underwater Image Enhancement *via* Minimal Color Loss and Locally Adaptive Contrast Enhancement," IEEE Transactions on Image Processing, vol. 31, pp. 3997–4010, 2022.

[38] W. Zhang et al., "Underwater Image Enhancement *via* Weighted Wavelet Visual Perception Fusion," IEEE Transactions on Circuits and Systems for Video Technology, vol. 34, no. 4, pp. 2469–2483, 2023.

[39] L. Denis et al., "A Review of Deep-learning Techniques for SAR Image Restoration," Proc. of the 2021 IEEE Int. Geoscience and Remote Sensing Symposium GARSS, pp. 411–414, 2021.

[40] R. Kumar et al., "A Review on Generative Adversarial Networks Used for Image Reconstruction in Medical Imaging," Proc. of the 2021 9th Int. Conf. on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO), pp. 1–5, Noida, India, 2021.

[41] W.-C. Sui, X. Cheng and H. A. Chan, "Critical Review on Deep Learning and Smart Technologies for Image Super-resolution," Proc. of the IEEE TENCON 2022-2022 IEEE Region 10 Conf. (TENCON), pp. 1–8, Hong Kong, Hong Kong, 2022.

[42] S. Hou, Y. Wang, K. Li, Y. Zhao, B. Lu and L. Fan, "Deep Learning for Screen-shot Image Demoiréing: A Survey," IEEE Access, vol. 10, pp. 108453–108468, 2022.

[43] M. Pandey, G. Rawat and P. Kanti, "Image Restoration Application and Methods for Different Images: A Review," Proc. of the 2022 IEEE Int. Conf. on Advances in Computing, Communication and Materials (ICACCM), pp. 1–4, Dehradun, India, 2022.

[44] X. Li, Y. Ren, X. Jin, C. Lan, X. Wang, W. Zeng, X. Wang and Z. Chen, "Diffusion Models for Image Restoration and Enhancement: A Comprehensive Survey," arXiv preprint, arXiv:2308.09388, 2023.

[45] G. P. Kumar et al., "A Comprehensive Review on Image Restoration Methods Due to Salt and Pepper Noise," Proc. of the 2023 2nd IEEE Int. Conf. on Automation, Computing and Renewable Systems (ICACRS), pp. 562– 567, Pudukkottai, India, 2023.

[46] Q. Feng et al., "GAN-based Image Deblurring: A Comparison," Proc. of the 2023 IEEE 2nd Int. Conf. on Electrical Engineering, Big Data and Algorithms (EEBDA), pp. 318–324, 2023.

[47] L. Zhai, Y. Wang, S. Cui and Y. Zhou, "A Comprehensive Review of Deep Learning-based Real-world Image Restoration," IEEE Access, vol. 11, pp. 21049–21067, 2023.

[48] N. Deluxni et al., "A Scrutiny on Image Enhancement and Restoration Techniques for Underwater Optical Imaging Applications," IEEE Access, DOI:10.1109/ACCESS.2023.3322153, 2023.

[49] S. Yu et al., "Review of Quality Assessment Algorithms on the Realistic Blurred Image Database (BID2011)," Proc. of the 2023 8th IEEE Int. Conf. on Signal and Image Process., pp. 450–454, 2023.

[50] K. Rajput et al., "An Enhanced Analysis of Machine Learning Techniques for Image Restoration and Enhancement," Proc. of the 2024 15th IEEE Int. Conf. on Computing Communication and Networking Technologies (ICCCNT), pp. 1–6, 2024.

[51] T. Manjunath et al., "Development of an Image Restoration Algorithm Utilizing Generative Adversarial Networks (GAN's) for Enhanced Performance in Engineering Applications: A Comprehensive

Approach to Improving Image Quality and Clarity through Advanced Machine Learning Techniques," Proc. of the 2024 IEEE Int. Conf. on Innovation and Novelty in Eng. and Tech., vol. 1, pp. 1–6, 2024.

[52] R. S. Jebur et al., "A Comprehensive Review of Image Denoising in Deep Learning," Multimedia Tools and Applications, vol. 83, no. 20, pp. 58181–58199, 2024.

[53] Z. Zhang et al., "NTIRE 2024 Challenge on Bracketing Image Restoration and Enhancement: Datasets, Methods and Results," Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition, pp. 6153–6166, Seattle, USA, 2024.

[54] T. Karras et al., "A Style-based Generator Architecture for Generative Adversarial Networks," Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition, pp. 4401–4410, 2019.

[55] Z. Liu, P. Luo, X. Wang and X. Tang, "Deep Learning Face Attributes in the Wild," Proc. of the IEEE International Conference on Computer Vision, pp. 3730–3738, 2015.

[56] R. Timofte et al., "NTIRE 2017 Challenge on Single Image Super-resolution: Methods and Results," Proc. of the Conf. on Comp. Vision and Pattern Recog. Workshops, pp.114–125, Honolulu, USA, 2017.

[57] E. Agustsson and R. Timofte, "NTIRE 2017 Challenge on Single Image Super-resolution: Dataset and Study," Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) Workshops, Honolulu, USA, July 2017.

[58] A. Foi et al., "Pointwise Shape-adaptive DCT for High-quality Denoising and Deblocking of Grayscale and Color Images," IEEE Transactions on Image Processing, vol. 16, no. 5, pp. 1395–1411, 2007.

[59] H. Sheikh, "Live Image Quality Assessment Database Release 2," [Online], Available: http://live.ece.utexas.edu/research/quality, 2005.

[60] A. Abdelhamed et al., "A High-quality Denoising Dataset for Smartphone Cameras," Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition, pp. 1692– 1700, Salt Lake City, USA, 2018.

[61] X. Zhang et al., "Single Image Reflection Separation with Perceptual Losses," Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition, pp. 4786–4794, Salt Lake City, USA, 2018.

[62] N. Silberman et al., "Indoor Segmentation and Support Inference from RGBD Images," Proc. of the Computer Vision–ECCV 2012: 12th European Conf. on Computer Vision, Part V12, pp. 746–760, Florence, Italy, 2012.

[63] D. Martin, C. Fowlkes, D. Tal and J. Malik, "A Database of Human Segmented Natural Images and Its Application to Evaluating Segmentation Algorithms and Measuring Ecological Statistics," Proc. of the 8th Int. Conf. Computer Vision, vol. 2, pp. 416–423, Vancouver, Canada, July 2001.

[64] R. Qian, R. T. Tan, W. Yang, J. Su and J. Liu, "Attentive Generative Adversarial Network for Raindrop Removal from a Single Image," Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition, pp. 2482–2491, Salt Lake City, USA, 2018.

[65] C. Li, J. Guo and C. Guo, "Emerging from Water: Underwater Image Color Correction Based on Weakly Supervised Color Transfer," IEEE Signal Processing Letters, vol. 25, no. 3, pp. 323–327, 2018.

[66] S. Nah, T. H. Kim and K. M. Lee, "Deep Multi-scale Convolutional Neural Network for Dynamic Scene Deblurring," Proc. of the 2017 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), pp. 257-265, July 2017.

[67] B. A. Research, "BSD68: Part of Berkeley Segmentation Dataset and Benchmark," [Online], Available: https://www2.eecs.berkeley.edu/Research/Projects/CS/vision/bsds/, 2018.

[68] A. Blau, K. Michaeli, et al., "The PIRM Challenge on Perceptual Image Enhancement on Smartphones," Proc. of the European Conf. Computer Vision Workshops, pp. 391–411, 2018.

[69] J. Zhang et al., "HAC Dataset: Benchmarking Adverse Condition Image Restoration," [Online], Available: https://github.com/jzbjyb/HAC-dataset, 2019.

[70] T. Karras et al., "FFHQ: Flickr-faces-HQ Dataset," [Online], Available: https://github.com/NVlabs/ffhq-dataset, 2019.

[71] H. Yue et al., "SCISR: Synthetic and Camera-based Image Super-resolution Dataset," [Online], Available: https://github.com/SCISR/dataset, 2019.

[72] S. Shen et al., "Hide: Human-aware Image Deblurring Dataset," [Online], Available: https://github.com/joeylitalien/hide-dataset, 2019.

[73] W. Qian et al., "Raindrop Dataset: Image Pairs with Raindrop Artifacts," [Online], Available: https://github.com/riddhishb/raindrop-removal, 2020.

[74] J. Wei et al., "UHDS: Ultra High-definition Synthetic Dataset for Rainy Image Restoration," [Online], Available: https://github.com/uhds/uhds-dataset, 2022.

[75] C. Program, "Sentinel-2 Satellite Images Dataset," [Online], Available: https://scihub.copernicus.eu/, 2022.

[76] Y. Yu et al., "TinyPerson Dataset for Tiny Object Detection," [Online], Available: http://github.com/TinyPerson/dataset, 2022.

[77] Z. Liu et al., "LSDIR: Large-scale Dataset for Image Restoration," [Online], Available: https://github.com/LSDIR/dataset, 2023.

[78] N. Corporation, "HQ-50K: High-quality Dataset for Image Restoration," [Online], Available: https://github.com/NVIDIA/ HQ-50K, 2023.

## ملخص البحث:

خلقت التّطورات السّريعة في تقنيات التّصوير الرّقمي –بما فيها استعادة الصّور- طلباً متزايداً على تقنياتٍ فعّالةٍ في مجال استعادة الصّور. ويتعيّن على تلك التّقنيات معالجة عددٍ من العيوب، مثل التّشويش والغَباش وانخفاض دقّة الصّور. ويُذكر أنّ استعادة الصّور تعدّ من الأمور المهمّة في العديد من التّطبيقات، مثل التّصوير الطّبّي والتّصوير المساحِي والاستشعار عن بُعد، حيث تُعدّ جودة الصّورة أمراً حاسماً لصّحة التّحليل والقرار.

هذه الورقة تقدّم مراجعةً تنطوي على مسحٍ شاملٍ لطُرق استعادة الصّور الّتي تناولتها أدبيات الموضوع، بما فيها الطّرق التّقليديّة والتّقنيات الحديثة القائمة على نماذج التّعلُّم العميق والنّماذج المستندة الى المحوّلات. وتعالج تقنيات استعادة الصّور التقليدية عدداً من عيوب الصّور؛ فتعمل على إزالة التّشويش والغَباش وزيادة تحليل الصُّور بناءً على نماذج رياضية وخوارزميات خاصّة. وعلى الرّغم من نجاعة تلك التّقنيات في إزالة بعض عُيوب الصّور المستعادة، فإنها لم تكنْ كذلك في بعض سيناريوهات العالم الحقيقي.

وقد نجحت التّطوّرات الحديثة في مجال تعلُّم الآلة بشكلٍ عامّ والتّعلُّم العميق بشكلٍ خاصّ باستخدام الشّبكات العصبية الالتفافية في إنجاز طرقٍ مدفوعةٍ بالبيانات يُمكنها التّعلّم مباشرةً من مجموعات البيانات الضّخمة وتتّسم بالكثير من الفعالية لدى مقارنتها بالطّرق التّقليدية. وحديثاً، فقد أظهرت النّماذج القائمة على المحوّلات القدرة على التقاط الاعتمادات الكامنة في الصّور، مؤدّيةً إلى تفوّقٍ واضح لتلك النّماذج في عددٍ من المهامّ المعقّدة المرتبطة بإزالة عُيوب الصّور المستعادة. وقد تبين أنّ تلك النّماذج كانت ناجعةً في معالجة مدىً واسع من مجموعات البيانات ذات العلاقة.

يتناول المسح الشّامل الّذي تقدّمه هذه الورقة التّحدّيات الّتي ينطوي عليها موضوع استعادة الصّور، إضافةً إلى كيفية معالجتها، ومنها تحدّي التكلفة الحوسبية وتحدّي التّعميم، ويفتح آفاقاً جديدة لبحوث مستقبلية في مجال معالجة عيوب الصّور المستعادة.

Jordanian Journal of Computers and Information Technology (JJCIT), Vol. 11, No. 02, June 2025.

# TOWARDS SECURE IOT AUTHENTICATION SYSTEM BASED ON FOG COMPUTING AND BLOCKCHAIN TECHNOLOGIES TO RESIST 51% AND HIJACKING CYBER-ATTACKS

Muwafaq Jawad[1], Ali A. Yassin[1], Hamid Ali Abed AL-Asadi[1], Zaid Ameen Abduljabbar[1,2], Vincent Omollo Nyangaresi[3], Zaid Alaa Hussien[4] and Husam A. Neamah[5]

*(Received: 7-Jan.-2025, Revised: 14-Mar.-2025, Accepted: 1-Apr.-2025)*

## ABSTRACT

*The Internet of Health Things (IoHT) is a network of healthcare devices, software and systems that enable remote monitoring and healthcare services by gathering real-time health data through sensors. Despite its significant benefits for modern smart healthcare, IoHT faces growing security challenges due to the limited processing power, storage capacity and self-defense capabilities of its devices. While blockchain-based authentication solutions have been developed to leverage tamper-resistant decentralized designs for enhanced security, they often require substantial computational resources, increased storage and longer authentication times, hindering scalability and time efficiency in large-scale, time-critical IoHT systems. To address these challenges, we propose a novel four-phase authentication scheme comprising setup, registration, authentication and secret-construction phases. Our scheme integrates chaotic-based public-key cryptosystems, a Light Encryption Device (LED) with a 3-D Lorenz chaotic map algorithm and blockchain-based fog computing technologies to enhance both efficiency and scalability. Simulated on the Ethereum platform using Solidity and evaluated with the JMeter tool, the proposed scheme demonstrates superior performance, with a computational-cost reduction of 40% compared to traditional methods like Elliptic Curve Cryptography (ECC). The average latency for registration is 1.25 ms, while the authentication phase completes in just 1.50 ms, making it highly suitable for time-critical IoHT applications. Security analysis using the Scyther tool confirms that the scheme is resistant to modern cyberattacks, including 51% attacks and hijacking, while ensuring data integrity and confidentiality. Additionally, the scheme minimizes communication costs and supports the scalability of large-scale IoHT systems. These results highlight the proposed scheme's potential to revolutionize secure and efficient healthcare monitoring, enabling real-time, tamper-proof data management in IoHT environments.*

## KEYWORDS

*Blockchain, Fog computing, IoHT, Authentication, Chaotic cryptography, Healthcare.*

## 1. INTRODUCTION

The Internet of Healthcare Things (IoHT) is a concept that integrates Internet of Things (IoT) technology with healthcare devices. Furthermore, the IoHT is predicted to be the cornerstone of future healthcare systems; every piece of healthcare equipment will be internet-connected and under the supervision of healthcare providers. As the IoHT grows, it can provide speedy and affordable healthcare [1]. Technological development over recent years has enabled the diagnosis of a multitude of illnesses and the monitoring of health through the utilization of compact devices, such as smartwatches, electrocardiography (ECG) machines and shoes.

Furthermore, the paradigm for healthcare has changed due to technology, moving from hospital-focused to patient-centred. For example, many clinical evaluations, such as blood pressure, blood glucose and pO2 readings, can now be performed at home without the need for direct medical

1. M. Jawad, A. Yassin, H. AL-Asadi and Z. Abduljabbar are with the Department of Computer Science, College of Education for Pure Sciences, University of Basrah, Basrah 61004, Iraq. Emails: pgs.muwafaq.abbas, ali.yassin, hamid.abed, zaid.ameen@uobasrah.edu.iq
2. Z. Abduljabbar is with the Department of Business Management, A-Imam University College, 34011 Balad, Iraq and with Shenzhen Institute, Huazhong University of Science and Technology, Shenzhen 518000, China. Email: zaid.ameen@uobasrah.edu.iq
3. V. Omollo is with the Department of Computer Science and Software Engineering, Jaramogi Oginga Odinga University of Science and Technology, Bondo 40601, Kenya; and with the Department of Applied Electronics, Saveetha School of Engineering, SIMATS, Chennai, Tamil Nadu, 602105, India. Email: vnyangaresi@jooust.ac.ke
4. Z. Alaa is with Management Technical College, Southern Technical University, Basrah 61004, Iraq. Email: zaid.alaa@stu.edu.iq
5. H. A. Neamah is with the Department of Electrical Engineering and Mechatronics, Faculty of Engineering, University of Debrecen, Ótemető u.4-5, Debrecen 4028, Hungary. Email: husam@eng.unideb.hu

assistance. Furthermore, advanced telecommunication technologies enable the transmission of clinical data from remote places to healthcare facilities [2]; the explosive growth highlights serious issues with user privacy and security, especially in the context of the IoHT and needs careful consideration and attention. Various vulnerabilities exist within healthcare systems that could lead to security and privacy breaches, including unauthorized access to vast amounts of sensitive patient data, encompassing personal and health records critical for making life-saving decisions [3]. As a result, in recent years, the protection of security and privacy in IoHT applications has gained attention. Confidentiality, non-repudiation, data integrity and the authentication and identification of IoHT devices and users are all critical security requirements. Since authentication is essential to maintaining the fulfilment of other security requirements, it stands out as a primary concern [4]-[5]. Authentication is the process of verifying and authenticating an entity's identification. Every entity should be able to recognize and verify every other entity in the system or the particular part of the system that it communicates with [6]. Due to the involvement of multiple applications and users in the monitoring, operation and management of healthcare devices, the potential for breaches in authentication and authorization schemes exists.

The authentication techniques described in the literature for the IoHT mostly belong to two architecture categories: centralized and decentralized. The centralization of authentication can be performed by distributing and managing login credentials through a single server or a reliable outside source. Moreover, it comprises three procedures. First, there is one-way authentication, which occurs when two parties want to communicate and only one party authenticates itself to the other, while the other party remains unauthenticated. Second, there is two-way authentication, also known as mutual authentication, where both entities authenticate each other. Lastly, there is three-way authentication, where a central authority authenticates each of the parties and assists them in mutually authenticating themselves [7]. Scalability problems with central-authentication systems could result in performance bottlenecks as user numbers increase. In addition, they are exposed to single points of failure, which can compromise the entire authentication process. Furthermore, the concentration of sensitive user credentials may give rise to privacy concerns [8].

Decentralized authentication solutions that employ blockchain technology are recommended more and more for IoHT systems because they are compatible with the scattered and heterogeneous nature of these systems [9]-[10]. Researchers have highlighted the basic properties of blockchains, which include consensus, immutability, decentralization and security [11]. They emphasised the benefits of using blockchain technology to improve big-data management and authentication in many areas, such as enhancing data integrity, promoting seamless data sharing, bolstering security and privacy measures and improving big-data overall quality [12]. As a result, several blockchain platforms, such as Multichain, Ethereum, Bitcoin and others, have emerged, each offering distinct advantages over the rest. These platforms operate on diverse consensus protocols, ensuring security and scalability at varying levels [13]. To strengthen the discussion and provide deeper insights into the computational complexity of cryptographic algorithms, consensus mechanisms and smart contracts, this study positions itself within the broader context of blockchain research. Blockchain-assisted systems are particularly relevant for IoHT due to their ability to address the limitations of centralized systems, such as scalability and single points of failure. By leveraging blockchain's inherent properties, such as decentralization and immutability, the proposed system ensures secure and efficient authentication while minimizing computational overhead and communication costs [14]-[15].

Smart and edge devices generate large amounts of data that are quickly transferred to the cloud via IoT devices. This can sometimes lead to network congestion [16]. Therefore, the fog-computing concept creates a decentralized computing environment by dispersing several fog nodes over various areas. It effectively handles data processing, solving computing constraints in cloud and IoT devices, by occupying the space between the edge and cloud layers [17]. This method improves cloud-based services by enabling quick data processing and data transfer from edge devices to the cloud. As a result, it lessens network congestion and the reliance of edge and IoT devices on direct cloud connection [18]. For this purpose, our devised work incorporates fog computing, extending cloud services to network edges and providing acceptable computational support for IoHT devices. To mitigate communication overhead during authentication, a chaotic-key cryptosystem is employed within our work that utilizes chaotic keys, is compact, minimises communication overhead and considers the limited computational capabilities inherent in IoHT devices [19]-[20]. In recent years, many authentication techniques have been suggested to enhance the security of the Internet of Human Things (IoHT) system. The present study proposes a decentralized authentication system that leverages fog computing and blockchain technology to contribute to these efforts. Multiple factors of authentication, including wallet address, password, OTP and fingerprint, are utilized. By leveraging the features of blockchain technology, such as peer-to-peer communication, cryptography, consensus mechanisms and smart contracts, it facilitates authentication through decentralized peer-to-peer communication among fog nodes. The suggested approach resists common attacks and modern threats like 51% attacks and hijacking. Furthermore, this work accomplishes authentication without relying on a central authority. Finally, to guarantee the security of parties interacting through public channels in a decentralized environment, our work combines an authentication mechanism with immutable blockchain technology. Additionally, decentralized node identification is supported by blockchain technology. Consequently, the following contributions are provided by this paper:

- We provide a lightweight authentication scheme over fog computing for a blockchain-based IoHT system. The proposed work employs blockchain in the fog-computing layer to allocate the IoHT into fog areas.
- The proposed scheme utilizes a chaotic-based cryptosystem to provide a higher level of scalability and efficiency. Furthermore, the chaotic cryptosystem offers remarkable efficiency and rapidity in encryption and decryption, particularly in the domain of image encryption.
- A comprehensive security evaluation is conducted using the well-regarded Scyther tool to showcase the robustness of the suggested design against common threats, such as replay attacks, man-in-the-middle attacks, 51% attacks and Hijacking. Moreover, it has been proven that our proposed approach is resistant to these malicious attacks. Preliminary security assessment is conducted to verify adherence to security requirements, including decentralization, identification, secrecy, non-repudiation and integrity.
- The proposed work is simulated and designed by the Ethereum blockchain platform to evaluate it for two main metrics, latency and throughput. We utilize Apache JMeter, which is a strong tool used for measuring evaluation metrics, like latency and throughput. Furthermore, the assessment results indicate that the suggested strategy is time-efficient (0.3201 ms), with latencies of 1.25 ms for registration and 1.50 ms for authentication.

The paper is ordered as follows: the related authentication schemes in the IoHT environment are presented in Section 2. Backgrounds are in Section 3. The network model is explained in Section 4. Moreover, the security model is shwon in Section 5. The proposed scheme and its phases are described in Section 6. The performance analysis, simulation, evaluation metrics, key-generation time, LED with 3-D Lorenz chaotic encryption and decryption time, computational cost and smart contract costs are detailed in Section 7. The formal and informal security analyses are presented in Section 8. Finally, the conclusion is presented in Section 9.

## 2. RELATED WORKS

In 2018, Almadhoun et al. [21] introduced an authentication system that utilizes blockchain-enabled fog nodes and Ethereum smart contracts to address the capacity constraints of the IoT, grant access to IoT devices and verify users. This method enables the system to expand its capacity by using fog nodes for computational operations. Although the scheme offers strong security, it does not align with the requirements of most IoT connectivity scenarios. This work has limitations, such as computational overhead, because the integration of the blockchain with a smart contract may not be suitable for all IoT devices, especially those with limited processing power. Moreover, in terms of scalability and security vulnerabilities, it is not entirely immune to attacks; there are potential vulnerabilities in smart contracts.

In 2018, Mehmood et al. [22] proposed a mutual-authentication method and key-agreement methodology utilizing chaotic maps and Diffie-Hellman key exchange. The suggested solution guarantees that only authorized healthcare professionals can retrieve patients’ health data collected through body sensors in the medical system. This paper has major limitations, particularly in terms of computational complexity. The technique used in this paper involves complex cryptographic operations, which can result in a longer processing time and increased energy consumption.

