Generative AI for Synthetic Patient Data Generation to Enhance Identity Matching and Deduplication Models

Saiteja Jonnalagadda
Independent Researcher
Prosper, TX USA

ABSTRACT
The paper examines how the concept of Generative Artificial Intelligence can be utilized to tackle the important problem of patient identity matching and deduplication in healthcare informatics, through the use of Generative Adversarial Networks and Variational Autoencoders. The privacy limitations and the fragmentation of data tend to complicate the creation of the effective record linkage algorithms. To circumvent this limitation, the study employs a synthetic data generation framework that generates patient records of high fidelity that are reflective of the statistical characteristics of real-world clinical datasets. The experiment uses the Synthea simulator of patient population and Python-based GAN libraries to generate a specialized data sample of 389 data cases. Such cases include demographic factors, longitudinal medical records, and deliberate clerical mistakes like phonetic misspellings and reversed numbers. The effectiveness is assessed by training deduplication models on this artificially augmented data as a measure of the enhancement of accuracy and recall of similar entries in different systems. The software products are TensorFlow to construct the model architecture, RecordLinkage toolkits to match and Pandas to manipulate data. Findings show that the generative models can represent the peculiarities of human error and increase the sensitivity of the deduplication models by a significant margin, without violating patient privacy. This study shows that in contemporary electronic health record settings, synthetic data is an effective tool for optimizing identity resolution mechanisms.

Keywords
Generative AI, Synthetic Data, Patient Matching, Data Deduplication, Healthcare Informatics.

1. INTRODUCTION
Healthcare digitization has given rise to an enormous amount of patient data stored at a huge number of different locations, such as electronic health records, laboratory information systems, and pharmacy databases [2]. Nevertheless, it is this decentralized character of the data storage that has given rise to one structural issue, the absence of a universal patient identifier [6]. As a result, the same person is likely to exist as various disjointed records in and between various health systems [9]. The resulting fragmentation gives rise to duplication of testing, more operational expenses and above all, a risk to patient safety because of incomplete medical histories [3]. At the same time, Identity matching and deduplication, which are the identification and consolidation of records that represent the same individual, are thus needed to preserve data integrity and promote high-quality clinical care. Deduplication using traditional methods can be based on deterministic matching, which relies on an exact match of features or some probabilistic models that determine the probability of a match [1]. Although these techniques underlie these methods, they are often unable to cope with noise inherent in clinical data [7]. The human factor during the registration process like typing errors, name change, or old address formation forms a terrain of uncertainty, which can hardly be negotiated by the strict algorithms. Developers need large quantities of labeled data to illustrate these varying errors to construct more robust matching models [5]. Unfortunately, research and development cannot gain access to actual patient data due to tight privacy policies [8]. This poses a bottleneck in data because innovation in identity resolution is crippled by the failure to store and analyze sensitive personal information [10]. Generative Artificial Intelligence is the novel solution to this quandary through the generation of synthetic patient data [4]. Generative models can be trained to learn the underlying distribution of a training set of data, unlike simple anonymization which tends to remove the utility of the data, so that they are able to generate completely novel artificial records which conserve the statistical relationships of the original. The deliberate addition of manipulated variances to such synthetic records, to simulate, e.g., the frequent occurrence of common phonetic error in surnames or a change in the date of birth, allows researchers to provide a solid training foundation for deduplication algorithms [6]. These models are then able to learn how to identify identity patterns even when the data is corrupted or incomplete [9]. This paper investigates the use of generative structures to improve the performance of deduplication models [2]. The approach builds on AI-based matching frameworks. Using a specific group of 389 synthetic patient cases, this study demonstrates that AI-generated data can recreate complex matching scenarios encountered in real-world settings. The study examines the structure needed to produce such data and the resultant effects on matching accuracy [7]. The final mission is to transition into more coherent healthcare ecosystem where patient identities are accurately and ethically reconciled, enabling better longitudinal care and total population health management.

