International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

Patient Identity Resolution in Healthcare Master Data Management Using Ensemble Machine Learning

Somnath Banerjee[1]

Submitted: 05/12/2025 **Revised:** 11/01/2026 **Accepted:** 20/01/2026

Abstract: Patient identity resolution is a cornerstone of healthcare Master Data Management (MDM), ensuring accurate linkage of records to the correct individual. Despite its importance for patient safety and care continuity, many organizations struggle with fragmented identities due to inconsistent data entry, the absence of a universal identifier, and increasing data heterogeneity across electronic health records (EHRs). Traditional deterministic and probabilistic matching methods, widely embedded in commercial Master Data Management tools, exhibit notable shortcomings such as high false positive and false negative rates, heavy reliance on manual stewardship, black-box implementations, and significant licensing costs. This paper examines machine learning techniques that address these challenges, including logistic regression, support vector machines, gradient boosting, bidirectional long short-term memory networks, and Siamese networks. Each model's strengths and limitations are compared with respect to matching accuracy, interpretability, and training data requirements. Comparative analyses suggest that while deep learning models, particularly Siamese networks, excel in text-rich identity resolution tasks, methods like gradient boosting strike a balance between accuracy and operational efficiency. To address the complexity of healthcare data, the paper proposes a multi-model patient matching solution. It incorporates data preprocessing techniques, including anomaly detection via isolation forests and autoencoders, and transformer-based natural language processing for feature extraction, to improve pre-match data quality. An ensemble learning architecture then integrates the complementary strengths of multiple machine learning models to achieve robust, scalable, and explainable patient identity resolution. The findings underscore machine learning's potential to reshape and modernize patient identity resolution across enterprise healthcare systems.

Keywords: *Master Data Management, MDM, Healthcare, Patient Identity Resolution, Machine Learning, Ensemble Classification, Siamese Network, Graph Neural Network, Anomaly Detection*

1. Introduction

Patient identity resolution – identifying and linking records that refer to the same individual – is a critical function in healthcare Master Data Management (MDM). Accurate patient matching ensures that all health records, test results, and treatments are correctly attributed, which is vital for patient safety and quality of care [1], [2]. However, achieving reliable patient matching is challenging due to data quality issues and the lack of a universal patient identifier in U.S. healthcare. Studies indicate that the average healthcare organization has around a 10% duplicate record rate (with some institutions up to 18%) [3] [4], meaning a significant fraction of patient records are fragmented across duplicates. Duplicate records can lead to missed diagnoses, treatment delays, and medical errors. Recognizing the severity of the issue, the U.S. Office of the National Coordinator (ONC) set a goal of 99.5% [4] patient record accuracy by 2020, yet many organizations have fallen short of this benchmark.

This paper examines the challenges underlying patient matching in healthcare MDM, critiques the limitations of current deterministic and probabilistic matching approaches, and surveys machine learning (ML) techniques – including Siamese neural networks, bidirectional LSTMs, gradient boosting, logistic regression, and SVMs – for improving patient identity resolution. The strengths and weaknesses of these models are compared and a composite multi-model patient matching solution is explored that integrates several techniques for robust MDM solutions. The goal is to showcase the state-of-the-art methods for patient identity resolution and how a custom ML-driven MDM solution can overcome the shortcomings of industry tools. The significance of this study lies in its potential to guide healthcare organizations toward higher matching accuracy, reduced manual data stewardship burden, and improved patient safety.

[1]*Independent Researcher, Senior MDM Architect, IEEE Senior Member, New Jersey, USA*
ORCID ID: 0009-0003-0132-4218

2. Challenges in Patient Matching in Healthcare MDM

Patient identity resolution in healthcare MDM faces multifaceted challenges that impede accurate linking of records and elevate clinical and operational risks.

2.1. Data Entry Variability

Healthcare data frequently suffers from inconsistencies due to manual data entry errors, variations in spelling, typographical mistakes, and incomplete information [3], [5]. Names, addresses, and dates of birth are often recorded differently across encounters. Patients may use nicknames or initials, and clerical staff under workload pressures may omit fields or enter approximate data, resulting in discrepancies that hinder deterministic matching. The lack of uniform data entry standards across institutions exacerbates this issue, making direct comparisons unreliable [6].

2.2. Absence of Universal Identifiers

Unlike some countries that utilize a national patient identifier, the United States lacks a single unifying key, compelling reliance on demographic fields that are neither unique nor stable [5]. Common names and birth dates create ambiguous scenarios; multiple patients often share identical demographic details, particularly within large health systems. This ambiguity is amplified in pediatric populations where Social Security numbers are often absent, and guardians may provide inconsistent demographic information [7].

2.3. Data Fragmentation

Patients routinely receive care from multiple providers, leading to data silos across different EHR systems and organizations. Each system may maintain its own record, generating fragmented identities [8], [9]. Moreover, mergers and acquisitions among healthcare entities introduce legacy data with disparate formats and duplicate patient registrations. This fragmentation is further complicated by patient mobility, such as address changes or varying insurance details over time, reducing the efficacy of traditional matching algorithms that depend on stable identifiers [10].