Furthermore, it experiences scalability challenges when it comes to managing a substantial number of users and devices.

Moreover, the scheme's objective of safeguarding user anonymity is compromised by the privacy hazards associated with relying on a centralized cloud. This includes the potential for data breaches and unauthorized access to critical health information. Additionally, there is vulnerability in having a single point of failure if the cloud server experiences a failure. In 2019, Liang et al. [23] developed a blockchain-powered system for managing and verifying identities. The system's goal is to enhance patient-data confidentiality while allowing more flexibility in accessing health records. This study has limitations in scalability due to the degradation of the blockchain performance as the number of transactions increases, resulting in significant implementation challenges in the healthcare sector. Furthermore, it poses data-privacy concerns. In 2020, Cheng et al. [24] created a blockchain-based multiple-identity authentication system for a safe medical-data exchange model that did not require a third party. This paper has limitations. The ability to scale large-scale blockchain applications faces a hurdle, as the performance of the technology can deteriorate with the growing volume of medical data. The complexity of integrating blockchain technology into current medical-data systems is an intricate process that necessitates substantial modifications to the existing-infrastructure issues with the protection of personal information. Despite the security aspects of the blockchain, ensuring complete data privacy remains a tough task.

In 2021, Wu et al. [25] examined the security of different authentication techniques. Their study showed that the examined schemes were susceptible to established attacks, such as session-specific temporary data, user impersonation and server impersonation. The examined scheme utilized formal and informal security studies, both of which verified its lack of security. However, as the numbers of servers and users increase, the scheme may face scalability issues, potentially affecting the overall efficiency and performance. In 2021, Guo et al. [26] provided FogHA, a lightweight cryptographic primitive for fog computing and an undetectable handover-authentication strategy. This system facilitates managing keys and mutual authentication among a mobile device and fog computing by removing redundant authentication messages. The method includes characteristics, like untraceability, anonymity and low latency, making it secure against attacks from insiders. Opponents can utilize the untraceability and anonymity characteristics to carry out attacks without being identified by the system. This paper has limitations. Scalability refers to the ability of a system or process to handle an increasing amount of work or data efficiently and effectively. The approach may face challenges when expanding to a significant number of devices due to the inherent computational and communication burdens associated with fog nodes. Furthermore, limitations on the available resources that fog nodes possess constrain the processing resources in comparison to cloud servers, potentially impacting the performance and efficiency of the authentication process. Moreover, in terms of security vulnerability, the scheme's objective is to offer reliable authentication; however, achieving a satisfactory equilibrium between security and performance can be difficult, especially in contexts with limited resources.

In 2021, Javed et al. [27] introduced blockchain-based decentralized identity control using smart contracts for electronic health records, having been the focus of various research investigations, such as Health-ID for remote healthcare and Health-ID for EHRs. Additionally, a blockchain-enabled authentication method was created to reduce the necessity of re-authentication across multiple hospitals, enhancing efficiency and reducing the time overhead for devices with constrained processing and memory capabilities. This paper has limitations, including challenges related to the ability of a system or process to handle increasing amounts of work or data efficiently. Although the suggested blockchain-based approach improves security and decentralization, the ability of blockchain networks to handle large amounts of data and transactions is still a matter of concern. The performance measures, such as the transaction gas cost and the transactions per second, suggest that as the numbers of users and transactions grow, the system may experience delays and incur larger operational expenses. In addition, the report acknowledges that although the blockchain has the potential to improve openness and trust, but it is challenging to ensure that all players comply with healthcare rules and privacy requirements. This is especially crucial in varied regulatory landscapes spanning multiple regions and nations.

In 2022, Chen et al. [28] proposed a method to shorten the time taken for authentication. The method consists of two parts; complete authentication and lightweight authentication. For complete authentication, they used CP-ABE to ensure confidentiality. For lightweight authentication, they utilized the hash function and XOR gate. This method enabled the creation of a physiological sensing device with a lower computing ability that can handle parameter calculations. They used the patients' information as seeds for a random-number generator. Finally, this method uses a third party and does not take advantage of the blockchain to make the security mechanism robust. In 2022, Umoren et al. [29] implemented blockchain smart contracts to tackle user authentication and other limitations in IoT and fog technology. The decentralized fog-computing framework incorporated scalability, immutability and secure authentication for fog devices. Additionally, it addressed issues of immutability and scalability in fog computing. The scheme provides robust security, but does not meet the needs of typical IoT connectivity scenarios. The proposed system's implementation is not sufficiently covered in the study. More precisely, the data structure and code offered lack clear explanations, which may impede the ability of other researchers to replicate and advance the work.

Moreover, the description of the experimental setup and performance measures is insufficient. In order to properly validate the findings, it is necessary to provide more comprehensive explanations of the simulation model and the results. The discussion lacks a thorough comparison between the suggested method and existing solutions. An extensive evaluation considering factors, such as the resilience to attacks, computational cost, calculation time and communication overhead would offer a more thorough understanding and verification of the suggested approach.

In 2023, H. Miriam et al. [30] introduced the LGE-HES algorithm to improve blockchain-based healthcare cybersecurity, focusing on securing medical-image data. Simulations show that the method achieves high PSNR (63 dB) and minimal MSE (0.003) while optimizing encryption and decryption times. Compared to standard approaches, it effectively identifies 94.9% of malicious communications. The results demonstrate superior image secrecy, suggesting future exploration of hybrid optimization techniques for enhanced security scalability. In 2024, Alsaeed et al. [20] introduced a method to address issues, like scalability and time; they proposed group authentication utilizing Shamir's secret-sharing (SSS) algorithm, ECC, fog-based computing and a multi-level blockchain to implement lightweight and scalable group authentication in the IoMT. The evaluation test shows good scalability and time efficiency, but although there are many good aspects to this method, one of the foundations of healthcare systems is missing: a robust authentication mechanism for users, particularly administrators and patients. In addition, handling the enormous number of devices and sensors presents difficulties for the ECC algorithm. Thus, we used the chaotic algorithm to solve this problem.

The next section examines the current authentication schemes and systems used in IoT and fog environments and explores how blockchain technology might be used to improve security and decentralization. However, the majority of the centralized systems are constrained by limits in terms of scalability, security and privacy. Additionally, some of the schemes rely on a centralized fog and IoT authentication system, which also has its own limitations. We provide a lightweight authentication scheme over fog computing for a blockchain-based IoHT system. Furthermore, the proposed work employs the blockchain in the fog-computing layer to allocate the IoHT into fog areas. Additionally, we utilized a 3-D chaotic cryptosystem to provide a higher level of scalability and efficiency. Finally, Table 1 provides a comparison of some different related schemes.

## 3. BACKGROUNDS

### 3.1 Blockchain Technology

In 2008, Satoshi Nakamoto introduced blockchain technology, as well as its distributed decentralized network which functions as a network of independent networks responsible for managing a collection of time-stamped documents. The blockchain's structure comprises interconnected blocks secured through fundamental cryptography. This technology operates on three core principles: transparency, decentralization and immutability [31]. Blockchain's decentralized nature allows secure, reliable data sharing in IoT, popular in mutual authentication. It serves as a dependable platform for authentication systems and secure storage. These advantages make use of blockchain technology in healthcare with several benefits [32]. It is a sensible decision, particularly because the healthcare sector has prioritized patient-data security due to technological advancements. Moreover, various experts have concluded that incorporating blockchain technology into the healthcare industry would be a feasible solution [33]. The blockchain is a secure method of exchanging information. It comprises a series of interconnected blocks that store encrypted data. Each block includes the data, its cryptographic hash and the hash of the preceding block [34], as illustrated in Figure 1.

Table 1. Comparison of different related schemes.

| Authors | Year | Problem | Contribution | Technique | Platform |
|---|---|---|---|---|---|
| Almadhoun [21] | 2018 | IoT devices are vulnerable to security breaches; centralized authentication systems are prone to single points of failure. IoT devices lack the capacity to secure themselves; high latency and communication overhead in IoT-cloud interactions. | Proposed a decentralized and scalable authentication mechanism using blockchain-enabled fog nodes and Ethereum smart contracts for authenticating user access to IoT devices; introduced a system where fog nodes handle authentication tasks, relieving IoT devices from heavy computational loads. | Blockchain,<br>Fog Computing,<br>Ethereum Smart Contracts,<br>Elliptic Curve Cryptography | Ethereum,<br>Remix IDE |
| Cheng et al. [24] | 2020 | Difficulty in secure sharing of medical data due to reliance on trusted third parties in Medical Cyber Physical Systems (MCPS). | Proposed a blockchain-based secure medical data sharing scheme that ensures data integrity, untraceability and secure authentication without relying on trusted third parties.<br>Utilized bilinear mapping and intractable problems for secure authentication. | Blockchain,<br>Bilinear Mapping,<br>Cloud Storage | Blockchain,<br>Cloud Storage |
| Guo et al. [26] | 2021 | High latency and security issues in handover authentication for mobile devices in fog computing. | Proposed FogHA, an efficient handover authentication scheme for mobile devices in fog computing, ensuring mutual authentication, key agreement and resistance to known attacks. | Lightweight cryptography,<br>Symmetric trivariate polynomials,<br>hash functions | Fog Computing |
| Javed et al. [27] | 2021 | Centralized identity management in eHealth restricts interoperability and security. | Proposed a decentralized identity-management system for remote healthcare using blockchain. | Blockchain,<br>Smart Contracts,<br>JSON Web Tokens (JWT) | Ethereum Blockchain |
| H. Miriam et al. [30] | 2023 | Ensuring the cybersecurity of blockchain-based healthcare systems is challenging due to vulnerabilities in medical-image data and the need for robust encryption mechanisms. | The proposed LGEHES algorithm enhances the cybersecurity of blockchain in healthcare by optimizing encryption and decryption processes while preserving medical-image quality and resisting malicious attacks. | The LGE-HES algorithm integrates Lionized Golden Eagle optimization with homomorphic encryption | Blockchain -Healthcare |
| N. Alsaeed [20] | 2024 | IoMT faces security challenges due to limited computational and storage capacities, making traditional authentication methods unsuitable for large-scale, time-sensitive systems. | proposed a lightweight and scalable group-authentication framework for IoMT systems using blockchain technology, enhances efficiency and scalability, achieving 0.5-second latency and 400 transactions per second. | ECC for lightweight and (SSS) algorithm for secure secret construction and group authentication | IoMT-Blockchain |

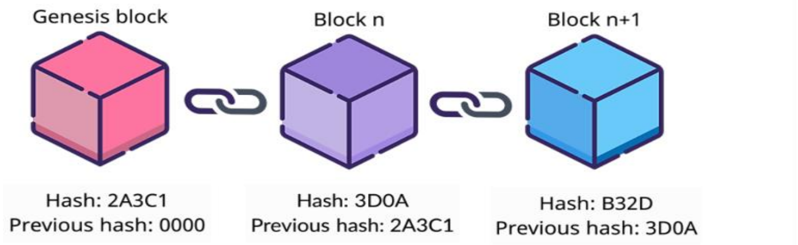


Figure 1. Blockchain.

### 3.1.1 Architecture

Let's use the following Figure 2, which illustrates the entire process of a transaction being sent from a user on the blockchain network, to better understand the blockchain architecture.

- Once a user initiates a transaction on a blockchain network, it is disseminated to all nodes within the network. Every node maintains a complete replica of the blockchain, which is instrumental in the verification process. All connected nodes collaborate to ensure that the block encompassing the user's transaction remains unaltered. If the validation process is successful, the nodes append that block to their version of the blockchain.
- To append a fresh block to the blockchain, consensus must be achieved amongst the network nodes regarding the validity of the blocks. This agreement is attained *via* a validation procedure that employs precise algorithms to authenticate the transaction and confirm the sender's membership in the network.
- Once the validation process is completed, the block is added to the blockchain.
- Subsequently, when the whole validation process has been completed, the transaction is considered finalized.

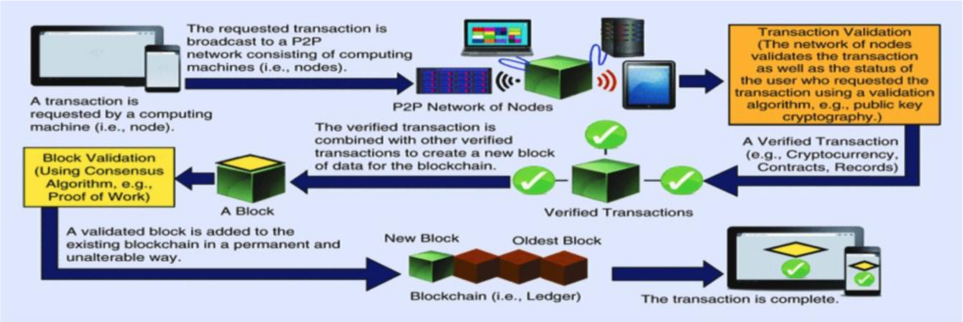

The requested transaction is broadcast to a P2P network consisting of computing machines (i.e., nodes).
A transaction is requested by a computing machine (i.e., node).
P2P Network of Nodes
Transaction Validation (The network of nodes validates the transaction as well as the status of the user who requested the transaction using a validation algorithm, e.g., public key cryptography.)
The verified transaction is combined with other verified transactions to create a new block of data for the blockchain.
A Verified Transaction (e.g., Cryptocurrency, Contracts, Records)
Block Validation (Using Consensus Algorithm, e.g., Proof of Work)
A Block
Verified Transactions
A validated block is added to the existing blockchain in a permanent and unalterable way.
New Block
Oldest Block
Blockchain (i.e., Ledger)
The transaction is complete.

Figure 2. An overview of blockchain architecture.

### 3.1.2 Consensus Algorithm

For a block to become a part of the blockchain, it must follow specific consensus guidelines. To ensure this, blockchain technology employs consensus algorithms. In the Bitcoin network, Nakamoto [31] introduced the Proof of Work (PoW) algorithm, which is now the most commonly used consensus method. The fundamental idea behind this algorithm is that since multiple nodes or users are present on a blockchain network, any transaction request made by a participating node must be computed before it can be added to the network. The nodes responsible for performing these calculations are called miners and this process is known as mining [35].

### 3.1.3 Key Features of Blockchain

1) **Decentralization:** Blockchain distributes information throughout the network as opposed to concentrating it in one place. Additionally, this means that information control will be dispersed and managed by consensus determined by the collective input of all connected nodes on the network. Nowadays, several reliable organizations handle the data that was previously centralized at one location [36].
2) **Data Transparency:** To achieve data transparency in any technology, relationships based on trust must exist between entities. The relevant data or record needs to be safe from heat and secure. Any data stored on the blockchain is dispersed throughout the network rather than being concentrated in one location or under the control of a single node. Since data ownership is now shared, it is transparent and protected from outside interference.
3) **Security and Privacy:** Blockchain technology employs cryptographic functions to provide security to the nodes connected to its network. It uses the SHA-256 algorithm for the hashes stored on the blocks, known as the "secure hash algorithm" (SHA), which ensures data integrity and adds security to the blockchain. Digital data is assigned checksums through strong one-way functions called cryptographic hashes, rendering them unusable for data extraction. This makes blockchain a decentralized and secure platform, using cryptographic techniques to safeguard user privacy, thus making it a reliable option for applications that require privacy protection [37].

### 3.2 Fog Computing

Industry first used the term fog computing to refer to the fundamental architectural concept of the technology: fog is a region that lies between the ground, where user devices are located and the cloud or data centers. In general, fog is referred to as a decentralized distributed computing system in which various fog devices are owned by various entities and organizations can interact with the system from various locations, including smart hubs, hospitals, schools and airports. [38] Fog computing's topology is the geographically dispersed nodes that carry out computation and providing network and storage services is its primary feature. In addition to standard network features, fog-computing resources can be incorporated into network gateways, routers and access points. Additionally, there might be specific fog-computing nodes, such as edge computing [39]. The following is a description of the main characteristics of fog computing [40]-[41].

1) Adaptability: This consists of multiple fog devices and network sensors that provide storage and perform computing tasks.
2) Reduced latency: Fog computing's proximity to edge devices shortens the time taken for information to be computed with those devices and helps the host-fog devices respond to position queries at multiple sites.
3) Physical distribution: Fog computing presents distributed applications and services that are hosted in various locations.
4) Compatibility: Fog modules can be used across a variety of platforms and service providers.

### 3.3 Chaotic Cryptography

Within systems, security is of the utmost importance. It is critical to ensure security, confidentiality, data-origin authentication, message integrity and non-repudiation of origin. The use of symmetric and asymmetric cryptographic algorithms is the foundation for improving message security over unsecured networks [42]. There has been a noticeable increase in the exploration of chaos-based cryptography in recent years, driven by a renewed interest in leveraging chaotic systems for various applications. Our work will utilize chaotic systems, such as the use of the logistic map for key pair generation, the beta-transform for key exchange and the integration of the Lorenz system for encryption and decryption [42].

## 4. NETWORK MODEL

Before delving into the intricate details of our proposed IoHT system, understanding the fundamental assumptions that form the foundation of this initiative is critical. These fundamental premises hold significance in blockchain-based authentication systems, serving as pivotal reference points.

- In the context of fog computing, the ecosystem comprises a wide range of both mobile and stationary devices, such as smartphones, sensors, embedded systems and stationary edge servers. These devices are intricately interconnected across a multitude of communication networks.
- The device used by registered users is adept at integrating and utilizing blockchain technology, thereby enhancing the system's functionality.
- To fulfill its role effectively, a fog server must meet specific pre-requisites, including the capability to host the blockchain and function as a server or node within the network architecture.
- The smart contracts are expected to execute the critical functions of device and user registration and authentication, playing a pivotal role in the seamless operation of the system.

The network model consists of four layers. These layers are shown in Figure 3 and explained below.

### 4.1 User Layer

A system user is a person who realizes the way to use system resources effectively. Users have distinct roles and attributes within the system that allow them to be identified. Patients, doctors, nurses, administrators and others are among those who interact with the system. Their primary responsibility is to interact with the system in order to perform essential functions, such as creating, reading, updating, deleting, accessing and managing medical records.

### 4.2 Edge-device Layer

In this layer, the IoHT ecosystem serves as a crucial and essential component, fulfilling diverse roles, such as gathering, managing data computation, secure storage and initial processing of IoHT sensor information. Its primary function involves meticulously safeguarding and organizing data, ensuring its integrity and security before transmission to the fog-computing infrastructure.

### 4.3 Fog-computing Layer

This layer includes more than one fog servers that act as blockchain nodes and devoted servers to support the decentralized blockchain infrastructure. These devices ensure secure information transmission from IoHT gadgets even as additionally retaining synchronized copies of the blockchain, ledger and smart contracts. The elaborate interplay of fog computing and blockchain enhances the latency, reduces the time cost and adds another layer of security.

### 4.4 Cloud Layer

A large quantity of data is generated inside IoHT sensors and gadgets. The cloud computing proposes as a robust actor with its extremely good computational skills, substantial storge capacity and strong bandwidth. It serves as an infrastructure that is specifically designed to store, compute and examine significant amounts of data. The cloud layer is responsible for the registration and approval of all of the fog servers, users and IoHT devices. Additionally, our scheme allowed the authorized users to communicate with other nodes which include cloud servers, fog servers, users and IoHT gadgets. Furthermore, the scheme employs a blockchain structure and smart contracts that hold an essential position in enhancing security and privacy.

- **Blockchain:** Operates as a decentralized authority for identifying and registration of all entity and IoHT devices. Authentication and identity procedures are controlled through smart contracts included in the blockchain infrastructure. Utilizing blockchain technology, every fog server authenticates IoHT gadgets and customers within its location. Significantly, the blockchain remains on hand throughout all layers of the system architecture.
- **Smart Contract:** It’s self-executing codes on blockchains, replacing centralized oversight in transactions. All contract executions are publicly documented, ensuring transparency across network nodes. Blockchain’s allotted records storage, secure protocols and consensus mechanisms extensively boost protection and streamline system gadget performance, decreasing time and charges. To register and manage the identities of all IoHT devices and users, the proposed scheme incorporates blockchain technology. Each fog server takes on the responsibility of authenticating IoHT devices in its network. Furthermore, the architecture is intended to address the scalability concerns inherent in blockchain-based authentication schemes, ensuring support for the IoHT system’s scalability requirements.

## 5. SECURITY MODEL

In the healthcare part, security is the predominant concern, exerting a profound influence on the reliability and confidentiality of devices and services. Consequently, there is a compelling necessity to initiate the proactive development of comprehensive solutions aimed at fortifying these systems against a wide array of potential threats, as substantiated by [43]. Subsequently, our discussion will pivot towards an examination of prevalent attacks directed at IoHT systems, thereby enhancing our understanding of the security challenges inherent in the healthcare sector.

Certain security requirements must be satisfied by authentication techniques used in wireless networks. These characteristics enhance the feasibility of implementing any proposed plan in the wireless body-area network (WBAN) environment. The Canetti-Krawczyk (CK) model allows the formal development and analysis of the suggested scheme. In addition, the proposed system must possess numerous crucial security attributes [44]-[45].

### 5.1 Security Requirements

Mutual Authentication: To safeguard sensitive information from potential interception by evil individuals, all parties must authenticate their identities before any data transfer [46].

- User Identity Anomaly and Untraceability: Anonymity is attained by the consolidation of a legitimate user's personal information in a manner that prevents an unauthorized individual from discovering or identifying the user.
- Forward Secrecy: It guarantees that the key used for the current session is distinct and will not be vulnerable to unauthorized access. Additionally, it prohibits the utilization of a primary session key for initiating a fresh session.
- Unlinkability: It is a private attribute that is effective when an attacker is unable to differentiate between two or more components of a system. Consequently, the attacker is unable to breach the system or improperly exploit it. This attribute is crucial in identifying systems. For example, an attacker may be unable to establish a connection between the contents of any communications, multiple sets of login credentials or multiple bank withdrawal transactions [47].
- Scalability: The constituents of an authentication system should possess the ability to adapt and evolve following alterations in the surrounding environment [48].

### 5.2 Threat Model

The attacker can execute the following threats:

- Distributed Denial-of-Service (DDoS) Attack: This attack represents a form of network manipulation characterised by the deliberate inundation of a targeted system with an overwhelming volume of network traffic, surpassing its operational capacity. As a consequence, these attacks cause a significant increase in the workload of the system [49].
- Man-in-the-Middle (MITM) Attack: This attack is a security threat that aims to compromise the privacy and integrity of data exchanged during a session. In this type of attack, an adversary strategically positions him/her self between two communicating hosts, intercepting and potentially modifying the data traffic, thereby breaching both confidentiality and integrity [50].
- Insider Attack: This attack occurs when an individual who possesses authorised access to an organization's systems, data or network deliberately compromises the privacy, accuracy or availability of sensitive information or resources for personal gain or malicious intent [51].
- Eavesdropping Attack: An attacker can access IoHT network traffic and read the contents of messages being transmitted across the network by using an eavesdropping attack. The payload and wireless session are passively observed by the attacker. If the communication is encrypted, the attacker may eventually be able to decrypt it [52].
- Impersonation Attack: To obtain the information that an attacker is not authorized to access, he/she assumes the identity of another person or impersonates a legitimate IoHT user (or group of users) or server [53].
- Replay Attack: This type of attack accesses the WLAN using phony authentication sessions and does not occur in real time. The attacker initially obtains a session's authentication. The attacker then replicates the initial session, changing or tampering with it [43].