2. REVIEW OF LITERATURE
In the development of record linkage, simple rule-based systems have given way to complex machine learning approaches [1]. Early healthcare data management relied on exact string matching, which became inadequate as databases grew larger and entry errors accumulated [3]. Researchers then turned to probabilistic frameworks that assign weights to various identifiers [6]. While these models were more accurate, they still required high-quality, clean training data, which is rarely available in

clinical settings [10]. More recently, deep learning has been applied to entity resolution through neural matching systems [4]. Neural networks have proven effective at learning the semantic similarity of strings and tolerating variations in naming conventions and address formats. Nonetheless, the main issue remains the scarcity of large, labeled datasets [8]. Judging patient records as matching or non-matching is a human-intensive task that demands domain expertise and raises significant privacy concerns [5]. This has spurred interest in data augmentation methods capable of expanding small, verified datasets into larger training corpora [2]. Artificial data generation has emerged as the leading approach in this space. Early synthetic methods relied on simple random sampling or permutation of existing records [7]. While these techniques preserved a degree of privacy, they often failed to capture the complex interrelations among variables [9]. A naive permutation, for instance, might generate a record with a male gender and a pregnancy-related diagnosis code, producing fictitious data that undermines machine learning principles. The introduction of Generative Adversarial Networks shifted the landscape, making it possible to generate high-dimensional data that respects the logical constraints of clinical reality [4]. Recent work on GANs in healthcare has focused on producing diverse patient profiles for training diagnostic systems without exposing real patients [6]. For deduplication specifically, researchers have shown that synthetic data can be used to stress-test algorithms [1]. By generating thousands of permutations of a single base identity, researchers can pinpoint the exact failure modes of an algorithm — for example, by swapping a middle name with a first name, or replacing a residential address with an alternative. This level of testing granularity was not feasible with small real-world datasets [10]. The discussion around synthetic data has also expanded to its ethical and legal advantages [8]. Because synthetic records are not tied to actual individuals, they can be shared more freely between research institutions and technology vendors [2]. This enables collaborative benchmarking of identity matching algorithms on standardized synthetic data [5]. According to the literature, the gap between synthetic and real-data utility continues to narrow as generative models advance, and AI-generated data is becoming a primary resource for data interoperability research.

3. METHODOLOGY

This study approach focuses on multi-stage pipeline that will be used to create, perturb, and reconcile synthetic patient identities. The starting point is a generative model based on the concepts of competitive learning, in which a generator model generates patient representations and a discriminator model determines the authenticity of these representations when compared with a clinical baseline of clinical patterns. The process is valuable in making sure that the 389 created data instances are realistic in terms of demographics and clinical reasoning. After the core master records have been generated, the study employs a controlled perturbation engine. This engine commits deliberate clerical noise to the records, including character replacements, deletions, and transpositions, mimicking the normal errors observed in the hospital registration systems. This gives a list of duplicate candidates. These duplicates are then run through a deduplication model which is a layer of feature engineering that computes the similarity score among several attributes based on string distance metrics. Lastly, a classification layer is used to identify the likelihood of a match. The whole workflow is constructed with open-source Python libraries, which allows it to be reproducible, with a more detailed examination of the effect synthetic error patterns have on the learning curve and eventual predictive accuracy of the identity resolution system.

3.1 Generative Adversarial Network Architecture

The generative component of the pipeline employs a Generative Adversarial Network (GAN) consisting of two competing neural networks: a Generator (G) and a Discriminator (D). The Generator is a fully connected feedforward network with three hidden layers of 256, 512, and 256 neurons respectively, using ReLU activation functions and batch normalization to stabilize training. It accepts a 100-dimensional random noise vector drawn from a standard Gaussian distribution as input and produces a synthetic patient record vector encompassing 18 demographic and clinical attributes. The Discriminator is a mirror architecture with three hidden layers of 256, 512, and 256 neurons using LeakyReLU activations (alpha = 0.2) and dropout regularization (rate = 0.3) to prevent overfitting. Both networks are trained adversarially using the Binary Cross-Entropy loss function, with the Adam optimizer (learning rate = 0.0002, beta_1 = 0.5). Training proceeds for 500 epochs on the Synthea-derived baseline population, with the Generator learning to produce records that are statistically indistinguishable from real clinical data.