2.4. Data Heterogeneity

The expanding scope of health data now includes IoT device outputs, social determinants of health (SDoH), and free-text clinical notes, which hold valuable context for identity resolution [11], [12]. However, these unstructured or semi-structured data types challenge traditional matching frameworks that primarily rely on structured demographic comparisons. For instance, device IDs or lifestyle indicators may strongly associate records to a patient but are not utilized by standard probabilistic models [13].

3. Matching Shortcomings in Industry MDM Tools

Many widely deployed MDM systems in healthcare use deterministic or probabilistic matching techniques. While these have served as the backbone of patient identity resolution for decades, they reveal substantial shortcomings when faced with the current scale and complexity of healthcare data.

3.1. Deterministic Matching Limitations

Deterministic methods use rigid rules, such as requiring exact matches on names or birth dates, making them highly vulnerable to minor typos or format differences [7], [10]. This sensitivity leads to frequent false negatives, while common demographic patterns can still produce false positives [5].

3.2. Shortcomings of Probabilistic Matching

Probabilistic systems compute similarity scores using weighted fields, offering more flexibility. Yet many industry MDM tools remain opaque "black boxes," limiting customization and adaptation to local data characteristics [8]. Pre-set weights often do not reflect real-world variations within a specific patient population, reducing matching quality [10]. Despite fuzzy logic, false negative rates keep duplicate records in the 10% to 19% range, while aggressive thresholds can cause false positives [3], [4]. Consequently, these systems still depend heavily on manual stewardship to resolve ambiguous matches, inflating operational costs [3].

3.3. Heavy Reliance on Manual Stewardship

Traditional MDM tools primarily process structured demographic fields, ignoring valuable unstructured data like clinical notes or IoT device logs that could improve matching [12]. As a result, large numbers of borderline cases require human review. This manual workload is costly and delays downstream processes.

3.4. Technical, Security and Manual Constraints

Vendor tools often impose fixed infrastructure requirements, limiting database or deployment flexibility. Some also require plaintext data to perform similarity checks, restricting encryption options and complicating compliance [6]. Additionally, high licensing and support costs make these systems expensive, particularly for organizations seeking to modernize or integrate diverse data sources [11], [13].

4. Methodology

4.1. Evaluating Machine Learning Models for Patient Matching in Healthcare MDM

Machine learning offers powerful alternatives to deterministic and probabilistic matching by learning complex patterns directly from data. In healthcare MDM, several ML approaches stand out for patient identity resolution, each suited to different aspects of the challenge.

4.1.1. Bidirectional LSTM (Bi-LSTM)

Bidirectional Long Short-Term Memory (BiLSTM) networks are a type of recurrent neural network that consists of two LSTM layers: one processes the input sequence from start to end (forward direction) and the other from end to start (backward direction). For each position in the sequence (each token, character, or data element), the BiLSTM generates two output vectors – one encoding the prefix (past context) and one encoding the suffix (future context) – which are then combined (concatenated or summed) into one representation [14]. This allows the model to consider both left and right context for any given position, which is critical for capturing order-sensitive text fields, such as, aligning "123 Main Street Apt 4" and "Apt 4, 123 Main St". As such, a BiLSTM can encode sequential data like names and addresses into vectors that capture the semantics more richly than simpler models. BiLSTMs are very effective at capturing contextual information in sequential data; however, that power comes with a high computational cost.

4.1.2. Siamese Networks

A Siamese neural network is a type of neural model particularly suited for comparing two inputs. It consists of two identical subnetworks (e.g., two instances of the same neural network with shared weights) that each process one of the two input records, producing vector representations, which are then compared to compute a similarity score [15],[16]. It is trained using matched and unmatched patient record pairs, minimizing distance for matches and maximizing it for non-matches. A common design is to use text embedding networks (like LSTMs or feed-forward layers) [17], [18]