## 6. PROPOSED SCHEME

Our work focuses on four main phases: Setup, Registration, Login and Authentication and Secure Construction. Furthermore, the environment of the proposed scheme consists of four main components: Health Cloud Server (HCS), Fog Service Provider (FSP), Blockchain (BC) and Wireless Body Area Network (WBAN) depending on three layers: IoHT sensors layer **($S_i = S_1, S_2, S_3, , S_n$),** Personal Devices Layer (***$Pd_i$***) (like mobile phone, Computer, tablet, …etc.) that use by (admin(***$ADM_i$***)), patient(***$P_i$***)) and doctor(***$Dr_i$***))), and Internet layer (such as gateway or FSP layer). Figure 3 explains the major components of our work.

### 6.1 Setup Phase

During this phase, *HCS* is regarded as the primary entity accountable for managing and enrolling users, *FSP* and *IoHT* devices. Each user and *IoHT* device obtains the shared key (*SK*) locally by implementing a key-exchange protocol based on a chaotic logistic system to guarantee security. The *HCS* employs a highly secure cryptographic hash function, known as the h(.) function, implemented through SHA-256 which is a member of the SHA-2 family and plays a main role in verification and anomaly for major parameters. Additionally, our proposed scheme uses $CTR_{mode}$ for the Encryption function (*Enc*()) and the Decryption function (*Dec*()).

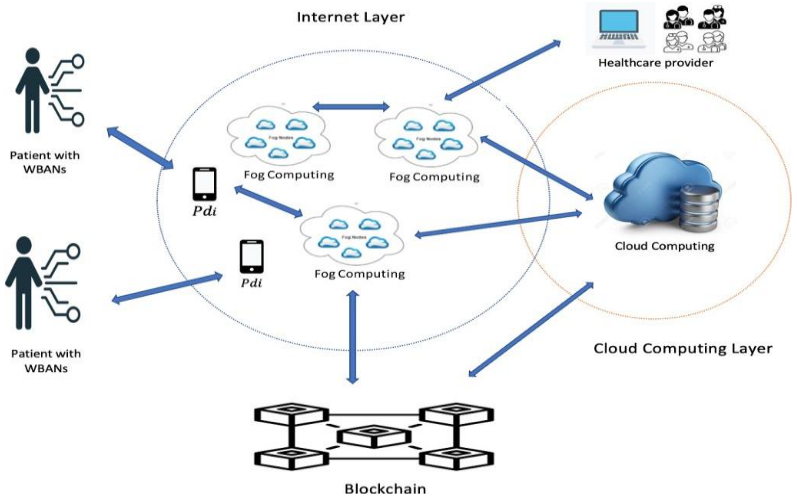


Figure 3. The proposed scheme's architecture.

## 6.2 Registration Phase

This sub-section describes the user-registration process of our proposed scheme; we focus on the FSP, IoHT devices and administrator, patient and doctor registration. Each user must provide valid information (username, password, wallet address, …etc.) one time. The user data is hashed and stored. Below is a description of the registration process.

### 6.2.1 Node$_i$ Registration

First, each $node_i$ (FSP, $Pd_i$, IoHT sensor $S_i$..........etc.) generates its own private key ($node_{i\_pr}$) and public key ($node_{i\_pu}$) using the chaotic system. The following step explains $node_i$ registration process.
**Step 1:** $node_i$ selects an identification $node_{iID}$ and time ($T_s$) and sends a registration request to the HCS{$node_{iID}$, $T_s$, $node_{ipu}$}.
**Step 2:** HCS checks the freshness of the received request by calculating $T'_s$ - $T_s \leq T_s$, where $T'_s$ denotes the request-receiving time and represents the acceptable difference between T' and T.
**Step 3:** HCS classifies the $node_{iID}$ to add to related list, such as (*FSP* list, $pd_i$ list, $S_i$ list) then, Save {$node_{iID}$, $node_{ipu}$, $HCS_{\ pu}$} in the BC.
**Step 4:** HCS checks if registered nodeiID its IoHT device (pdi), then need to assign to the FSP, then send {$HCS_{\ pu}$, $Pd_{ipu}$, $FSP_{\ pu}$} to BC.
**Step 5:** After assigning $pd_i$ to the FSP, finally the registration of $node_i$ is successful.

### 6.2.2 Administrator Registration

The administrator ($ADM_i$) is in charge of controlling the system components in the healthcare domain. As a result, the administrator must register specific details, such as (username ($Un_{ADMi}$ ), address ($Ad_{ADMi}$), phone number ($Pn_{ADMi}$ ), password ($Pw_{ADMi}$ ), wallet address ($Wa_{ADMi}$) and fingerprint $Fn_{ADMi}$)) in the HCS once and generates the private key ($ADM_{ipr}$ ) and public key ($ADM_{ipu}$ ) based on the chaotic system as described in sub-section 6.1. Furthermore, $ADM_i$ computes the shard key ($SK_{ADMi}$). Then, HCS submits the following set of steps.

**Step 1**: The HCS computes the following anonymous parameters based on the following:

1) $Un'_{ADMi}$ = h ($Un_{ADMi}$).
2) $Pw'_{ADMi}$ = h($Pw_{ADMi} \parallel Un_{ADMi}$)
3) $F\ 'n_{ADMi}$ = FEX-ADM ($Fn_{ADMi}$ ), where FEX-ADM is the function used for fingerprint pre-processing and feature extraction and then returns the feature-extraction vector; as a result, it refers to level 2 of the administrator's fingerprint-feature extraction.

**Step 2**: The HCS generates the shared key ($SK_{ADMi}$) to encrypt Enc (.) / decrypt Dec (.) data based on symmetric key encryption Counter (CTR) based on the chaotic system.

**Step 3**: The HCS assigns admin ($ADM_i$) information to the fog server $FSP_{ID}$ that connects within the same area.
**Step 4**: The HCS sends ADMi information {$Pw'_{ADMi}$, $Un'_{ADMi}$ $Fn'_{ADMi}$} to the Blockchain by calling the smart-contract registration method.

### 6.2.3 Patient Registration

In this part, the patient ($P_i$) who wishes to register in the system must do the following steps.

**Step 1**: The patient should register his/her information such as (username ($Un_{Pi}$ ), address ($Ad_{Pi}$ ), phone number ($Pn_{Pi}$), password ($Pw_{Pi}$), wallet address ($Wa_{Pi}$), type of disease $Td_{Pi}$)) in the *HCS* computed $HP_{Pi}$ anomaly by calculating $HP_{\ Pi}$ = H ($Un_{Pi}$ || $Pw_{Pi}$), then stores it in the *BC* through a smart contract.
**Step 2:** The patient generates the private key ($Pi_{pr}$ ) and public key ($Pi_{pu}$ ) based on the chaotic system.
**Step 3**: The patient computes a shared key ($SK_{Pi}$), ensuring that the encryption (Enc(.)) and decryption (Dec(.)) processes for safeguarding Si sensitive health information data are carried out with a robust key based on the chaotic logistic system.
**Step 4**: HCS creates an Electronic Health Record ($EHR_{Pi}$) with all of the aforementioned medical information associated with a new patient.
**Step 5**: HCS assigns Patient ($P_i$) information to the fog server $FSP_{ID}$ that connects within the same area.
**Step 6**: HCS sends the patient ($P_i$) information ($HP_{\ Pi}$) to the BC by calling the smart contract.

### 6.2.4 Doctor Registration

At this time, the doctors ($Dr_i$) send a registration request to HCS with their personal information, such as (username ($Un_{Dri}$ ), address ($Ad_{Dri}$), phone number ($Pn_{Dri}$), password ($Pw_{\mathrm{Dri}}$), wallet address ($Wa_{\mathrm{Dri}}$) and specialization ($Sp_{Dri}$)) in the HCS once, which then generates the private key ($Dr_{ipr}$) and public key ($Dr_{ipu}$) based on the chaotic system as described in sub-section 6.1. HCS impalement $HD_{Dri}$ = h ($Un_{Dri}$ || $Pw_{Dri}$), after that, HCS sends the information ($HD_{Dri}$) above to the Blockchain to register the new doctor.

## 6.3 Login and Authentication Phase

In this phase, once all entities are registered, the login and authentication phase of user, like Administrator, Patient, Doctor, is describe below.

### 6.3.1 Administrator Login and Authentication

Here, the main interaction occurs between the two basic parts (*FSP*) and the system administrator ($ADM_i$). Through this structure, all system privileges are linked with $ADM_i$ accessed and overseed critical processes and system components overseen; the phase is defined as follows:

**1:** ADMi inputs $\mathrm{Un}_{ADMi}$, $\mathrm{Pw}_{ADMi}$ and selects a $r_i \in Z^*$. Then, $ADM_i$ computes A = h ($\mathrm{Un}_{ADMi}$) and $HA_{ADMi}$=h($Pw_{ADMi}$‖$Un_{ADMi}$‖h($r_i$)).
**2:** $ADM_i$ encrypts ($r_i$) using the $SK_{ADMi}$, $E = Enc_{SK_{ADMi}}(r_i)$.
**3:** $ADM_i$ submits the login request $HA_{ADMi}$, E, A to the FSP as the first factor.
**4:** When the FSP obtains the login credentials from $ADM_i$, it performs the following verifications:
  **a.** FSP checks if A =? Un′ADMi. If true, the FSP retrieves $r'_i$, where $r_i' = Dec_{SK_{ADMi}}(E)$.
  **b.** The FSP fetches $Pw'_{ADM_i}$ from the BC, calculates $HA'_{ADM_i} = h(Pw'_{ADM_i} \| h(r_i'))$, and verifies if $HA_{ADM_i =} ?\, HA'_{ADM_i}$. If the verification passes, the FSP sends a challenge vc *via* email to Admin.
**5:** Upon receiving vc′ from FSP, $ADM_i$ evaluates L= h(FEX-ADM ($Fn'_{ADMi}$)⊕vc′⊕h(ri)) and transmits L to FSP.
**6:** When FSP obtains L from $ADM_i$, it retrieves $Fn_{ADMi}$ from the BC and computes L′ = h( $Fn'_{ADMi}\oplus$ $vc'\oplus$ h($r_i'$ ). The FSP then compares L and L′. If L == L′, the FSP confirms the successful authentication of ADMi, granting access to the system’s resources and services. Otherwise, the login process is denied.

**Remark 1:** Even though our work used fingerprints, it can also deal with biometric-based authentication methods, such as facial recognition, iris scanning, keystrokes and voice authentication.

### 6.3.2 Users' (Patients' and Doctors') Login and Authentication

At this stage, the user (*$U_i$*) requests access to the system's resources and services by providing valid credentials, outlining the procedural steps as follows:

**1:** $U_i$ inputs $Un_{Ui}$, $Pw_{Ui}$ and selects a $r_i \in Z^*$. Additionally, it computes $A = h(Un_{U_i})$ and $HU_{U_i} = \mathrm{H}(Pw_{U_i} \| Un_{U_i} \| \mathrm{h}\ (r_i))$.
**2:** $U_i$ encrypts $r_i$ using $SK_{Ui}$ , $E = Enc_{SK_i}(r_i)$ *via* symmetric encryption.
**3:** $U_i$ submits the login parameters $\{HA_{U_i}, \mathrm{E}, A\}$ to FSP as the first factor for authentication.
**4:** Upon receiving login parameters from $U_i$, FSP validates:
 **a.** The FSP checks if $A =?\, Un'_{U_i}$. If matched, the FSP retrieves $r'_i$, where $r'_i = Dec_{SK_{U_i}}(E)$.
 **b.** FSP fetches $Pw'_{U_i}$ from BC, computes $HA'_{U_i} = \mathrm{h}(Pw'_{U_i} \| \mathrm{h}\ (r'_i))$ and verifies whether $HA_{U_i} = ?\, HA'_{U_i}$.
 If the confirmation is positive, the FSP sends a challenge vc *via* email to Admin.
**5:** Upon receiving vc′, $U_i$ performs $L = \mathrm{h}(Wa_{U_i} \oplus vc' \oplus \mathrm{h}(r_i)$ and transmits L toward FSP.
**6:** When FSP obtains L from Ui, it retrieves $Wa_{U_i}$ from the BC, evaluates $L' = \mathrm{h}(Wa_{U_i} \oplus vc \oplus h(r'_i))$, h (r′)), The FSP then compares L and L′. If L == L′, FSP verifies Ui authentication and grants access to system resources. Otherwise, the login request is denied.

**Remark 2:** The login and authentication process for doctors follows procedure for $P_i$. $Dr_i$, necessity inputs a valid credential to gain an access to system services, allowing to review $EHR_{P_i}$ and making updates as permitted by roles and privileges assigned by the administrator.

## 6.4 Secure Construction Phase

The initiation of this phase includes the assignment of the responsibility for formulating a construction group secret (*GS*) to the *FSP*. The procedural steps are outlined as follows:

**1:** *HCS* picks $T_s$ and calls the smart-contract method of the BC to assign *FSP* by creating the following transaction (TR):

$$TR\ \{HCS_{Pu},\ FSP_{Pu},\ T_s\}.$$

**2:** The smart-contract mechanism verifies HCSP u and assesses the transaction's freshness using the criterion $T'_s - T_s \leq T_s$. Subsequently, it validates whether $FSP_{pu}$ corresponds to the designated fog server and, if affirmative, activates *FSP*.
**3:** $Node_{Fog}$ initiates a request to *FSP* for the creation of $Node_{Fog_{GS}}$ dedicated to its group members. Following this, $Node_{Fog}$ selects a time ($T_s$) and employs $FSP_{pu}$ to encrypt a message. Subsequently, $Node_{Fog}$ transmits the encrypted message {E(*Msg* || $T_s$)} to *FSP*.
**4:** FSP decrypts the encrypted message {E(M sg || Ts)} through the utilization of FSPP r. Subsequently, FSP assesses the timeliness of the message by verifying $T'_s - T_s \leq T_s$.
**5:** Utilizing the smart contract, *FSP* employs a transaction to retrieve $Node_{D_{pu}}$, where ($Node_D$ represents a medical device), associated with the medical device owned by *FSP* from the *BC*.

$$TR\ \{FSP_{Pu},\ Node_{Fog_{pu}}\}.$$

**6:** *BC* responds by sending the public key of the medical device $Node_{D_{pu}}$to *FSP*.
**7:** *FSP* selects the present timestamp ($T_s$) and signs {$S_i$ ||$T_s$} using $FSP_{Pr}$. Consequently, *FSP* encrypts E $\{(Si\ ||Ts)\}$ utilizing the public key of $Node_{D_{pu}}$. Then E $\{(Si\ ||Ts)_{signed}\}$ signed distributed across the medical devices ($D_1$, $D_2$, ..., $D_n$ ) affiliated with $Node_{Fog}$.
**8:** $Node_D$ decrypts $(Si\ ||Ts)$ using $Node_{D_{pu}}$, subsequently validating $Si\ ||Ts$ through $FSP_{Pu}$. Following this, it performs calculating $T'_s - T_s \leq T_s$. $Node_D$ and saving $S_i$ for future utilization.
**9:** FSP picks $T_s$ and then uses the method in the SM of *BC* to calculate:

$$TR\ \{FSP_{Pu},\ Node_{Fog},\ Node_{GS},\ T_s)\}$$

**10:** *BC* verifies the validity of $FSP_{Pu}$ and evaluates the transaction's timeliness by applying the condition $T'_s - T_s \leq T_s$. Subsequently, it stores $H(Node_{Fog_{GS}})$. In conclusion, the construction of the secret is completed.

## 7. PERFORMANCE ANALYSIS

In this section, we examine the outcome of the simulation process and the evaluation metrics used in our work, with particular attention to the performance assessment of the generation time of private, public, encryption/decryption and shared keys. This analysis considers equivalent key sizes for several 3-D map chaotic public-key cryptosystems and elliptic-curve cryptography (DHECC). Furthermore, we explore the computational costs and smart-contract costs.

### 7.1 Simulation

Simulation was conducted using the widely recognized Ganache simulator tool. Ganache enables the local deployment of the Ethereum blockchain in a controlled environment, offering developers a practical platform for testing and evaluation. The Truffle framework was utilized to test and deploy the smart contracts on the blockchain. Node.js was also utilized in the development of the proposed framework implemented on a Mac OS 476.0.0.0.0 LTS 64-bit platform, equipped with 8 GB of RAM and powered by a Dual-core Intel Core i5 processor operating at 2.7 GHz.

### 7.2 Results and Analysis

To assess the practical effectiveness of the proposed framework across various user roles and functionalities, a comprehensive performance evaluation was conducted using Apache JMeter version 5.6.3. As a widely recognized and robust performance-testing tool, Apache JMeter facilitated an in-depth analysis of the framework's capabilities, simulating real-world scenarios to ensure its reliability and efficiency.

#### 7.2.1 Key-generation Time

The proposed work followed the DHEC key-generator algorithm of Mohammed et al. [54]; the proposed chaotic cryptosystem involves a two-part initiation time, encompassing private-key and public-key generation times. The private key is derived from a logistic chaotic map and after private-key generation, the public key is created using a modified three-dimensional beta transform system. Figure 4 illustrates a comparative analysis of DHEC key-generation times alongside one-dimensional and three-dimensional chaotic key-generation methods. Notably, the DHEC key generator exhibits a significantly slower performance than the chaotic cryptosystem's key generator at equivalent key sizes.

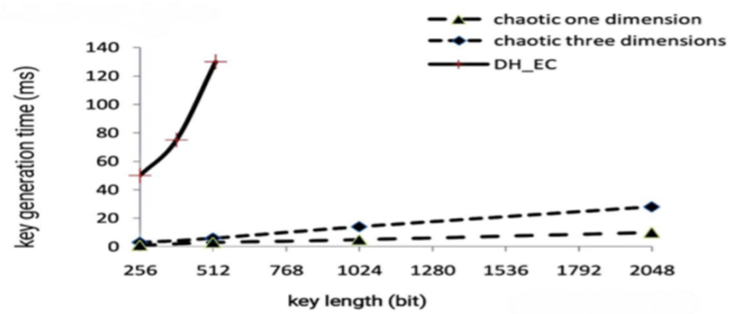
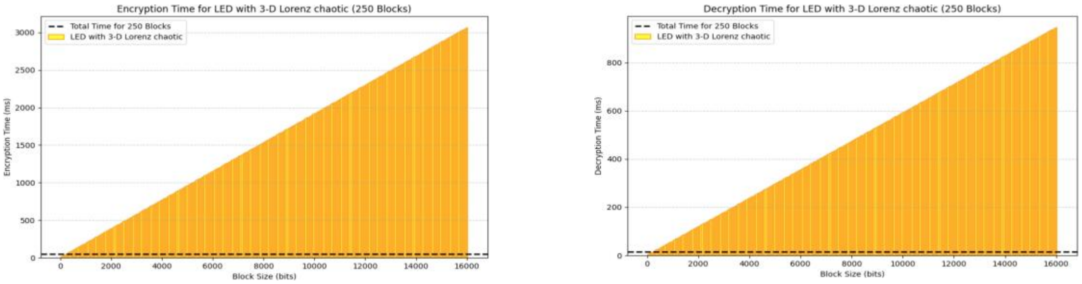


Figure 4. Comparison of DHEC key generation and one-dimensional and three-dimensional chaotic key generation [54].

#### 7.2.2 LED with 3-D Lorenz Chaotic Encryption and Decryption Time

In the context of block-cipher encryption and decryption, the utilization of chaotic maps and the LED algorithm, as proposed by Hussain et al. [55], is explored. The shared key, generated through the chaotic Lorenz map, undergoes a double XOR operation with the state during the encryption process, contributing to the creation of ciphertext for a data block. The user-friendly nature of block ciphers is acknowledged, with an emphasis on the pivotal role of the key-generation process in determining their strength. The enhancement of the LED algorithm through integration with a 3-D Lorenz chaotic map amplifies both diffusion and randomization aspects. This augmentation results in the creation of an unexpectedly robust key, thwarting potential attacks, such as MITM and scan-based attacks. After calculating the completion time for 250 blocks, each block with a size of 64 bits takes 19.1400 ms for the encryption process and 14.7750 ms for the decryption process. Figure 5 illustrates the amount of time taken by the process along the 250 blocks.



(a) Encryption time (b) Decryption time

Figure 5. (a) Encryption time. (b) Decryption time.

### 7.2.3 Computational Cost

The computational cost serves as a metric for measuring the temporal complexity of the proposed methodology within this research paper. Moreover, Table 2 and Figure 6 compare our technique with other relevant research endeavors to thoroughly evaluate its computational efficiency. Within the scope of our investigation, the proposed protocol delineates four distinct phases: setting, registration, login and authentication and secret-construction phases. Our focus will be directed towards analyzing the computational requisites specifically associated with the registration and authentication of the proposed system, as this is the most frequently accessed and utilized in the context of our research. To streamline computational analysis, we establish a clear framework by defining the computational pre-requisites associated with a verification $T_v$, one-way hash function $T_h$, symmetric key encryption and decryption $T_{sym}$, exclusive-or operations as $T$, the paring operation $T_p$ , signature time $T_{sign}$, the exponential operation $T_e$, the one-point addition $T_a$ , the concatenation operation $T$ and one-point multiplication as $T_m$ [56]-[57]. The performance evaluation of the proposed procedure includes a comprehensive comparison with contemporary state-of-the-art schemes, similar to those published in prominent publications, such as Arun et al. [58], Wu et al. [25], Jia et al. [59] and Nora et al. [20]. The proposed solution clearly outperforms them, except that there is a slight difference in the computational-time consumption between our proposed scheme and that of Wu et al. [25]. However, Wu et al. failed to meet security features, such as multi-factor authentication, and to provide a lightweight and distributed model. Moreover, their method does not use blockchain technology.

The time duration for various cryptographic operations is summarized as follows:

The time required for one-point multipli¦ation ($T_m$) is 2.226 ms, while the pairing operation ($T_p$) takes 2.91 ms. The time to generate a signature ($T_{sign}$) is 0.085 ms and the exponential operation ($T_e$) takes 3.85 ms. The concatenation operation ($T||$) is highly efficient, requiring only 0.001 ms. Verification ($T_v$) is performed in 0.09 ms and a one-way hash function ($T_h$) takes 0.0023 ms. The time taken for encryption/decryption ($T_{sym}$) is 0.14 ms, whereas the exclusive OR operation ($T$) takes just 0.001 ms. Finally, one-point addition ($T_a$) is also completed in 0.001 ms.

### 7.2.4 Smart-contract Costs

Our approach leverages Remix as the tool for smart-contract development, formulating the contract through the Solidity language and deploying the compiled contract *via* the Ethereum Ganache tool. To determine the authentic gas costs for individual functions within the smart contract, we employed Meta- mask and Ether-scan. Within the Ethereum blockchain paradigm, fees are defined as the gas required, aligning with the payment or value essential for the successful completion of each transaction or contract execution. A user cannot execute any service and the transaction is deemed illegitimate if he/she does not have an active balance on his/her account. The deployment and expenditures of our proposed contract occur within the Remix IDE and the Meta-mask software cryptocurrency wallet, along with the corresponding blockchain block in Ganache Ethereum. The detailed outcomes of the smart contract costs are systematically documented in Figure 5 and Table 3. It is obvious from the findings presented in Table 3 that our suggested contract entails reduced costs for both deployment and function requests.

