Journal of Theoretical and Applied Information Technology
15st September 2026. Vol.104. No.17

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

CONDITIONAL TABULAR GENERATIVE ADVERSARIAL NETWORK WITH RECURRENT ATTENTION NETWORK FOR DATA IMBALANCE IN MEDICAL DATA

JAYANTHI MUDDAPPA[1,3], JAYASUDHA KALIANNAN[1], M. JAGADEESAN[2]

[1]Department of Computer Science, Thiruvalluvar Govt.Arts College, Rasipuram, Namakkal, India

Affiliated to Periyar University, Salem

[2]Department of Master of Computer Applications, Kongu Engineering College, Perundurai, Erode, India

[3]CMR University, Bangalore, India.

E-mail: [1]jai_yadav@hotmail.com, [2]jayasudhakaliannan@gmail.com, [3]jagadeesankec@gmail.com

ABSTRACT

The class imbalance problem is critical in medical diagnosis due to large-scale clinical datasets exhibiting an imbalanced class distribution. Data balancing algorithms are employed to address this problem by oversampling minority class or undersampling majority class. In this manuscript, Conditional Tabular Generative Adversarial Network (CT-GAN) is developed in preprocessing phase to balance the data and improve classification performance. CT-GAN learns the feature distributions and generates samples that preserve feature correlation, which enhances the classification model's performance. In the classification phase, the Recurrent Attention Network (RAN) is developed to classify the performance of balanced data instances in diabetes. The RAN integrates a Recurrent Neural Network (RNN) with an attention mechanism for focusing more on essential features of data. The developed CT-GAN with RAN obtains 97.85% accuracy on the PIMA dataset, 99.75% accuracy on heart dataset, and 99.00% accuracy on Statlog dataset.

Keywords: *Class Imbalance, Conditional Tabular Generative Adversarial Network, Feature Correlation, Feature Distributions, And Recurrent Attention Network.*

1. INTRODUCTION

Diabetes is chronic disease which occurs when blood sugar is not sufficiently performed [1]. Over time, that affects several organs, leading to long-term damage and malfunction. Diabetes causes serious complications like heart attack, stroke, kidney failure, and nerve damage [2, 3]. Hyperglycemia is condition where high blood glucose levels damage the cardiovascular system and vital organs such as nerves, kidneys, and eyes [4]. Class imbalance is an issue in that amount of samples in all classes is not same. Classes with many samples are represented as the majority class, while those with fewer instances are called the minority class [5]. The imbalance of any dataset is measured by the Class Imbalance Ratio (CIR), which is the ratio of minority instances to majority ones [6]. Data skewness often compromises Machine Learning (ML) techniques effectiveness. The ML models trained on majority samples often fails to recognize minority class, typically overfitting issue [7]. Balancing skewed data pre-training is a primary remedy. To address pre-training data disparities data augmentation used for generate supplementary information.

Random sampling is a simple technique for data augmentation. This method involves the random generation of new instances based on the minority class [8]. Oversampling and undersampling are two primary random sampling algorithms. The undersampling approach randomly eliminates samples from majority-class samples [9]. Therefore, amount of samples in both majority and minority class labels are equalized, though the loss of information remains a significant disadvantage of these methods [10]. The oversampling approach replicates minority class instances to ensure its volume matches that of the majority class. These duplicate samples improve classifier's accuracy compared to training on original imbalanced dataset [11, 12]. Data balancing is essential to enhancing accuracy and reliability of ML algorithms when the target class is unevenly distributed [13]. In imbalanced datasets, the method becomes biased toward majority class, resulting in poor and unreliable predictions for the minority class. This

causes inaccurate and unreliable outcomes, specifically for the minority class [14]. Data balancing algorithms are utilized for addressing the problem by oversampling minority class or undersampling majority class. This process supports the development of well-balanced data, leading to more precise and reliable outcomes.

Sugendran and Sujatha [15] introduced the Enhanced Genetic Algorithm (EGA) enabled Fuzzy Weight Updating Support Vector Machine (FWSVM). Abdellatif et al. [16] suggested an efficient algorithm based on Synthetic Minority Oversampling Technique (SMOTE) to handling imbalance problem. Shokouhifar et al. [17] presented the Ensemble Heuristic-Metaheuristic Feature Fusion Learning (EHMFFL) to diagnose heart disease by tabular data. In EHMFFL, an ensemble method was developed, which diverse feature subsets across heterogeneous base learner, including SVM, K-Nearest Neighbor (KNN), Logistic Regression (LR), Naive Bayes (NB), Random Forest (RF), and XGBoost algorithms.