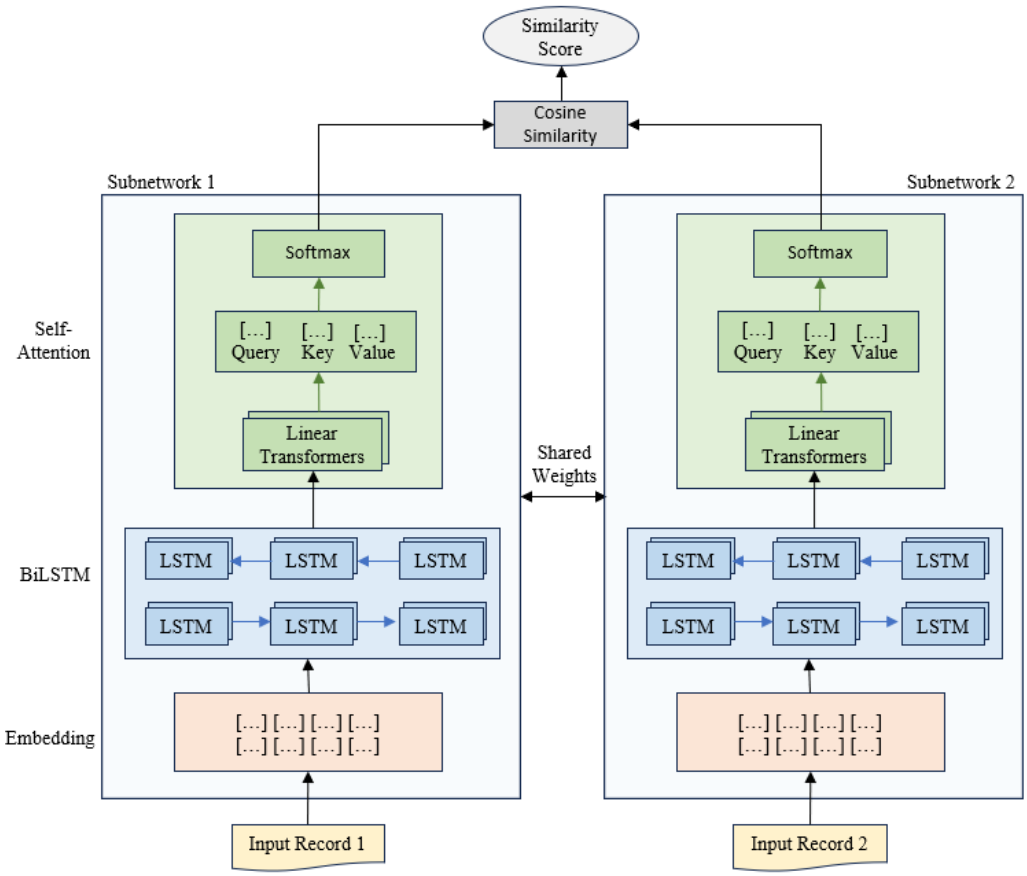

Fig. 1. Siamese Network Implementation with a BiLSTM layer for text embedding and a Self-Attention layer to improve accuracy

to encode each record and then use a distance metric (such as cosine similarity) to indicate how close the records are. In healthcare, this architecture excels at comparing names, addresses, and free-text identifiers, learning to recognize nuances such as nicknames or common misspellings.

Siamese networks excel at pairwise similarity. They do not require manual feature engineering of record comparisons; instead, they automatically learn the most informative features for distinguishing matches vs non-matches. Their main drawback is the requirement for laborious training data.

4.1.3. Gradient Boosting Models

Gradient boosting refers to an ensemble machine learning technique where multiple "weak learners" (often shallow decision trees) are trained in sequence, each one correcting the errors of the previous [19]. The final model is a weighted sum of the outputs of all the trees. Popular implementations include XGBoost, LightGBM [20], and CatBoost. In patient matching, gradient boosting typically consumes a set of engineered features derived from two records, for example phonetic matches, edit distances, and demographic comparisons [21]. The boosting algorithm can capture non-linear interactions between these features and produce a match probability.

Gradient boosting decision trees (GBDTs) are known for their high predictive accuracy on structured data. They can handle a mix of feature types (binary flags, categorical encodings, and numeric similarities) and automatically model interactions between features. Their primary limitation is their reliance on feature engineering. They are only as good as the comparison features provided.

4.1.4. Logistic Regression

Logistic regression is a linear classification model traditionally used in many record linkage systems. It computes a weighted sum of input features and applies a logistic (sigmoid) function to produce a probability between 0 and 1 [22]. Logistic regression has long been used in probabilistic record linkage under the Fellegi-Sunter framework [23]. Some older MDM tools effectively use a form of logistic regression under the hood – the user sets weights for name, DOB, etc., which are then summed to compute a match score, often with some tuning to approximate probabilities [24].

The primary advantage of logistic regression lies in its simplicity and interpretability; however, it cannot capture interactions unless explicitly modelled.

4.1.5. Support Vector Machine (SVM)

Support Vector Machines (SVMs) aim to find a hyperplane that separates matches from non-matches with maximal margin. In record matching, an SVM would ingest a feature vector comparing two records and output a match vs a non-match. SVMs can use various kernel functions to capture non-linear relationships in the input features [25]. In practice, linear and RBF (Gaussian) kernels are common.

SVMs, especially with non-linear kernels, can be quite powerful for binary classification with complex decision boundaries. They often handle high-dimensional feature spaces well and can model interactions implicitly. On the other hand, scalability is a major drawback for SVMs. Training an SVM, especially with a non-linear kernel, on tens or hundreds of thousands of record pairs can be very slow and memory-intensive.

4.2. Optimized Patient Matching using Ensemble Classification

Given the complexity of healthcare data, the most effective system should combine multiple models and rules, leveraging the strength of each. A conceptual Multi-Model Patient Matching Solution is described here. The solution begins with data pre-cleaning and feature extraction as essential preparatory steps, ensuring input quality before any matching logic is applied. The cleaned data then feeds into an ensemble model that integrates the machine learning models discussed above, alongside graph-based clustering and traditional deterministic methods, to achieve a more robust and scalable outcome. The underlying principle is to mimic the decision-making of an expert human matcher, one who intuitively draws on different strategies to confirm, dispute, or defer a match based on the evidence at hand.

4.2.1. Pre-Match Data Cleaning and Wrangling Steps using Machine Learning

A robust patient identity resolution pipeline begins not with matching itself, but with rigorous data preparation. Machine learning can significantly strengthen this stage by detecting anomalies, standardizing disparate inputs, and extracting structured features from unstructured text, thereby enhancing downstream match accuracy.

A. **Anomaly Detection using Isolation Forests and Autoencoders**

One common barrier to accurate matching is the presence of anomalous records. These might involve improbable values—such as future birth dates, implausibly old ages, or phone numbers with incorrect digit lengths—that confound both deterministic and probabilistic systems [26].

