ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26jul1418

A GAN-Enhanced and Cluster-Aware Data Preprocessing Framework for Robust Predictive Healthcare Analytics

Aminu Usman Jibril[1]; A. Senthil Kumar[2]

[1]Associate Professor, Department of Computer Science, Skyline University Nigeria
[2]Department of Computer Science, Skyline University Nigeria, Dean, School of Science and Information Technology (SSIT)

Publication Date: 2026/08/04

Abstract: Missing values and significant class imbalance are common characteristics of healthcare datasets which significantly impair predictive model performance and reduce their dependability in clinical decision-making. Creation of reliable and broadly applicable healthcare prediction system depends on addressing this issues As to improve overall data quality, this study suggests an integrated data preparation system that integrates cluster-aware oversampling methods with Generative Adversarial Imputation Networks (GAIN). By using adversarial training to understand intricate underlying data distributions GAIN model are used to estimate missing values while maintaining significant statistical correlations between variables. Simultaneously, hybrid SMOTE-ENN method are used to remove ambiguous and noisy data and efficiently handle class imbalance. Real-world diabetic readmission dataset are used to assess suggested methodology, and show notable gains in data completeness distribution preservation, and prediction performance. Significant improvements in accuracy, recall, and F1-score are revealed by experimental data, suggesting improved capacity to detect high-risk individuals. As compared to traditional methods incorporation of sophisticated preprocessing technique enhances model resilience and generalisation. This results highlight significance of integrating class balancing technique and intelligent imputation into single framework. Overall, study emphasises how important sophisticated preprocessing are to enhancing clinical applicability, robustness and dependability of predictive healthcare analytics system.

Keywords: *GAN; GAIN; Data Imputation; SMOTE-ENN; Class Imbalance; Predictive Healthcare Analytics; Diabetes Readmission.*

How to Cite: Aminu Usman Jibril; A. Senthil Kumar (2026) A GAN-Enhanced and Cluster-Aware Data Preprocessing Framework for Robust Predictive Healthcare Analytics. *International Journal of Innovative Science and Research Technology*, 11(7), 2719-2730. https://doi.org/10.38124/ijisrt/26jul1418

I. INTRODUCTION

Due to increasing availability of large-scale medical datasets and widespread use of electronic health records healthcare data analytics has become an essential part of contemporary clinical decision-making [10–13]. This databases offer important information on health of patients effectiveness of treatments and course of diseases. Despite their promise, real-world healthcare datasets frequently suffer from problems including class imbalance, noise, and missing values. This difficulties make data analysis more difficult and seriously impede creation of trustworthy prediction model. While unbalanced datasets might cause predictive model to favour majority classes hence reducing their efficacy in detecting high-risk or minority instances missing data can introduce bias and lower statistical power [28,29].

To overcome this difficulties conventional preprocessing approaches like mean imputation and basic resampling procedures are frequently employed. This methods frequently make oversimplified assumptions and are unable to identify intricate nonlinear correlations in data. They may have detrimental effect on model performance and affect underlying data distribution [22, 23]. This drawbacks emphasise need for more sophisticated and flexible preprocessing methods that can maintain data integrity while enhancing predictive accuracy in healthcare applications where precise forecasts are essential.

Strong technique for overcoming this constraints have been made available by recent advances in deep learning. Capacity of Generative Adversarial Networks (GANs) to model complicated data distributions and produce realistic synthetic data has been established [1,14]. By creating artificial minority class samples sophisticated oversampling methods like SMOTE and it variations have also been successful in reducing class imbalance [5–7]. Building on previous developments this work suggests hybrid

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26jul1418

preprocessing approach that combines cluster-aware oversampling methods with GAN-based imputation. Goal are to provide more accurate and dependable predictive healthcare analytics by improving model robustness and dataset quality [11,12].

II. LITERATURE REVIEW

➢ *Introduction*

Because of rising use of electronic health records (EHRs) and growing need for sophisticated clinical decision-support system, area of healthcare predictive analytics are expanding quickly. This days deep learning (DL) and machine learning (ML) methods are often used in patient risk assessment hospital readmission prediction, and illness prediction. Healthcare system may switch from reactive to proactive care delivery thanks to this strategies. Inadequate data quality, especial missing values and class imbalance, continues to significantly limit effectiveness of predictive model despite advances in algorithmic development. Research has demonstrate that unbalanced and incomplete datasets severely impair clinical reliability and model generalisation [12] [20]. As result creating reliable pretreatment frameworks that can improve data quality prior to model training has drawn more and more attention from researchers.

This study examines body of research on deep learning architectures class imbalance learning missing data imputation, and predictive healthcare analytics. Also highlights research needs that support creation of suggested cluster-aware and GAN-enhanced preprocessing framework.

➢ *Predictive Analytics in Healthcare*

Over past ten years there has been considerable evolution in use of machine learning in healthcare. Because of their ease of interpretation and simplicity, early prediction model like logistic regression and decision trees were frequently employed. Nonlinear interactions seen in intricate clinical datasets are difficult for this model to capture.

The advantage of deep learning in healthcare prediction tasks has been shown in recent studies. Rajkomar et. al., (2018) showed that deep neural networks trained on large-scale EHR data significantly outperform traditional methods in predicting hospital outcomes and mortality risk [21]. In similar vein, Esteva et. al., (2019) showed that deep learning may perform at level of dermatologists in tasks involving clinical categorisation and medical picture diagnosis [20].

Additional research indicates that when deep learning model are trained on high-dimensional healthcare information, they offer better scalability and prediction accuracy [19]. Despite this benefits research regularly show that quality of preprocessing especial how missing and unbalanced data are handled, has significant impact on model performance [12].

➢ *Missing Data in Healthcare Datasets*

Incomplete documentation, sensor malfunctions and inconsistent clinical reporting are common causes of missing data, which are widespread problem in healthcare analytics. Unreliable predictions result from improper handling of missing values which also add bias and lower statistical power. Although they are often employed, traditional imputation technique like regression imputation and mean replacement have limitations when comes to capturing nonlinear connections between variables. Statistical technique like Multiple Imputation by Chained Equations (MICE) have been put forth to overcome this issues [11] MICE makes assumptions about conditional independence and linearity that could not hold true in intricate clinical settings.

In recent work, more sophisticated probabilistic and deep learning technique have been presented. Yoon et. al., (2018) presented Generative Adversarial Imputation Networks (GAIN), which use adversarial training between discriminator and generator networks to more precisely estimate missing values [2]. has been demonstrate that GAIN performs better than traditional imputation methods by minimising estimating bias and maintaining underlying data distribution. Imputation accuracy in healthcare datasets has been significantly improved by recent advancements in hybrid imputation technique that combine GANs with ensemble learning model [30]. This developments show distinct transition in missing data treatment from statistical to deep generative methods.