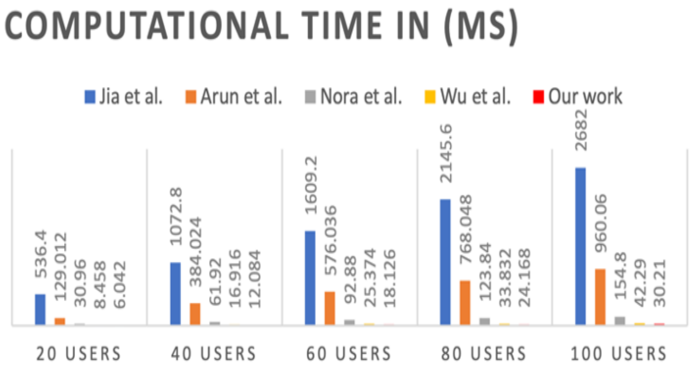


Figure 6. Computational time compared with those of other schemes.

Table 2. Relation for the calculation of computational time for registration and login and authentication phases.

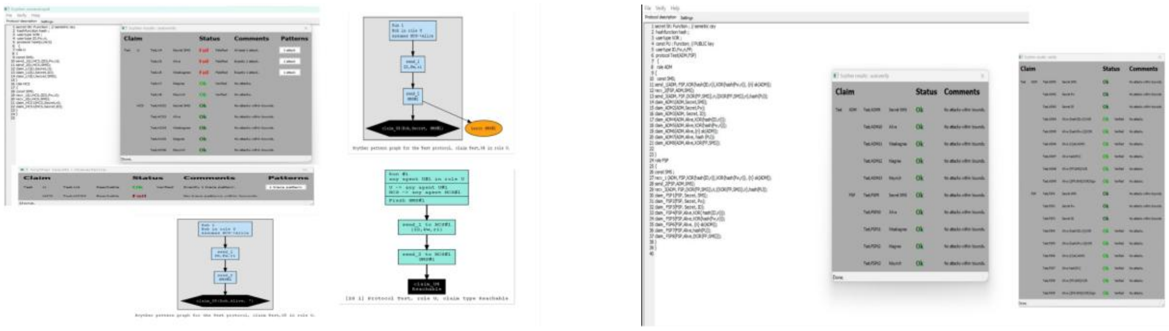
| Schemes | Registration phase | Authentication phase | Total Time (ms) |
|---|---|---|---|
| Jia et al. [59] | $4T_m+T_e+5T_h$ | $T_p+5T_m+(2n+1)T_a+5T_h$ | 26.82 |
| Wu et al. [25] | $8T_h+3T+7T_{\|\|}$ | $35T_h+11T+30T_{\|\|}+2T_{sym}$ | 0.4229 |
| Arun et al. [58] | $2T_m+T_h+4T$ | $T_m+3T+T_h+T_p+T_a$ | 9.6006 |
| Nora et al. [20] | $4T_v+T_{\|\|}$ | $5T_v+2T_{sign}+4T_{sym}+7T_{\|\|}$ | 1.548 |
| Our work | $2T_h+T_{\|\|}$ | $5z_h+2T+2T_{sym}+2T_{\|\|}$ | 0.3021 |

## 8. Security Analysis

The security analysis and experimental results are explained in this section. Furthermore, the security analysis is shown in two ways: the first is a formal analysis using Scyther and the second is an informal analysis using the CK threat model [60]-[61]; after that, we determined that the proposed protocol achieves greater privacy and security than the alternatives. The GUI is intended for anyone who wants to verify or comprehend a protocol. We implemented the proposed system without utilizing security functions in the same traditional systems. Figure 7 refers to the traditional system and explains its shortcomings.

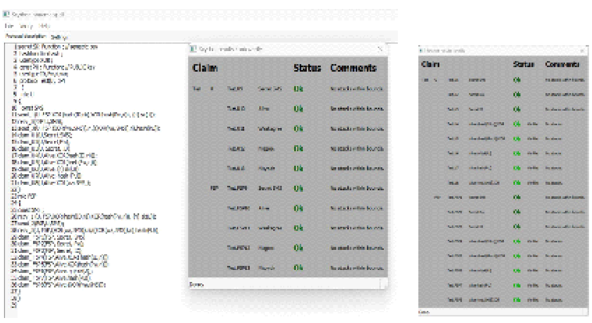
### 8.1 Formal Analysis

In this sub-section, we officially analyze the proposed system and demonstrate the system's data security against various attacks. By utilizing Symmetric Key Encryption, the crypto-hash function and the encryption and decryption function based on a chaotic system, we created a secure system that over-comes the disadvantages of traditional methods. Furthermore, as illustrated in Figure 7, the results of the proposed system are resistant to well-known harmful attacks.



(a)

(b)



(c)

Figure 7. (a) Weakness of the traditional system of user, (b) Admin-verification protocol and automatic climes and (c) User-verification protocol and automatic climes.

Table 3. Smart-contract gas cost.

| **Gas Cost** | **Contact Functions** |
|---|---|
| 711699 | Deploy contract |
| 24765 | Create_User |
| 26621 | Update_User |
| 26621 | Delete_User |
| 175029 | Add_Document |

## 8.2 Informal Analysis

**Therom1:** The proposed work provides mutual authentication
**Proof:** This safety measure indicates that an attacker should be unable to impersonate components of the legal system, such as $U_i$ (Admin, Patient and Doctor). The following six steps were used in this work to authenticate:
**1:** $U_i$ inputs their $Un_{Ui}$, $Pw_{Ui}$ and selects a $r_i \in Z_n^*$. Additionally, it computes A=h($Un_{Ui}$) and $HU_{Ui}$=H($Pw_{Ui}$||$Un_{Ui}$||h ($r_i$)).
**2:** $U_i$ encrypts $r_i$ using $SK_{Ui}$, $E = Enc_{SK_{Ui}}(r_i)$ *via* symmetric encryption.
**3:** $U_i$ submits the login parameters $\{H\ A_{Ui}, E, A\}$ to FSP as the first factor for authentication.
**4:** Upon receiving login parameters from $U_i$, FSP validates:

- **a.** The FSP checks if A =? $Un'_{Ui}$ . If matched, the FSP retrieves $r'_i$, where $r_i' = Dec_{SK_{Ui}}E$.
- **b.** FSP fetches $Pw_{Ui}'$ from BC, computes H A′Ui = h($Pw_{Ui}'$ ||h($r_i'$)) and verifies whether $H\ A_{Ui}$ =? $H\ A'_{Ui}$. If the confirmation is positive, FSP issues a challenge VC to $U_i$ *via* email.

**5:** Upon receiving VC′, $U_i$ evaluates L = h($Wa_{Ui}$⊕vc′⊕h($r_i$)) and transmits L toward FSP.
**6:** When FSP obtains L from $U_i$, it retrieves $Wa_{Ui}$ from the BC, evaluates L′ = h($Wa_{Ui}$⊕vc⊕h ($r_i'$)), the FSP then compares L and L′. If L == L′, FSP verifies $U_i$ authentication and grants access to system resources. Otherwise, the login request is denied.

As a result, our proposed scheme accomplishes mutual authentication between the two entities (Ui, FSP). Otherwise, the current phase is rejected.

**Therom2:** Our proposed work aims to provide support for user anonymity.
**Proof:** Using C.K. adversary's perspective, an adversary has difficulty revealing the user's identity/password.
To reflect anonymity, checking the identity of login information transmitted among system components is currently required. Because the crypto hash function is integrated with $r_i$, which the attacker cannot identify, he/she cannot decipher the user's identity if he/she eavesdrops on the login request. Furthermore, the system generates a unique hash for every login request made by a user depending on the random number $r_i$. During the period of login and authentication phase, $U_i$ sends the login request $\{H\ A_{Ui}, E, A\}$ to the *FSP* as a first authentication factor. Thus, it has been encrypted using a shared key that is known by Ui and FSP only.

An attacker finds it challenging to identify the user and is unable to recover the shared key, which is created just once for each login attempt. This suggests that our proposed scheme can support user anonymity.

**Therom 3:** Our proposed work can provide unlikability.

**Proof:** This feature confirms that an individual can make many login attempts to the FSP to access resources and services without anybody else being able to link the logins together and identify the individual. Under the suggested plan, whenever he/she wants to access the system, he/she sends $\{H\,A_{Ui}, E, A\}$ to FSP. Thus, the basic elements of $\{H\,A_{Ui}, E, A\}$ are constructed once using the following set of points:

- **a.** The FSP verifies whether A =? $Un_{Ui}$. If they match, the FSP restores the $r_i'$, where $r_i' = Dec_{SK_{Ui}}(E)$.
- **b.** The FSP obtains PwU′    from the BC, calculates $HA'_{Ui}$ = h($Pw'_{Ui}$||h($r_i'$)  and checks if $HA_{Ui}$= $HA'_{Ui}$. If the values match, the FSP sends a challenge, which is typically delivered *via* email.
- **c.** After receiving VC′, $U_i$ evaluate L = h($Wa_{Ui}$⊕VC′⊕h($r_i$)) and transmits L back to the FSP.
- **d.** Upon receiving L from $U_i$, the FSP retrieves $Wa_{Ui}$ from the BC, performs L′ = h($WaUi$⊕VC⊕h($r_i'$)), the FSP then compares L and L′. If L == L′, FSP performs $U_i$ authentication and grants access. Otherwise, the login process is denied.

**Therom 4:** Our suggested work can guarantee forward secrecy.

**Proof:** During the login and authentication phase, the widely used session key relies on $SK_{Ui}$. Even if the shared key is revealed or leaked, our suggested system protects the password. The shared key $SK_{Ui}$ is only generated once based on VC, so even if an attacker discloses it, the system's authentication remains secure during subsequent login attempts. It is very difficult for an opponent to determine the random number and password, as well as the characteristic of the crypto one-way hash function $HU_{Ui}$ = h ($Pw_{Ui}$ || $Un_{Ui}$ || h($r_i$)). Furthermore, this is the case if a malicious party can intercept all messages that are sent $\{H\,A_{Ui}, E, A\}$, since these parameters are created just once for each user's login request, so he/she won't be able to use them again to log into the system. Consequently, absolute forward secrecy is guaranteed by our suggested scheme.

**Therom 5:** Our suggested work can resist MITM attacks.

**Proof:** A Man-in-the-Middle attacker intercepts, alters and resends all information during a conversation, without the knowledge of the participants. We presume that the attacker has obtained $\{H\,A_{Ui}, E, A\}$and changed it as $\{HA^*_{Ui}, E^*, A^*\}$. The modified settings are ineffective and do not work because the FSP verifies A and finds (A≠A*), where A represents user identity.  Additionally, the request $\{H\,A_{Ui}, E, A\}$ is generated once for each login. Thus, our suggested work does not allow MITM attacks.

**Therom 6:** Our proposed scheme is resistant to replay attacks.

**Proof:** As per our recommended plan, any new login attempt must precisely match the FSP parameters $\{H\,A_{Ui}, E, A\}$. These parameters are generated only once for every user's login request based on $r_i$ and cannot be obtained by the user again. Therefore, this prevents any replayed message from being sent for verification, making it impossible for an attacker to launch such an attack. Hence, this technique ensures that the enemy cannot use this type of strike.

**Therom 7:** Our recommended scheme is resistant to eavesdropping.

**Proof:** This is the process for deciphering communications to find information. Each parameter shared between the user and the FSP is used only once $\{H\,A_{Ui}, E, A, VC, SK_{Ui}\}$. Consequently, if these variables are intercepted, the attacker will be unable to access the system. The user sends $\{H\,A_{Ui}, E, A\}$to FSP, then FSP decrypts $r_i$ and sends VC to the user. Finally, the user sends L = h ($Wa_{Ui}$ ⊕VC′⊕h($r_i$)) to the FSP. As we notice, these parameters are generated once. Accordingly, the recommended scheme is resistant to eavesdropping.

**Therom 8:** Our proposed scheme affords key management.

**Proof:** For every login request, the principal parties have consented to generate a shared key using chaotic key management and public-key cryptography to ensure the security of the shared key (*SK*) between the user and the *FSP* based on ($r_i$, $SK_{Ui}$). Once the patient checks in successfully, the following actions are carried out by the primary parties ($U_i$, *FSP*) to carry out this phase:

- **a.** $U_i$ calculates $SK_{Ui}$  = $SK_{Ui}$ ⊕ $r_i$.
- **b.** *FSP* side computes $SK_{Ui}$= $SK_{Ui}$ ⊕$r_i'$.

**Therom 9:** Our offered scheme withstands an insider attack.
**Proof:** Here, instead of sending these parameters ($Pw_{Ui}$, $Un_{Ui}$), users provide {$HA_{Ui}$, E, A} when they register with FSP, where $HU_{Ui}$ = h ($Pw_{Ui} \parallel Un_{Ui}$), $E = Enc_{SK_{Ui}}(r_i)$, A = h ($Un_{Ui}$ ). It's difficult for an attacker to use a one-way hash function to get the user's password from the hashed result. Also, to pretend to be a real user, the attacker must create a genuine login-request parameter. However, the attacker will be unable to obtain the user's shared key ($SK_{Ui}$) or forge such parameters.

**Theorem 10:** Our scheme withstands a 51% attack.
**Proof:** A 51% attack in a blockchain network refers to the situation where an entity acquires more than a half of the network's mining power, enabling it to alter transactions. To resist such attacks, a distributed network is maintained where no single authority has control over the network's computer capacity. Resistance attack calculation: Let N represent the overall hashing power of the network. Let H denote the hashing power held by the attacker in order to execute a 51% attack effectively. To carry out such an attack, the attacker must have control over more than 51% of the total hashing power, where H is less than 0.5 times N and greater than 0.5 times N. Thus, the ability to prevent a 51% attack is determined by the level of decentralization in the network, where no single entity possesses the majority of the hashing power.

**Theorem 11:** Our scheme withstands a hijacking attack.
**Proof:** Blockchain technology utilizes robust cryptographic methods to safeguard data and transactions. By employing methods, such as digital signature and encryption, one may effectively check the legitimacy of both the user and the data. This, in turn, thwarts any efforts at hijacking by implementing multi-factor authentication. It enhances security by implementing an additional layer that necessitates users to submit several forms of verification, such as passwords, biometrics and OTP, before gaining access to the system. This enhances the level of complexity for potential attackers attempting to seize control of user accounts.

## 9. CONCLUSION

This research elucidates significant concerns about privacy and security within the IoHT industry. Our developed system integrates fog computing, extends cloud services to network peripheries and offers enough computational support for IoHT devices, so that there is less communication needed for authentication. Our work uses a chaotic key cryptosystem that works with random keys, is small, reduces communication needs and is the right size for IoHT devices' limited processing power. In addition, to protect people and organizations that use public channels in a decentralized setting, our research combines an authentication system with blockchain technology that can't be changed. Furthermore, blockchain technology facilitates decentralized node identification. According to the results of the evaluation, the proposed approach is very reliable and scalable, which means that it will provide strong security and be resistant to common attacks. In addition, it has less latency than current blockchain-based authentication systems, which shows how useful and efficient it is in the real world. We conducted a simulation of the proposed project using the Ethereum platform, Ganache and the Solidity programming language for the deployment and testing of smart contracts. Additionally, we used the Apache JMeter tool to assess both latency and throughput, with a time cost of 0.3021 ms, an average registration delay of 1.25 ms and an authentication time of 1.50 ms. A security study of the suggested method was conducted using the Scyther tool. The formal and informal security assessments demonstrated that the proposed method is secure and resilient against possible assaults. Furthermore, our research bolstered the scalability of the IoHT system. Future development plans also include the use of quantum cryptography as an alternative to existing technology and the utilization of 6G networks to enhance speed and efficiency.

## REFERENCES

[1] A. H. Ameen, M. A. Mohammed and A. N. Rashid, "Dimensions of Artificial Intelligence Techniques, Blockchain and Cyber Security in the Internet of Medical Things: Opportunities, Challenges and Future Directions," Journal of Intelligent Systems, vol. 32, no. 1, p. 20220267, 2023.

[2] B. Pradhan, S. Bhattacharyya and K. Pal, "IoT-based Applications in Healthcare Devices," Journal of Healthcare Engineering, vol. 2021, no. 1, p. 6632599, 2021.

[3] N. Garg et al., "BAKMP-IoMT: Design of Blockchain Enabled Authenticated Key Management Protocol for Internet of Medical Things Deployment," IEEE Access, vol. 8, pp. 95956–95977, 2020.

[4] Y. Aydin, G. K. Kurt, E. Ozdemir and H. Yanikomeroglu, "A Flexible and Lightweight Group Authentication Scheme," IEEE Internet of Things Journal, vol. 7, no. 10, pp. 10277–10287, 2020.

[5] S. M. Umran et al., "Secure Data of Industrial Internet of Things in a Cement Factory Based on a Blockchain Technology," Applied Sciences, vol. 11, no. 14, p. 6376, 2021.

[6] M. El-Hajj, A. Fadlallah, M. Chamoun and A. Serhrouchni, "A Survey of Internet of Things (IoT) Authentication Schemes," Sensors, vol. 19, no. 5, p. 1141, 2019.

[7] G. Apruzzese et al., "The Role of Machine Learning in Cyber-security," Digital Threats: Research and Practice, vol. 4, no. 1, pp. 1–38, 2023.

[8] S. Matsumoto et al., "Authentication Challenges in a Global Environment," ACM Transactions on Privacy and Security (TOPS), vol. 20, no. 1, pp. 1–34, 2017.

[9] S. C. Seak et al., "A Centralized Multi-modal Unified Authentication Platform for Web-based Application," Proc. of the World Congress on Engineering and Computer Science (WCECS 2014), vol. 1, San Francisco, USA, 2014.

[10] D. Nkomo and R. Brown, "Hybrid Cyber Security Framework for the Internet of Medical Things," Blockchain and Clinical Trial: Securing Patient Data, Part of the Book Series: Advanced Sciences and Technologies for Security Applications (ASTSA), pp. 211–229, 2019.

[11] M. Jmaiel et al., "The Impact of Digital Technologies on Public Health in Developed and Developing Countries," Proc. of the 18th Int. Conf. on Smart Homes and Health Telematics (ICOST 2020), vol. 12157, Hammamet, Tunisia, 2020.

[12] I. Purdon and E. Erturk, "Perspectives of Blockchain Technology, Its Relation to the Cloud and Its Potential Role in Computer Science Education," Engineering, Technology & Applied Science Research, vol. 7, no. 6, 2017.

[13] N. Deepa et al., "A Survey on Blockchain for Big Data: Approaches, Opportunities and Future Directions," Future Generation Computer Systems, vol. 131, pp. 209–226, 2022.

[14] S. M. Umran et al., "Secure and Privacy-preserving Data-sharing Framework Based on Blockchain Technology for Al-Najaf/Iraq Oil Refinery," Proc. of the 2022 IEEE Smartworld, Ubiquitous Intelligence & Computing, Scalable Computing & Communications, Digital Twin, Privacy Computing, Metaverse, Autonomous & Trusted Vehicles (Smart-World/UIC/ScalCom/DigitalTwin/PriComp/Meta), pp. 2284–2292, 2022.

[15] H. Sheth and J. Dattani, "Overview of Blockchain Technology," Asian J. For Convergence in Technology (AJCT), vol. 5, no. 1, 2019.

[16] S. M. Umran et al., "Multi-chain Blockchain Based Secure Data-sharing Framework for Industrial IoTs Smart Devices in Petroleum Industry," Internet of Things, vol. 24, p. 100969, 2023.

[17] S. Tuli, R. Mahmud, S. Tuli and R. Buyya, "FogBus: A Blockchain-based Lightweight Framework for Edge and Fog Computing," Journal of Systems and Software, vol. 154, pp. 22–36, 2019.

[18] H. Reffad, A. Alti and A. Almuhirat, "A Dynamic Adaptive Bio-inspired Multi-agent System for Healthcare Task Deployment," Engineering, Technology & Applied Science Research, vol. 13, no. 1, pp. 10192–10198, 2023.

[19] O. Umoren, R. Singh, S. Awan, Z. Pervez and K. Dahal, "Blockchain-based Secure Authentication with Improved Performance for Fog Computing," Sensors, vol. 22, no. 22, p. 8969, 2022.

[20] N. Alsaeed, F. Nadeem and F. Albalwy, "A Scalable and Lightweight Group Authentication Framework for Internet of Medical Things Using Integrated Blockchain and Fog Computing," Future Generation Computer Systems, vol. 151, pp. 162–181, 2024.

[21] R. Almadhoun et al., "A User Authentication Scheme of IoT Devices Using Blockchain-enabled Fog Nodes," Proc. of the 2018 IEEE/ACS 15th Int. Conf. on Computer Systems and Applications (AICCSA), pp. 1–8, Aqaba, Jordan, 2018.

[22] A. Mehmood et al., "Anonymous Authentication Scheme for Smart Cloud Based Healthcare Applications," IEEE Access, vol. 6, pp. 33552–33567, 2018.

[23] Y. Liang, "Identity Verification and Management of Electronic Health Records with Blockchain Technology," Proc. of the 2019 IEEE Int. Conf. on Healthcare Informatics (ICHI), pp. 1–3, Xi'an, China, 2019.

[24] X. Cheng, F. Chen, D. Xie, H. Sun and C. Huang, "Design of a Secure Medical Data Sharing Scheme Based on Blockchain," Journal of Medical Systems, vol. 44, no. 2, p. 52, 2020.

[25] T.-Y. Wu et al., "Improved ECC-based Three-factor Multiserver Authentication Scheme," Security and Communication Networks, vol. 2021, no. 1, p. 6627956, 2021.

[26] Y. Guo and Y. Guo, "FogHA: An Efficient Handover Authentication for Mobile Devices in Fog Computing," Computers & Security, vol. 108, p. 102358, 2021.

[27] I. T. Javed et al., "Health-ID: A Blockchain-based Decentralized Identity Management for Remote Healthcare," Healthcare, vol. 9, no. 6, p. 712, MDPI, 2021.

[28] I.-T. Chen, J.-M. Tsai, Y.-T. Chen and C.-H. Lee, "Lightweight Mutual Authentication for Healthcare IoT," Sustainability, vol. 14, no. 20, p. 13411, 2022.

[29] O. Umoren, R. Singh, Z. Pervez and K. Dahal, "Securing Fog Computing with a Decentralized User Authentication Approach Based on Blockchain," Sensors, vol. 22, no. 10, p. 3956, 2022.

[30] H. Miriam et al., "Secured Cyber Security Algorithm for Healthcare System Using Blockchain Technology," Intelligent Automation & Soft Computing, vol. 35, no. 2, 2023.

[31] S. Nakamoto, "Bitcoin: A Peer-to-peer Electronic Cash System," [Online], Available: https://www.ussc.gov/sites/default/files/pdf/training/annual-national-training-seminar/2018/Emerging_Tech_Bitcoin_Crypto.pdf, 2008.

[32] A. A.-N. Patwary et al., "FogAuthChain: A Secure Location-based Authentication Scheme in Fog Computing Environments Using Blockchain," Computer Communications, vol. 162, pp. 212–224, 2020.

[33] W. J. Gordon et al., "Blockchain Technology for Healthcare: Facilitating the Transition to Patient-driven Interoperability," Computational and Structural Biotechnology J., vol. 16, pp. 224– 230, 2018.

[34] P. P. Ray, D. Dash, K. Salah and N. Kumar, "Blockchain for IoT-based Healthcare: Background, Consensus, Platforms and Use Cases," IEEE Systems Journal, vol. 15, no. 1, pp. 85–94, 2020.