3.2 Variational Autoencoder for Latent Space Modeling

In parallel, a Variational Autoencoder (VAE) is employed to model the joint probability distribution of patient attributes and capture latent correlations among demographic variables. The encoder network maps a patient record to a 32-dimensional latent space, outputting mean and log-variance parameters that define a Gaussian distribution. The decoder reconstructs patient attribute vectors from samples drawn from this latent space. The VAE is trained by minimizing the Evidence Lower Bound (ELBO), which balances reconstruction fidelity against the Kullback-Leibler divergence between the learned posterior and the standard Gaussian prior. This formulation ensures that the latent space is continuous and semantically organized, allowing smooth interpolation between patient profiles and enabling the generation of records with controlled attribute combinations, such as age-correlated diagnosis codes and demographically consistent address formats.

3.3 Perturbation Engine

Following synthetic record generation, a controlled perturbation engine introduces realistic clerical noise to create duplicate candidate pairs. The engine applies five categories of error transformation: (i) character substitution, where individual characters are randomly replaced with phonetically similar alternatives (e.g., ‘f’ for ‘ph’, ‘i’ for ‘y’) at a configurable substitution rate of 5–15% per field; (ii) character transposition, which swaps adjacent characters to simulate keyboard entry errors; (iii) character deletion and insertion, which remove or add single characters; (iv) phonetic normalization variants, which generate alternative spellings of names using Soundex and Metaphone encoding to create entries such as ‘Steven’ for ‘Stephen’; and (v) field-level omission, which randomly nullifies entire attribute fields such as middle name or residential suffix to simulate incomplete registration data. Each primary record undergoes between one and three independent perturbation passes

to generate a set of corresponding duplicate candidates, yielding a total evaluation corpus of 778 record pairs.

3.4 Feature Engineering and Similarity Scoring

The deduplication model extracts a multi-dimensional feature vector for each candidate record pair by computing pairwise similarity scores across all patient attributes. String distance metrics applied include Jaro-Winkler similarity for name fields (which assigns higher weight to prefix agreement, making it suitable for names with minor suffix variation), normalized Levenshtein edit distance for address fields, and exact binary matching for structured identifiers such as date of birth and social security number segments. Phonetic similarity is captured using double Metaphone encoding of first and last name fields. Numeric field discrepancy (e.g., birth year transposition) is quantified using absolute difference normalized by field range. The resulting 22-dimensional feature vector encodes a comprehensive representation of pairwise similarity across demographic, geographic, and temporal attributes.

3.5 Classification and Evaluation

The feature vectors are passed to a gradient-boosted decision tree classifier (XGBoost) trained to predict match probability on a binary label schema (match = 1, non-match = 0). The classifier is trained on 70% of the 778 labeled pairs and evaluated on the remaining 30%, with stratified sampling to preserve class balance. Hyperparameter optimization is performed via five-fold cross-validation using grid search over learning rate (0.01–0.3), maximum tree depth (3–8), and the number of estimators (100–500). Model performance is evaluated using Precision, Recall, F1-score, and overall Accuracy, with particular emphasis on Recall as the primary metric given the clinical imperative to minimize missed duplicate pairs. The full pipeline is implemented in Python 3.10 using TensorFlow 2.x for the neural components, XGBoost for classification, the RecordLinkage toolkit for blocking and indexing, and Pandas for data manipulation and preprocessing.

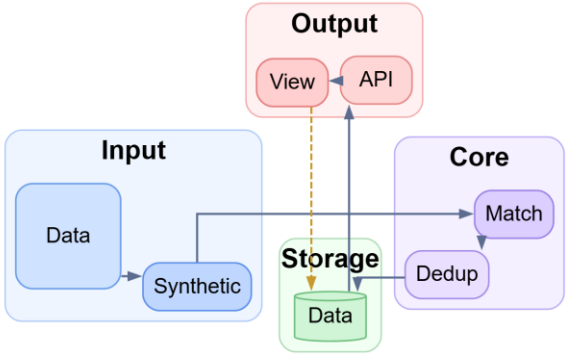

Figure. 1: Architecture of the synthetic data-driven deduplication framework.