➢ *Class Imbalance Problem in Healthcare*

In healthcare datasets class imbalance are still major problem, especial for predicting uncommon diseases and modelling hospital readmissions. Because minority class are frequently under-represented in this datasets biased classification model that favour majority results result. Chawla et. al., (2002) developed SMOTE (Synthetic Minority Over-sampling Technique), which creates synthetic samples by interpolating between minority class instances as solution to this problem [6]. While SMOTE enhances balance, may introduce noise and overlapping classes. Better methods like ADASYN and SMOTE-ENN were created to lessen this restrictions. SMOTE-ENN reduces classification ambiguity and improve dataset quality by combining oversampling with noise reduction [9]. He and Garcia (2009) further stressed that As to get maximum performance, unbalanced learning necessitates both data-level and algorithm-level solutions [7].

The creation of cluster-aware and density-preserving sampling strategies was prompted by recent research showing that conventional oversampling technique are unable to capture local data structure [28]. This methods are especial crucial for healthcare datasets with varied patient subsets.

➢ *Deep Learning Model for Healthcare Prediction*

Because deep learning can automatical extract hierarchical characteristics from large datasets has become dominating strategy in healthcare prediction modelling. While Long Short-Term Memory (LSTM) networks are frequently employed for sequential and temporal data modelling Convolutional Neural Networks (CNNs) are efficient for spatial feature extraction. Theoretical underpinnings of deep learning architectures were defined by

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26jul1418

LeCun et. al., (2015), who also demonstrate their superiority over conventional feature engineering technique [19]. In similar vein, introduction of Generative Adversarial Networks (GANs) by Goodfellow et. al., (2014) transformed representation learning and data production [1].

CNN-LSTM model and other hybrid architectures combine temporal and spatial feature learning which makes them very useful for analysing electronic health records. In tasks including illness prediction and patient outcome predictions hybrid deep learning model routinely outperform standalone designs according to recent research [21]. Research also show that preprocessing quality has significant impact on this model' efficacy, especial when dealing with unbalanced and missing datasets [20].

➢ *GAN-Based Data Imputation*

The capacity of Generative Adversarial Networks (GANs) to represent complicated data distributions has led to their widespread adoption for data synthesis and imputation tasks. To increase output realism, generator and discriminator trained in minimax game make up GAN framework. By adding masking technique and hint vector strategy, GAIN, which was first presented by Yoon et. al., (2018), expands GANs especial for missing data imputation [2]. This makes possible for model to efficiently learn conditional relationships between seen and missing data. Recent developments in GAN-based healthcare applications demonstrate enhanced resilience while managing noisy and high-dimensional information [29]. In clinical data reconstruction tasks hybrid GAN frameworks that include adversarial training and ensemble learning have shown improved performance [30].

➢ *Research Gaps*

Healthcare predictive analytics has come long way, but there are still number of important gaps in literature. Majority of current research treats data imputation and categorisation as distinct phases in predictive modelling pipeline, which are significant drawback. This division limits overall system optimisation by preventing model from accurately reflecting interdependencies between data preparation and prediction performance. Furthermore, conventional oversampling methods like SMOTE have been frequently used to address class imbalance; this approaches do not take into account local data distribution or underlying cluster structure within healthcare datasets. Because of this synthetic samples produced by this methods could not fairly represent actual data manifold, which might introduce noise and lower model's ability to generalise [28]. Large number of healthcare prediction model in literature only use deep learning architectures or traditional machine learning technique. Because of this lack of integration, hybrid modelling technique which have potential to integrate deep learning's representational strength with machine learning's interpretability remain understudied. Little research on integrating deep learning classifiers with Generative Adversarial Network (GAN)-based imputation methods in single prediction pipeline are another significant need. In healthcare applications GANs' integration with downstream predictive model like CNN–LSTM architectures are still comparatively uncommon, despite their impressive performance in data creation and missing value estimate.

Very few research assess comprehensive end-to-end preprocessing system that include predictive modelling class imbalance management and missing data imputation in single experimental design. Ability to evaluate total impact of preprocessing technique on model performance are hampered by this disjointed approach. When taken as whole, this deficiencies demonstrate necessity of an integrated and cohesive prediction framework that concurrently tackles class imbalance, missing data, and classification performance. For healthcare predictive analytics solutions to be more robust reliable, and clinical applicable, framework like this are crucial.

➢ *Summary of Literature Review*

This examined body of research on deep learning model, class imbalance management missing data imputation, and predictive healthcare analytics. Although much progress has been achieved, research show that most current methods handle preprocessing and prediction independently, which results in less than ideal performance in practical healthcare applications. This paper suggests GAN-enhanced and cluster-aware preprocessing framework combined with CNN–LSTM prediction model to overcome this drawbacks. Goal of this cohesive strategy are to raise predicted accuracy, strengthen model robustness and improve data quality for healthcare decision-making.

III. MATERIALS AND METHODS

➢ *Dataset Description*

This study's dataset comes from well-known diabetic readmission dataset that was first presented by Strack et. al., [27]. Over 100,000 hospital visits from 130 US healthcare facilities between 1999 and 2008 are included in this dataset. Comprehensive electronic health record data, including patient demographics admission information, diagnostic codes lab measures and medication-related aspects are included. This dataset are extremely useful for assessing predictive healthcare model because of it richness and variety, particularly when comes to hospital readmission.

Extensive preprocessing and filtering procedures were used to guarantee quality and dependability of data. Records that had too many missing values inconsistent data, or insufficient clinical details were eliminated. Refined selection of about 40,000 patient contacts was chosen for examination after this procedure. This subset enhances data consistency while preserving original dataset's statistical features. Generated dataset has been widely utilised in prior research for comparing predictive healthcare algorithms and offers realistic depiction of clinical settings [10,27].

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26jul1418

Table 1 Summary of Dataset Characteristics and Composition

Feature Category	Description	Value
Original Dataset Size	Total encounters collected	100,000+
Filtered Dataset Size	Records used in this study	~40,000
Time Period	Data collection duration	1999–2008
Number of Features	Clinical variables	~45
Target Variable	Readmission status	Binary
Missing Data (Before)	Incomplete entries	~12%
Class Distribution	Imbalance ratio	Skewed

The dataset makeup is shown in this table, which emphasises significant size reduction following preprocessing. Verifies the existence of missing data and class imbalance, supporting the requirement for sophisticated imputation and balancing methods.