Feng et al. [18] developed an ML-enabled diabetes classification framework optimized with a Generative Adversarial Network (GAN). Nandakumar and Narayan [19] implemented the Hamming feature selection algorithm for data preprocessing and cleaning procedures in various cardiac disease datasets. A Deep Learning (DL) algorithm, such as Deep Belief Networks (DBN), was utilized with a cuckoo search bio-inspired approach to identify precise predictions of cardiac disease. Medical datasets often suffer from class imbalance, where amount of majority class samples significantly outweighs minority class. This imbalance led to biased ML-based algorithms that favored the majority class, reducing the accuracy and reliability of predictions for minority class instances. Conventional oversampling and undersampling methods resulted in data loss or redundancy, limiting their effectiveness. In the context of essential medical diagnosis, namely, diabetes and heart disease, misclassification of minority classes can lead to severe consequences. The primary objective of this article is to develop integrated framework combining Conditional Tabular Generative Adversarial Network (CT-GAN) and Recurrent Attention Network (RAN) to handle class imbalance in medical datasets. CT-GAN is employed in the preprocessing phase for generating synthetic minority class samples, which preserves the feature relationship, thereby balancing the dataset. RAN is utilized for classifying the balanced data through focusing attention on most appropriate features. This algorithm aims to improve overall classification performance, particularly for minority classes, over different medical datasets such as PIMA, Heart, and Statlog.

Several ML and hybrid algorithms exist for medical diagnosis, but these models still face limitations. Conventional oversampling algorithms, namely SMOTE and its variants, generate redundant or noisy synthetic samples without preserving complex feature dependencies in tabular medical data. Similarly, conventional GAN-based models struggle with mixed data types and highly imbalanced categorical distributions that are common in clinical datasets. Additionally, many classification models focus primarily on accuracy improvement without incorporating attention mechanisms to dynamically prioritize clinically significant features. Hence, there is a need for an integrated model that efficiently handles severe class imbalance while preserving feature correlations in tabular data and improves classification performance by attention-based learning to enhance minority class detection and the overall model generalization. The main issue addressed is severe class imbalance in medical tabular data, which leads to biased learning and poor minority class detection. Although generative models, such as CT-GAN, are utilized for synthetic sample generation, existing algorithms primarily focus on data balancing and do not include adaptive feature learning or attention-based classification, resulting in suboptimal minority sensitivity and limited generalization. Moreover, conventional classifiers fail to capture complex inter-feature dependencies present in medical attributes. To address these drawbacks, the CT-GAN-RAN model is introduced, where CT-GAN performs realistic minority data augmentation, and RAN learns contextual feature interactions with attention-based prioritization. This enables balanced learning, enhanced minority detection, and table performance across datasets. Developing a robust, integrated model for medical data classification that utilizes CT-GAN to mitigate class imbalance while leveraging RAN to improve predictive accuracy is the primary objective of this research. The framework aims to improve minority class recognition, preserve feature relationships in synthetic data generation, and enhance overall classification reliability across benchmark datasets such as PIMA, Heart, and Statlog.

The significant contributions of the manuscript are described below. The CT-GAN method is developed to address severe issue of data imbalance in medical data. Unlike conventional oversampling or

undersampling models, CT-GAN learns the joint feature distribution of minority and majority classes and generates synthetic samples, which preserve complex feature correlations. This process prevents redundancy and data loss, ensuring a balanced dataset that represents realistic medical data patterns. The RAN method is introduced in the classification stage to enhance the detection accuracy of minority classes. By integrating a Recurrent Neural Network (RNN) with an attention mechanism, RAN dynamically assigns higher importance to clinically relevant features. This helps a model to extract long-range dependencies, and focusing on critical patterns in medical data leads to enhanced interpretability and prediction precision. The standard scalar method is applied to normalize dataset attributes into a uniform scale, minimizing bias caused by varying feature magnitudes. This ensures stable convergence in model training and improves the overall performance of the CT-GAN and RAN models.

The rest of the article is organized as follows: Section 2 provides the details of a proposed approach. Section 3 presents results and comparitive analysis of a proposed algorithm, and Section 4 concludes an article.

2. PROPOSED METHOD

In this manuscript, CT-GAN and RAN are developed to address data imbalance and enhance the classification performance. The PIMA, heart, and Statlog datasets are used in this manuscript. Data are balanced by using the developed CT-GAN, and they are normalized by using the standard scalar technique. Subsequently, the classification is performed by using RAN with higher classification performance. Figure 1 describes the process of data imbalance in medical data.