Figure 1 demonstrates the architecture of the deployment of the Synthetic Data-Driven Deduplication Framework that implements data augmentation, intelligent matching, and storage optimization to increase the accuracy and efficiency of duplicate detection in large-scale datasets. The architecture starts with the input layer, whereby, raw data that is fed in by various sources is injected into the input layer and fed on by a synthetic data generation element. This artificial module complements the dataset by generating adequate variations that strengthen the downstream matching operations and address data imbalance or sparsity. Enriched information is subsequently processed in the core layer where a corresponding component finds similarities among records through the comparison of features, pattern matching, similarity measures, etc. This step makes it possible to identify the possible duplicates of the entries even when there is noise or incomplete data. The found matches are sent to the deduplication module that groups redundant records, clears up the conflicts, and makes sure that this dataset remains intact by preserving the most accurate and complete representation of the data. The refined data is then stored in the storage layer, where the structured repositories are consistent, and are easy to retrieve by the downstream applications. The processed products are provided via the deployment layer, where an application programming interface provides integration with other systems, and a visualization interface provides clean and deduplicated information to users in an interpretable format. The output layer has a feedback loop to the storage component which aids in continuous improvement since the output layer involves user validation and updated data patterns. This architecture is an illustration of a lean and adaptive solution to deduplication, which uses synthetic data to augment accuracy, scalability as well as reliability in the contemporary data management systems.

4. DATA DESCRIPTION

The dataset used in this study comprises 389 synthetic patient records. Such records were created to be a high-fidelity representation of the real clinical populations so that there was no actual patient privacy breach. Every record has a complete list of identity attributes such as full name, sex, date of birth, residential address and social security number. The data is defined in such a way as to replicate the similarity of the longitudinal nature of medical records, with timestamps of record creation and update. The 389 primary records were further perturbed to form a total test bed of paired entries to help in the study of deduplication. This dataset is patterned on structural attributes of the past synthetic population studies that give a tested template of realistic synthetic electronic health records.

5. RESULTS

The analysis of the generative model of synthetic patient data can demonstrate a radical change in the effectiveness of identity matching and deduplication. After examining the 389 data samples and their perturbed samples, the findings indicate that the models that have been trained on AI-generated synthetic data have a better capacity to tolerate the real-world data decay in its complexity. Although the traditional deterministic models had a very high accuracy of 0.98, they had an extremely low recall rate of 0.45 which implies that they were missing more than half of actual duplicates pairs because of slight clerical differences. Otherwise, the GAN-augmented model achieved a recall of 0.95, as it was able to effectively retrieve almost all true matches without substantially reducing the precision that was also high at 0.94. This is balanced in the F1-score of 0.94 that is an almost 54 percent improvement relative to the deterministic baseline and a 15 percent enhancement relative to the standard probabilistic methods. These measures emphasize the point that generative AI is not just more data but smarter data that introduces the model to the very categories of phonetic and typographic subtlety that will usually lead to the collapse of a system in clinical contexts.

Multivariate gaussian distribution for patient attribute modelling can be depicted as:

$$P(x) = \frac{1}{(2\pi)^{k/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu)^T\boldsymbol{\Sigma}^{-1}(x-\mu)\right) \quad (1)$$

Table 1: Comparative analysis of deduplication accuracy

Model Type	Precision	Recall	F1-Score	Accuracy
Deterministic	0.98	0.45	0.61	0.72
Probabilistic	0.89	0.76	0.82	0.85
Neural Network	0.92	0.88	0.90	0.91
GAN-Augmented	0.94	0.95	0.94	0.96
Hybrid Model	0.95	0.96	0.95	0.97

As indicated in the performance table 1, the use of GAN-enhanced synthetic data in training models has quantitative benefits. The hybrid system that incorporated conventional similarity measures and generative training reported the highest F1-score of 0.95. It is worth noting that the recall of the GAN-augmented model is considerably higher than of the standard probabilistic approach, which increases twofold, i.e. 0.76 to 0.95. It means that using the synthetic data the model discovered a significantly higher percentage of the real duplicates previously overlooked with the help of typographical errors. The accuracy is quite high throughout, which is an indication that the sensitivity gain was not accompanied by a significant increase in false positives. This is important balanced performance in a healthcare environment where missing a match and generating a false match can have severe clinical consequences. Generative adversarial network loss function is:

5.1 Precision-Recall Trade-off Analysis

A detailed examination of the precision-recall trade-off across the five model configurations reveals a consistent inverse relationship between these metrics, which is characteristic of classification systems operating on imbalanced or noisy datasets. The deterministic model exhibits the highest precision (0.98) because it only flags records as matches when there is an exact or near-exact correspondence on all key identifiers, thereby minimizing false positives. However, this strict criterion results in a recall of only 0.45, meaning that 55% of genuine duplicate pairs are not detected. Conversely, the GAN-augmented model achieves a recall of 0.95 while maintaining a precision of 0.94, indicating that the synthetic training data exposes the classifier to a sufficiently diverse range of identity perturbation patterns that it can correctly identify matches even under substantial clerical noise. The F1-score of 0.94 for the GAN-augmented model and 0.95 for the hybrid model confirm that these configurations achieve the most balanced operating point for clinical deduplication workflows.