[35] Z. Zheng et al., "An Overview of Blockchain Technology: Architecture, Consensus and Future Trends," Proc. of the 2017 IEEE Int. Congress on Big Data (BigData Congress), pp. 557–564, Honolulu, USA, 2017.

[36] N. Z. Benisi, M. Aminian and B. Javadi, "Blockchain-based Decentralized Storage Networks: A Survey," Journal of Network and Computer Applications, vol. 162, p. 102656, 2020.

[37] R. Zhang, R. Xue and L. Liu, "Security and Privacy on Blockchain," ACM Computing Surveys (CSUR), vol. 52, no. 3, pp. 1–34, 2019.

[38] F. Bonomi, R. Milito, J. Zhu and S. Addepalli, "Fog Computing and Its Role in the Internet of Things," Proc. of the 1st Edition of the MCC Workshop on Mobile Cloud Comp. (MCC’12), pp. 13–16, 2012.

[39] Y. C. Hu et al., "Mobile Edge Computing: A Key Technology Towards 5G," ETSI White Paper, vol. 11, no. 11, pp. 1–16, 2015.

[40] H. Sabireen and V. Neelanarayanan, "A Review on Fog Computing: Architecture, Fog with IoT, Algorithms and Research Challenges," ICT Express, vol. 7, no. 2, pp. 162–176, 2021.

[41] M. Yannuzzi et al., "Key Ingredients in An IoT Recipe: Fog Computing, Cloud Computing and More Fog Computing," Proc. of the 2014 IEEE 19th Int. Workshop on Computer-aided Modeling and Design of Communication Links and Networks (CAMAD), pp. 325–329, Athens, Greece, 2014.

[42] M. Mohammed et al., "Chaotic-based Public Key Cryptosystem for PGP Protocol," Proc. of the Int. Conf. on Aerospace Sciences and Aviation Technology, vol. 15, pp. 1–17, The Military Technical College, Cairo, Egypt, 2013.

[43] B. Bai, S. Nazir, Y. Bai and A. Anees, "Security and Provenance for Internet of Health Things: A Systematic Literature Review," J. of Software: Evolution and Process, vol. 33, no. 5, p. e2335, 2021.

[44] V. O. Nyangaresi, "Biometric-based Packet Validation Scheme for Body Area Network Smart Healthcare Devices," Proc. of the 2022 IEEE 21st Mediterranean Electrotechnical Conf. (MELECON), pp. 726–731, Palermo, Italy, 2022.

[45] X. Li, J. Ma and S. Moon, "On the Security of the Canetti-Krawczyk Model," Proc. of the Int. Conf. on Computational and Information Science, Part of the Book Series: Lecture Notes in Computer Science, vol. 3802 pp. 356–363, 2005.

[46] S. Shamshad et al., "An Enhanced Scheme for Mutual Authentication for Healthcare Services," Digital Communications and Networks, vol. 8, no. 2, pp. 150–161, 2022.

[47] Z. Bao et al., "A Group Signature Scheme with Selective Linkability and Traceability for Blockchain-based Data Sharing Systems in E-health Services," IEEE Internet of Things Journal, vol. 10, no. 23, pp. 21115–21128, 2023.

[48] A. A. Mazlan et al., "Scalability Challenges in Healthcare Blockchain System: A Systematic Review," IEEE Access, vol. 8, pp. 23663– 23673, 2020.

[49] G. Somani et al., "DDoS Attacks in Cloud Computing: Issues, Taxonomy and Future Directions," Computer Communications, vol. 107, pp. 30–48, 2017.

[50] N. Sivasankari and S. Kamalakkannan, "Detection and Prevention of Man-in-the-Middle Attack in IoT Network Using Regression Modeling," Advances in Engineering Software, vol. 169, p. 103126, 2022.

[51] C.-M. Chen, Z. Chen, S. Kumari and M.-C. Lin, "LAP-IoHT: A Lightweight Authentication Protocol for the Internet of Health Things," Sensors, vol. 22, no. 14, p. 5401, 2022.

[52] W. Yang et al., "Security Analysis of a Distributed Networked System under Eavesdropping Attacks," IEEE Transactions on Circuits and Systems II: Express Briefs, vol. 67, no. 7, pp. 1254–1258, 2019.

[53] X. Xiang, M. Wang and W. Fan, "A Permissioned Blockchain-based Identity Management and User Authentication Scheme for E-health Systems," IEEE Access, vol. 8, pp. 171771–171783, 2020.

[54] M. T. Mohammed et al., "Chaotic Based Key Management and Public-key Cryptosystem," Int. J. of Computer Science and Telecommunications, vol. 3, no. 11, pp. 35–42, 2012.

[55] H. M. Al-Saadi and I. Alshawi, "Provably-secure Led Block Cipher Diffusion and Confusion Based on Chaotic Maps," Informatica, vol. 47, no. 6, pp. 105-114, 2023.

[56] S. Majumder et al., "Wearable Sensors for Remote Health Monitoring," Sensors, vol. 17, no. 1, p. 130, 2017.

[57] D. Formica and E. Schena, "Smart Sensors for Healthcare and Medical Applications," Sensors, vol. 21, no. 2, p. 543, 2021.

[58] A. S. Rajasekaran et al., "Blockchain Enabled Anonymous Privacy-preserving Authentication Scheme for Internet of Health Things," Sensors, vol. 23, no. 1, p. 240, 2022.

[59] X. Jia, D. He, N. Kumar and K.-K. R. Choo, "A Provably Secure and Efficient Identity-based Anonymous Authentication Scheme for Mobile Edge Computing," IEEE Systems Journal, vol. 14, no. 1, pp. 560–571, 2019.

[60] M. Burrows, M. Abadi and R. Needham, "A Logic of Authentication," DEC System Research Centre Report, ACM Transactions on Computer Systems, vol. 8, no. 1, pp. 18-36, February 1990.

[61] M. N. Aman, K. C. Chua and B. Sikdar, "A Light-weight Mutual Authentication Protocol for IoT Systems," Proc. of GLOBECOM 2017-2017 IEEE Global Communications Conf., pp. 1–6, Singapore, 2017.

## ملخص البحث:

إنّ إنترنت الأشياء الصّحّية عبارة عن شبكةٍ من أجهزة الرّعاية الصّحّية والبرمجيات والأنظمة الّتي تمكّن من الرّصد عن بُعد وتقديم خدمات الرّعاية الصّحّية عبر جمع بياناتٍ تتعلّق بالصّحّة في الزّمن الحقيقي من خلال المجسّات. وعلى الرّغم من فوائدها الجمّة للرّعاية الصّحّية الحديثة الذّكية، فإن إنترنت الأشياء الصّحّية تواجه تحدّياتٍ ترتبط بالأمان ترجع إلى القدرة المحدودة على المعالجة وسعة التّخزين وإمكانيات الدّفاع عن النّفس لأجهزتها. وبينما تمّ تطوير حلول مصادقة قائمة على سلاسل الكُتل ومستندة إلى تحسين أمن البيانات، فإنها تتطلّب مصادر حوسبة عديدة وزيادة في إمكانات التّخزين وتحتاج إلى أزمان طويلة من أجل إنجاز المصادقة، الأمر الّذي يعيق إمكانية توسيعها وزيادة فاعليتها في أنظمة إنترنت الأشياء الصّحّية الموسّعة والّتي يلعب الزّمن فيها دوراً حاسماً.

وللتعامل مع هذه التّحدّيات، نقترح في هذه الورقة نظام مصادقةٍ مكوّناً من أربع مراحل تشمل الإعداد، والتّسجيل، والمصادقة، وإنشاء السّرّية. ويدمج النّظام المقترح بين أنظمة ترميز المفاتيح العامّة بناءً على "الفوضى"، وجهاز ترميز مع خوارزمية خريطة لورنس الفوضوية ثلاثية الأبعاد، وتقنيات الحوسبة الضّبابية القائمة على سلاسل الكُتل؛ من أجل تحسين كلٍّ من الفاعلية وإمكانية التّوسيع.

وبتقييم النّظام المقترح عبر مقارنته بأنظمة مصادقةٍ تقليديةٍ وردت في أدبيات الموضوع، أبدى النّظام المقترح أداءً متفوّقاً مع تخفيض في تكلفة الحوسبة. وقد بلغ معدّل زمن التّأخير اللّازم للتّسجيل (1.25) ميلي ثانية، بينما تكتمل عملية المصادقة في (1.50) ميلي ثانية، الأمر الذي يجعل النّظام المقترح ملائماً لتطبيقات إنترنت الأشياء للرّعاية الصّحية الّتي يُعدّ فيها الزّمن عاملاً حاسماً. وأثبت تحليل الأمان أن النّظام المقترح مقاوم للهجمات الإلكترونية بما فيها هجمات 51% وهجمات الاختطاف، بينما يتمّ الحفاظ على تكامل البيانات وسرّيتها. علاوة على ذلك، فإنّ النّظام المقترح يخفضّ تكلفة الاتّصال ويدعم إمكانية توسيع أنظمة إنترنت الأشياء الصّحّية ذات الحجم الكبير.

والجدير بالذّكر أنّ هذه النّتائج تبين أنّ النّظام المقترح يمتلك الإمكانية لإحداث ثورةٍ في مجال الرّصد الآمن والفعّال لبيانات أنظمة إنترنت الأشياء الصّحّية، الأمر الّذي يمكّن من إدارة البيانات في الزّمن الحقيقي في بيئات إنترنت الأشياء الصّحّية بأمانٍ وفاعلية.

# GROUND-TO-SATELLITE FSO COMMUNICATION: EVALUATING MODULATION TECHNIQUES UNDER CLOUD AND TURBULENCE EFFECTS

Mouna Garai[1], Maha Sliti[1] and Abdelrahman Elfikky[2]

*(Received: 28-Dec.-2024, Revised: 26-Mar.-2025 and 3-Apr.-2025, Accepted: 5-Apr.-2025)*

## *ABSTRACT*

*Free-space optical (FSO) communication is a vital solution to meet the growing demand for high-bandwidth satellite-to-ground communication, offering advantages, such as higher data rates and security compared to traditional RF systems. However, its performance is significantly affected by meteorological conditions, particularly cloud formations (e.g. cirrus, cumulus and stratocumulus) and atmospheric turbulence, which cause signal attenuation, scattering and phase distortions. Addressing these challenges through better understanding and mitigation strategies is essential to ensure reliable and efficient performance of FSO systems under various atmospheric conditions. In this study, we evaluated the performance of ground- to-satellite FSO systems under varying atmospheric turbulence and cloud conditions using the OptiSystem simulator. We analyze multiple modulation techniques, including Quadrature Phase Shift Keying (QPSK), 8-Phase Shift Keying (8PSK), 16PSK and 16-Quadrature Amplitude Modulation (16QAM), to assess their resilience based on link range, bit-error rate (BER), quality factor, optical signal-to-noise ratio (OSNR) and error-vector magnitude (EVM). The results demonstrate that QPSK outperforms higher- order modulation schemes in high-attenuation environments, maintaining the lowest BER and highest quality factor, making it the most suitable choice for FSO communication in satellite networks. These findings provide critical insights into the optimization of modulation strategies for robust and reliable ground-to-satellite optical links.*

## *KEYWORDS*

*Atmospheric turbulence, Cloud attenuation, FSO, Gamma-Gamma turbulence model, Modulation techniques, Optical ground-to-satellite link.*

## 1. INTRODUCTION

Free-space Optical (FSO) communication has emerged as a cornerstone technology in modern telecommunications, widely recognized for its ability to transmit data at exceptional speeds with minimal latency [1]-[4]. This capability makes FSO an indispensable component in telecommunication networks, where it is crucial to enable high-speed Internet connectivity and support data-intensive applications. Beyond its conventional uses, FSO demonstrates remarkable adaptability in various domains, including real-time surveillance, high-definition video broadcasting and bridging the digital divide in underserved and rural regions [5]-[6]. These diverse applications underscore the technology's potential to address complex communication challenges while offering scalable and efficient solutions.

FSO-communication links are highly sensitive to environmental factors, such as fog, rain and snow, which can significantly attenuate the optical signal [7]-[19]. Furthermore, atmospheric turbulence causes beam scintillation and wavefront distortions, leading to deterioration in link quality [20]-[22]. Thus, the adoption of FSO technology in satellite-to-ground communications is hindered by significant challenges posed by atmospheric conditions, including cloud cover and turbulence [23]-[24].These meteorological factors alter the reliability of the communication link, leading to attenuation and signal degradation. This occurs primarily because of photon absorption and scattering caused by the presence of dense cloud formations and water droplets. Moreover, atmospheric turbulence, induced by random fluctuations in temperature and pressure along the signal's propagation path, further compounds these issues. Turbulence introduces random phase distortions, scintillations and beam wanders, severely impacting the system's overall performance and dependability. Overcoming these atmospheric hurdles is vital to achieve the full potential of FSO systems in challenging operational environments.

[1] M. Garai and M. Sliti are with University of Carthage, Higher School of Communication of Tunis (SUP'COM), LR11TIC04, Communication Networks and Security Research Lab. & LR11TIC02, Green and Smart Communication Systems Research Lab, Tunisia. Emails: mouna.garai@gmail.com and {slitimaha@gmail.com; maha.sliti@istic.ucar.tn}.

[2] A. Elfikky is with Arab Academy for Sci. and Technol. and Maritime Support, Alexandria, Egypt. Email: afikky@ucsc.edu

This study examines the influence of turbulence and various types of cloud, such as stratocumulus, cumulus and cirrus clouds, on ground-to-satellite FSO systems. We also use OptiSystem to simulate various modulation schemes, such as Quadrature Phase Shift Keying (QPSK), 8-Phase Shift Keying (8PSK), 16-PSK and 16-Quadrature Amplitude Modulation (16-QAM), which provide important insights into the system's performance under different cloud types in the presence of turbulence. We then compare their robustness using the bit-error rate (BER), quality factor, optical signal-to-noise ratio (OSNR) and error-vector magnitude (EVM). The findings show that QPSK modulation is the optimal choice for FSO communication in satellite networks with cloud-induced attenuation levels. The study revealed that QPSK regularly exceeds alternative modulation techniques, delivering the lowest Bit Error Rate (BER) even in challenging scenarios. QPSK is particularly effective in evaluating the influence of transmitted power, attaining significantly low bit-error rate (BER) values even at low power levels. QPSK and 16-PSK offer improved performance and higher quality-factor values, particularly at shorter distances. QPSK consistently has lower EVM values, demonstrating greater noise tolerance even at low OSNR levels. QPSK is recommended as the modulation technology for FSO communication over satellite links, due to its ability to maintain reliable communication even in unfavorable conditions.

We can summarize the current study's contributions as follows:

- We develop a comprehensive FSO-channel model that integrates the effects of both atmospheric turbulence and cloud attenuation. Unlike conventional models, our approach distinguishes between different cloud types (e.g. cirrus, cumulus, stratocumulus), enabling a more precise characterization of signal degradation.
- Our study systematically compares multiple modulation schemes (QPSK, 8-PSK, 16-PSK and 16-QAM) under identical and realistic atmospheric conditions. This analysis highlights the trade-offs between spectral efficiency and robustness, thereby providing clear insights into optimal modulation strategies for ground-to-satellite links.
- The simulation results obtained through the OptiSystem platform are rigorously validated against analytical models derived from established channel theories. This dual validation confirms the accuracy and reliability of key performance metrics, such as Bit-Error Rate, Optical Signal-to-Noise Ratio and Error-Vector Magnitude.

The following sections are organized as follows: Section 2 examines existing mitigation approaches to reduce the impact of cloud formation on the performance of FSO systems. Section 3 explores the impact of attenuation caused by various types of clouds, as well as an examination of the known models and methodologies used to measure these effects. Section 4 provides a comprehensive description of the design of the proposed ground-to-satellite FSO system, taking into account different cloud conditions and presents the different modulation approaches considered for the proposed ground-satellite FSO system. Section 5 presents the obtained simulation results and provides the analytical validation of the simulation results. Section 6 shows a list of abbreviations used in the study. Finally, Section 7 concludes the paper.

## 2. Related Work

Cloud-induced attenuation in satellite-based free-space optical communications is an important research area [25]-[31]. Numerous studies have provided valuable information on modeling, understanding and mitigation of the impact of clouds on FSO links.

Using the meteorological ERA-Interim database, the authors in [32] examined the availability of links in various regions of Japan. They suggested a site-diversity strategy to enhance system availability and offer practical advice for resolving cloud-induced issues in FSO communications. The authors of [33] investigated cloud-induced attenuation in satellite-based free-space optical communications, with a particular emphasis on regions of Japan. The FSO channel model is selected based on the log-normal distribution, which enhances the understanding of the probabilistic nature of cloud-induced attenuation. The authors of [34] simulated a 30 Gbps ground-to-geostationary satellite-FSO communication link that accommodates a variety of cloud states and atmospheric impacts. The results illustrate the efficacy of a 2x2 MIMO system that employs coherent detection and QPSK modulation, with a particular emphasis on the minimal occurrence of symbol errors in a variety of cloud and weather conditions. An analytical model for estimating the probability of cloud-free line-of-sight

(CFLOS) in optical satellite links is introduced in [35]. This model is predicated on the assumption of a lognormal distribution for Integrated Liquid Water Content (ILWC). This approach offers a practical approach to effectively anticipating and resolving cloud-related challenges. The authors of [36] suggested the development of hybrid free-space optical/radio frequency (FSO/RF) systems to relay satellite communication from high-altitude platforms. This approach uses rate adaptation to adjust the data rates in response to channel-condition oscillations. Additionally, an examination of perfect inter-symbol interference (ISI) caused by cloud effects is conducted, which offers valuable insights into the constraints and restrictions that clouds place on FSO communication lines. The authors of [37] investigated the impact of cloud attenuation on the performance of optical wireless networks. They evaluated various cloud varieties. The research underscores the significance of wavelength-dependent attenuation and illustrates that stratocumulus clouds have the most significant influence on signal transmission. The meteorological ERA-Interim database is employed by the authors in [38] to determine the monthly average cloud attenuation over Japan. Based on the simulation results, the selection of a diverse array of sites from the proposed pool of options leads to a high level of system availability, which assists optical satellite-communication systems in mitigating the effects of cloud-induced attenuation. A comprehensive performance analysis of the atmospheric influence on visibility in FSO systems is conducted in [39], with a particular emphasis on ground station-satellite communications. The objective of the investigation is to establish correlations between visibility and climatic events, including factors, such as precipitation and snowfall. A study conducted in [40] examined the probability of the failure of integrated ground-air-space free-space optical communication lines when various models of atmospheric turbulence are employed. The article provides closed-form equations for cumulative distribution functions (CDFs) and probability density functions (PDFs). These equations take into account the zenith angle, atmospheric disturbances, the channel condition and the configuration of the links. The study offers a comprehensive analysis of the effectiveness of integrated FSO links, accounting for atmospheric attenuation, turbulence, angle of arrival fluctuations and targeting error. The evaluation of cloud-induced optical attenuation is the subject of [41]. This article suggested a comprehensive model that takes into account the influence of clouds on transmitted optical beams. Cloud-induced dispersed optical power is evaluated in this study through the application of modified gamma-particle size distributions (PSDs) and Mie theory. Cirrus clouds and their likelihood of formation, as well as their impact on FSO transceivers situated at elevated altitudes, are the primary focus. This research makes a substantial contribution to our comprehension of the impact of various atmospheric conditions on FSO communication systems, which encompasses both deep-space and near-Earth scenarios. Two distinct models for the scheduling of space-to-ground optical communication that incorporate uncertainty are presented in [42]. The first model employs robust optimization with a moment-based ambiguity set, while the second model employs robust optimization with a polyhedral uncertainty set. The study illustrates the efficacy of formulations that consider uncertainty when employing computational analysis on a real-world communication system, particularly in the context of cloud-cover predictions. The models that have been presented offer critical insights for the scheduling of space-to-ground optical communication systems by acknowledging the dynamic and variable character of the cloud cover.

Several studies have investigated the performance of various modulation schemes in FSO communication systems under atmospheric turbulence. For example, [43] compared several formats, including OOK, BPSK, DPSK, QPSK and 8-PSK and found that BPSK generally provides the lowest BER under severe turbulence. [44] evaluated the performance of schemes, such as BPSK-SIM, DPSK, DPSK-SIM, Polarization Shift Keying and M-ary Pulse Position Modulation and concluded that DPSK often achieves the best outage probability and higher channel capacity under turbulent conditions. In another study, [45] compared PPM, OOK, Differential Pulse Interval Modulation and Dual Header Pulse Interval Modulation in various weather scenarios and observed that PPM and OOK-NRZ generally deliver better BER performance. Recent research in FSO satellite networks has addressed key system-level issues. For example, [46] conducted a link-budget analysis for FSO satellite networks, providing valuable insights into power allocation under various atmospheric conditions. Furthermore, Liang et al. [47] performed a performance analysis of FSO satellite networks that examined transmission power and latency, while Liang et al. [48] explored the trade-off between latency and transmission power in networks with multiple intercontinental connections. These studies provide a strong foundation for the design and optimization of FSO satellite networks.

Although much of the existing literature focused on terrestrial FSO links, our work specifically addresses the challenges of long-range ground-to-satellite communication, including high free-space path loss and atmospheric variability across different altitudes. Using advanced simulation tools, such as OptiSystem, we evaluated detailed performance metrics, such as BER, Q factor and EVM for various modulation schemes. This component-level analysis complements the network-level evaluations found in the literature, providing additional insights necessary for optimizing FSO system performance in practical deployment scenarios.

Table 1. Comparative overview of research in satellite-based free-space optical (FSO) communications under atmospheric conditions.

| Ref. | Methodology | Modeling Technique | Parameters Considered | Data Source |
|---|---|---|---|---|
| [32] | Investigation of site-diversity scheme for enhanced system availability. | Not specified | Clouds, atmospheric turbulence | Meteorological ERA-Interim database [49] |
| [33] | Presentation of a novel distribution model of monthly cloud attenuation for several regions in Japan. | Log-normal distribution | Monthly cloud attenuation, CLWC | ERA-Interimal Meteorologic database |
| [34] | Simulation of a 30 Gbps satellite-FSO communication link. Consideration of atmospheric effects, like different cloud types.<br>Use of a 2×2 MIMO system with QPSK modulation and coherent detection. | Not specified | Haze, fog, cloud types (stratus, cumulus, cumulonimbus), atmospheric turbulence, intensity scintillation | Not specified |
| [35] | Prediction of visibility/range of FSO link due to different cloud conditions. | Not specified | Link length, transmitted power, data rate, cloud types | Not specified |
| [36] | Presentation of analytical models for Cloud-Free Line-of-Sight (CF- LOS) probability. Evaluation of joint CFLOS statistics. | Not specified | Elevation angle, ground-station altitude, spatial variability of clouds | Not specified |
| [56] | Addressing the design of hybrid FSO/RF systems for high-altitude platform (HAP)-aided relaying satellite communication. | Rate-adaptation design | Beam-spreading loss, cloud attenuation, atmospheric turbulence, pointing misalignment | Not specified |
| [37] | Investigation of bit-error rate performance of intensity-modulated FSO with direct detection (IM/DD) in single-input single- output (SISO) due to beam broadening at the receiver caused by cloud. | IM/DD with direct detection & Perfect ISI due to beam broadening by clouds | Not specified | Not specified |
| [38] | Study of the influence of cloud attenuation on the performance of optical wireless links. | Not specified | Received power, visibility range, cloud type | Not specified |
| [39] | Study of a placement method of optical ground stations (OGSs) to realize site diversity in optical satellite-to-ground communications under cloud attenuation. | Greedy heuristic method | Monthly average cloud attenuation | Meteorological ERA-Interim database |
| [40] | Performance analysis for atmospheric influence on visibility in Free Space Optical Communications (FSOC). | Not specified | Atmospheric events (rain, snow), relationships to visibility | Not specified |
| [41] | Analysis of the outage probability of integrated ground-air-space FSO communication links for different atmospheric turbulence channel models. | Lognormal, Gamma exponentiated Weibull distributed channel models | Zenith angle, channel state, deviations, altitude, beam waist, …etc. | Not specified |
| [42] | Estimation of cloud-induced optical attenuation over near-Earth and deep-space FSO-communication systems. | Optical thickness parameter, modified gamma PSD, Mie theory | Type of atmospheric clouds, ground-space FSO link distances | Not specified |
| [50] | Provision of two alternative models of uncertainty for cloud-cover predictions. A robust optimization model with a polyhedral uncertainty set and a distributionally robust optimization model with a moment-based ambiguity set. | Robust optimization, distributionally robust optimization | Cloud-cover predictions, satellite operation scheduling | Official weather forecasts |

Furthermore, while existing studies primarily emphasized BPSK and DPSK for their robustness, our work focuses on QPSK modulation for ground-to-satellite FSO links. Although QPSK may be slightly more sensitive to atmospheric disturbances than BPSK, it offers a substantial advantage in terms of spectral efficiency, an important factor for high-data-rate applications. The literature indicates that, while BPSK achieves the lowest BER under severe turbulence, its lower spectral efficiency limits its utility in bandwidth-constrained environments. DPSK, on the other hand, offers a balanced trade-off, but often requires more complex receiver designs. Our results demonstrate that QPSK delivers an acceptable BER while significantly increasing data throughput, making it a compelling option for space-to-ground communications. In addition, our detailed comparative evaluation of multiple modulation techniques under realistic channel conditions, combined with rigorous simulation and analytical validation, clearly demonstrates the superior performance of QPSK. This distinct integration of simulation and analytical validation not only reinforces the reliability of our findings, but also sets our research apart from the existing literature.

