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:

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

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:

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

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*.

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

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.

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

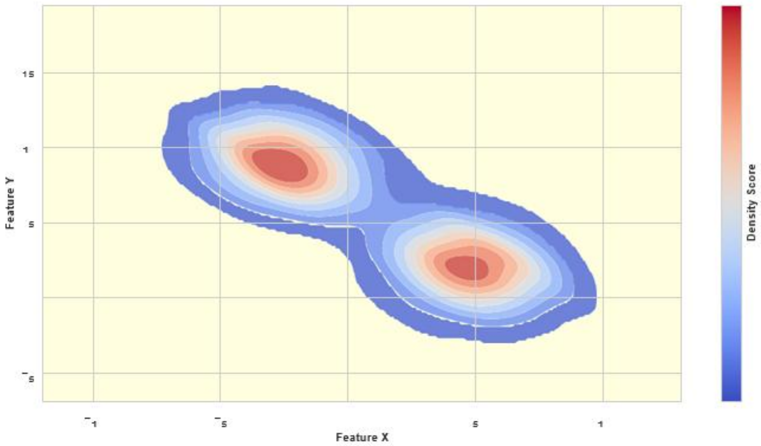


Figure 1. A 2D density plot of data distribution.

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).

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

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 α, which is a weighting factor between 0 and 1, to combine the informativity I and representativity R scores. The formula is as follows:

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

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

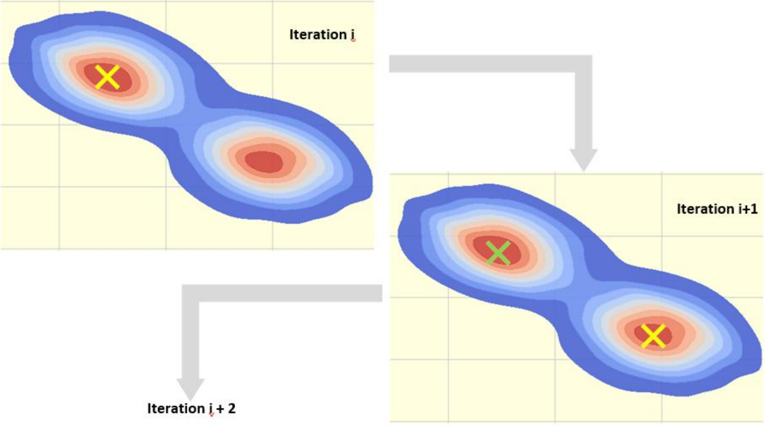


Figure 2. Active ML process.

- 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.

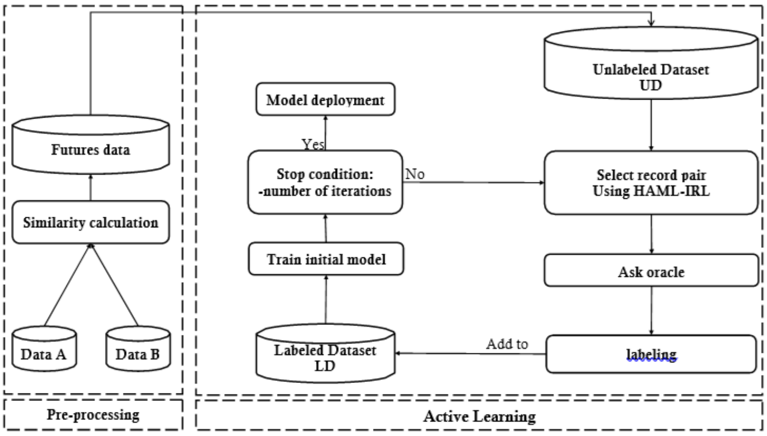


Figure 3. Workflow diagram.

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
$\|D_1\|$	1363	4345	533	6907	2554	1081	1114	6907	2554	51	51
$\|D_2\|$	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] (0.628 [27])
	Informativity	0.560		
	HAML-IRL	0.679		
Amazon-Google	Representativity	0.637	0.480 [13]	0.699 [8] (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.