Isolation Forest: An Isolation Forest is an unsupervised machine learning algorithm that attempts to find anomalies by leveraging the power of decision trees. It operates based on the principle that anomalies are few and numerically different from normal instances. The tree-based algorithm randomly splits a variable until every single data point is isolated into a leaf node. Anomalies are determined by the number of partitions required to isolate a data point into its leaf node. The lower the number of partitions (i.e., a leaf node positioned closer to the root of the tree), the more anomalous the data point [27], [28]. For example, an Isolation Forest might detect a record where all fields are blank or default values, or one with a combination of demographics that is extremely rare compared to the population (e.g., improbable age combined with unlikely address).

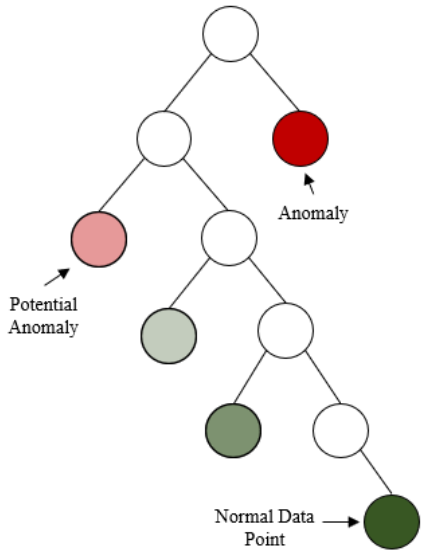

Fig. 2. Isolation Forest - A data point with leaf node closer to the root of the tree indicates an anomaly.

Autoencoders: Autoencoders are neural networks trained to reconstruct the input data. Records that reconstruct poorly (high error) are considered anomalies [29]. For patient identity data, autoencoders trained on common formats for names, dates, and contact details can thus surface unusual or corrupted entries that require manual review or exclusion.

Those anomalous records identified by these models could be set aside for clerical review rather than being fed into the matching process, where they may cause problematic comparisons. By employing these models before matching, organizations reduce noise, improving both precision (by avoiding false positive links due to corrupted data) and recall (by preventing legitimate records from being unfairly treated as distant).

B. **Feature Extraction using Transformer-Based Natural Language Processing**

Modern healthcare datasets increasingly include unstructured or semi-structured elements: physician notes, intake forms, or patient narratives. Ignoring such rich data overlooks signals crucial for resolving identities, such as unique descriptors in notes that indirectly corroborate matches.

Transformer-based models, such as BERT and its clinical derivatives (Clinical BERT), have proven highly effective at mining structured attributes from free text [30]. For instance, a BERT model fine-tuned to extract contact references can pull out phone numbers or family relations embedded in clinical notes, enriching patient profiles. This extracted data then becomes additional inputs for matching models, enabling links that traditional systems, relying purely on demographic tables, would miss.

4.2.2. Ensemble Classifier Using Stacked Machine Learning Models

A. Siamese Networks for Pairwise Text Similarity

A Siamese neural network (possibly with BiLSTM or other text encoders) is ideally suited for comparing patient name strings, addresses, and other textual identifiers. They learn to compute similarity scores that are sensitive to misspellings, cultural variants, and common abbreviations—areas where edit distances or deterministic rules frequently fail [15], [16]. By producing direct similarity probabilities, these networks form a foundational layer for nuanced text comparison.

Within the Siamese network, BiLSTMs encode longer sequential data such as multi-line addresses or clinical notes. For instance, consider multi-line address with street, city, state, ZIP – a BiLSTM can encode the entire address and output an embedding. Two address embeddings can then be compared via cosine similarity. Processing input forward and backward, BiLSTMs capture context-rich embeddings that improve pairwise similarity judgments on fields where order and phrasing matter [17], [18].

B. Graph Neural Networks for Cluster-Level Inference

One limitation of pairwise models is that they don't inherently handle transitive matches. If suppose Record A matches Record B, and Record B matches Record C according to pairwise scores, then A, B, C should all refer to the same patient. A Graph Neural Network (GNN) can capture these higher-order connections. In a graph representation, each patient record is a node, and edges connect records that are deemed likely matches (based on initial pairwise comparisons). A GNN can then be used to classify or cluster the nodes such that those belonging to the same real patient are grouped together. Recent research suggests GNNs are promising for modeling patient linkage [31], [32].

The Graph Neural Network (GNN) essentially can learn entity resolution as a graph partitioning problem, where certain patterns (like two records sharing a phone number and one of them sharing an address with a third record) indicate they all belong to one entity. A GNN might output features like cluster confidence or directly adjust the pairwise probabilities

C. Traditional Deterministic Rules

Alongside advanced machine learning (ML) models, deterministic rules, such as exact matches on Social Security Numbers or close proximity in birth dates, are also invaluable. Many of these deterministic comparison results can either be turned into features for ML (e.g., a binary feature "same SSN" or an edit distance score), or could be used as blocking filters (e.g., do not consider pairs that share no common values in any identifier field to reduce comparisons), thereby reducing the computational load by filtering out obviously unrelated pairs before invoking resource intensive ML scoring.