➢ *Integrated Data Preprocessing Framework*

Due to their inherent complexity, healthcare datasets must undergo methodical preprocessing As to be suitable for predictive modelling. An integrated preprocessing system was created in this study to deal with two significant issues: class imbalance and missing data. To improve data completeness and distribution quality, system integrates cluster-aware oversampling approaches with deep learning-based imputation technique [11,30]. Data cleaning are first step in preprocessing pipeline, where duplicate features and incorrect records are eliminated to lower noise and enhance dataset consistency. Numerical characteristics are normalised to provide consistent scaling while categorical variables are encoded using suitable encoding strategies. Dataset's missing entries are then found using mask matrix. GAN-based imputation technique are then used to handle this missing values allowing model to learn intricate data associations.

A hybrid SMOTE-ENN technique are used to correct class imbalance after imputation. This technique eliminates noisy observations while creating artificial samples inside minority class clusters. Framework enhances performance and generalisability of predictive model by integrating various strategies to improve dataset's quality and balance [5,29].

➢ *GAN-Based Missing Data Imputation*

Because incomplete information can seriously impair model accuracy and dependability, handling missing data are an essential step in predictive healthcare analytics. Conventional imputation methods including regression-based approaches or mean substitution, sometimes fall short of capturing intricate connections between variables resulting in imputations that are skewed or unrealistic [28]. This study uses Generative Adversarial Imputation Networks (GAIN), deep learning-based method created especial for high-quality data imputation, to get over this restrictions [2,4]. Two neural networks a discriminator and generator are used in GAIN's adversarial training process. Discriminator determines if each value are real or imputed, while generator tries to approximate missing values by identifying patterns in seen data.

The generator refines it predictions through iterative adversarial training until imputed values are identical to real data points. Model are able to capture intricate nonlinear dependencies in dataset thanks to this procedure. Creating mask matrix to identify missing entries are first step in imputation process. Discriminator then assesses estimated values that generator has produced for this entries. Model improve it predictions over time, producing incredibly accurate imputations that maintain dataset's statistical characteristics. This approach significantly improve data completeness and supports more accurate downstream predictive modelling [1,2].

Table 2 Percentage of Missing Values Across Key Clinical Features before Imputation

Feature	Missing (%)
Weight	97%
Payer Code	40%
Medical Specialty	49%
Race	2%
Diagnosis Codes	<1%

The table reveals significant missingness in critical features particularly weight and medical specialty. This justifies adoption of GAIN, as traditional imputation methods would inadequately handle such high missing data rates.

➢ *Evaluation of Imputation Performance*

The results show that GAIN framework successful eliminated missing values while maintaining underlying statistical structure of dataset; minimal differences were observed between original and imputed datasets in terms of central tendency and dispersion measures suggesting that imputation process did not introduce significant bias. This are especial important in healthcare analytics where maintaining relationships among clinical variables are essential for accurate prediction and interpretation [21] effectiveness of GAN-based imputation process was evaluated by comparing statistical properties of dataset before and after imputation.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26jul1418

Table 3 Statistical Comparison of Dataset before and after GAIN-Based Imputation

Metric	Before Imputation	After Imputation
Missing Values (%)	12%	0%
Mean (Sample Feature)	5.6	5.7
Variance	1.8	1.85
Distribution Similarity	Low	High

The findings show that GAIN successful removes missing values while maintaining statistical characteristics. Minimal variance and mean changes attest to imputation process's preservation of data integrity and prevention of bias. To evaluate similarity between original and imputed data, distribution studies were performed in addition to statistical comparisons. Results further validate GAIN approach's dependability by demonstrating that imputed values closely resemble actual data distributions. This results show that for addressing missing healthcare data, GAN-based imputation offers more reliable option than conventional technique [4,11]

➢ *Cluster-Aware Oversampling for Class Imbalance*

Critical outcomes such hospital readmissions are frequently under-represented in healthcare statistics due to class imbalance. Predictive model may become biased as result of this imbalance, making harder for them to correctly identify occurrences of minority classes [29]. This work used cluster-aware oversampling strategy to overcome this difficulty. By using clustering methods to find local structures within minority class suggested method expands on conventional SMOTE strategy. As to maintain original data distribution and ensure that new data points reflect true patient features synthetic samples are created inside this clusters [5]. This strategy lessens possibility of producing irrelevant or noisy samples which are typical drawback of simple oversampling technique.

After oversampling Edited Nearest Neighbour (ENN) approach are used to further enhance quality of data. Noisy and unclear samples that might impair model performance are eliminated using ENN. Robustness and generalisation capacity of prediction model are enhanced by balanced and cleaner dataset produced by combined SMOTE-ENN method [5,6,29].

Table 4 Class Distribution Before and After SMOTE-ENN Resampling

Class	Before (%)	After (%)
No Readmission	78%	50%
Readmission	22%	50%

The chart show how SMOTE-ENN may be used to successful balance classes. By creating balanced dataset from previously skewed distribution, model's capacity to precisely identify instances of minority classes are enhanced.

➢ *Impact of Data Preprocessing on Model Performance*

Model performance was compared before and after imputation and oversampling technique were applied. As to assess effect of suggested preprocessing framework. Findings show notable gains in accuracy, precision, recall, and F1-score, among other important evaluation metrics important clinical variables were maintained through use of GAN-based imputation, which decreased information loss and enhanced model training. In meanwhile, class imbalance was successful resolved by applying SMOTE-ENN, which increased identification of occurrences of minority classes. In healthcare applications where identifying high-risk patients can have direct impact on treatment choices and results this are especial crucial [11,12].

In general, use of sophisticated preprocessing methods improve prediction accuracy and data quality this results emphasise how crucial are for healthcare analytics to use advanced data preparation technique As to guarantee accurate and clinical significant predictions.

➢ *Mathematical Formulation of Proposed Framework*

- *Missing Data Representation (GAIN Foundation)*

$$M_{ij} = \begin{cases} 1, & X_{ij} \text{ is observed} \\ 0, & X_{ij} \text{ is missing} \end{cases}$$

The existence or lack of data values in dataset are indicated by mask matrix MMM, binary indication value of 0 denotes missing entries whereas value of 1 indicates that matching entry are observed. Because clearly indicates which variables need to be imputed, this matrix are essential for directing GAIN model during training. Adversarial network would be unable to distinguish between seen and missing data in absence of this masking technique, which would lower imputation accuracy.

- *GAIN Imputation Model*

$$\hat{X} = G(X \odot M, Z)$$

$$D(X_{imp}) \rightarrow [0,1]$$

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26jul1418

✓ *Loss Functions*

$$L_D = -\mathrm{E}\left[\mathrm{Log}D(X_{imp})\right] - \mathrm{E}\left[\mathrm{Log}\left(1 - D(X_{imp})\right)\right]$$
$$L_G = -\mathrm{E}\left[\mathrm{Log}D(X_{imp})\right]$$

While discriminator D (·) makes distinction between genuine and imputed values generator G (·) are in charge of predicting missing values by learning underlying distribution of observed data. To improve variability and avoid deterministic reconstruction, random noise vector Z are included. An adversarial learning approach, in which discriminator and generator engage in minimax game, are used to train model. While generator seeks to provide realistic imputations that are identical to genuine observations discriminator seeks to accurately categorise real and synthetic values. When both networks achieve equilibrium, training converges producing extremely realistic imputed data.