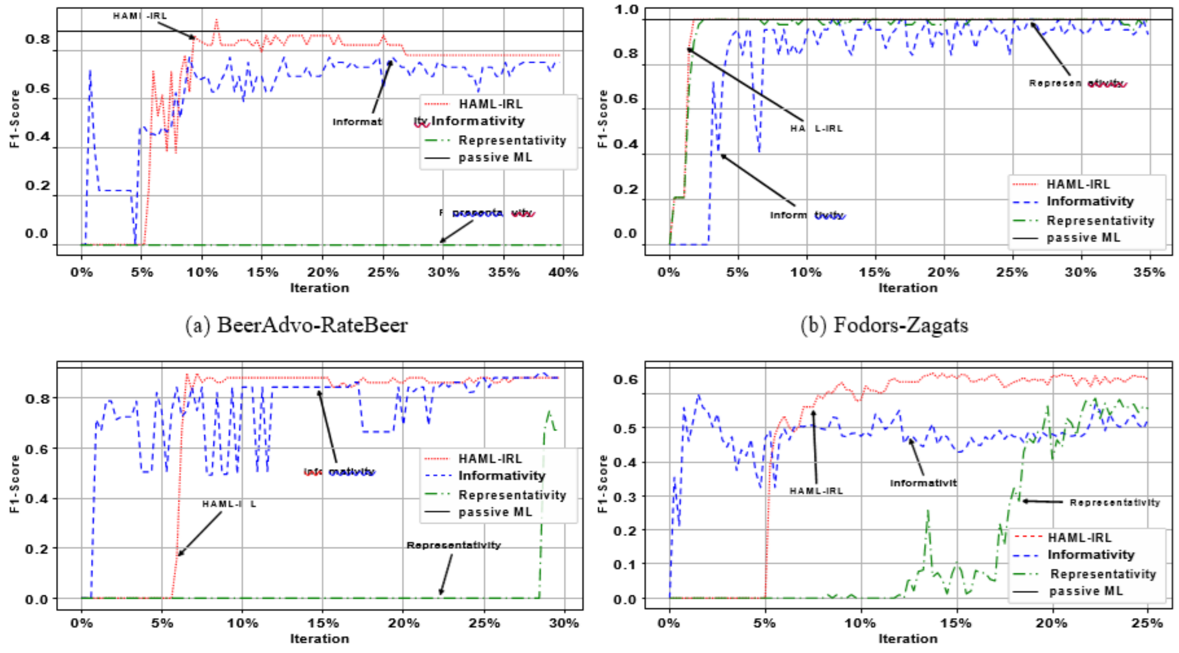
Table 5. Comparative analysis on dirty datasets.

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] (0.849 [24])
	Informativity	0.527		
	HAML-IRL	0.825		
WDC-headphones	Representativity	0.899	0.738 [8]	0.966 [8] (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