D. Meta-Classifiers for Final Decision

As a final step, all the generated features – Siamese similarity scores, BiLSTM embedding distances, GNN cluster features, and rule-based features – feed into a meta-classifier. The meta-classifier could be a logistic regression model that outputs the final probability of match, or a gradient boosted model that can capture interactions among the input features. For example, logistic regression might assign a high weight to the SSN match feature, a moderate weight to the name similarity, and so on, and produce a probability. Gradient boosting might notice non-linear patterns, such as when the address similarity and phone similarity are both moderate, together they indicate a match (even if individually they might not).

The ensemble step learns how to balance the different sources of evidence. Ensemble model methodologies have also been reported in industry literature [21]. Combining multiple models helps counteract the blind spots of individual models. As a result, the ensemble can achieve higher accuracy than any single model alone.

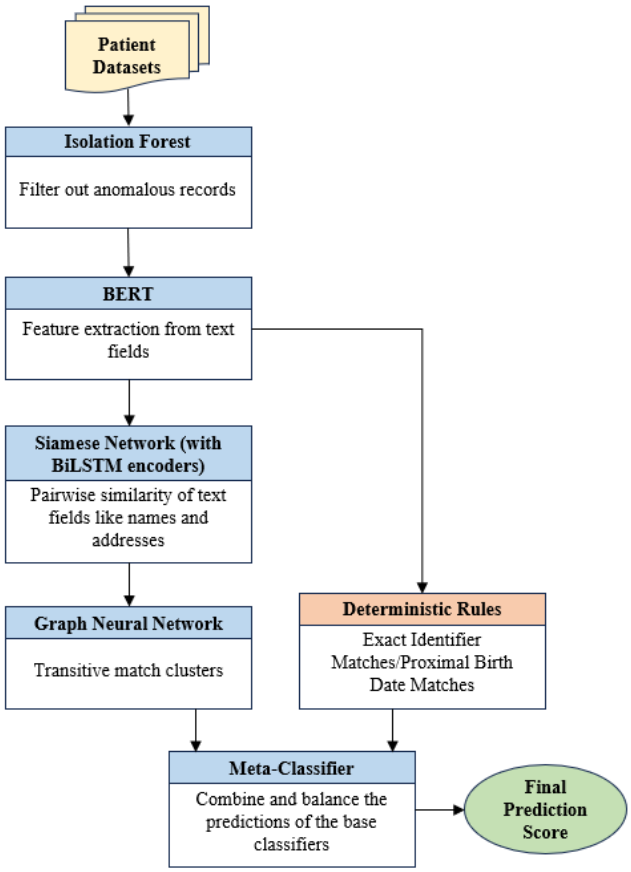

Fig. 3. A conceptual multi-model patient matching solution that uses a multi-step ensemble of machine learning models.

5. Results and Discussion

5.1. Benefits of Using Machine Learning (ML) Models in Custom MDM Solutions

Custom ML-based MDM systems offer several strategic and operational advantages over conventional industry tools.

5.1.1. Adaptation of Local Data Characteristics

Unlike vendor systems that rely on static, generic rules, custom ML models can be trained directly on an organization's own data. This enables learning the unique patterns of data variability and common discrepancies specific to that institution. For example, ML models can discover frequent local misspellings, cultural name variations, or typical errors in demographic fields, leading to improved matching sensitivity without arbitrary thresholds [11]. This tailored learning ensures that the match logic genuinely reflects the statistical properties of the patient population served.

5.1.2. Improved Matching Accuracy and Reduced Manual Oversight

Several studies have demonstrated that ML approaches, especially ensemble methods and deep learning models, achieve higher accuracy in entity resolution tasks compared to conventional probabilistic techniques. For instance, ML-optimized configurations have been shown to improve true positive from approximately 90% to nearly 100%, significantly reducing false negatives [11]. This directly lowers the burden on manual data stewardship teams by minimizing the volume of ambiguous cases requiring human intervention.

5.1.3. Enhanced Flexibility and Feature Scope

ML-based solutions readily incorporate a broader array of features, including unstructured and semi-structured data such as free-text clinical notes, device identifiers, or SDoH metrics. Advanced models can process these diverse inputs without rigid schema constraints, improving match reliability in scenarios where structured demographic data alone is insufficient [12], [13]. This capability is critical as healthcare data grows in heterogeneity with increasing IoT adoption and integration of social and behavioral health indicators.

5.1.4. Cost Efficiency and Elimination of Licensing Barriers

Custom solutions circumvent recurring vendor licensing fees and proprietary upgrade costs. While the initial investment in ML expertise, infrastructure, and iterative tuning can be substantial, organizations retain full ownership of the system, yielding long-term cost savings. This autonomy also allows integration with existing enterprise data platforms without the technical constraints that often accompany vendor products [8].

5.1.5. Transparency and Explainability Options

Developing an in-house ML system grants the organization control over model architecture and feature contributions, improving interpretability. For example, logistic regression models offer direct insight into the importance of each feature through weight coefficients, while gradient boosting frameworks can report feature importances and decision paths [11]. Even for more complex deep learning models, organizations can integrate explainability techniques such as SHAP values or attention heatmaps, enhancing trust and facilitating regulatory compliance

5.1.6. Security and Privacy Customization

Building the MDM system internally enables full alignment with an organization's security policies. This includes deploying advanced encryption, de-identification techniques, and secure multi-party computation strategies to maintain compliance with data protection standards, even during similarity calculations. Such granular privacy controls are often not feasible in off-the-shelf platforms that depend on plaintext for edit distance and phonetic calculations [8].