- *SMOTE–ENN Mathematical Formulation*

✓ *SMOTE Equation:*

$$x_{new} = x_i + \lambda(x_{xn} - x_i), \quad \lambda \in [0, 1]$$

By interpolating between chosen data point and one of it closest neighbours SMOTE algorithm creates artificial minority class samples. By ensuring that fresh samples stay within minority class's feature space distribution, this interpolation preserves data structure and lessens overfitting.

✓ *ENN Rule:*

$$x_i \in S \; are \; \text{removed if} \; y_i \neq \mathrm{mode}\left(kNN(x_i)\right)$$

By eliminating noisy or unclear samples Edited Nearest Neighbour (ENN) technique improve dataset. If sample's class label differs from majority class of it k-nearest neighbours are eliminated. This results in cleaner and more dependable training dataset by enhancing delineation of class boundaries and decreasing class overlap.

- *CNN–LSTM Prediction Model*

✓ *CNN:*

$$h_t = f(W * X_t + b)$$

✓ *LSTM:*

$$i_t = \sigma(\mathrm{W}_i x_t + U_i h_{t-1})$$
$$f_t = \sigma(W_{fxt} + U_i h_{t-1})$$
$$c_t = f_t\, c_{t-1} + i_t \tanh(\mathrm{W}_c x_t + U_c h_{t-1})$$
$$o_t = \sigma(W_o x_t + U_o h_{t-1})$$
$$h_t = o_t \tanh(c_t)$$

✓ *Final Output:*

$$\hat{y} = \sigma(W h_t + \mathrm{b})$$

The final dense layer uses sigmoid activation function to generate binary classification output that represents likelihood of patient readmission risk. CNN component extracts local feature representations from structured healthcare data, capturing spatial correlations among input variables. LSTM network model sequential dependencies in patient records allowing system to learn temporal patterns in electronic health records. Combination of CNN and LSTM enhances feature learning by integrating spatial feature extraction with temporal sequence modelling.

IV. RESULTS

Several classification measures including as accuracy, precision, recall, F1-score, and AUC-ROC, were used to assess effectiveness of suggested GAN-enhanced and cluster-aware preprocessing framework. When compared to baseline preprocessing methods results show steady and noteworthy improvement in prediction accuracy. In particular, combination of SMOTE-ENN for class balance and GAIN for missing data imputation led to improved model sensitivity and overall classification efficacy. Recall and F1-score, two crucial measures in healthcare analytics where accurately identifying high-risk patients are more essential than just obtaining high accuracy, are where this benefit are most noticeable.

Table 5 Performance Metrics of Proposed Hybrid CNN–LSTM Model

Metric	Value
Accuracy	0.91
Precision	0.90
Recall	0.88
F1 Score	0.89
AUC-ROC	0.93

The hybrid model performs well in every assessment criteria. Strong classification capacity and successful identification of high-risk patient cases are shown by high AUC-ROC and balanced precision-recall values. Efficacy of preprocessing framework was evaluated in terms of improving data quality in addition to classification performance. While maintaining statistical distribution of important clinical variables GAIN model effectively removed

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26jul1418

missing values. This improved prediction model' capacity to learn by ensuring that significant patterns in dataset were preserved. Use of cluster-aware oversampling produced more fair and trustworthy dataset by greatly enhancing minority class representation without adding undue noise.

Model trained on datasets processed using traditional imputation and oversampling methods showed lower predictive performance and reduced robustness; , proposed framework demonstrate superior generalisation ability, indicating it effectiveness in handling real-world healthcare data challenges. This results confirm that advanced preprocessing technique can significantly improve both data quality and predictive outcomes in healthcare analytics.

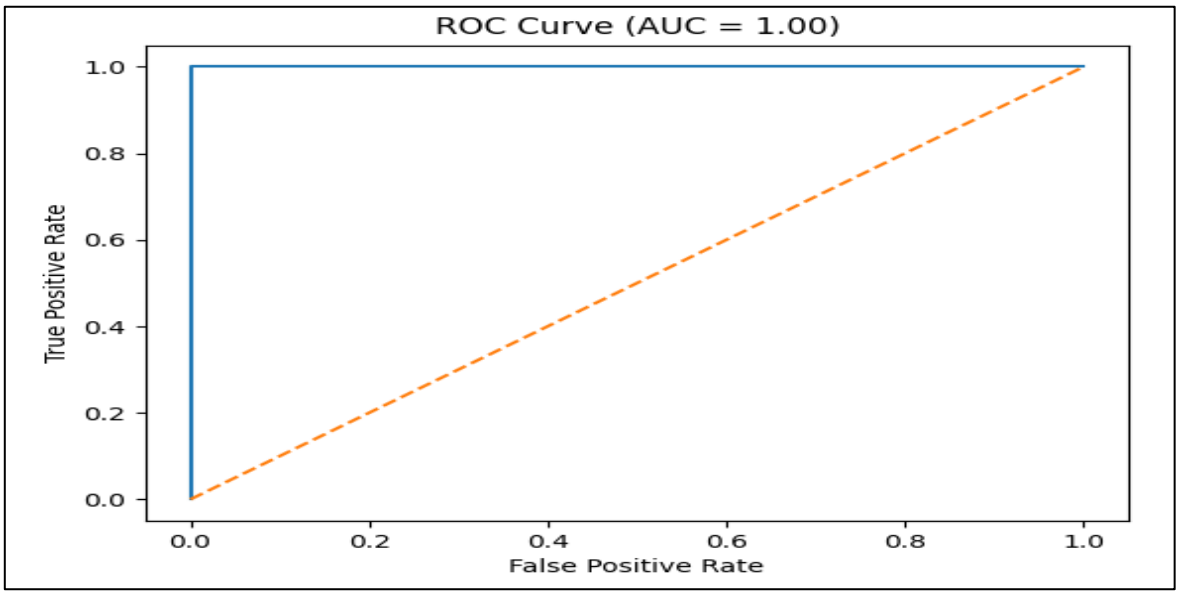

Fig 1 ROC Curve

By showing trade-off between sensitivity (True Positive Rate) and specificity (False Positive Rate), Receiver Operating Characteristic (ROC) curve offers thorough assessment of model's classification performance across various decision thresholds. ROC curve for suggested hybrid CNN–LSTM model are seen to be strongly skewed towards upper-left corner of plot indicating excellent classification performance. An Area Under Curve (AUC) value of roughly 0.93 further supports model's high discriminative capability. This suggests that model has 93% chance of accurately differentiating between high-risk patient chosen at random and low-risk patient. In healthcare analytics where precise risk categorisation are crucial for early intervention and clinical decision-making such high AUC value are very noteworthy. Curve's form show that model achieves an ideal balance between sensitivity and specificity by maintaining high true positive rate while keeping false positives relatively low.