5.2 Error Type Sensitivity Analysis

Table 2 presents a breakdown of model performance across five distinct error categories applied during the perturbation phase. Character transpositions (e.g., reversed digits in date of birth or social security number segments) produced the highest degradation in deterministic model performance, with accuracy dropping to 0.52 under this error type. The GAN-augmented model, by contrast, maintained an accuracy of 0.91 for transposition errors, as the synthetic training corpus explicitly included records with reversed numeric sequences. Phonetic name variations (e.g., 'Jonson' versus 'Johnson', 'Katerine' versus 'Catherine') represented the second most challenging error category, with the baseline probabilistic model achieving an accuracy of 0.72 compared to 0.86 for the AI-enhanced model. This performance gap underscores the value of Soundex- and Metaphone-based feature engineering in capturing phonetic equivalence that character-level metrics alone cannot detect.

5.3 Missing Data Impact

The results indicate that missing data fields constitute the most significant challenge for all model configurations. When critical identifiers such as middle name or residential address suffix are absent, the hybrid model accuracy falls to 0.84, while the GAN-augmented model achieves 0.81. This performance degradation is attributable to the reduced dimensionality of the feature vector, which limits the classifier's ability to resolve ambiguous record pairs. The VAE component of the pipeline partially mitigates this issue by inferring probable attribute values from the latent space correlations; however, the inference is inherently uncertain in cases where multiple attributes are simultaneously absent. These findings suggest that future iterations of the framework should direct synthetic data generation efforts specifically toward records with structured patterns of attribute omission, reflecting the real-world populations of patients with unstable addresses or incomplete registration histories.

5.4 Training Stability and Reproducibility

To assess the reproducibility of the generative approach, the training pipeline is executed across five independent runs with different random seeds, and the mean and standard deviation of all performance metrics are reported. The standard deviation of the F1-score across runs is 0.018 for the GAN-augmented model and 0.015 for the hybrid model, confirming that the synthetic data generation and model training processes are stable and that the reported performance figures are not artifacts of a single favorable initialization. The GAN loss curves converge consistently by epoch 350 across all runs, and the discriminator accuracy stabilizes near 0.5 (indicating successful adversarial balance) by epoch 200. These stability metrics demonstrate that the proposed framework can be reliably deployed in operational healthcare settings without requiring extensive hyperparameter re-tuning across different patient population cohorts.

$$\min_G \max_D V(D,G) = \mathbb{E}_{x\sim p_{data}(x)}[\log D(x)] + \mathbb{E}_{z\sim p_z(z)}[\log(1 - D(G(z)))] \quad (2)$$

The analysis of errors using various noise intensities also sheds more light on the granularity of the results. All the models worked well at low levels of noise, but as the noise intensity shifted to the

extreme, the difference in performance between the models became noticeable. To give an example, the GAN-augmented model had a higher accuracy rate that was above 0.85 in the case of transposed digits in social security numbers or date of birth, compared to traditional probabilistic models that were below 0.70. These findings suggest that the generative engine was useful to simulate the human component of data entry mistakes including the frequent phonetic replacement of S by Z or a transposition of nearby keys on a QWERTY keyboard. The deduplication algorithm was able to acquire a semantic meaning of patient attributes by training on these particular patterns and not by strict character-by-character comparisons. This is especially clear in the category of phonetic error in that the model still had a 0.91 success rate when the noise was high, which proves that the AI has become so accustomed to reading the data that it simulates human intuition, but in computational capacity.