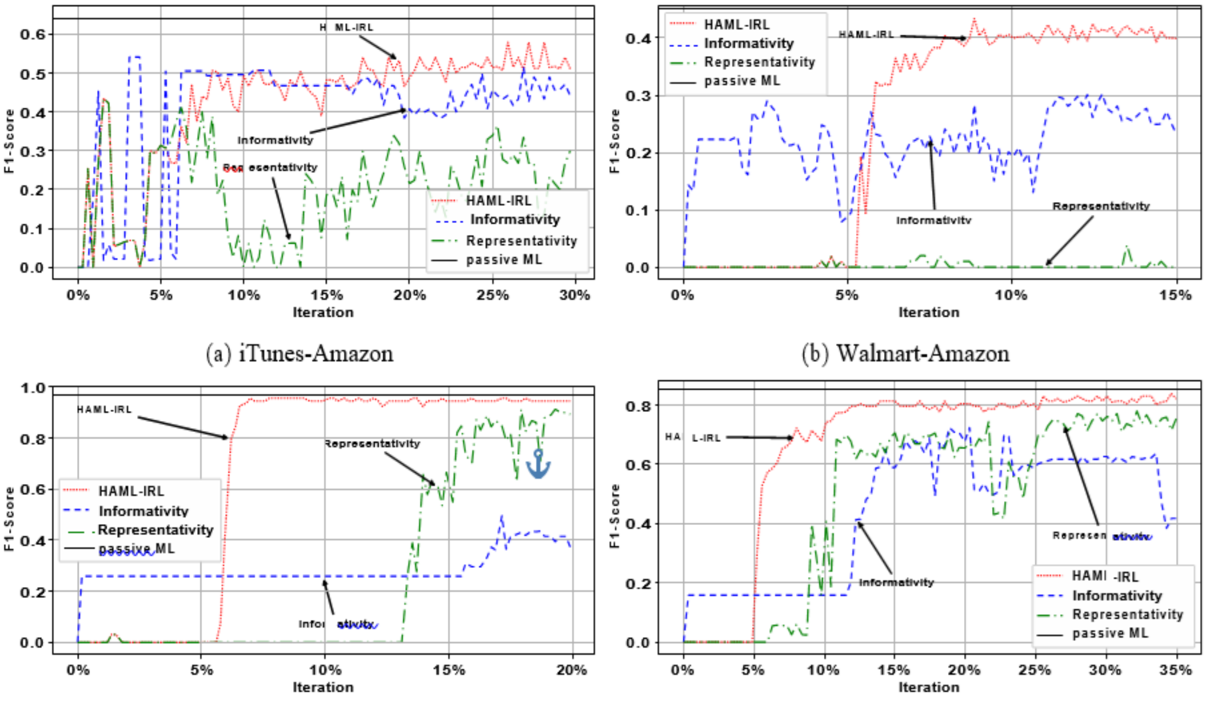
Figure 4. F1-score per AML iteration - structured datasets.

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

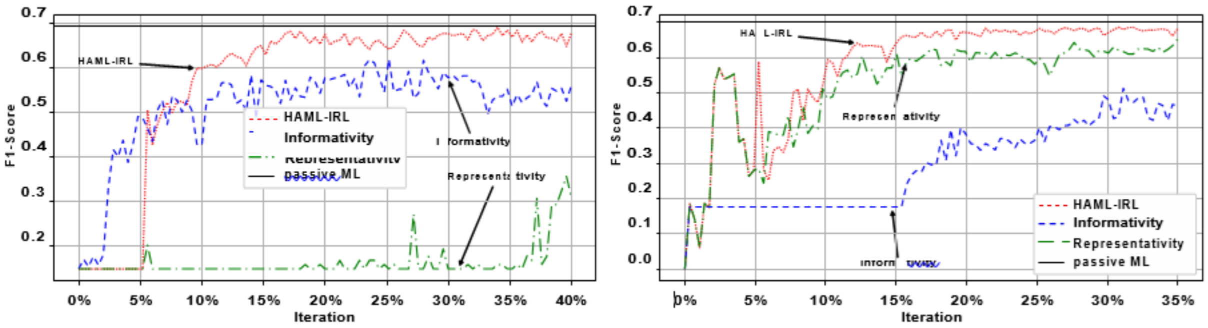
(b) Walmart-Amazon

(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.

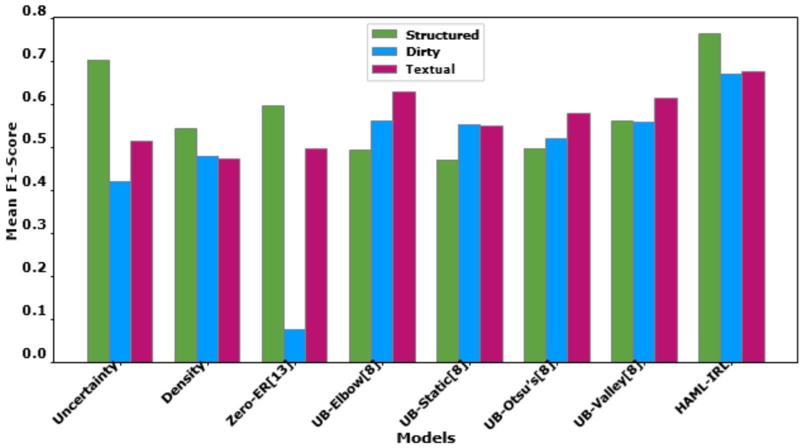


Figure 7. The mean F1-score over structured, dirty and textual datasets.