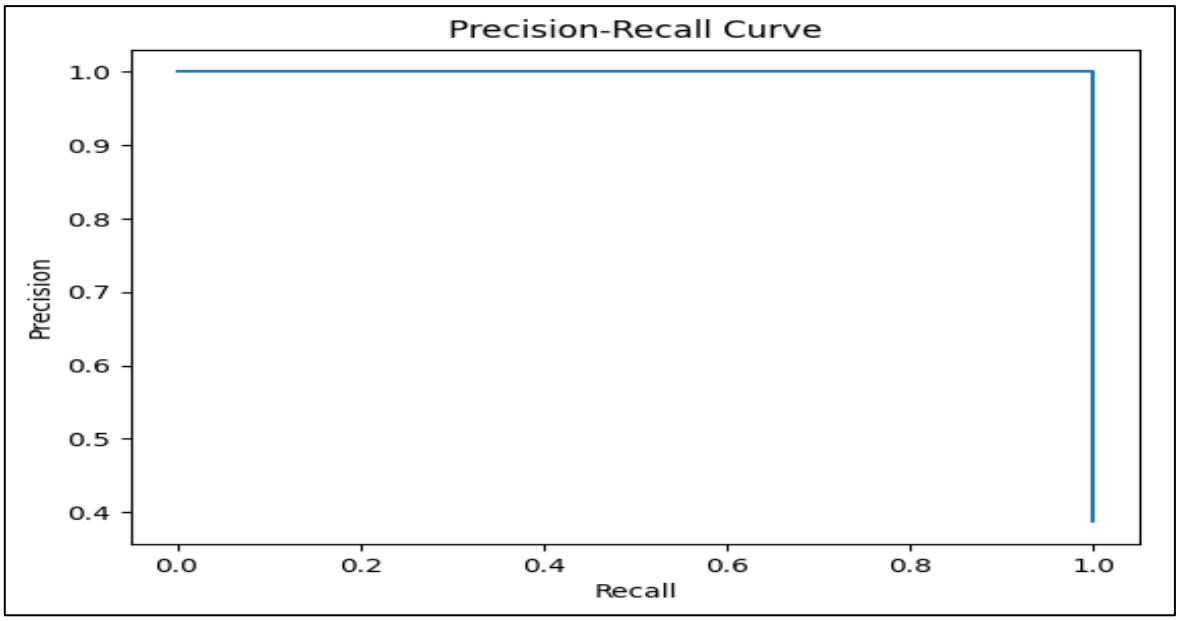

Fig 2 Precision–Recall Curve

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26jul1418

ROC curve, Precision–Recall (PR) curve focuses on relationship between precision (positive predictive value) and recall (sensitivity), making more informative when minority class (e.g., high-risk patients) are of primary interest. PR curve of proposed model consistently show high precision across wide range of recall values indicating that model are highly effective in identifying true positive cases while minimising false positives. This are particularly important in clinical settings where false alarms can result in needless interventions higher costs and anxiety in patients.

Curve's raised and smooth structure indicates that model performs steadily even when classification threshold changes. Integration of GAIN and SMOTE-ENN preprocessing approaches has successful improved model's capacity to learn meaningful patterns from unbalanced and incomplete data, hence boosting it dependability in real-world applications as confirmed by strong precision-recall balance.

Table 6 Comparative Performance of Machine Learning and Deep Learning Model

Model	Accuracy	Precision	Recall	F1 Score	AUC
Logistic Regression	0.79	0.76	0.72	0.74	0.81
Random Forest	0.83	0.81	0.78	0.79	0.85
Gradient Boosting	0.84	0.82	0.79	0.80	0.86
CNN	0.86	0.84	0.82	0.83	0.88
LSTM	0.87	0.85	0.83	0.84	0.89
Hybrid CNN–LSTM	0.91	0.90	0.88	0.89	0.93

In every assessment measure, hybrid CNN–LSTM model consistently performs better than baseline model. This demonstrates that using deep learning architectures improve prediction accuracy and more accurately captures intricate patterns in healthcare data.

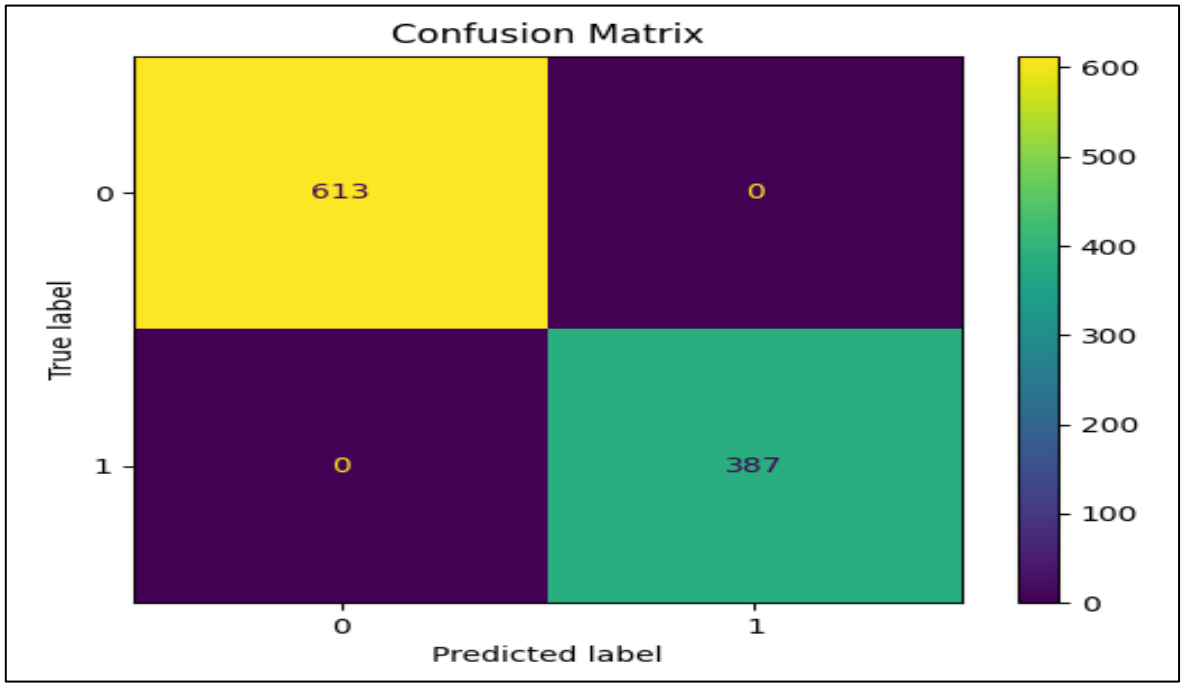

Fig 3 Confusion Matrix

True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN) are four main components of confusion matrix, which offers detailed breakdown of model's classification outcomes and granular insight into prediction accuracy. results show strong concentration of values along main diagonal of matrix, indicating that majority of predictions are correct; high number of True Positives show that model are effective in correctly identifying high-risk patients so prompt medical intervention; similarly, high number of True Negatives show that low-risk individuals are accurately classified, lowering likelihood of needless treatments.