Table 1. Comparative Summary of Industry MDM Tools Versus Custom ML-Based MDM Solution

Aspect	Industry MDM Tool (Rules/Probabilistic)	Custom ML-based MDM Solution
Customization	Limited (black-box logic)	Highly customizable to data and use-case needs
Accuracy	Moderate accuracy; prone to false negatives/positives	Higher accuracy with ML (learns data nuances)
Manual Effort	Ongoing data stewardship to handle mismatches	Reduced manual review due to improved precision/recall
Explainability	Often opaque; proprietary logic	Model can be opaque, but techniques (feature weights, SHAP) can aid explainability
Data Types	Mostly structured fields (name, DOB, etc.)	Can include unstructured data (notes via Natural Language Processing - NLP) and complex features
Security	May require plaintext data for matching (limited encryption)	Full control to implement encryption, de-identification
Installation	Vendor-specific stack	Free choice of tech stack; can align with org's IT standards
Cost	High licensing and maintenance fees	Initial development cost, but no license fees
Support	Vendor support (with contract)	In-house expertise required for maintenance and updates

5.2. Comparing Various Machine Learning Models for Patient Matching

In general, deep learning models (like Siamese networks, BiLSTMs) tend to push the highest accuracy, especially on unstructured or noisy data, but require more data and entail more complexity. Tree-based models (like gradient boosting) offer a good balance of high accuracy with more manageable complexity and are currently popular in many industrial solutions for matching [21]. Logistic regression is simpler and can perform well with the right features, but may not capture all patterns and interactions present in the data.

A comparison of the various models in terms of accuracy, AUC, training data requirements and interpretability has been summarized in the following table and the Receiver Operating Characteristic (ROC) curve that follows [22], [33]-[35].

Table 2. Comparative Overview of Machine Learning Models

Model	Typical Accuracy	Training Data Requirements	Interpretability
Siamese Network with BiLSTM	Very High (~97%–98%)	Requires a large number of labeled pairs	Low (deep features)

Gradient Boosting	High (~90%–98%)	Works with moderate training datasets.	Medium (feature importances available)
Logistic Regression	Moderate (~87%–92%)	Works with smaller training datasets.	High (weights directly show impact)
SVM	Moderate (~80%–95%)	Not ideal for large datasets. Needs careful kernel tuning.	Low (especially with kernels – no clear way to explain individual decisions)

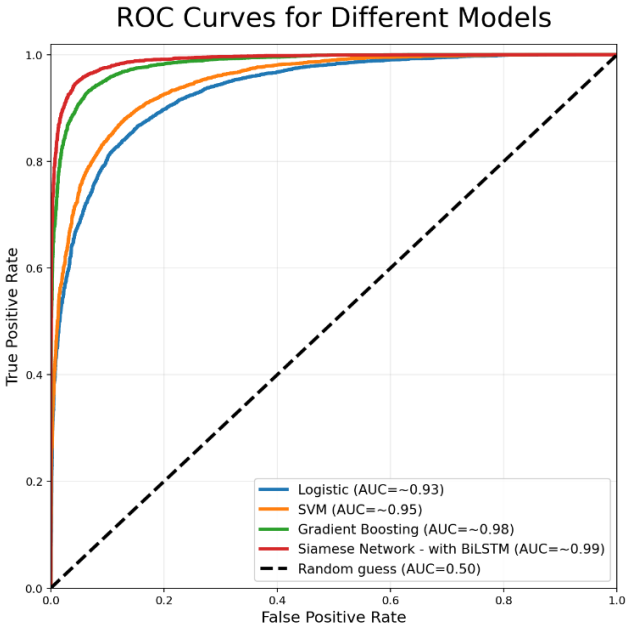

Fig. 4. Receiver Operating Characteristic (ROC) curves for various models, each plotting true positive rate against false positive rate

5.3. Advantages of the Multi-Model Approach

- **Transitive Consistency:** The Graph Neural Network ensures that indirect links—such as Record A matching Record B, and Record B matching Record C—are captured, reducing fragmented clusters that manual stewards frequently have to reconcile.
- **Resilience to Missing Data:** If certain features are unavailable (for example, phone numbers), other components like NLP-derived similarities or graph-based relational cues compensate.
- **Explainable Hybrid Decisions:** While deep learning modules capture complex patterns, the final logistic or boosting layer still permits feature importance analysis, satisfying regulatory demands for transparency.
- **Operational Flexibility:** By tuning ensemble thresholds, organizations can emphasize precision (to minimize false merges) or recall (to minimize duplicate records), aligning the system to specific clinical or billing priorities.

6. Conclusion

Patient identity resolution remains a cornerstone of healthcare Master Data Management, yet traditional approaches, rooted in deterministic rules and probabilistic weights, struggle to meet the precision demanded by modern clinical and regulatory standards. Machine learning techniques learn directly from institutional data patterns and improve matching precision by capturing complexities that static rules cannot. Comparative analyses show that while deep learning and ensemble methods often achieve the highest accuracy, simpler models retain value for interpretability and deployment flexibility. By adopting a multi-model architecture, healthcare organizations can substantially reduce false positives and false negatives while minimizing manual data stewardship.