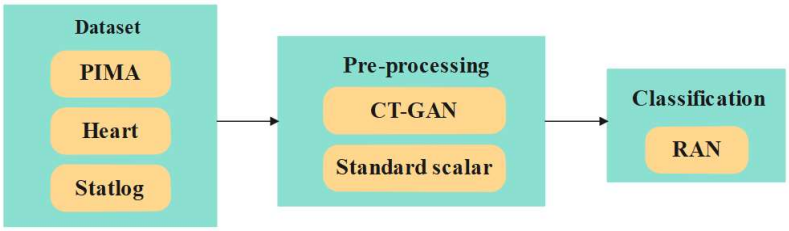

Figure 1: Process of Data Imbalance in Medical Data

2.1 Dataset

The datasets used in this manuscript are the PIMA, heart, and Statlog datasets, which contain imbalanced data. The detailed description of these datasets is given below.

2.1.1 PIMA

This study utilizes the PIMA Indian Diabetes Dataset [20], which was gathered and compiled by National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK). The dataset includes 768 patients, comprising 268 individuals with diabetes and 500 individuals without diabetes. For each patient, eight physiological attributes are recorded, including blood pressure, skin thickness, pregnancies, insulin, glucose, Diabetes Pedigree Function (DPF), Body Mass Index (BMI), and age. These attributes are used for predicting the presence of diabetes in each patient.

2.1.2 Heart and statlog dataset

Here, two popular datasets sourced from University of California's Irvine (UCI) Machine Learning Repository (ML Repository), such as the heart and Statlog datasets, are used [21]. The attributes are general for both datasets, with final feature acting as indicator of person's heart disease status. There is a relationship between highest heart rate and age, and distribution of remaining 13 attributes.

2.2 Preprocessing

The dataset is preprocessed by using CT-GAN for balancing the data and the standard scalar to normalize the attributes into a uniform format. A detailed explanation of these processes is given below.

2.2.1 GAN

Adversarial techniques are utilized by GANs to facilitate the training of generative frameworks. Two distinct components, a generator G and discriminator D, are simultaneously optimized by the architecture. Generator by taking random noise as its primary input, the generator learns the underlying distribution of the training data. It aims to produce fake data which is indistinguishable from actual data in order to deceive discriminator. Both types of data are simultaneously fed into the discriminator. During

GAN training, G and D simultaneously learn from each other by adversarial interaction. After generator learns real data distribution and produces fake data, which nearly resembles real one, discriminator is not able to correctly distinguish the validity of input instances, indicating that training is complete.

The GAN optimization process is significant, where the aim is to find an extremum and obtain an equilibrium. While $P_{data}(x)$ denotes the distribution of real data, $P_z(z)$ serves as the representation for noise distribution. The $D(x)$ is probability which sample is obtained from actual data, with the values between 0 and 1, and $G(z)$ is data produced through G using learning process. For obtaining an equilibrium between G and D mathematical expression for fitness function of GAN is given as Equation (1).

$$\min_G \max_D V(D,G) = E_{x\sim Pdata(x)}[\log D(x)] + E_{x\sim pz(z)}\left[\log\left(1 - D\left(G(z)\right)\right)\right] \quad (1)$$

Where the first component of the equation signifies the discriminator's likelihood for training data, whereas the subsequent term reflects its prediction for synthetic data.

2.2.2 CT-GAN

In this dataset, each column is taken as random variable followed by an unidentified distribution, where every row represents an individual sample of the combined probability distribution across all variables. Therefore, dataset represents different attributes, which pose difficulty in GAN model. This includes existence of heterogeneous data formats, distributions that are non-Gaussian or multi-modal, and significant skewness in categorical feature., and highly imbalanced categorical columns. Hence, this manuscript introduced the CT-GAN for addressing these issues, which incorporates a normalization strategy tailored to specific modes and constructs a conditional generator optimized via sampling techniques. A variational Gaussian mixture model is utilized to estimate the count of patterns and align a Gaussian mixture to each individual column C_i, and its mathematical expression is given as Equation (2).

$$P_{C_i}(c_{i,j}) = \sum_{k=1}^{m} \mu_k N(c_{i,j}; \eta_k, \phi_k) \quad (2)$$

In Equation (2), $c_{i,j}$ represents value of ith column and jth row, and μ_k, η_k, and ϕ_k represent weights and standard deviations of the kth modes. Hence, possibility of every value $c_{i,j}$ being taken from every mode is measured through possibility densities $\rho_k = \mu_k N(c_{i,j}; \eta_k, \phi_k)$ of each pattern. The $c_{i,j}$ is a one-hot encoded vector, $\beta_{i,j}$ represents a particular pattern, and $\alpha_{i,j}$ represents the scalar for a particular value in the modes. The mathematical expression for $\alpha_{i,j}$ is given as Equation (3).