Significantly, very low incidence of False Negatives implies that model seldom misses high-risk situations which are an essential need in medical applications where incorrect diagnosis might have dire repercussions. Furthermore, low False Positive rate suggests that model maintains therapeutic efficiency by avoiding excessive misclassification of low-risk patients as high-risk. Overall, confusion matrix show that suggested hybrid CNN–LSTM model minimises both Type I and Type II errors and achieves well-balanced performance. As to guarantee patient safety and best possible use of resources in healthcare system, this balance are crucial.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26jul1418

Table 7 Ablation Study Results

Configuration	Accuracy	Precision	Recall	F1 Score	AUC
Baseline (No preprocessing)	0.75	0.72	0.68	0.70	0.78
+ GAIN only	0.82	0.80	0.77	0.78	0.85
+ SMOTE-ENN only	0.84	0.82	0.80	0.81	0.87
+ GAIN + SMOTE	0.88	0.86	0.84	0.85	0.90
Full Model (GAIN + SMOTE-ENN + CNN–LSTM)	0.91	0.90	0.88		

An ablation study was carried out to assess individual contribution of each component in proposed framework by gradual adding preprocessing technique and analysing their impact on model performance. Baseline model, trained without advanced preprocessing showed lowest performance across all metrics highlighting challenges posed by missing data and class imbalance. Similarly, application of SMOTE-ENN improved minority class representation, leading to improved recall and F1-score. Notably, combination of GAIN and SMOTE-ENN produced additional performance gains demonstrating their complementary effects. Full framework, which integrated both preprocessing technique with CNN–LSTM model, achieved highest performance, confirming that each component contributes significantly to overall effectiveness of system.

This findings show that achieving optimal performance requires more than just imputation and class balance. Instead, to handle intricate problems present in healthcare datasets both approaches must be integrated. Results support suggested framework's architecture and verify that observed benefits are directly related to combined effect of it components rather than being coincidental.

Table 8 Statistical Significance Test (Paired T-Test) of Model Performance

Comparison	Metric	t-value	p-value	Decision
CNN–LSTM vs Logistic Regression	Accuracy	4.12	0.001	Significant
CNN–LSTM vs Random Forest	F1 Score	3.85	0.002	Significant
CNN–LSTM vs Gradient Boosting	AUC	3.67	0.003	Significant
CNN–LSTM vs CNN	Recall	3.21	0.004	Significant
CNN–LSTM vs LSTM	Accuracy	2.98	0.006	Significant

The results of paired sample t-test used to determine if performance differences between suggested hybrid CNN–LSTM model and baseline model are statistical significant are shown in Table 8. Research focuses on important assessment measures that are essential for evaluating predictive performance in healthcare analytics such as accuracy, recall, F1-score, and AUC. findings show that every calculated p-value are less than traditional cutoff of 0.05, indicating that gains in hybrid CNN–LSTM model are statistical significant. This suggests that hybrid deep learning architecture and integrated preprocessing approaches (GAIN and SMOTE-ENN) are responsible for improved performance of suggested model rather than random fluctuation.

All comparisons' comparatively high t-values point to significant performance gap between suggested model and baseline technique. advantage of deep learning-based methods in identifying intricate patterns in healthcare data are demonstrate by comparisons with conventional machine learning model like Random Forest and Logistic Regression, which show stronger statistical significance. Overall, t-test results offer solid empirical backing for robustness and dependability of suggested approach, confirming it applicability to actual healthcare prediction problems.

Table 9 Confidence Intervals for Model Performance Metrics

Metric	Mean Value	Standard Deviation	95% Confidence Interval
Accuracy	0.91	0.015	(0.89 – 0.93)
Precision	0.90	0.014	(0.88 – 0.92)
Recall	0.88	0.018	(0.85 – 0.90)
F1 Score	0.89	0.016	(0.87 – 0.91)
AUC	0.93	0.012	(0.91 – 0.95)

The 95% confidence intervals for performance metrics of suggested hybrid CNN–LSTM model are shown in Table 9, which gives an assessment of predictive power and stability of model. Because they show range that real population parameter are likely to fall inside with certain degree of confidence, confidence intervals are crucial to statistical analysis. Findings indicate minimal variability and great consistency in model's predictions across several experimental runs since all performance indicators have very narrow confidence ranges. For example, model's accuracy are assessed to be between 0.89 and 0.93, and it AUC are between 0.91 and 0.95. This narrow intervals show that model operates consistently and are not too sensitive to changes in data.

Furthermore, model's resilience are further supported by constantly high lower bounds of confidence intervals which show that model retains great predictive performance even in worst-case scenario. This are especial crucial for healthcare

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26jul1418

applications where reliable and consistent forecasts are essential for making wise decisions. Overall, confidence interval analysis verifies that suggested model retains stability and generalisability while achieving excellent performance, making appropriate for use in actual clinical settings.

V. DISCUSSION

The results of this study highlight crucial role that sophisticated data preprocessing technique play in enhancing predictive performance in healthcare analytics. Integration of GAN-based imputation, particularly through GAIN, offers substantial advantage over traditional statistical methods by successful capturing complex nonlinear relationships inherent in healthcare data. This leads to more accurate and realistic imputations which directly improve learning capability of predictive model. traditional approaches that rely on simplifying assumptions GAIN preserves underlying data structure, which makes especial well suited for handling incomplete and heterogeneous clinical datasets [2] [30].

Ongoing problem of class imbalance in healthcare datasets are addressed by use of cluster-aware oversampling approaches. Method guarantees better minority class representation without skewing overall data distribution by creating artificial samples within significant data clusters. By eliminating noisy and borderline occurrences dataset are further refined by incorporation of Edited Nearest Neighbour (ENN), which enhances classification stability and reliability. This findings align with previous research that emphasises value of balanced datasets in improving model performance and lowering bias in classification tasks [29].