Future research will be vital to advance privacy-preserving techniques, develop explainable AI tailored to patient matching, leverage active learning to optimize labeling, and create richer multi-modal patient embeddings. As healthcare data ecosystems continue to grow in volume and complexity, such innovations promise to enhance the safety, reliability, and ethical stewardship of patient identity data.

Machine learning presents a transformative opportunity to modernize patient identity resolution in healthcare MDM. By thoughtfully combining these technologies, healthcare organizations can move closer to achieving near-perfect patient matching—thereby reducing medical errors, improving care continuity, streamlining operations, and ultimately safeguarding patient well-being in increasingly interconnected health ecosystems.

Conflicts of interest

The author declares no conflicts of interest.

References

[1] O. Bess, “Why Duplicate and Mismatched Patient Records Are a Bigger Problem Than You Think,” Medical Economics, vol. 100, no. 11, pp. 12–15, Oct. 2023. Accessed on: Jan. 14, 2026. [Online]. Available: https://www.medicaleconomics.com/view/why-duplicate-and-mismatched-patient-records-are-a-bigger-problem-than-you-think

[2] C. Reifsnyder and A. Weinberg, “How Duplicate Patient Records Can Harm Your Practice—and How to Prevent Them”, Veradigm Blog, Aug 2024. Accessed on: Jan. 14, 2026. [Online]. Available: https://veradigm.com/veradigm-news/prevent-duplicate-patient-records/

[3] J. Sultan, “Patient Matching: Obstacles And Solutions For Critical Patient Data Requirements,” Healthcare Business Today, Aug. 2021. Accessed on: Jan. 15, 2026. [Online]. Available: https://www.healthcarebusinesstoday.com/patient-matching-obstacles-and-solutions-for-critical-patient-data-requirements/

[4] Verato Blog, “Achieving the ONC’s mandated 0.5% duplicate rate with Referential Matching” Verato Blog, n.d. Accessed on: Jan. 16, 2026. [Online]. Available: https://verato.com/blog/achieving-onc-mandated-duplicate-rate-referential-matching/

[5] G. Church, “The deadly cost of duplicate patient records | Viewpoint” Chief Healthcare Executive, Nov. 2023. Accessed on: Jan. 16, 2026. [Online]. Available: https://www.chiefhealthcareexecutive.com/view/the-deadly-cost-of-duplicate-patient-records-viewpoint

[6] AHIMA White Paper, “A Realistic Approach to Achieving a 1% Duplicate Record Error Rate”, AHIMA Report, n.d. Accessed on: Jan. 16, 2026. [Online]. Available: https://ahima.org/media/m1pldevh/ahima-pim-whitepaper.pdf

[7] Hospital Access Management, "Training and Tools Can Stop Duplicate Medical Records," Clinician.com, Sep. 2017. Accessed on: Jan. 16, 2026. [Online]. Available: https://www.clinician.com/articles/141259-training-and-tools-can-stop-duplicate-medical-records

[8] IBM Knowledge Center, "Probabilistic Matching in IBM InfoSphere Master Data Management" IBM White Paper, Apr. 2022. Accessed on: Jan. 16, 2026. [Online]. Available: https://www.ibm.com/support/pages/probabilistic-matching-ibm-infosphere-master-data-management

[9] Verato Blog, "The impact of duplicate medical records: How to prevent overlaps and ensure patient safety" Verato Blog, n.d. Accessed on: Jan. 16, 2026. [Online]. Available: https://verato.com/blog/duplicate-medical-records/

[10] C. Leahy et al., "Matching Patient Data with Machine Learning (Part 1: The Problem with Rules)," Included Health Tech Blog, Aug. 2022. Accessed on: Jan. 16, 2026. [Online]. Available: https://includedhealth.com/blog/tech/matching-patient-data-with-machine-learning-part-1/

[11] W. Nelson et al., "Optimizing Patient Record Linkage in a Master Patient Index Using Machine Learning: Algorithm Development and Validation," JMIR Formative Research, vol. 7, e44331, Jun. 2023. DOI: https://doi.org/10.2196/44331.

[12] F. Alafari et al., "Advances in natural language processing for healthcare: A comprehensive review of techniques, applications, and future directions" Computer Science Review, vol. 56, 100725, May 2025. DOI: https://doi.org/10.1016/j.cosrev.2025.100725

[13] T.S. Brisimi et al., "Federated learning of predictive models from federated Electronic Health Records" Int J Med Inform., 112:59-67. Apr. 2018. DOI: 10.1016/j.ijmedinf.2018.01.007

[14] GeeksforGeeks, "Bidirectional LSTM in NLP," GeeksforGeeks Tutorials. May 2025. Accessed on: Jan. 17, 2026. [Online]. Available: https://www.geeksforgeeks.org/nlp/bidirectional-lstm-in-nlp/

[15] A. Jurek-Loughrey, "Deep learning based approach to unstructured record linkage", International Journal of Web Information Systems, vol. 17, no. 2, pp. 607-621, 2021. Accessed on: Jan. 17, 2026. [Online]. Available: https://pureadmin.qub.ac.uk/ws/files/273397384/IJWIS_manuscript.pdf

[16] M. Loster et al., "Knowledge Transfer for Entity Resolution with Siamese Neural Networks", Journal of Data and Information Quality, vol. 13, no. 1, pp. 1-25. Jan. 2021. DOI: https://doi.org/10.1145/3410157