$$\alpha_{i,j} = \frac{c_{i,j} - \eta_k}{4 \cdot \phi_k} \quad (3)$$

To handle imbalance issues in categorical columns, vector $cond$ is employed to encode category-specific conditions. The different columns in the dataset are represented as D_1 and D_2, and d_1 and d_2 represent the one-hot vectors. The mask vector m_i is used to represent integrated one-hot vector d_i. Additionally, cross-entropy among m_i and d_i is incorporated to penalize their loss and ensuring conditional generator produces $d_i = m_i$. A fully connected architecture is utilized by the CT-GAN generator G and discriminator D to capture the full range of potential inter-column dependencies.. The network's structural design incorporates a pair of fully connected hidden layers. The generator G makes use of the Rectified Linear Unit (ReLU) activation function alongside batch normalization.

CT-GAN alone is primarily a data-level imbalance handling technique that focuses on generating synthetic minority samples by learning the joint feature distribution. However, it does not incorporate task-specific discriminative learning, feature dependency modeling, or class-aware decision optimization during classification. As a result, although CT-GAN improves class balance, the downstream classifier may still exhibit bias towards the majority of the patterns and fail to capture the complex inter-feature relationships in the medical tabular data. To address this limitation, the proposed framework integrates CT-GAN with an RAN, where CT-GAN performs realistic minority sample augmentation, and the RAN learns contextual feature interactions with attention-driven weighting of clinically significant attributes. Additionally, imbalance-aware learning is achieved through stratified training, minority-sensitive loss optimization, and attention-based feature prioritization, ensuring improved recall for minority classes and balanced decision boundaries beyond what CT-GAN alone can provide.

2.2.3 Standard scalar

This technique is used to scale data through removing mean and scaling unit variance by standardizing a feature. By computing the required statistics on samples in the set, features are centered and independently scaled [22]. The mean and standard deviation are stored to further use on data. Standardization of the dataset is needed since the individual features are not normally distributed, which generates expected outcomes. Figure 2

represents the architecture of the CT-GAN-RAN model for class imbalance.

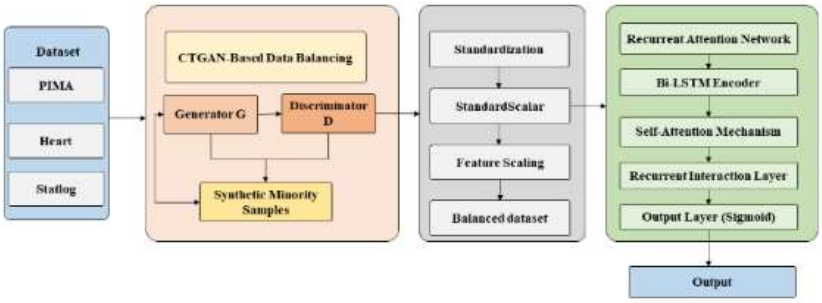

Figure 2: Architecture of the CT-GAN-RAN Model for the Class Imbalance Process

2.3 RAN

In RAN, initially, an encoder layer uses a Long Short-Term Memory (LSTM) and self-attention for embedding label definitions and facts into lower-dimensional representation space. Next, recurrent layer models the process of repeatedly learning facts and label descriptors to obtain mutual data. Finally, output layer produces the final prediction outcome of the network.

The selection of RNN within the proposed RAN framework is theoretically justified by its capability to model sequential dependencies and contextual relationships among input features. Although medical tabular data are not time-series in nature, the ordered feature interactions can be modeled as sequential patterns in which inter-feature dependencies influence prediction outcomes. RNN, particularly Bi-directional LSTM (Bi-LSTM), effectively captures long-range contextual information through gated memory mechanisms, enabling the model to retain clinically relevant feature interactions. Compared to feedforward networks, RNNs provide dynamic feature representation learning, which becomes more powerful when integrated with an attention mechanism to emphasize discriminative attributes, thereby improving minority class sensitivity and overall classification robustness.