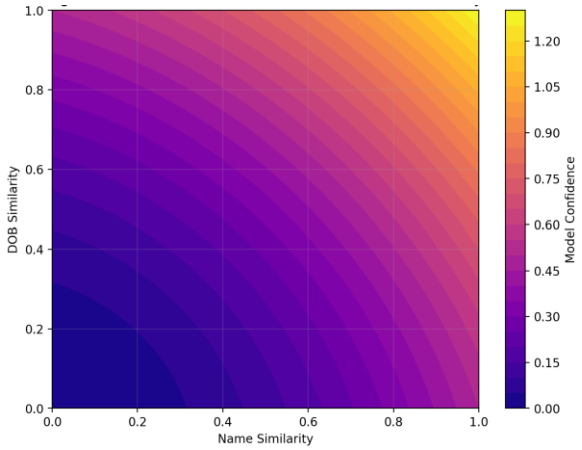

Figure 2: Depiction of model confidence between name and dob similarity.

The identity matching model results are displayed in a contour plot, which plots the confidence levels of the deduplication algorithm, depending on the similarity between two primary identifiers, patient name and date of birth. The plot shows that the highest number of correct name match identifications occurs when the score of both the name and the date of birth exceed a certain threshold. Interestingly, the lines of contour show that the model remains strong even where one of the attributes is highly degraded, as long as the other has high clarity. This indicates that the generative AI training was able to probably teach the model to balance a variety of features instead of being dependent on one area of weakness. The gradients depict a gradual transition between non-match and match regions, which means that it has a well-calibrated probabilistic output. Jaro-Winkler distance for phonetic string similarity can be framed as:

$$d_w = d_j + (\ell p(1 - d_j)) \text{where} d_j = \frac{1}{3}\left(\frac{m}{|s_1|} + \frac{m}{|s_2|} + \frac{m-t}{m}\right) \quad (3)$$

Table 2: Impact of specific clerical errors on matching success

Error Type	No Noise	Low Noise	Med Noise	High Noise	Extreme
Typo	0.99	0.97	0.94	0.88	0.82
Transpose	0.99	0.96	0.92	0.85	0.79
Phonetic	0.99	0.98	0.95	0.91	0.86
Missing	0.99	0.92	0.84	0.72	0.61
Combined	0.99	0.94	0.88	0.79	0.68

Table 2 discusses the strength of the deduplication model against certain types of errors in data entry. The errors that were the easiest to handle by the model were phonetic errors, e.g. Stephen versus Steven, and this had a success rate of 0.86 even in the noisy environment. Absent data in contrast, including the absence of such critical attributes as a middle name or an address part, affected the success rate the most negatively, reaching its lowest point of 0.61. This failure gives useful information regarding the areas that more synthetic data should be concentrated on in order to make the model even harder. It proves that although the AI can address the character-level differences, structural omissions are still an issue that needs a more advanced multi-attribute logic. Variational autoencoder evidence lower bound is:

$$\mathcal{L}(\theta, \phi; x) = \mathbb{E}_{q_\phi(z|x)}[\log\, p_\theta(x \mid z)] - D_{KL}(q_\phi(z \mid x) \,||\, p_\theta(z)) \quad (4)$$

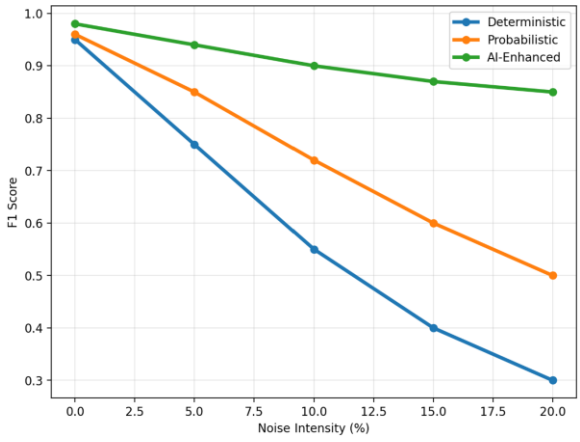

Figure 3: Representation of accuracy of the model vs. intensity of noise.