[17] D. Fernández-Llaneza et al., "Siamese Recurrent Neural Network with a Self-Attention Mechanism for Bioactivity Prediction", ACS Omega, 6(16):11086-11094. Apr. 2021.doi: 10.1021/acsomega.1c01266

[18] Y. H. Park et al., "Key Intrinsic Connectivity Networks for Individual Identification With Siamese Long Short-Term Memory," Front. Neurosci., vol. 15, 2021. DOI: https://doi.org/10.3389/fnins.2021.660187

[19] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in Proc. 22nd ACM SIGKDD. Aug. 2016, pp. 785–794. DOI: https://doi.org/10.1145/2939672.2939785

[20] G. Ke, Q. Meng et al., "LightGBM: A Highly Efficient Gradient Boosting Decision Tree", Advances in Neural Information Processing Systems 30 (NIPS 2017), 2017. Accessed on: Feb. 12, 2026. [Online]. Available: https://papers.nips.cc/paper/2017/file/6449f44a102fde848669bdd9eb6b76fa-Paper.pdf

[21] C. Leahy et al., "Matching Patient Data with Machine Learning (Part 2: Leaving Rules Behind)," Included Health Tech Blog, Aug. 2022. Accessed on: Feb. 16, 2026. [Online]. Available: https://includedhealth.com/blog/tech/matching-patient-data-with-machine-learning-part-2/

[22] P. Röchner and F. Rothlauf, "Using machine learning to link electronic health records in cancer registries: On the tradeoff between linkage quality and manual effort", International Journal of Medical Informatics, vol. 185, May 2024. DOI: https://doi.org/10.1016/j.ijmedinf.2024.105387.

[23] I. Fellegi and A. Sunter, "A Theory for Record Linkage," J. Amer. Stat. Assoc., vol. 64, pp. 1183–1210, 1969. Accessed on: Feb. 11, 2026. [Online]. Available: https://www.cs.cornell.edu/~shmat/courses/cs6434/fellegi-sunter.pdf

[24] W. Nelson, N. Khanna et al., "Optimizing Patient Record Linkage in a Master Patient Index Using Machine Learning: Algorithm Development and Validation", JMIR Form Res. Jun. 2023; 7:e44331. DOI: 10.2196/44331

[25] S.T. Chen, Y.H. Hsiao YH et al., "Comparative analysis of logistic regression, support vector machine and artificial neural network for the differential diagnosis of benign and malignant solid breast tumors by the use of three-dimensional power Doppler imaging", Korean J Radiol. Sep. 2009 Sep-Oct. DOI: 10.3348/kjr.2009.10.5.464

[26] P. Christen, "Data Pre-Processing", Data Matching: Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection. Heidelberg, Germany: Springer, 2012, pp. 39-66

[27] B. Bentson, "Using Anomaly Detection to Validate Data Quality," Medium.com, Jul. 2021. Accessed on: Feb. 11, 2026. [Online]. Available: https://bentson-brian.medium.com/using-anomaly-detection-to-validate-data-quality-856944c5e6cd

[28] D. Ribeiro et al., "Isolation Forests and Deep Autoencoders for Industrial Screw Tightening Anomaly Detection", Computers 11, no. 4: 54. DOI: https://doi.org/10.3390/computers11040054

[29] V.Churová V, R. Vyškovský et al., "Anomaly Detection Algorithm for Real-World Data and Evidence in Clinical Research: Implementation, Evaluation, and Validation Study", JMIR Med Inform. May. 2021; 9(5):e27172. doi: 10.2196/27172

[30] K. Yuan et al., "Transformers and large language models are efficient feature extractors for electronic health record studies," Communications Medicine, Mar. 2025. DOI: https://doi.org/10.1038/s43856-025-00790-1

[31] J. G. D. Ochoa and F. E. Mustafa, "Graph neural network modelling as a potentially effective method for predicting and analyzing procedures based on patients' diagnoses", Artif. Intell. Med., vol. 131, Sep. 2022. doi: 10.1016/j.artmed.2022.102359, Sep. 2011.

[32] R. A. Barton, T. Neiman, C. Yuan, "Graph neural networks for inconsistent cluster detection in incremental entity resolution", Amazon Science Publication, 2021. Accessed on: Feb. 11, 2026. [Online]. Available: https://www.amazon.science/publications/graph-neural-networks-for-inconsistent-cluster-detection-in-incremental-entity-resolution

[33] A. T. McNutt et al., "Comparison of Supervised Machine Learning and Probabilistic Approaches for Record Linkage", AMIA Informatics Summit Proc., 2020. Accessed on: Feb. 11, 2026. [Online]. Available: https://scholarworks.indianapolis.iu.edu/bitstreams/decaeb02-42a9-4539-9d82-be4663e479de/download

[34] E.D. Omar, H. Mat et al., “Comparative Analysis of Logistic Regression, Gradient Boosted Trees, SVM, and Random Forest Algorithms for Prediction of Acute Kidney Injury Requiring Dialysis After Cardiac Surgery”, Int J Nephrol Renovasc Dis. Jul. 2024; 17:197-204. doi: 10.2147/IJNRD.S461028

[35] N. Guttenberg and R. Kanai, “Learning to generate classifiers”, arxiv.org. Accessed on: Feb. 11, 2026. [Online]. Available: https://arxiv.org/pdf/1803.11373