Encoder layer – Every fact and label is represented as sequence of words. One-hot representation is used as input, and every word is embedded into continuous vector space. Consider the $x = \{w_1, w_2, w_3, \dots, w_n\}$ with m words; label definitions with n words are represented as $l = \{w_1, w_2, w_3, \dots, w_n\}$, and w_i represents the ith word. Consider $V^I = \{v_t^I \in R^{D_v} | t = 1, \dots, N\}$ representing the complete set of continuous embeddings. For every fact and label definition, word vector is integrated to develop the fact and label representation. Bi-LSTM is employed to calculate hidden states for each token at every time step t, as given in Equations (4) and (5).

$$\vec{h}_t = \overrightarrow{LSTM}(\vec{h}_{t-1}, w_t) \quad (4)$$

$$\overleftarrow{h}_t = \overleftarrow{LSTM}(\overleftarrow{h}_{t-1}, w_t) \quad (5)$$

With Bi-LSTM, the hidden representation of the ith sample is obtained through concatenating latent representations from both directions $h_t = [\vec{h}_t; \overleftarrow{h}_t]$, and then x fact and label l are marked for distributed vector representations $H_c = [h_1, h_2, \dots, h_m]$, $H_a = [h_1, h_2, \dots, h_n]$. Since various words have different significance in one frame, self-attention is applied to derive weighted embeddings for facts and labels H_{cs} and H_{as}.

2.3.1 Recurrent layer

The layer learns to identify meaningful data and selects relevant features as potential candidates. Then, a detailed semantic analysis across facts and potential candidate articles is employed to determine the final outcome. this procedure is iteratively executed to achieve a definitive conclusion. Rather than applying basic cross-attention between the source and target characteristics, the encoder layer derives the desired fact and label encoding vectors H_{cs} and H_{as}. The aggregation process for obtaining the final representation of all labels is given in Equation (6).

$$H_m = \left[c\left(H_{as}^{(1)}\right), c\left(H_{as}^{(2)}\right), \dots, c\left(H_{as}^{(C)}\right)\right] \quad (6)$$

In Equation (6), c represents aggregation process that averages the final representation to develop the representation of every label. Subsequently, an alignment score matrix comparing the fact and label representations is derived, and its mathematical expression is given in Equation (7).

$$M(j,k) = H_m(j) \cdot H_{cs}(k) \quad (7)$$

In Equation (7), $M \in R^{|c| \times |x|}$, where each value represents the interaction value between the fact and

every label. After obtaining matrix M, a softmax operation is performed column-wise to generate probability distributions. In this setup, each column represents a distinct attention vector, where $\alpha(t)$ signifies the label-specific attention for each fact token at time step t , its mathematical expression is given in Equations (8) and (9).

$$\alpha(t) = softmax\big(M(1,t), M(2,t), \dots, M(|C|,t)\big) \quad (8)$$

$$\alpha = [\alpha(1), \alpha(2), \dots, \alpha(|x|)] \quad (9)$$

Next, all $\alpha(t)$ are averaged to obtain average label-level attention $\alpha' \in R^{|C|}$, where averaging process does not break normalizing condition, and its mathematical expression is given in Equation (10).

$$\alpha' = \frac{1}{|x|}\sum_{i=1}^{|x|} \alpha(t) \quad (10)$$

Similarly, a row-wise softmax is used for obtaining label attention $\beta' \in R^{|x|}$ where the attention scores α' and β' are obtained. Within the developed recurrent architecture, this process is iteratively repeated to learn significant mutual semantic information, and its mathematical expression is given in Equations (11) and (12).

$$H_{cs} = H_{cs} + H_{cs}W^{\alpha'}\alpha' \quad (11)$$

$$H_m = H_m + H_m W^{\beta'}\beta' \quad (12)$$

In Equations (11) and (12), $W^{\alpha'}$ and $W^{\beta'}$ represent the dimension transformation matrices. This includes multi-level semantic interaction data to help with accurate label matching.

2.3.2 Output layer

Fact and label-specific features are leveraged to integrate global label information and forecast results within the final layer. The probability distribution across entire set of categories is computed using Equations (13) to (16).

$$v_{cr} = g(H_{cr}) = \frac{1}{n}\sum_{t=1}^{n} v_{cr}(t) \quad (13)$$

$$v_{ar} = g(H_{ar}) = \frac{1}{n}\sum_{t=1}^{n} v_{ar}(t) \quad (14)$$

$$v_f = v_{cr} \oplus v_{ar} \quad (15)$$

$$v_o = W^o v_f + b^o \quad (16)$$

In Equations. (13) to (16), g represents the averaging process, v_{cr} is context representation, v_{ar} is averaged representation that represents global label features, $\oplus$ is the concatenation process, and W^o and b^o serving as the trainable weights.