Table 1 compares the different approaches considered in the literature to mitigate the impact of clouds on FSO links.

## 3. FSO Channel Model

FSO communication systems are extremely sensitive to atmospheric conditions, which can dramatically reduce signal quality and system performance. This section gives a complete FSO-channel model that includes both atmospheric-turbulence and cloud-attenuation effects.

### 3.1 Atmospheric Turbulence

Atmospheric turbulence is a random phenomenon produced by differences in temperature and pressure along the transmission channel. Turbulence causes fluctuations in the temperature, pressure and, most importantly, the refractive index of the air. These changes in the refractive index affect the transmission of optical signals, resulting in scintillation, which is the fluctuation of the intensity of transmitted light.

In free-space optical satellite ground communication, the gamma-gamma distribution is commonly used to characterize the fluctuating received optical power (*I*) due to atmospheric turbulence. This model is recommended over others, because it successfully combines both small- and large-scale turbulence effects, making it applicable to a wide variety of turbulence situations, from moderate to high. The gamma-gamma model is particularly useful in satellite-ground communication because of the various turbulence levels experienced as the optical signal passes through several atmospheric layers, each with specific turbulence characteristics.

The gamma-gamma distribution's probability density function (PDF) is as follows ([52]):

$$p_I(I) = \frac{2\alpha^{\alpha}\beta^{\beta}}{\Gamma(\alpha)\Gamma(\beta)}\frac{1}{I_0}\left(\frac{1}{I_0}\right)^{\alpha-1}\exp(-\beta\frac{1}{I_0}) \tag{1}$$

where $I$ is the optical power received, α and β are shape parameters related to the small- and large-scale turbulence effects, Γ(.) denotes the gamma function and $I_0$ is a normalization constant related to the average optical power received. Parameters α and β can be determined based on the strength of atmospheric turbulence, characterized by the refractive index structure parameter $C_n^2$, the propagation distance $L$ and the wavelength $\lambda$.

The expressions for the parameters $\alpha$ and $\beta$ are given by:

$$\alpha = \left(exp\left(\frac{0.49\sigma_R^2}{\left(1+1.11\sigma_R^{12/5}\right)^{7/6}}\right)-1\right)^{-1} \tag{2}$$

$$\beta = \left(exp\left(\frac{0.51\sigma_R^2}{\left(1+0.69\sigma_R^{12/5}\right)^{5/6}}\right)-1\right)^{-1} \tag{3}$$

The strength of atmospheric turbulence is often quantified by the Rytov variance, denoted by $\sigma_R^2$. For a plane wave propagating through the atmosphere, the Rytov variance is given by [53]:

$$\sigma_R^2 = 1.23C_n^2k^{7/6}L^{11/6} \tag{4}$$

where $C_n^2$ is the refractive index structure parameter, indicating the strength of turbulence, $k = \frac{2\pi}{\lambda}$ is the number of optical waves, λ is the wavelength of the optical signal and $L$ is the length of the propagation path.

In vertical FSO communication, the calculation of atmospheric turbulence becomes more complicated due to the varying refractive index with height. The refractive index structure parameter ($C_n^2$) measures the intensity of refractive-index fluctuations in the atmosphere and varies both spatially and temporally. Higher $C_n^2$ values indicate more turbulence, leading to increased scintillation and signal loss in optical communication.

Using the Hufnagel-Valley model, the fluctuation of the refractive-index structure parameter $C_n^2(h)$ with altitude $h$ is expressed as ([54]):

$$C_n^2(h) = 0.00594\left(\frac{v_{wind}}{27}\right)^2 (10^{-5}h)\exp\left(-\frac{h}{1000}\right) + 2.7\times10^{-16}\exp\left(-\frac{h}{1500}\right) + C_n^2(0)\exp(-\frac{h}{100}) \quad (5)$$

where, $C_n^2(0)$ is the level of ground turbulence, ranging from $10^{-17}$ $\mathrm{m}^{-2/3}$ (weak turbulence) to $10^{-13}$ $\mathrm{m}^{-2/3}$ (strong turbulence) and $\upsilon_{wind}$ (m/s) is the mean squared root wind speed, typically around 21 m/s.

### 3.2 Cloud Attenuation Analysis

The presence of liquid-water particles in clouds has a significant impact on FSO communication with atmospheric conditions. When laser beams pass through the Earth's atmosphere, these small-scale components cause light waves to scatter in multiple directions, inhibiting coherent transmission. Cloud-induced scattering has a significant impact on visibility, which is a crucial factor in evaluating the effectiveness of optical-communication systems. Visibility, measured in kilometers, is a quantifiable measure of the clarity and transparency of the atmosphere. The loss of signal power in FSO networks caused by cloud scattering's effect on vision indicates that the optical transmission conditions are insufficient. In [35], the authors provided many types of cloud with variable attenuation effects, as depicted in Figure 1. Table 2 compares their impacts on the FSO signal.

- Cirrus, cirrostratus and cirrocumulus are high-altitude clouds. These clouds, which exist at high elevations and low temperatures, help to reduce solar radiation and maintain a delicate balance of thermal exchanges in the Earth's atmosphere.
- Altocumulus and altostratus are mid-level clouds that migrate at high and low altitudes. Their role in controlling solar radiation is crucial, as they have a transitional effect on thermal dynamics that determines surface temperatures and atmospheric conditions.
- Cumulus, stratus and stratocumulus are the most common low-level clouds. These clouds, located closer to the Earth's surface, play a crucial role in temperature regulation and atmospheric stability. Cumulus clouds, with their puffy and distinct appearances, denote fair weather, whereas stratus clouds, with their blanket-like patterns, typically imply cloudy conditions. Stratocumulus clouds have both stratus and cumulus features and act as transitory elements in meteorological processes.
- Raining clouds are multi-layered clouds that appear at all levels of the atmosphere. This classification includes nimbostratus and cumulonimbus clouds. Nimbostratus clouds generate huge, featureless strata linked by persistent precipitation, contributing to the replenishment of the Earth's water resources. Cumulonimbus clouds, popularly known as the "king of clouds," spread across multiple atmospheric strata, encapsulating dynamic convective processes that create severe weather events, such as thunderstorms and heavy rainfall.

Equation (6) shows how to determine visibility based on the concentration of the number of cloud droplets (Nc) and the content of liquid cloud water (CLWC).

$$V = \frac{1.002}{(L\times Nc)^{0.6473}} \quad (6)$$

where $L$ (g/m$^3$) represents the mean CLWC and $Nc$ (cm$^{-3}$) denotes the concentration of cloud droplet number [55].

Figure 2 shows the impact of different types of cloud on visibility, including stratus ($Nc$ = 250 cm$^{-3}$), altostratus (Nc = 400 cm$^{-3}$) and nimbostratus ($Nc$ = 200 cm$^{-3}$). Figure 2 also shows a rapid decline in visibility as CLWC increases.

For a mono-dispersed droplet distribution, Equation (7) can be used to calculate the concentration of cloud droplets.

$$Nc = \frac{L}{\frac{4}{3}\Pi r^3\rho\times10^{-6}} \quad (7)$$

where ($\rho$ = 1 g/cm3) signifies the density of liquid water and $r(\mu m)$ represents the average radius of cloud droplets. The value varies with cloud type, such as stratus ($r$ = 3.33$\mu m$), nimbostratus ($r$ = 4.7$\mu m$) and cumulus ($r$ = 6.0$\mu m$) ([57]).

Equation (8) expresses the total cloud attenuation in the cloud layers considered, indicated as $A_c$ (dB),

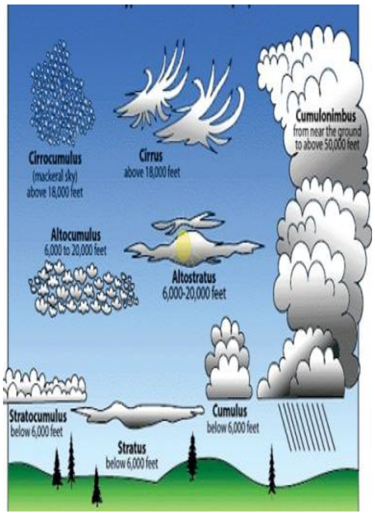


Figure 1. Cloud types.

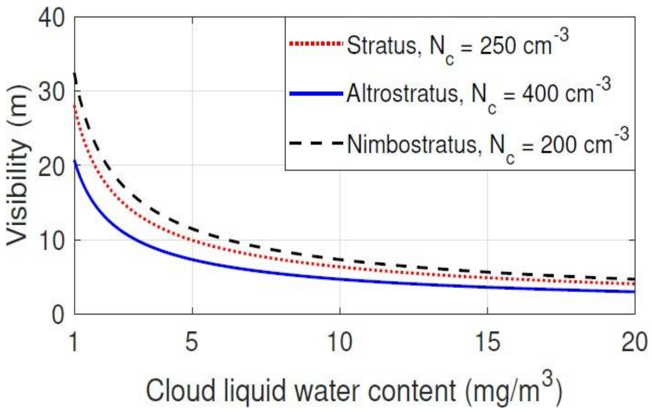


Figure 2. Visibility for several cloud types [56].

using visibility V (km) in Equation (6) and the Kim model [51] to represent the attenuation due to Mie scattering.

$$A_c = \sum_{k=1}^{M} 4.34 \left(\frac{3.91}{V_k}\left(\frac{\lambda}{550}\right)^{-\delta_k}\right)\frac{\Delta h_k}{\sin(\theta)} \tag{8}$$

where $\lambda$ is the optical wavelength, θ is the elevation angle of the satellite, M is the total layers of the cloud investigated and $\Delta h_k$ is the vertical extent of the layer of liquid clouds k^th ([58]). Furthermore, the coefficient $\delta$, which depends on the size distribution of the scattering particles, is estimated using empirical models [59] and provided by the Kim model as follows.

$$\delta_k = \begin{cases} 1.6 & if\ V > 50km \\ 1.3 & if\ 6km < V < 50km \\ 0.16V + 0.34 & if\ 1km < V < 6km \\ V - 0.5 & if\ 0.5km < V < 1km \\ 0 & if\ V < 0.5km \end{cases} \tag{9}$$

Table 2. Cloud-type comparison.

| Cloud Type | Height Range | FSO Attenuation | CLWC (g/$m^3$) | Optical Thickness |
|---|---|---|---|---|
| **Stratus** | Low altitude | Moderate attenuation | 0.28 [35] | Moderate to high |
| **Cumulus** | Moderate altitude | Low attenuation | 0.26 [35] | Low to moderate |
| **Cumulonimbus** | Low to high altitude | High attenuation | 1 [35] | High |
| **Stratocumulus** | Low to middle altitude | Moderate attenuation | 0.44 [35] | Moderate |
| **Cirrus** | High altitude | Low attenuation | 0.03 [35] | Low |

### 3.3 Combining Atmospheric Turbulence and Cloud Attenuation

The performance of Free-Space Optical (FSO) communication systems is significantly influenced by atmospheric turbulence and cloud attenuation, which collectively degrade the quality and reliability of optical links. Understanding the combined effect of these environmental factors is essential for designing robust FSO systems capable of maintaining reliable communication under varying weather conditions.

The combined effect of atmospheric turbulence and cloud attenuation on the FSO channel can be mathematically expressed as follows:

$$P_{\text{received}} = P_{\text{transmitted}} \cdot e^{-\alpha L} . e^{-\beta C_n^2 L^{5/3}} \tag{10}$$

where, $P_{\text{received}}$ is the received optical power after propagation through the atmosphere, accounting for both cloud attenuation and atmospheric turbulence, while $P_{\text{transmitted}}$ is the optical power transmitted from the FSO transmitter. The coefficient α represents the attenuation due to clouds, which is influenced by the optical depth, density and type of the cloud (e.g. cirrus, cumulus) [51]. The term $\beta$ is the turbulence coefficient that quantifies the effect of atmospheric turbulence on the received signal, $C_n^2$ being the refractive-index structure parameter [52]. Finally, $L$ denotes the propagation distance through the atmosphere.

In this equation, the term $e^{-\alpha L}$ accounts for the exponential attenuation of the optical signal due to clouds along the propagation path. The coefficient α depends on the optical depth and scattering properties of the clouds, which affect the amount of incident light that reaches the FSO receiver [51].

The term $e^{-\beta C_n^2 L^{5/3}}$ represents the Rytov-based model of turbulence-induced fading in the presence of atmospheric turbulence. The coefficient β characterizes the strength of turbulence, while $C_n^2$ quantifies the spatial and temporal variations in the refractive index caused by turbulence [52]. The exponent $L^{5/3}$ reflects the scaling of turbulence effects with propagation distance $L$, highlighting the increasing impact of turbulence on signal degradation over longer distances [51].

Cloud attenuation and air turbulence can have a substantial impact on the reliability and quality of FSO communication systems. Because clouds can significantly reduce signal quality, real-time monitoring and adaptive techniques are critical for increasing the overall system performance. Advanced modulation algorithms, error-correction coding and dynamic link adaptability are crucial to improving the resilience of FSO systems under a variety of weather conditions.

## 4. GROUND TO SATELLITE FSO SYSTEM

### 4.1 System Design

In this sub-section, we use Optisystem software to develop a ground-satellite FSO system that operates under turbulence and various cloud circumstances. OptiSystem was selected because of its ability to provide a comprehensive and realistic simulation of free-space optical (FSO) communication systems by incorporating key physical impairments. These include atmospheric turbulence effects, which are modeled using the gamma–gamma turbulence distribution and cloud-induced attenuation, integrated through empirical models based on Mie scattering theory. In addition, the software offers accurate optical component modeling, enabling a precise evaluation of modulation techniques and signal-quality metrics. In our simulation, we investigate various modulation methods for reducing cloud and turbulence impacts in optical ground-to-satellite communication. Table 3 summarizes the simulation parameters and their respective values.

Our simulations target ground-to-satellite links over extreme distances (up to 35,000 km), where loss of free-space path, atmospheric attenuation and turbulence impose severe challenges. We selected a link distance range from 2000 km to 36000 km to focus on the most challenging conditions in long-range FSO communications. Although shorter links (e.g. 500 km) experience significantly lower losses and reduced atmospheric effects, our study targets scenarios where severe free-space path loss, atmospheric attenuation and turbulence dominate. Established literature confirms that performance degradation becomes critical above 2000 km, justifying our focus on this range to ensure that our system design, particularly in terms of transmitter power and receiver aperture, is validated under the most extreme conditions encountered in ground-to-satellite and inter-satellite links. Under such harsh conditions, a higher transmit power is essential to maintain an acceptable signal-to-noise ratio (SNR) and achieve a low bit-error rate (BER). This design choice is supported by established link-budget analyses; for example, [60] demonstrated that long-range FSO links require elevated power levels to overcome significant path losses and [46] provided a detailed analysis that reinforces the need for higher transmission power in ground-to-satellite transmissions. Although lower transmit power might suffice for terrestrial FSO systems, the unique challenges of space-to-ground links justify our approach, ensuring that our simulation accurately reflects the operational demands of these systems.

Figure 3 shows a ground-to-satellite FSO system with an uplink from a ground station to a geostationary satellite. FSO links provide high-speed communication between satellites and ground stations. However, certain meteorological circumstances encountered during uplink transmission, including cloud attenuation and atmospheric turbulence, can have a major impact on the FSO link's performance. Cloud attenuation occurs when clouds absorb, scatter and refract laser beams, causing signal attenuation. Another key difficulty is atmospheric turbulence, which occurs when air masses move irregularly, creating temperature and pressure variations. These fluctuations cause variations in the refractive index of the atmosphere, resulting in scintillation effects on laser beams. Mitigating the effects of atmospheric turbulence is critical for stabilizing FSO lines and maintaining constant signal quality.

Table 3. Simulation parameters.

| **Parameters** | **Value** |
|---|---|
| Pre-amplification gain | 30 (dB) |
| Post-amplification gain | 30 (dB) |
| Noise-figure pre-amplification | 5 (dB) |
| Noise-figure post-amplification | 3 (dB) |
| Transmit power | 0-50 (dB) |
| Transmitter diameter | 5 (cm) |
| Receiver diameter | 30 (cm) |
| Data rate | 30 (Gb/s) |
| Link distance | 2000-36000 (km) |
| Photo-detector responsitivity | 1 (A/W) |
| Laser wavelength | 1550 (nm) |
| Tropopause height | 9.4 (km) |
| Turbulence model | Hafmagel Valley Model |
| Refractive-index structure | $C_n^2(0)$ Turbulence close to ground (strong turbulence: $C_n^2(0) = 2.10^{-13} m^{-2/3}$) |

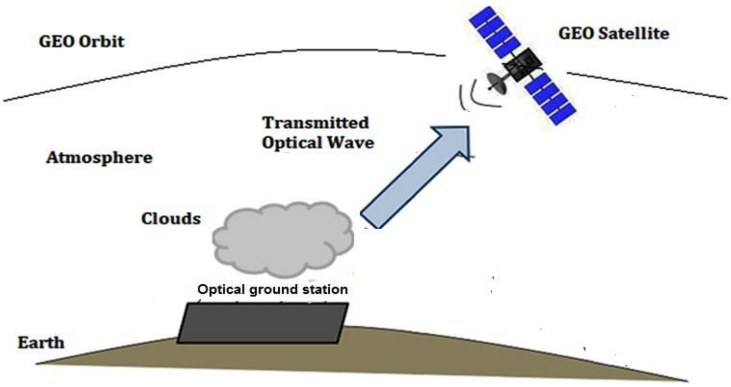


Figure 3. Satellite-to-ground FSO system.

In addition, 30 cm receiver diameter was selected to overcome the substantial free-space loss and atmospheric attenuation inherent in ground-to-satellite FSO links over long distances. Although terrestrial FSO systems commonly operate with receiver diameters in the 5–10 cm range, space-based optical ground stations often employ larger apertures to collect much weaker signals from vast distances. For example, NASA’s Lunar Laser Communication Demonstration (LLCD) used a receiver telescope with an aperture of the order of 40 cm to ensure sufficient signal collection from the Moon [61]. Furthermore, detailed link-budget analyzes for free-space optical satellite networks indicate that larger receiver diameters are necessary to achieve the signal-to-noise ratio (SNR) required over such long distances [46]. Therefore, our choice of a 30 cm receiver diameter is both practical and consistent with established approaches in long-range space optical communications.

The ground-to-satellite FSO channel consists of two serially coupled channels. The first FSO channel that simulates the atmosphere is 12 kilometers long. The length of the atmospheric channel is determined by averaging the tropopause elevation, which is 9 km at the pole and 15 km at the equator ([62]). The tropopause marks the boundary between the troposphere and the stratosphere. The troposphere accounts for 75% of the atmosphere’s mass and 99% of the total mass of water vapor and aerosols ([62]). This FSO channel represents the atmosphere and simulates the attenuation induced by various types of cloud and atmospheric turbulence. The second FSO channel depicts a space channel with a length of 35,988 kilometers. Because it depends solely on beam divergence and transmitter-receiver aperture diameter, it is only useful for geometric loss, which has a fixed value.

## 4.2 Modulation Approaches for the Proposed Ground-Satellite FSO System

The choice of modulation method is crucial to maintaining the reliability and efficiency of data transmission in a high-speed FSO communication system operating at 30 Gbps, especially when dealing with cloud interference. This sub-section presents advanced modulation techniques, such as QPSK, 8-PSK, 16-PSK and 16-QAM. We selected QPSK, 8-PSK, 16-PSK and 16-QAM for the following reasons:

- QPSK (Quadrature Phase Shift Keying): QPSK is widely recognized for its robustness against signal degradation, particularly in high-attenuation environments caused by clouds and atmospheric turbulence. It offers a balance between data rate and resistance to errors, making it optimal for low- to-moderate signal-to-noise ratio conditions. This makes QPSK particularly suitable for satellite communication links, where maintaining low bit error rates (BER) is crucial.
- Higher-order Modulation Schemes (8-PSK, 16-PSK, 16-QAM): These modulation techniques provide higher spectral efficiency than QPSK, allowing for greater data throughput within the same bandwidth. 8-PSK and 16-PSK use phase variations to encode data, while 16-QAM adds both amplitude and phase variations, allowing the encoding of even more information in a given symbol. While these schemes offer higher data rates, they are also more susceptible to noise and signal impairments, especially in turbulent or attenuated atmospheric conditions.

### 4.2.1 M-PSK Modulation

M-PSK divides the signal's phase into M distinct states, each corresponding to a specific bit combination. Phases are often uniformly distributed throughout the signal constellation, creating a distinct pattern. In binary phase-shift keying (BPSK), there are two phase states at 0 and 180 degrees, while in quadrature phase-shift keying (QPSK), there are four phase states spread at 90-degree intervals. M-ary Phase Shift Keying (M-PSK) modulation is more advanced than binary methods, because it uses higher-order modulation methods, such as 8-PSK and 16-PSK to effectively encode digital data onto a carrier signal. Figure 4 shows the design of the M-PSK system. 8-PSK divides the phase of the carrier signal into eight equidistant states, each representing a distinct three-bit pattern. Phase states are usually 45 degrees apart, enabling the transmission of three bits per symbol. This increase in spectral efficiency is especially beneficial in situations where a greater data rate is necessary. 8-PSK achieves a compromise between enhanced capacity and controllable complexity, making it a pragmatic option in digital communication systems. 16-PSK increases complexity by segmenting the phase of the carrier signal into sixteen states, each separated by 22.5 degrees. Higher data-transmission rates are achieved as a result of being able to convey a distinct four-bit pattern with every phase state.