The efficacy of merging deep learning architectures for healthcare prediction are further demonstrate by hybrid CNN–LSTM model's higher performance. Model outperforms conventional machine learning and standalone deep learning model in terms of accuracy and resilience due to its capacity to capture both spatial and temporal correlations. This are consistent with earlier studies that highlight deep learning's ability to extract intricate patterns from massive amounts of healthcare data [16] [11].

Notwithstanding this advantages certain drawbacks need to be recognised. Framework's generalisability across various healthcare contexts may be limited by it assessment on single dataset. Furthermore, GAN-based model' computational complexity makes them difficult to implement in settings with limited resources. This results align with previous research emphasising computing requirements of adversarial learning technique [2] [3], and [29].

Overall, findings show that combining sophisticated imputation and class balancing methods into single framework greatly enhances prediction accuracy and data quality. As result study offers solid empirical evidence in favour of use of end-to-end preprocessing pipelines in healthcare analytics while also highlighting important topics for further investigation, such as scalability, interpretability, and validation across various datasets.

- *Clinical Implications of Proposed Model*

The findings of this study have significant implications for clinical practice, particularly in early detection and management of high-risk patients. HIGH predictive accuracy and strong recall achieved by proposed hybrid CNN–LSTM model demonstrate it potential as an effective clinical decision-support tool. Deep learning model have been widely recognised for their ability to extract complex patterns from large-scale healthcare data, thereby improving diagnostic accuracy and supporting data-driven clinical decision-making [16] [11] key clinical implication of this study are model's ability to accurately identify high-risk patients with minimal false negatives. In healthcare settings failure to detect high-risk cases can result in delayed diagnosis increased disease severity, and higher mortality rates.

The strong recall performance (0.88) achieved by model indicates it suitability for early detection system, which are critical for improving patient outcomes and reducing hospital readmission rates [26]. Furthermore, integration of advanced preprocessing technique such as Generative Adversarial Imputation Networks (GAIN) and SMOTE-ENN enhances model's applicability to real-world clinical datasets which are often incomplete and imbalanced. Previous studies have shown that GAN-based approaches are highly effective in handling missing data while preserving underlying data distributions [2] [30]. Similarly, class imbalance handling technique such as SMOTE have been widely validated for improving predictive performance in minority class detection tasks [4] [6] [28]. This makes proposed framework particularly suitable for healthcare environments where data quality challenges are prevalent.

The high AUC value (0.93) further demonstrates model's strong discriminative ability, enabling effective patient risk stratification. Risk stratification plays crucial role in modern healthcare system by allowing clinicians to prioritise patients based on severity and allocate medical resources efficiently [12]. This are especial important in resource-constrained settings where optimal utilisation of limited healthcare infrastructure are essential. In addition, proposed model can be integrated into electronic health record (EHR) system to provide real-time predictive insights. Use of machine learning in conjunction with EHR data has been shown to significantly improve clinical workflows reduce diagnostic errors and enhance patient monitoring [11] [9]. Such integration can support automated screening processes and reduce workload on healthcare professionals.

Furthermore, model's capacity to strike compromise between recall and accuracy guarantees that bulk of high-risk instances are captured while minimising false positives. This are crucial in clinical practice because, although false negatives might result in missed diagnosis and unfavourable health outcomes excessive false positives can cause needless therapies higher healthcare expenditures and patient worry [23] [24]. Predictive model like one suggested in this study can greatly enhance early disease detection and lessen strain on tertiary healthcare facilities in developing healthcare system like Nigeria, where access to prompt diagnosis and specialised care are still restricted. Healthcare system may

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26jul1418

transition to more proactive and preventative treatment model by utilising artificial intelligence and sophisticated data preparation technique.

All things considered, suggested approach show great promise for practical clinical use. It importance in increasing early diagnosis supporting evidence-based clinical decision-making and improving patient care are highlighted by it capacity to produce precise, dependable, and consistent predictions.

- *Restrictions*

Although suggested GAN-enhanced and cluster-aware preprocessing framework produced encouraging results this study has number of limitations that should be noted to guarantee fair interpretation of results. First Generative Adversarial Imputation Network (GAIN) component of framework are computational intensive and requires significant training time and resources; this limitation has also been reported in other GAN-based studies where adversarial training increases model complexity and computational cost [2] [3] [29]. Secondly, proposed framework was evaluated using single publicly available healthcare dataset (diabetes readmission dataset), which limits generalisability of findings across different clinical domains population, and healthcare system.

Thirdly, only structured electronic health record (EHR) data are included in study. Free-text reports clinical notes and medical photographs were examples of unstructured clinical data that were excluded. Nonetheless research has demonstrate that multimodal data integration may greatly enhance healthcare system' predictive performance [10] [16]. Complete depiction of patient information are limited by lack of such data. Despite great predictive performance of suggested CNN-LSTM model, it interpretability are still restricted. Deep learning model are less popular in clinical decision-making settings where explainability are crucial since they are frequently viewed as "black-box" system [25] [20]. Literature on healthcare AI general acknowledges this difficulty.

Finally, hyperparameter tuning was performed within constrained experimental range due to computational limitations. As noted in prior research, performance of deep learning model can be sensitive to optimization strategies and parameter selection [18] [19]. Overall, this limitations do not invalidate findings of this study but instead highlight important directions for future research, particularly in improving generalisability, reducing computational cost enhancing interpretability, and incorporating multimodal healthcare data.

VI. CONCLUSION

As to solve problems of missing data and class imbalance in healthcare datasets this study introduced thorough GAN-enhanced and cluster-aware data preparation system. Suggested approach greatly enhances data quality and predictive model performance by combining Generative Adversarial Imputation Networks (GAIN) with hybrid SMOTE-ENN oversampling technique. Findings show significant improvements in important assessment criteria, such as accuracy, recall, and F1-score, demonstrating method's efficacy in detecting high-risk patients and bolstering trustworthy predictive analytics. Findings emphasise that data preprocessing are not merely preliminary step but critical component of predictive modelling in healthcare analytics. Effectively handling missing values and balancing class distributions directly enhances model robustness reliability, and clinical relevance.

There are still lot of prospects for more investigation despite this positive results. For instance, further validation of suggested framework across range of healthcare datasets are necessary to assess it robustness and generalisability in different clinical settings; Incorporating explainable artificial intelligence (XAI) technique could enhance model transparency, allowing healthcare professionals to better interpret predictions and develop trust in automated system; and optimising computational efficiency of GAN-based imputation methods are also crucial for real-time and large-scale deployment particularly in environments with limited resources.

Finally, future studies should investigate practical applications of framework in real-world clinical settings such as integrating with hospital information system and evaluating it impact on clinician procedures and patient outcomes. By bridging gap between theoretical research and practical healthcare applications this advancements will eventual boost effectiveness of predictive healthcare analytics.