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.

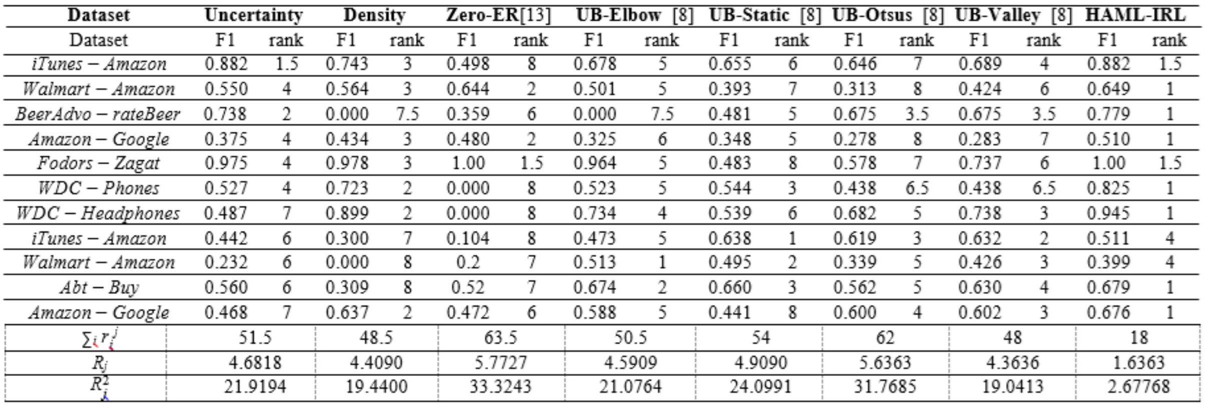
The UB-Valley model shows strong performance in both dirty and textual datasets, with relatively high mean F1- scores. This suggests that UB-Valley is effective at managing both noise and complexities of text-based record linkage. Although its performance in structured datasets is also good, it is slightly 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 χ_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.

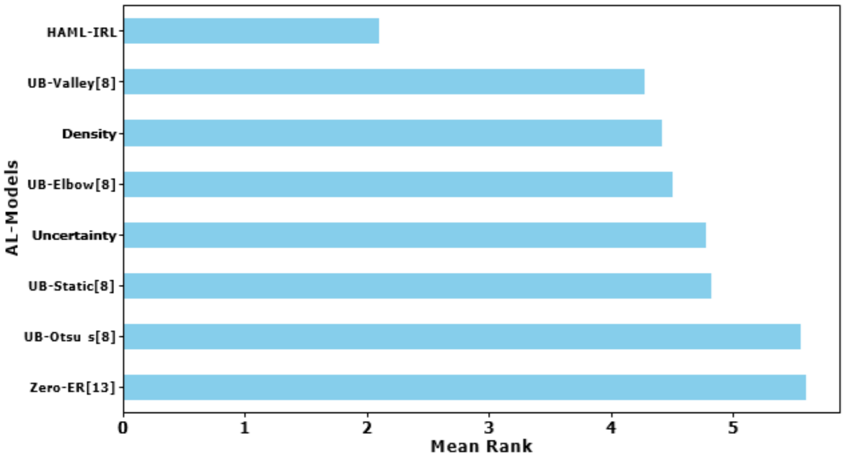


Figure 8. The mean rank over all datasets.

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 (χ^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_α 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	χ^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.

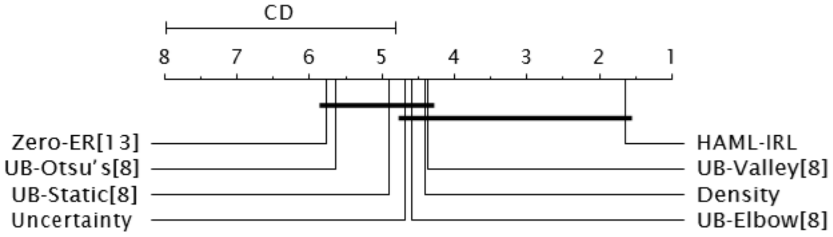


Figure 9. Comparison of all methods against each other using the Nemenyi test.

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.