2.3.3 Learning and predictions

Binary cross-entropy loss and the stochastic gradient descent method are utilized throughout the training phase to optimize RAN parameters. Its mathematical expression is given in Equation (17).

$$l = -\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{|C|}[G^{(i)}(j)\log\left(\sigma\left(v_o^{(i)}(j)\right)\right) + (1 - G^{(i)}(j)\log\left(1 - \sigma\left(v_o^{(i)}(j)\right)\right)] \quad (17)$$

In Equation (17), σ represents the sigmoid function, and $G^{(i)}(j) \in \{1,0\}$ is binary variable indicating if the $x^{(i)}$ instance results in violation of label $y^{(i)}$. By applying the optimized weights, a probability distribution across all labels is obtained for each label. Using a threshold, a label set of instances $x^{(i)}$ is obtained, and its mathematical expression is given in Equation (18).

$$Y^{(i)}(j) = \begin{cases} 1, & if\ I\left(\sigma\left(v_o^{(i)}(j)\right)\right) > t \\ 0, & else \end{cases} \quad (18)$$

In Equation (18), I represents the indication function, and $Y^{(i)}(j)$ is the jth element of $Y^{(i)}$, considering label $y^{(i)}$ with a result probability higher than t as relevant label for $x^{(i)}$. The CT-GAN-RAN model is proposed to handle class imbalance in medical datasets by effectively integrating generative data augmentation with attention-based classification. CT-GAN preserves complex feature correlations when generating synthetic minority samples by overcoming the drawbacks of conventional oversampling and undersampling models. RAN improves predictive performance by capturing temporal dependencies and dynamically focusing on clinically significant features with its attention mechanism.

Algorithm 1 – CT-GAN-RAN

Initialize metric storage.

for fold t=1 to k, do

Initialize generator G and discriminator D.

for epoch = 1 to E_{GAN}, do

for every batch (x_{real}, c) in D_{train}, do

Sample noise z

$z_{fake} \leftarrow G(z, c)$

Compute discriminator loss.

$L_D = -\left[\log D(x_{real}) + \log\left(1 - D(x_{fake})\right)\right]$

Update discriminator parameters.

Calculate generator loss.

$L_G = -\log D(x_{fake})$

Update generator parameters.

end for

end for

Generate synthetic minority samples.

for each feature column j, do

$$\mu_j \leftarrow mean(X_j \text{ in } D_{balanced})$$

$$\sigma_j \leftarrow Standard\ Deviation\ (X_j)$$

Normalize

$$X_j \leftarrow (X_j - \mu_j)/\sigma_j$$

end for

Initialize RAN parameters.

for epoch=1 to E_{RAN}, do

for every batch (x, y) in $D_{balanced}$, do

$$h \leftarrow BiLSTM(x)$$

$$\alpha \leftarrow Softmax(score(h))$$

$$C \leftarrow \sum(\alpha \times h)$$

$$y_{hat} \leftarrow Sigmoid$$

$$L \leftarrow -[y \log(y_{hat}) + (1-y) \log(1 - y_{hat})$$

Update parameters by Stochastic Gradient Descent (SGD).

end for

end for

Return the final performance results.

3. EXPERIMENTAL ANALYSIS

The performance of the developed CT-GAN and RAN algorithm is developed in the Python 3.7 environment and is analyzed with the performance metrics of accuracy, precision, F1-score, recall, and Area Under Curve (AUC) on PIMA, heart, and Statlog datasets. All experiments are conducted with a fixed random seed of 42 to ensure reproducibility across data splitting, CT-GAN sample generation, and model initialization. The models are implemented in Python 3.7 using TensorFlow 2 and Scikit-learn libraries. The hyperparameters of the proposed CT-GAN-RAN model are tuned to ensure stable training and optimal performance across all datasets. In the CT-GAN module, the generator and discriminator include two fully connected hidden layers with ReLU activation and batch normalization employed in generator. The vector dimension is set to 100, batch size to 64, and training is performed for 100 epochs by Adam optimizer with learning rate of 0.0002. The penalty coefficient of 10 and conditional vector sampling are used for improving stability. In the RAN classifier, a two-layer Bi-LSTM with 128 hidden units and a dimension of 64 is used with a dropout rate of 0.5 to prevent overfitting. Binary cross-entropy loss and a classification threshold of 0.5 are applied during model training, which utilizes the Adam optimizer with a 0.001 learning rate and a batch size of 32. The datasets are split using a stratified 80:20 train-test ratio. Stratified sampling preserves the original class distribution in subsets to ensure fair evaluation under imbalanced conditions. Moreover, k-fold cross-validation is performed on training set to validate model's performance. To ensure fair evaluation and prevent data leakage, CT-GAN is trained using training set after applying a stratified 80:20 split. Synthetic minority samples are generated from training data, and test data remains unseen during data balancing and model training.