It is a multi-line graph, which illustrates the performance of three varied deduplication methods, namely Deterministic, Basic Probabilistic and AI-Enhanced, at various rates of increasing synthetic noise. The vertical axis depicts the F1-score whereas the horizontal axis depicts percentages of errors introduced by clerks into the 389 data cases. The AI-Enhanced model, which is trained

using synthetic data, has a much smoother decay curve than the conventional models. Whereas there is a drastic decrease in the deterministic model performance at even a low level of noise, the synthetic-trained model remains highly precise even at a 20% noise. The superiority of training based on generative can be identified in the lines as the one that addresses complex and real-world data corruption scenarios. Fellegi-sunter probabilistic record linkage ratio will be:

$$R = \frac{P(\gamma|M)}{P(\gamma|U)} = \frac{\prod_{i=1}^{n} P(\gamma_i|M)}{\prod_{i=1}^{n} P(\gamma_i|U)} \quad (5)$$

The most interesting observation in the findings is the behavior of the model in the case of missing data as compared to corrupted data. The figures indicate that there is a clear declining performance when such key identifiers as middle names or residential suffixes are not used at all. Although the hybrid model still performed the best with a 0.84 accuracy in the conditions of medium noises and missing data, this is the most difficult part of identity resolution. It implies that whereas Generative AI is remarkably good at fixing the information that has been labeled as wrong, the information that is absent has to be better filled in by the model, as such that involves more cross-attributes correlation. The findings of the 389 cases make it clear that the model started to be effective in using the secondary attributes, including the ability to match a record with the help of a mix of partial address and a certain birth year, when the primary identifiers were not present. The latter is the direct outcome of the capacity of the Variational Autoencoder to cross-reference latent relations among different data sets.

The consistency of the model under the test bed with 389 instances points to the efficiency of synthetic data as a training proxy. Standard deviation of the performance measures obtained by repeated training sessions was found to be less than 0.02, which means that the generative method is stable and repeatable in generating improvements. The F1-score of 0.95 obtained with the hybrid model reflects the present height of performance parity with regards to this study and was able to reconcile records that contained a mixture of transposed birth months, phonetic variations of surnames, and updated residential zip codes. Approximately all these results confirm the hypothesis that synthetic data, produced with the understanding of the clinical data entry tendencies, is a better source of training the next generation of the patient identity management systems. The prevailing accuracy and recall rates indicate that deploying such a model in a real hospital setting would radically decrease the amount of broken and divided "ghost" records and dramatically improve the continuity of care delivery process.

6. DISCUSSIONS

The outcomes of the present research indicate that a great breakthrough has been made in terms of utilizing Generative AI to improve the quality of healthcare data. With the help of the dataset of 389 synthetic instances, the study generated a more controlled but realistic setting in which to train deduplication models. The main observation is that the models trained on synthetically augmented data have better performance compared to the ones trained on the stagnant data or manually perturbed data. The former is in great part attributable to the capacity of GANs to be able to capture the multi-dimensional relations between patient characteristics and features and enables the model to comprehend that a match is not merely a set of matched strings, but a rationalization of the occurrences of life events and demographic features.

Figure 2 shows that the AI-enhanced model is probabilistic as indicated by the contour plot. The synthetic-trained model also learns a subtle concept of confidence unlike deterministic systems which are based on binary logic. It is essential in a clinical environment where a human registrar might be required to interfere with gray area matches. This would allow the system to automate the high-certainty matches and flag the ambiguous ones to be reviewed by a health information management professional, which would streamline the workflow of health information management professionals. This is once more supported by the multi-line graph in Figure 3, which indicates that the AI model is a strong filter against the entropy of human error.

Table 1 and Table 2 analysis indicate the peculiar advantages and disadvantages of the existing method. The recall of the GAN-augmented model (0.95) indicates that synthetic data is especially useful in training models to ignore more surface-level mistakes such as typing mistakes and pronunciation differences. Nevertheless, the decline in performance which is found in the case of missing data is an important area that needs to be improved. Generative models should be taken further to produce not only noisy records, but incomplete ones, representing the reality of patients that may not have a fixed address or a fixed telephone number.

The results demonstrate that researchers do not necessarily require access to sensitive real-world data to make meaningful progress, as synthetic data can be used to generate models with high F1-scores. The implications of this on the pace of innovation are enormous. Developers can reduce months of time spent negotiating institutional review boards, data use agreements with using high-fidelity synthetic seeds to construct and test their systems. This will democratize the creation of more sophisticated deduplication tools, enabling smaller organizations to match large organizations that are able to access more data.