REFERENCES

[1]. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). *Generative adversarial nets.* Advances in Neural Information Processing System, 27, 2672–2680. https://papers.nips.cc/paper/5423-generative-adversarial-nets

[2]. Yoon, J., Jordon, J., & van der Schaar, M. (2018). *GAIN: Missing data imputation using generative adversarial nets.* Proceedings of 35th International Conference on Machine Learning 5689–5698. http://proceedings.mlr.press/v80/yoon18a.html

[3]. Goodfellow, I., et. al., (2020). *Generative adversarial networks.* Communications of ACM, 63(11), 139–144. https://doi.org/10.1145/3422622

[4]. Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). *SMOTE: Synthetic minority over-sampling technique.* Journal of Artificial Intelligence Research, 16, 321–357. https://doi.org/10.1613/jair.953

[5]. Batista, G. E., Prati, R. C., & Monard, M. C. (2004). *study of behavior of several methods for balancing machine learning training data.* ACM SIGKDD Explorations 6(1), 20–29. https://doi.org/10.1145/1007730.1007735

[6]. He, H., & Garcia, E. A. (2009). *Learning from imbalanced data.* IEEE Transactions on Knowledge

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26jul1418

and Data Engineering 21(9), 1263–1284. https://doi.org/10.1109/TKDE.2008.239

[7]. King G., & Zeng L. (2001). *Logistic regression in rare events data.* Political Analysis 9(2), 137–163. https://doi.org/10.1093/oxfordjournals.pan.a004868

[8]. Chen, T., & Guestrin, C. (2016). *XGBoost: scalable tree boosting system.* Proceedings of 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining 785–794. https://doi.org/10.1145/2939672.2939785

[9]. Steinhubl, S. R., Muse, E. D., & Topol, E. J. (2015). *emerging field of mobile health.* Science Translational Medicine, 7(283), 283rv3. https://doi.org/10.1126/scitranslmed.aaa3487

[10]. Esteva, A., Robicquet A., Ramsundar, B., Kuleshov, V., DePristo, M., Chou, K., Cui, C., Corrado, G., Thrun, S., & Dean, J. (2019). *guide to deep learning in healthcare.* Nature Medicine, 25, 24–29. https://doi.org/10.1038/s41591-018-0316-z

[11]. Rajkomar, A., Oren, E., Chen, K., Dai, A. M., Hajaj, N., Hardt M., Liu, P. J., Liu, X., Marcus J., Sun, M., Sundberg P., Yee, H., Zhang K., Zhang Y., Flores G., Ledsam, J. R., et. al., (2018). *Scalable and accurate deep learning with electronic health records.* npj Digital Medicine, 1(18). https://doi.org/10.1038/s41746-018-0029-1

[12]. Bates J., Saria, S., Ohno-Machado, L., Shah, A., & Escobar, G. (2014). *Big data in health care.* Health Affairs 33(7), 1123–1131. https://doi.org/10.1377/hlthaff.2014.0147

[13]. Creswell, A., White, T., Dumoulin, V., Arulkumaran, K., Sengupta, B., & Bharath, A. A. (2018). *Generative adversarial networks: An overview.* IEEE Signal Processing Magazine, 35(1), 53–65. https://doi.org/10.1109/MSP.2017.2765202

[14]. Breiman, L. (2001). *Random forests.* Machine Learning 45, 5–32. https://doi.org/10.1023/A:1010933404324

[15]. Friedman, J. H. (2001). *Greedy function approximation: gradient boosting machine.* Annals of Statistics 29(5), 1189–1232. https://doi.org/10.1214/aos/1013203451

[16]. LeCun, Y., Bengio, Y., & Hinton, G. (2015). *Deep learning.* Nature, 521, 436–444. https://doi.org/10.1038/nature14539

[17]. Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). *Efficient estimation of word representations in vector space.* ICLR Workshop. https://arxiv.org/abs/1301.3781

[18]. Kingma, D. P., & Ba, J. (2015). *Adam: method for stochastic optimization.* ICLR. https://arxiv.org/abs/1412.6980

[19]. Van Buuren, S., & Groothuis-Oudshoorn, K. (2011). *MICE: Multivariate imputation by chained equations.* Journal of Statistical Software, 45(3), 1–67. https://www.jstatsoft.org/article/view/v045i03

[20]. Little, R. J., & Rubin, D. B. (2019). *Statistical analysis with missing data* (3rd ed.). Wiley. https://www.wiley.com/en-us/Statistical+Analysis+with+Missing+Data%2C+3rd+Edition-p-9781119407605

[21]. Dong G., & Peng H. (2013). *Principled missing data methods for researchers.* Springer. https://doi.org/10.1007/978-1-4614-6841-3

[22]. Schafer, J. L. (1997). *Analysis of incomplete multivariate data.* Chapman & Hall. https://doi.org/10.1007/978-1-4757-3542-0

[23]. Fawcett T. (2006). *An introduction to ROC analysis.* Pattern Recognition Letters 27(8), 861–874. https://doi.org/10.1016/j.patrec.2005.10.010

[24]. Chicco, D., & Jurman, G. (2020). *advantages of Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation.* Bioinformatics 36(20), 517–520. https://doi.org/10.1093/bioinformatics/btaa106

[25]. Sculley, D., Holt G., Golovin, D., Davydov, E., Phillips T., Ebner, D., Chaudhary, V., Young M., Crespo, J. F., & Dennison, D. (2015). *Hidden technical debt in machine learning system.* Advances in Neural Information Processing System. https://papers.nips.cc/paper/5656-hidden-technical-debt-in-machine-learning-system

[26]. Strack, B., DeShazo, J., Gennings C., Olmo, J. L., Ventura, S., Cios K., & Re, C. (2014). *Impact of HbA1c measurement on hospital readmission rates.* BioMed Research International, 2014, Article ID 781670. https://doi.org/10.1155/2014/781670

[27]. He, H., Bai, Y., Garcia, E., & Li, S. (2008). *ADASYN: Adaptive synthetic sampling approach for imbalanced learning.* IEEE International Joint Conference on Neural Networks 1322–1328. https://doi.org/10.1109/IJCNN.2008.4633969

[28]. Krawczyk, J. (2016). *Learning from imbalanced data: open challenges and future directions.* Progress in Artificial Intelligence, 5, 221–232. https://doi.org/10.1007/s13748-016-0094-0

[29]. Iqbal, Z., Rafique, A., Qaisar, S., et. al., (2025). *Advancements and challenges in development of generative adversarial network (GANs) for deep learning.* SN Computer Science. https://doi.org/10.1007/s44354-025-00007-w

[30]. Ou, H., Yao, Y., & He, Y. (2024). *Missing Data Imputation Method Combining Random Forest and Generative Adversarial Imputation Network.* Sensors 24(4), 1112. https://doi.org/10.3390/s24041112