Figures 3 and 4 illustrate the class distribution before and after balancing on the PIMA dataset. Figures 5 and 6 depict the class distribution before and after balancing on the heart dataset. Figures 7 and 8 illustrate the class distribution before and after balancing on the Statlog dataset.

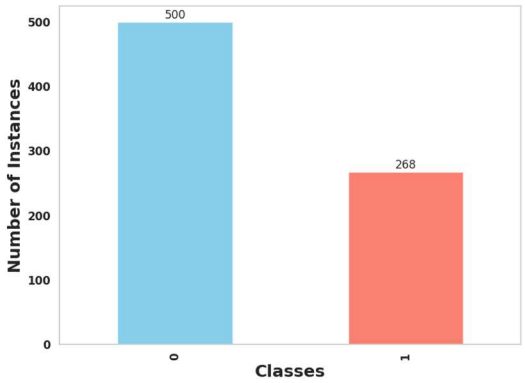

Figure 3: Before Addressing Class Imbalance on the PIMA Dataset

Journal of Theoretical and Applied Information Technology
15st September 2026. Vol.104. No.17

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

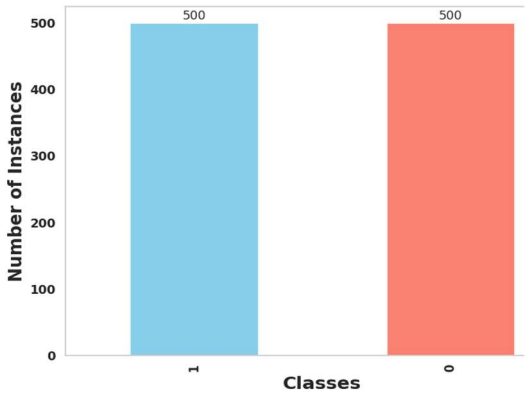

Figure 4: After Class Balance on the PIMA Dataset

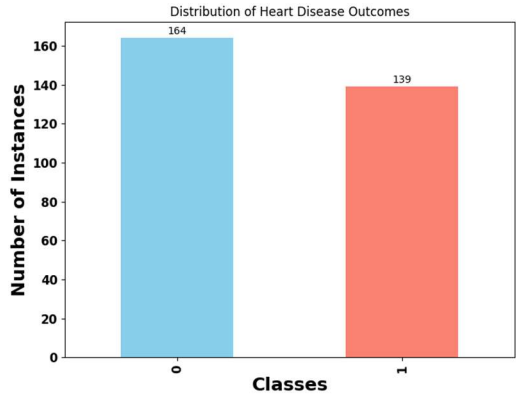

Figure 5: Before Addressing Class Imbalance on the Heart Dataset

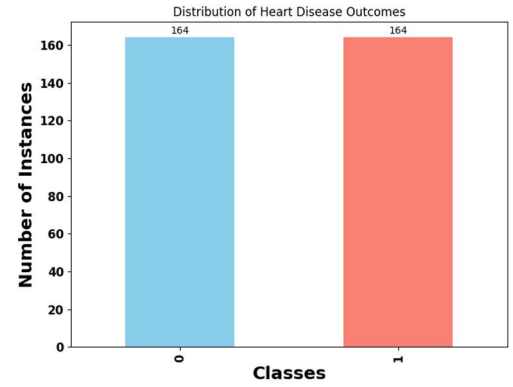

Figure 6: After Class Balance on the Heart Dataset.

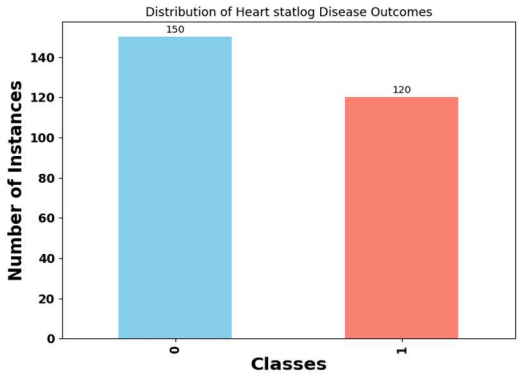

Figure 7: Before Addressing Class Imbalance on the Statlog Dataset.