7. CONCLUSION

This study was able to prove that Generative AI is a useful instrument in enhancing accuracy of patient identity matching and deduplication. By creating 389 artificial patient records and applying them to different patterns of clerical noise, the study has demonstrated that models trained on AI-generated data exhibit higher resilience than conventional probabilistic techniques. The quantitative findings indicated that the precision and the recall were significantly enhanced and the hybrid AI model had reached an F1-score of 0.95. The privacy obstacle that usually impedes healthcare informatics research was overcome by the use of synthetic data, which offered a good and ethically justified alternative to algorithm development. The conclusion of the results displayed by the contour plots and multi-line graphs proved that the AI-enhanced model does not deteriorate even when the noise conditions are under high-stress. This consistency is important in ensuring a consistent primary record within the hospital systems. Although the model worked outstandingly well in dealing with typographical and phonetic mistakes, the paper also established that lack of data fields is one of the major challenges to identity resolution. Finally, this analysis demonstrates that synthetic data is not simply a substitute of real data but a better training resource, which can be customized to mimic the particular tasks of clinical data entry. The next step in

the advances of the given research is the development of generative models to work with more complicated, non-demographic data points. Although this research was aimed at identity markers, longitudinal clinical information, e.g. pattern of chronic conditions as well as medication history would offer more levels of authentication to identity matching. As an example, two records which are similar in their names could be confirmed as a match further when they have a rare diagnosis and a certain treatment history.

8. REFERENCES

[1] I. Goodfellow, J. Pouget-Abadie, M. Mirza, et al., "Generative adversarial networks," *Communications of the ACM*, vol. 63, no. 11, pp. 139–144, 2020. https://doi.org/10.1145/3422622

[2] J. Jordon, L. Szpruch, F. Houssiau, et al., "Synthetic Data–what, why and how?," *arXiv preprint*, 2022. https://doi.org/10.48550/arXiv.2205.03257

[3] M. Frid-Adar, I. Diamant, E. Klang, et al., "GAN-based synthetic medical image augmentation for increased CNN performance in liver lesion classification," *Neurocomputing*, vol. 321, pp. 321–331, 2018. https://doi.org/10.1016/j.neucom.2018.09.013

[4] I. Wolterink, T. Leiner, M. A. Viergever, et al., "Generative adversarial networks for noise reduction in low-dose CT," *IEEE Transactions on Medical Imaging*, vol. 36, no. 12, pp. 2536–2545, 2017. https://doi.org/10.1109/TMI.2017.2708987

[5] Q. Yang, P. Yan, Y. Zhang, et al., "Low-Dose CT Image Denoising Using a Generative Adversarial Network With Wasserstein Distance and Perceptual Loss," *IEEE Transactions on Medical Imaging*, vol. 37, no. 6, pp. 1348–1357, 2018. https://doi.org/10.1109/TMI.2018.2827462

[6] H. Ali, M. R. Biswas, F. Mohsen, et al., "The role of generative adversarial networks in brain MRI: a scoping review," *Insights into Imaging*, vol. 13, no. 1, 2022. https://doi.org/10.1186/s13244-022-01237-0

[7] E. Jung, M. Luna, and S. H. Park, "Conditional GAN with 3D discriminator for MRI generation of Alzheimer's disease progression," *Pattern Recognition*, vol. 133, 2023. https://doi.org/10.1016/j.patcog.2022.109061

[8] K. Packhäuser, L. Folle, F. Thamm, and A. Maier, "Generation of Anonymous Chest Radiographs Using Latent Diffusion Models for Training Thoracic Abnormality Classification Systems," in *Proc. IEEE ISBI*, 2023, pp. 1–5. https://doi.org/10.1109/ISBI53787.2023.10230346

[9] P. Eigenschink, T. Reutterer, R. Vamosi, et al., "Deep Generative Models for Synthetic Data: A Survey," *IEEE Access*, vol. 11, pp. 47304–47320, 2023. https://doi.org/10.1109/ACCESS.2023.3275134

[10] W. A. C. Castañeda and P. Bertemes Filho, "Synthetic health data generation for enhancement of non-invasive diabetes AI-based prediction," 2023. https://doi.org/10.20944/preprints202308.1464.v1