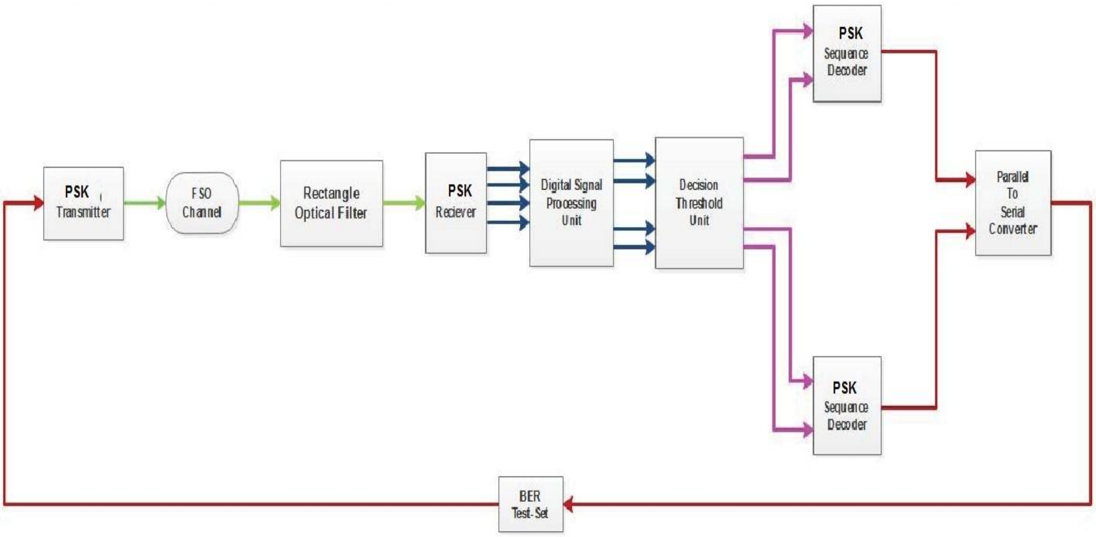


Figure 4. M-PSK system design.

### 4.2.2 M-QAM Modulation

Optical M-QAM (M-ary Quadrature Amplitude Modulation) is a modulation scheme used in optical-communication systems to encode digital information onto an optical carrier wave. Figure 5 illustrates the design of the M-QAM system. The concept combines both amplitude and phase modulation to achieve higher spectral efficiency. In M-QAM, the amplitude and phase of the optical signal are modulated simultaneously, allowing for the transmission of multiple bits per symbol. The "M" in M-QAM denotes the number of different states in the signal constellation, representing unique combinations of amplitude and phase. For example, in 16-QAM, there are 16 different states, enabling the representation of 4 bits per symbol. The optical signal's amplitude and phase states are typically arranged in a square or rectangular constellation, with the states positioned at specific points in the complex plane.

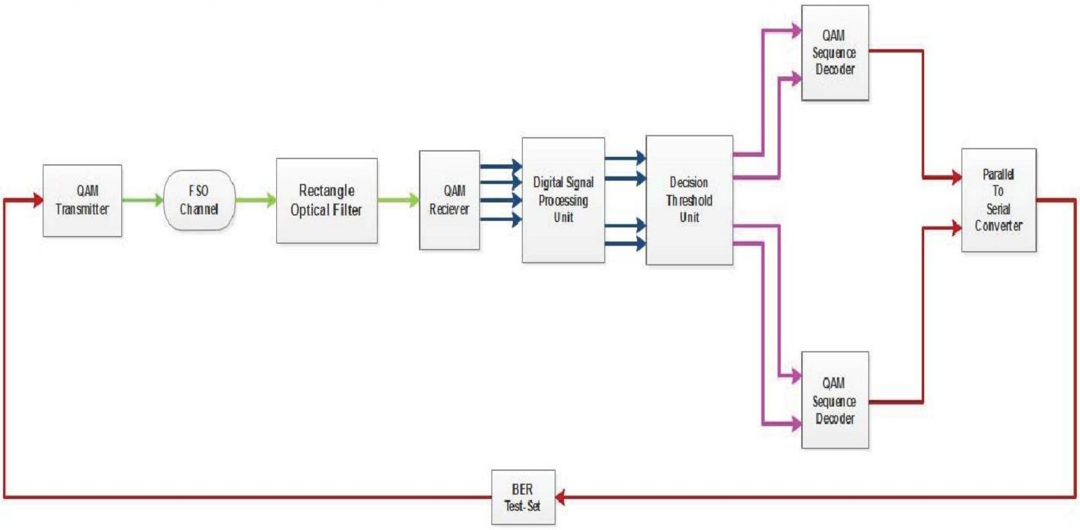


Figure 5. M-QAM system design.

## 5. SIMULATION RESULTS

In this section, we evaluate the performance of the proposed ground-satellite-ground FSO communication system under different cloud conditions and under strong turbulence.

Figure 6 illustrates the quality factor for different modulation schemes (QPSK, 8-PSK, 16-PSK and 16- QAM) at varying link distances. At a link distance of 2000 km, all modulation schemes exhibit relatively high quality-factor values. In particular, QPSK and 16-PSK demonstrate robust performance, possessing higher quality factors compared to 8-PSK and 16-QAM in this short range.

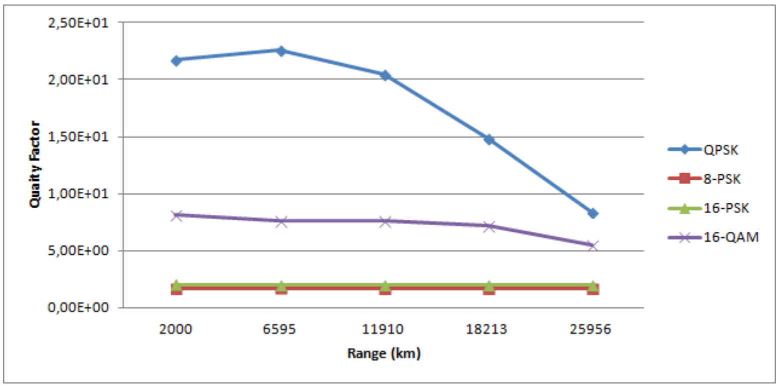


Figure 6. Quality factor *vs*. link distance (km).

The superior performance of QPSK and 16-PSK can be attributed to their enhanced tolerance to noise and channel impairments, resulting from larger symbol separations. This is contrasted with 8-PSK and 16-QAM, the denser constellation arrangements of which lead to higher susceptibility to noise even when SNR is high. These findings underscore the trade-off between spectral efficiency and robustness, reinforcing our conclusion that QPSK, in particular, is the most robust modulation scheme for ground-to-satellite FSO links under adverse atmospheric conditions. As the link distance increases, a general trend of decreasing quality-factor values emerges across all modulation schemes. The decline in the quality factor indicates the sensitivity to distance-induced signal degradation and the varying rates of decrease among modulation schemes highlight differences in their resilience to longer link distances.

Figure 7 shows the BER obtained for different modulation schemes (QPSK, 8-PSK, 16-PSK and 16-QAM) at different levels of attenuation. Each attenuation value corresponds to a cloud type, as described in Table 4.

Table 4. Attenuation at 1550nm of different types of cloud.

| Cloud Type | Attenuation (dB/km) |
|---|---|
| Stratus | 0.035 |
| Cumulus | 0.037 |
| Cumulonimbus | 0.011 |
| Stratocumulus | 0.026 |
| Cirrus | 0.134 |

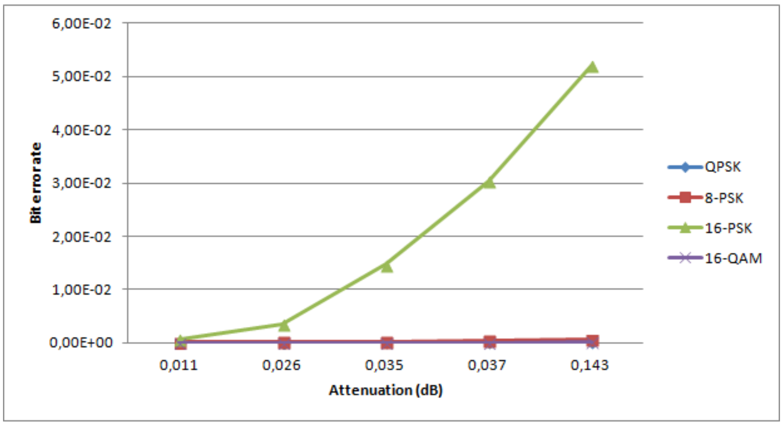


Figure 7. BER *vs*. attenuation (dB).

QPSK consistently exhibits lower BER values, making it more robust and less susceptible to errors induced by signal attenuation. 16-QAM experiences a comparatively higher BER compared to QPSK, 8-PSK and 16-PSK at an attenuation level of 0.143, indicating that it may be more sensitive to higher levels of attenuation, making it less robust under challenging communication conditions. 8-PSK and 16-PSK have moderate sensitivity to attenuation, indicating a trade-off between higher data rates and increased susceptibility to signal degradation.

To statistically characterize the fluctuations induced by turbulence in the intensity received *I*, we employ the gamma–gamma distribution. Its probability density function is given by:

$$p_I(I) = \frac{2(\alpha\beta)^{\frac{\alpha+\beta}{2}}}{\Gamma(\alpha)\Gamma(\beta)} I^{\frac{\alpha+\beta}{2}} K_{\alpha-\beta}(2\sqrt{\alpha\beta\frac{I}{I_0}}) \tag{11}$$

where, $\alpha$ and $\beta$ are parameters related to the small-scale and large-scale turbulence effects, $I_0$ is the average received power, $\Gamma(\cdot)$ denotes the gamma function and $K_v$ $(\cdot)$ is the modified Bessel function of the second kind.

This model allows us to derive the effective signal-to-noise ratio (SNR), $\gamma_{eff}$, which is used to compute the bit-error rate (BER) through the Q function:

$$\text{BER} = Q(\sqrt{\gamma_{eff}}) \tag{12}$$

Figure 8 illustrates the BER for various modulation schemes, including QPSK, 8-PSK, 16-PSK and 16- QAM, at different communication ranges. We note that at shorter ranges, such as 2000 km and 6595km, both QPSK and 16-QAM exhibit BER of 0, indicating reliable and error-free performance within these distances. On the other hand, 8-PSK and 16-PSK, while maintaining relatively low BER values, show slightly higher error rates compared to QPSK and 16-QAM in these short ranges. The QPSK modulation technique demonstrates robustness throughout the entire range, consistently delivering lower BER values compared to the other modulation schemes. This underscores the suitability of QPSK for scenarios that require reliable communication over extended distances.

Thus, at shorter distances, such as 2000 km and 6595 km, the free-space path loss and turbulence-induced fading are minimal, resulting in a higher received signal power and a correspondingly high signal-to-noise ratio. Under these favorable conditions, the system operates in an optimal regime, where the channel impairments are negligible. This shows that when the channel is not significantly stressed by attenuation or noise, even more complex modulation schemes can achieve near-perfect decoding.

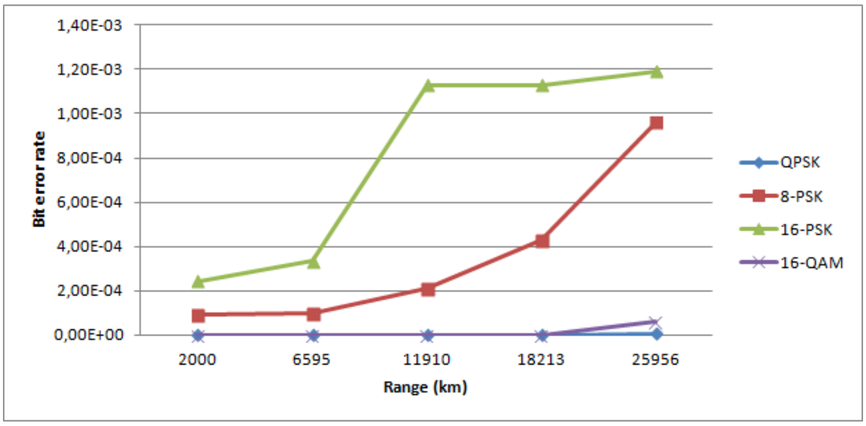


Figure 8. BER *vs*. link distance (km).

In addition to the BER and OSNR, we evaluate the performance of the FSO system using the quality factor Q. The quality factor quantifies the separation between the signal and noise levels and is defined as:

$$Q = \frac{\mu_{\text{signal}} - \mu_{\text{noise}}}{\sigma_{\text{signal}} + \sigma_{\text{noise}}} \tag{13}$$

where $\mu_{\text{signal}}$ and $\mu_{\text{noise}}$ denote the mean values of the signal and noise, respectively, and $\sigma_{\text{signal}}$ and $\sigma_{\text{noise}}$ represent their corresponding standard deviations. A higher $Q$ value implies a clearer distinction between the signal and noise, which generally corresponds to a lower bit-error rate (BER).

Figure 9 describes the BER *versus* the transmitted power for various modulation schemes, including QPSK, 8-PSK, 16-PSK and 16-QAM. At a transmitted power of 0, all schemes exhibit BER values around 0.5, except for 16-QAM, which has a slightly higher BER of 0.51. As the transmitted power increases to 10, the BER decreases, with QPSK showing a significant improvement, at a transmitted power of 15, QPSK achieves an exceptionally low BER of 3.00E-05, demonstrating its resilience to errors. At higher transmission-power levels, the BER values consistently approach zero, indicating that higher transmission-power levels contribute to a stronger and more reliable signal, resulting in significantly reduced error rates. QPSK consistently achieves lower BER values across the range of transmitted powers, while 16-QAM appears to be more sensitive to variations in transmitted power, especially at lower levels.

The observed differences in BER performance between modulation schemes can be attributed to the inherent characteristics of their constellation designs and the robustness to noise. QPSK utilizes a four-point constellation with wider Euclidean distances between symbols, making it more resilient to noise and channel impairments. In contrast, 16-QAM compresses 16 symbols into the same signal space, resulting in closer constellation points and increased susceptibility to errors in the presence of noise or power variations. Although higher-order modulation schemes, such as 16-QAM, offer superior spectral efficiency, they suffer from reduced robustness under adverse conditions. This trade-off is evident in our simulations (as seen in Figure 9), where QPSK maintains a lower BER across various transmitted power levels due to its larger decision regions, while 16-QAM exhibits a higher BER in low-power scenarios. These results highlight the fundamental balance between spectral efficiency and signal robustness in FSO-communication systems.

In our model, the received optical power $P_{\text{received}}$ is expressed as:

$$P_{\text{received}} = P_{\text{transmitted}} . e^{-\alpha L} . e^{-\beta C_n^2 L^{5/3}} \tag{14}$$

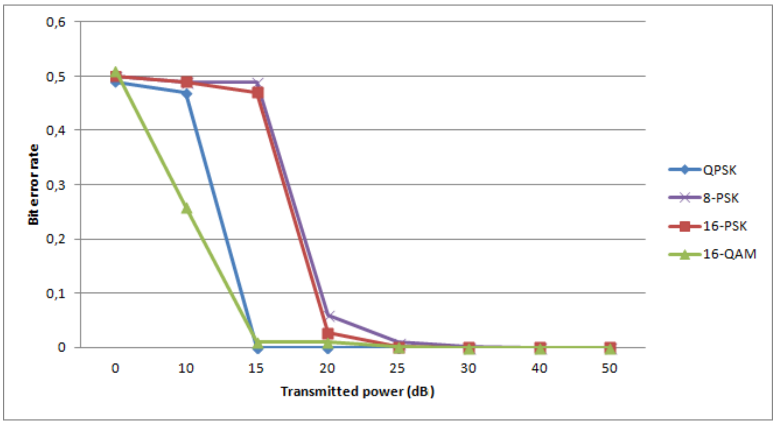


Figure 9. BER *vs*. transmitted power (dB).

where, $P_{\text{transmitted}}$ is the transmitted optical power, $L$ is the link distance, $\alpha$ represents the attenuation coefficient due to clouds (derived from empirical models, such as those based on Mie scattering theory), $C_n^2$ is the refractive-index structure parameter, characterizing the strength of atmospheric turbulence and $\beta$ is a scaling constant that captures turbulence-induced fading.

The degradation in received power due to atmospheric turbulence is further characterized by the Rytov variance:

$$\sigma_R^2 = 1.23 C_n^2 k^{7/6} L^{11/6} \tag{15}$$

with $k = \frac{2\pi}{\lambda}$ being the optical wave number and $\lambda$ the operating wavelength. Under strong turbulence, the factor $e^{-\beta C_n^2 L^{5/3}}$ effectively models the fading effect that affects the received signal.

Figure 10 illustrates the BER in different modulation schemes, including QPSK, 8-PSK, 16-PSK and 16-QAM, at different levels of the OSNR. The OSNR is defined as the ratio of the received optical signal power to the noise power spectral density:

$$\text{OSNR} = \frac{P_{\text{signal}}}{N_0 . B} \tag{16}$$

where $P_{\text{signal}}$ represents only the power of the transmitted signal (without the noise contribution) contrary to $P_{\text{received}}$ which is the optical power detected in the receiver after accounting for losses due to atmospheric and cloud attenuation. $N_0$ represents the spectral density of the noise power. $B$ is the bandwidth of the channel over which the optical signal and noise are measured. The data shows that at an OSNR of 0, higher BER values are observed, with QPSK demonstrating a lower BER, indicating its robustness in maintaining signal quality. Higher-order modulation schemes, like 16-PSK and 16-QAM, exhibit higher BERs, indicating increased sensitivity to noise. As OSNR increases to 5, 10 and 15, the BER decreases consistently, with QPSK demonstrating superior performance with extremely low BER values. At higher OSNR levels (20 and 25), the BER values approach zero, indicating a substantial reduction in errors. QPSK maintains its advantage with consistently low BER, but even higher-order modulation schemes achieve highly reliable communication under these favorable conditions.

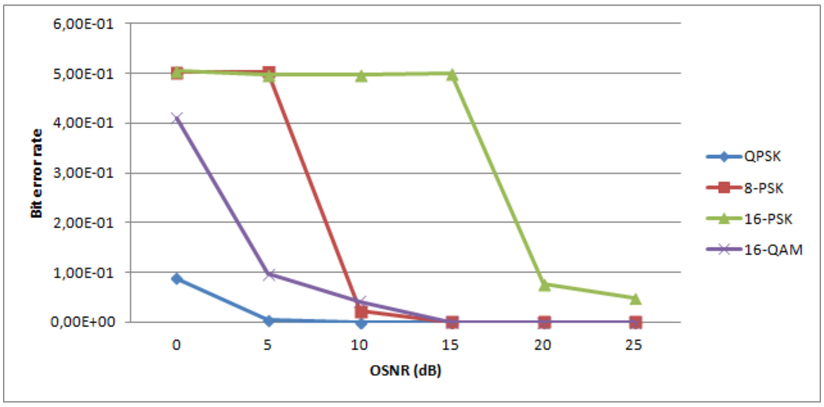


Figure 10. BER *vs*. OSNR (dB).

The trade-off between spectral efficiency and noise robustness plays a critical role in the performance of different modulation schemes in FSO communication. QPSK, with its four-symbol constellation, benefits from larger Euclidean distances between symbols, making it inherently resilient to noise and channel impairments. This ensures a consistently low BER even in challenging conditions. However, higher-order modulation schemes, such as 16-QAM, achieve greater spectral efficiency by packing more symbols into the same signal space. Although this increases data throughput, it also reduces the separation between constellation points, making the modulation more sensitive to noise and signal distortions. However, under optimal conditions, such as high optical signal-to-noise ratio, minimal attenuation and low turbulence, 16-QAM and other higher-order schemes can maintain low BER values due to the reduced impact of noise. This highlights the fundamental trade-off: QPSK remains a robust choice across a wide range of conditions, whereas higher-order modulations can achieve similar performance when the channel provides a high- quality, low-noise environment.

Figure 11 shows the values of the quality factor for different modulation schemes, QPSK, 8-PSK, 16-PSK and 16-QAM, at varying levels of OSNR. At an OSNR of 0, the quality-factor values are relatively high, indicating decent signal quality. However, 16-QAM has the lowest quality factor at this level, suggesting that it is more susceptible to noise. As OSNR increases to 5 and 10, the quality factor improves consistently, with QPSK and 8-PSK exhibiting higher quality factors, while 16-QAM shows a significant improvement, highlighting its ability to benefit from increased OSNR for enhanced signal quality. At OSNRs of 15 and 20, the quality factor continues to increase, with QPSK and 8-PSK maintaining strong performances, while 16-QAM shows a substantial improvement, indicating its ability to achieve higher signal quality at elevated OSNR levels. At the highest OSNR level of 25, QPSK and 8-PSK maintain high quality factors, while 16-QAM achieves a notable improvement.

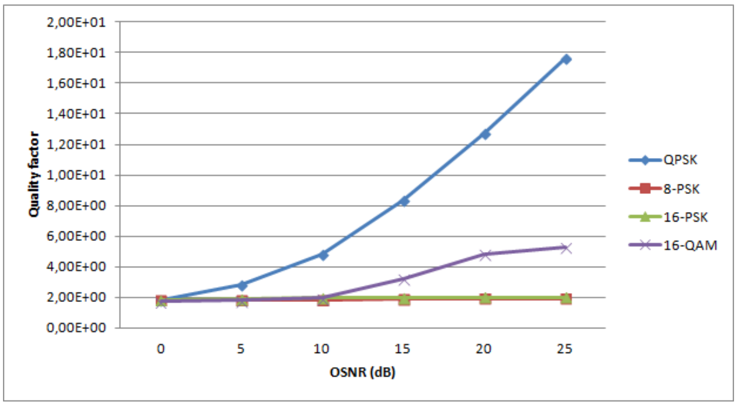


Figure 11. Quality factor vs OSNR (dB).

Figure 12 illustrates the EVM values obtained for different modulation schemes, QPSK, 8-PSK, 16-PSK and 16-QAM, at different levels of OSNR. EVM is calculated by comparing the received optical signal’s constellation points to the expected or ideal points based on the modulation scheme used. Lower percentages of EVM indicate a higher fidelity of the received signal to the ideal signal. The relationship between EVM and OSNR provides insight into how signal distortion varies with changes in signal quality for each modulation scheme. At an OSNR of 0, all modulation schemes exhibit relatively high EVM values, indicating significant signal distortion due to low signal-to-noise ratios. Our simulation results indicate that, under optimal atmospheric conditions, 16-QAM can achieve lower EVM values. This suggests that, with effective receiver processing and in low-noise environments, higher-order modulation schemes can also deliver high signal fidelity. As OSNR increases to 5, 10 and 15, the EVM values decrease for all modulation schemes, contributing to improved signal quality and reduced distortion. QPSK and 8-PSK consistently exhibit lower EVM values, demonstrating their resilience even under optimal conditions. Higher-order modulation schemes, such as 16-PSK and 16-QAM, still exhibit higher distortion compared to simpler modulation schemes. Although QPSK typically benefits from a larger decision region, our simulation results indicate that under favorable OSNR conditions, 16-QAM can achieve lower EVM values. This suggests that, with effective receiver processing and in low-noise environments, higher-order modulation schemes can also deliver high signal fidelity.

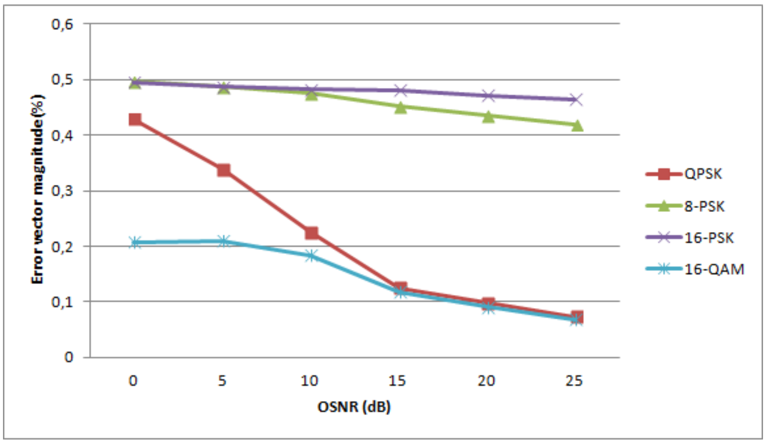


Figure 12. EVM (%) *vs* OSNR (dB).

In conclusion, our study provides a comprehensive evaluation of ground-to-satellite FSO communication systems under realistic atmospheric conditions. The simulation results demonstrate that as the link length increases, the adverse effects of atmospheric turbulence and cloud attenuation become more pronounced, leading to deteriorated performance metrics. Among the modulation schemes analyzed, QPSK emerges as the most robust option, consistently providing lower BER and higher quality factors in high-attenuation scenarios. These insights are critical for designing resilient FSO systems, particularly for satellite-communication applications where maintaining link reliability under extreme conditions is paramount. Future work will explore adaptive modulation strategies and real-world experimental validation to further optimize system performance.

## 6. List of Abbreviations

Table 5. List of abbreviations.

| Abbreviation | Full Form | Abbreviation | Full Form |
|---|---|---|---|
| FSO | Free Space Optical | DWT | Discrete Wavelet Transform |
| BER | Bit Error Rate | MIMO | Multiple-Input Multiple-Output |
| OSNR | Optical Signal-to-Noise Ratio | ADC | Analog-to-Digital Converter |
| SNR | Signal-to-Noise Ratio | DSP | Digital Signal Processing |
| QPSK | Quadrature Phase Shift Keying | LEO | Low Earth Orbit |
| 8-PSK | 8-Phase Shift Keying | GEO | Geostationary Orbit |
| 16-PSK | 16-Phase Shift Keying | WDM | Wavelength Division Multiplexing |
| 16-QAM | 16-Quadrature Amplitude Modulation<br>OOK On-Off Keying | RF | Radio Frequency |
| DPSK | Differential Phase Shift Keying | HT | Hilbert Transform |
| M-PSK | M-ary Phase Shift Keying | $C_n^2$ | Refractive Index Structure Parameter |
| M-QAM | M-ary Quadrature Amplitude Modulation | MZM | Mach–Zehnder Modulator |
| LOS | Line of Sight | AWGN | Additive White Gaussian Noise |
| CWT | Continuous Wavelet Transform | Q | Quality Factor |

## 7. Conclusion

FSO communication is crucial to improve resilience and adaptive capacity to climate-related hazards and natural disasters, with high data rates and low latency. Its applications include telecommunications, disaster response, military communications, scientific research and aerospace. FSO's rapid deployment capabilities and versatility in addressing diverse communication challenges make it essential in regions where traditional infrastructure is impractical or disrupted.

The study examines the challenges faced by FSO technology, particularly in satellite-to-ground communications amid atmospheric obstructions, such as clouds and turbulence. It reveals that QPSK modulation is the optimal choice for FSO communication in satellite networks facing cloud-induced attenuation. QPSK consistently maintains the lowest BER even in adverse conditions and exhibits exceptional resilience in the face of varying transmitted power and noise levels.

Future research will focus on refining mitigation strategies for cloud effects on FSO system performance, exploring innovative modulation techniques and developing advanced models to quantify atmospheric impacts more accurately. Practical implementation and field testing of the proposed FSO system under real-world conditions would provide valuable insights into its performance and robustness.

## REFERENCES

[1] A. U. Chaudhry and H. Yanikomeroglu, "Free Space Optics for Next-Generation Satellite Networks," IEEE Consumer Electronics Magazine, vol. 10, pp. 21-31, 2020.

[2] A. U. Chaudhry and H. Yanikomeroglu, "When to Crossover from Earth to Space for Lower Latency Data Communications?" IEEE Trans. on Aerospace and Electronic Syst., vol. 58, pp. 3962-3978, 2022.

[3] A. U. Chaudhry et al., "Laser Intersatellite Link Range in Free-space Optical Satellite Networks: Impact on Latency," IEEE Aerospace and Electronic Systems Magazine, vol. 38, pp. 4-13, 2023.

[4] S. Magidi and A. Jabeena, "Free Space Optics, Channel Models and Hybrid Modulation Schemes: A Review," Wireless Personal Communications, vol. 119, pp. 2951 – 2974, 2021.

[5] T. Ahmmed et al., "The Digital Divide in Canada and the Role of LEO Satellites in Bridging the Gap," IEEE Communications Magazine, vol. 60, pp. 24-30, 2022.

[6] Z. Ali et al., "Enhanced Learning-based Hybrid Optimization Framework for RSMA-aided Underlay LEO Communication with Non-collaborative Terrestrial Primary Network," IEEE Transactions on Communications, DOI: 10.1109/TCOMM.2024.3465375, 2024.

[7] M. Mrabet and M. Sliti, "Performance Analysis of FSO Communications in Desert Environments," Optical and Quantum Electronics, vol. 56, no. 4, DOI: 10.1007/s11082-024-06315-9, 2024.

[8] M. Sliti and M. Garai, "Performance Analysis of FSO Communication Systems under Different Atmospheric Conditions," Proc. of the 2022 27th Asia Pacific Conf. on Communications (APCC), DOI: 10.1109/apcc60132.2023.10460727, 2023.

[9] S. Magidi and A. Jabeena, "Analysis of Multi-pulse Position Modulation Free Space Optical Communication System Employing Wavelength and Time Diversity over Malaga Turbulence Channel," Scientific African, vol. 12, p. e00777, DOI: 10.1016/j.sciaf.2021.e00777, 2021.

[10] C. a. B. Dath and N. a. B. Faye, "Resilience of Long Range Free Space Optical Link under a Tropical Weather Effects," Scientific African, vol. 7, p. e00243, DOI: 10.1016/j.sciaf.2019.e00243, 2019.

[11] S. Magidi and A. Jabeena, "Parallel Relay-assisted Free-space Optical Communication Using Multi-pulse Position Modulation over the Generalized Turbulence Channel Model," Journal of Optics, vol. 51, no. 1, pp. 133–141, 2021.

[12] J. O. Bandele et al., "Multiple Transmitters for Gain Saturated Pre-amplified FSO Communication Systems Limited by Strong Atmospheric Turbulence and Pointing Error," IEEE Access, vol. 11, pp. 110985-110994, 2023.

[13] M. T. Mbezi et al., "Temperature and Wind Velocity Effects on Bit Error Rate during Free Space Optical Link under Matrix Málaga Turbulence Channel," Sādhanā, vol. 48, pp. 1-6, 2023.

[14] M. T. Mbezi et al., "Variable Antennas Positions Solution to Reduce Pointing Errors Due to Wind Speed and Temperature Coupled Effects During Free Space Optical Link Using Matrix Rician Pointing Error Model," Optica Applicata, vol. 3, pp. 393-406, DOI:10.37190/oa230305, 2023.

[15] E. E. Elsayed, "Performance Analysis and Modeling: Atmospheric Turbulence and Crosstalk of WDM-FSO Network," Journal of Optics, DOI: 10.1007/s12596-024-02434-4, 2021.

[16] E. E. Elsayed, "Performance Enhancement of Atmospheric Turbulence Channels in DWDM-FSO PON Communication Systems Using M-ary Hybrid DPPM-M-PAPM Modulation Schemes under Pointing Errors, ASE Noise and Interchannel Crosstalk," J. Optics, DOI: 10.1007/s12596-024- 01908-9, 2024.

[17] A. Elfikky, M. Soltani and Z. Rezki, "End-to-End Learning Framework for Space Optical Communications in Non-differentiable Poisson Channel," IEEE Wireless Communications Letters, vol. 13, no. 8, pp. 2090-2094, DOI: 10.1109/LWC.2024.3401692, Aug. 2024.

[18] A. Elfikky et al., "Spatial Diversity-based FSO Links under Adverse Weather Conditions: Performance Analysis," Optical and Quantum Electronics, vol. 56, p. 826, DOI: 10.1007/s11082-024-06625-y, 2024.

[19] E. E. Elsayed et al., "Coding Techniques for Diversity Enhancement of Dense Wavelength Division Multiplexing MIMO-FSO Fault Protection Protocols Systems over Atmospheric Turbulence Channels," IET Optoelectronics, DOI: 10.1049/ote2.12111, 2024.

[20] J. O. Bandele, "Performance of Cascaded Gain Saturated and Fixed Gain Optical Amplifier FSO Communication Systems Limited by Scintillation and Pointing Error," Scientific African, vol. 20, p. e01687, DOI: 10.1016/j.sciaf.2023.e01687, 2023.

[21] M. T. Mbezi et al., "Designing of a Quantum TIA to Improve FSO Signal Reception," Scientific

African, vol. 19, p. e01539, DOI: 10.1016/j.sciaf.2022.e01539, 2022.

[22] E. E. Elsayed and B. B. Yousif, "Performance Enhancement of M-ary Pulse-position Modulation for a Wavelength Division Multiplexing Free-space Optical Systems Impaired by Inter-channel Crosstalk, Pointing Error and ASE Noise," Optics Communications, vol. 475, p. 126219, 2020.

[23] H. Kaushal and G. Kaddoum, "Optical Communication in Space: Challenges and Mitigation Techniques," IEEE Communications Surveys and Tutorials, vol. 19, no. 1, pp. 57–96, 2017.

[24] D. Giggenbach and F. Moll, "Scintillation Loss in Optical Low Earth Orbit Data Downlinks with Avalanche Photodiode Receivers," Proc. of the 2017 IEEE Int. Conf. on Space Optical Systems and Applications (ICSOS), pp. 115-122, Naha, Japan, 2017.

[25] L. Luini et al., "Investigation and Modeling of Ice Clouds Affecting Earth-Space Communication Systems," IEEE Trans. on Antennas and Propagation, vol. 66, no. 1, pp. 360–367, 2018.

[26] N. K. Lyras et al., "Cloud Free Line of Sight Prediction Modeling for Optical Satellite Communication Networks," IEEE Communications Letters, vol. 21, no. 7, pp. 1537–1540, 2017.

[27] N. K. Lyras et al., "Cloud Attenuation Statistics Prediction from Ka-band to Optical Frequencies: Integrated Liquid Water Content Field Synthesizer," IEEE Transactions on Antennas and Propagation, vol. 65, no. 1, pp. 319–328, 2017.

[28] T. Nguyen et al., "TCP over Hybrid FSO/RF-based Satellite Networks in the Presence of Cloud Coverage," IEICE Communications Express, vol. 11, no. 10, pp. 649-654, 2022.

[29] P. B. Bhatt et al., "Designing and Simulation of 30Gbps FSO Communication Link under Different Atmospheric and Cloud Conditions," Int. J. of Eng. Trends and Technology, vol. 69, pp. 228-234, 2021.

[30] D. Ko et al., "Cloud Shape and Attenuation Based UAV Trajectory Optimization for FSO Communication," IEEE Transactions on Vehicular Technology, vol. 73, pp. 9911-9926, 2024.

[31] T. V. Nguyen et al., "On the Design of RIS–UAV Relay-assisted Hybrid FSO/RF Satellite–Aerial–Ground Integrated Network," IEEE Trans. on Aerospace and Elect. Syst., vol. 59, pp. 757-771, 2023.

[32] T. V. Nguyen et al., "Link Availability of Satellite-based FSO Communications in the Presence of Clouds and Turbulence," IEICE Communications Express, vol. 10, no. 5, pp. 206–211, 2021.

[33] H. D. Le, T. V. Nguyen and A. T. Pham, "Cloud Attenuation Statistical Model for Satellite-based FSO Communications," IEEE Antennas and Wireless Propagation Letters, vol. 20, no. 5, pp. 643–647, 2021.

[34] P. Bhatt et al., "Designing and Simulation of 30Gbps FSO Communication Link under Different Atmospheric and Cloud Conditions," Int. J. of Eng. Trends and Tech., vol. 69, no.5, pp. 228–234, 2021.

[35] N. Tabassum et al., "Performance Analysis of Free Space Optics Link for Different Cloud Conditions," Proc. of the 2018 4th Int. Conf. on Computing Communication and Automation (ICCCA), DOI: 10.1109/ccaa.2018.8777546, Greater Noida, India, 2018.

[36] N. K. Lyras et al., "Cloud Free Line of Sight Prediction Modeling for Optical Satellite Communication Networks," IEEE Communications Letters, vol. 21, no. 7, pp. 1537–1540, 2017.

[37] M. M. Haque, A. Jahid, M. M. Hasan and P. Das, "Performance of a FSO Link in Presence of Cloud," ULAB Journal of Science and Engineering, vol. 6, no. 1, 2015.

[38] K. Rammprasath and S. Prince, "Analyzing the Cloud Attenuation on the Performance of Free Space Optical Communication," Proc. of the 2013 Int. Conf. on Communication and Signal Processing, DOI: 10.1109/iccsp.2013.6577165, Melmaruvathur, India, 2013.

[39] T. V. Pham et al., "A Placement Method of Ground Stations for Optical Satellite Communications Considering Cloud Attenuation," IEICE Communications Express, vol. 12, no. 10, pp. 568–571, 2023.

[40] J. G. JOlmedo et al., "Visibility Framework and Performance Analysis for Free Space Optical Communications in Satellite Links," IEEE Access, vol. 11, pp. 68897–68911, 2023.

[41] Y. Ata and M. S. Alouini, "Performance of Integrated Ground-Air-Space FSO Links over Various Turbulent Environments," IEEE Photonics Journal, vol. 14, no. 6, pp. 1–16, 2022.

[42] H. Ivanov et al., "Estimation of Cloud-induced Optical Attenuation over Near-Earth and Deep-space FSO Communication Systems," Proc. of the 2021 Int. Conf. on Software, Telecomm. and Computer Networks (SoftCOM), DOI: 10.23919/softcom52868.2021.9559077, Split, Croatia, 2021.

[43] A. K. Sharoar Jahan Choyon and R. Chowdhury, "Performance Comparison of Free-space Optical (FSO) Communication Link under OOK, BPSK, DPSK, QPSK and 8-PSK Modulation Formats in the Presence of Strong Atmospheric Turbulence," J. of Optical Comm., vol. 44, pp. s763 - s769, 2020.

[44] K. Prabu, D. S. Kumar and T. Srinivas, "Performance Analysis of FSO Links under Strong Atmospheric Turbulence Conditions Using Various Modulation Schemes," Optik, vol. 125, pp. 5573-5581, 2014.

[45] M. H. Ibrahim et al., "Effect of Different Weather Conditions on BER Performance of Single-channel Free Space Optical Links," Optik, vol. 137, pp. 291-297, 2017.

[46] J. Liang et al., "Link Budget Analysis for Free-space Optical Satellite Networks," Proc. of the 2022 IEEE 23rd Int. Symposium on a World of Wireless, Mobile and Multimedia Networks (WoWMoM), pp. 471-476, Belfast, UK, 2022.

[47] J. Liang et al., "Free-space Optical (FSO) Satellite Networks Performance Analysis: Transmission Power, Latency and Outage Probability," IEEE Open J. of Vehicular Techn., vol. 5, pp. 244-261, 2021.

[48] J. Liang et al., "Latency *versus* Transmission Power Trade-off in Free-space Optical (FSO) Satellite

Networks with Multiple Inter-continental Connections," IEEE Open Journal of the Communications Society, vol. 4, pp. 3014-3029, 2023.

[49] ECMWF, "ERA-interim Database," Available: https://apps.ecmwf.int/datasets/data/interimfull-daily.

[50] M. Polnik et al., "Scheduling Space-to-Ground Optical Communication under Cloud Cover Uncertainty," IEEE Trans. on Aerospace and Electronic Systems, vol. 57, no. 5, pp. 2838–2849, 2021.

[51] L. C.Andrews and R. L. Phillips, Laser Beam Propagation through Random Media, ISBN: 9780819478320, DOI: 10.1117/3.626196, 2005.

[52] M. A. Al-Habash, "Mathematical Model for the Irradiance Probability Density Function of a Laser Beam Propagating through Turbulent Media," Optical Engineering, vol. 40, no. 8, p. 1554, 2001.

[53] L. C. Andrews, R. L. Phillips and C. Y. Young, "Laser Beam Scintillation with Applications," SPIE eBooks, DOI: 10.1117/3.412858, 2001.

[54] H. Kaushal and G. Kaddoum, "Optical Communication in Space: Challenges and Mitigation Techniques," IEEE Communications Surveys and Tutorials, vol. 19, no. 1, pp. 57–96, 2017.

[55] I. Gultepe, "Fog and Boundary Layer Clouds: Fog Visibility and Forecasting," Birkhäuser Basel eBooks, DOI: 10.1007/978-3-7643-8419-7, 2007.

[56] T. V. Nguyen et al., "On the Design of Rate Adaptation for Relay-Assisted Satellite Hybrid FSO/RF Systems," IEEE Photonics Journal, vol. 14, no. 1, pp. 1–11, 2022.

[57] J. F. Yin et al., "An Investigation into the Relationship between Liquid Water Content and Cloud Number Concentration in the Stratiform Clouds over North China," Atmospheric Research, vol. 139, pp. 137–143, DOI: 10.1016/j.atmosres.2013.12.004, 2014.

[58] M. Alzenad et al., "FSO-based Vertical Backhaul/Fronthaul Framework for 5G+ Wireless Networks," IEEE Comm. Magazine, vol. 56, no. 1, pp. 218–224, DOI: 10.1109/mcom.2017.1600735, 2018.

[59] I. I. Kim et al., "Comparison of Laser Beam Propagation at 785 nm and 1550 nm in Fog and Haze for Optical Wireless Communications," Proc. of SPIE, the Int. Society for Optical Engineering, DOI: 10.1117/12.417512, 2001.

[60] M. A. Khalighi and M. Uysal, "Survey on Free Space Optical Communication: A Communication Theory Perspective," IEEE Communications Surveys & Tutorials, vol. 16, pp. 2231-2258, 2014.

[61] B. S. Robinson et al., "The Lunar Laser Communications Demonstration," Proc. of the 2011 Int. Conf. on Space Optical Systems and Applications (ICSOS), DOI:10.1109/icsos.2011.5783709, 2011.

[62] D. F. Rex, "Climate of the Free Atmosphere," Elsevier eBooks, vol. 4, [Online], Available: http://ci.nii.ac.jp/ncid/BA0097239X, 1969.

## ملخص البحث:

يُعدّ الاتّصال الضّوئي في الفضاء الحرّ حلاً لتلبية الطّلب المتزايد على الاتّصالات من القمر الصّناعي؛ فهو يحقّق فوائد منها معدّلات بيانات أعلى وأمان أفضل. لكنّ أداء الاتّصال الضّوئي في الفضاء الحرّ يتأثر بالظّروف المناخية، مثل تشكُّل الغيوم، والاضطراب الجوّي، والتي تتسبّب في إضعاف الإشارة وتشتُّتها إضافةً الى تشوّهات الطّور. ومن الأمور الأساسية أنّ الفهم الأفضل لتأثيرات الظّروف الجوّية على أداء أنظمة الاتّصال الضّوئي في الفضاء الحرّ إلى جانب تبنّي استراتيجيات ناجعة للحدّ من هذه التّأثيرات تعدّ جوهرية وحاسمة لضمان أداءٍ آمنٍ وفعّال لتلك الأنظمة.

في هذه الدّراسة، قمنا بتقييم أداء أنظمة الاتّصال الضّوئي في الفضاء الحرّ تحت ظروف جوّية مختلفة من حيث الاضطراب الجوّي وتشكُّل الغيوم وأنواعها، مع التّركيز على أنظمة الاتّصال من الأرض إلى القمر الصّناعي. وقد عملنا على تحليل العديد من تقنيات التّعديل والمقارنة بينها لمعرفة التّطبيقات التي تلائمها كل تقنية من تقنيات التّعديل. وتسهم النّتائج الّتي تمّ الحصول عليها في توفير رؤىً حاسمة من أجل تحسين استراتيجيات التعديل؛ من أجل الحصول على روابط ضوئية تتّسم بالمتانة والموثوقية بين الأرض والأقمار الصّناعية.

# المجلة الأردنية للحاسوب وتكنولوجيا المعلومات

المجلة الأردنية للحاسوب وتكنولوجيا المعلومات (JJCIT) مجلة علمية عالمية متخصصة محكمة تنشر الأوراق البحثية الأصيلة عالية المستوى في جميع الجوانب والتقنيات المتعلقة بمجالات تكنولوجيا وهندسة الحاسوب والاتصالات وتكنولوجيا المعلومات.

تحتضن وتنشر جامعة الأميرة سمية للتكنولوجيا (PSUT) المجلة الأردنية للحاسوب وتكنولوجيا المعلومات، وهي تصدر بدعم من صندوق دعم البحث العلمي في الأردن. وللباحثين الحق في قراءة كامل نصوص الأوراق البحثية المنشورة في المجلة وطباعتها وتوزيعها والبحث عنها وتنزيلها وتصويرها والوصول إليها. وتسمح المجلة بالنسخ من الأوراق المنشورة، لكن مع الإشارة إلى المصدر.

## الأهداف والمجال

تهدف المجلة الأردنية للحاسوب وتكنولوجيا المعلومات (JJCIT) إلى نشر آخر التطورات في شكل أوراق بحثية أصيلة وبحوث مراجعة في جميع المجالات المتعلقة بالاتصالات وهندسة الحاسوب وتكنولوجيا المعلومات وجعلها متاحة للباحثين في شتى أرجاء العالم. وتركز المجلة على موضوعات تشمل على سبيل المثال لا الحصر: هندسة الحاسوب وشبكات الاتصالات وعلوم الحاسوب ونظم المعلومات وتكنولوجيا المعلومات وتطبيقاتها.

## الفهرسة

المجلة الأردنية للحاسوب وتكنولوجيا المعلومات مفهرسة في كل من:

## فريق دعم هيئة التحرير

## عنوان المجلة

الموقع الإلكتروني: www.jjcit.org
البريد الإلكتروني: jjcit@psut.edu.jo
العنوان: جامعة الاميرة سمية للتكنولوجيا، شارع خليل الساكت، الجبيهة، عمان، الأردن.
صندوق بريد: 1438 عمان 11941 الأردن
هاتف: +962-6-5359949
فاكس: +962-6-7295534

Princess Sumaya University for Technology
جامعة الأميرة سميّة للتكنولوجيا

صندوق دعم البحث العلمي والابتكار
Scientific Research and Innovation Support Fund

# المجلة الأردنية للحاسوب وتكنولوجيا المعلومات

ISSN 2415 - 1076 (Online)
ISSN 2413 - 9351 (Print)

العدد ٢ | المجلد ١١ | حزيران ٢٠٢٥

JJCIT

jjcit@psut.edu.jo
www.jjcit.org

مجلة علمية عالمية متخصصة تصدر
بدعم من صندوق دعم البحث العلمي والابتكار