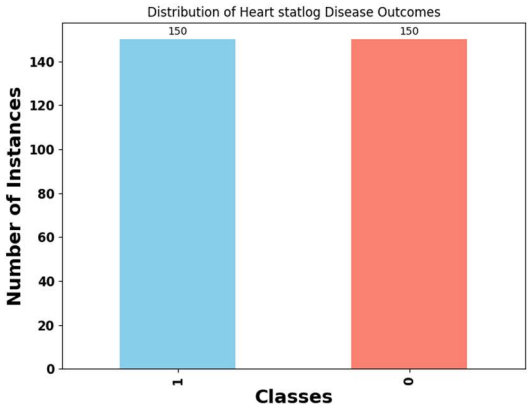

Figure 8: After Class Balance on the Statlog Dataset.

Table 1: Performance of the Developed CT-GAN with Different Data Balancing Methods.

Methods	Dataset	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	AUC (%)
DCGAN	PIMA	95.20	94.85	95.00	94.92	95.10
	Heart	94.50	94.00	94.20	94.10	94.40
	Statlog	94.80	94.60	94.70	94.65	94.85
CGAN	PIMA	96.00	95.50	95.80	95.65	96.10
	Heart	96.30	96.00	96.10	96.05	96.20
	Statlog	96.50	96.30	96.40	96.35	96.50
GAN	PIMA	97.30	97.00	97.20	97.10	97.40
	Heart	98.00	97.90	98.00	97.95	98.10
	Statlog	97.80	97.60	97.70	97.65	97.80
	PIMA	97.85	97.12	97.34	97.23	98.00

CT-GAN-RAN	Heart	99.75	99.85	99.75	99.80	99.56
	Statlog	99.00	99.20	99.35	99.27	99.00

Table 1 presents the performance of the developed CT-GAN-RAN with different data balancing methods on the PIMA, heart, and Statlog datasets. The data balancing methods, such as Deep Convolutional GAN (DCGAN), Conditional GAN (CGAN), and conventional GAN, are considered. The developed CT-GAN-RAN obtains 97.85% accuracy on the PIMA dataset, 99.75% accuracy on heart dataset, and 99.00% accuracy on Statlog dataset.

Table 2 evaluates the performance of the developed RAN without the CT-GAN algorithm with different classification methods on the PIMA, heart, and Statlog datasets. The different classification methods, such as Forward Neural Network (FNN), RNN, and LSTM, are considered. The developed RAN obtains 96.25% accuracy on PIMA dataset, 97.85% accuracy on heart dataset, and 98.10% accuracy on Statlog dataset.

Table 3 evaluates the performance of the developed RAN with the CT-GAN algorithm with different classification methods on the PIMA, heart, and Statlog datasets. The different classification methods, such as FNN, RNN, and LSTM, are considered. The developed CT-GAN with RAN obtains 97.85% accuracy on the PIMA dataset, 99.75% accuracy on the heart dataset, and 99.00% accuracy on the Statlog dataset.

Table 2: Performance of the RAN without the CT-GAN Algorithm.

Methods	Dataset	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	AUC (%)
FNN	PIMA	90.50	90.30	90.40	90.35	90.50
	Heart	92.00	91.80	91.90	91.85	92.00
	Statlog	93.50	93.30	93.40	93.35	93.50
RNN	PIMA	91.80	91.60	91.70	91.65	91.80
	Heart	93.20	93.00	93.10	93.05	93.20
	Statlog	94.00	93.80	93.90	93.85	94.00
LSTM	PIMA	94.50	94.30	94.40	94.35	94.50
	Heart	95.20	95.00	95.10	95.05	95.20
	Statlog	95.80	95.60	95.70	95.65	95.80
RAN	PIMA	96.25	96.30	96.45	96.37	96.00
	Heart	97.85	97.90	98.00	97.95	97.00
	Statlog	98.10	98.20	98.35	98.25	98.00

Table 3: Performance of the RAN with the CT-GAN Algorithm.

Methods	Dataset	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	AUC (%)
FNN	PIMA	94.20	93.80	94.00	93.90	94.10
	Heart	94.80	94.50	94.60	94.55	94.70
	Statlog	95.10	94.90	95.00	94.95	95.05
RNN	PIMA	95.50	95.10	95.30	95.20	95.40
	Heart	96.00	95.80	95.90	95.85	96.10
	Statlog	96.40	96.20	96.30	96.25	96.50
LSTM	PIMA	96.80	96.50	96.60	96.55	96.75
	Heart	97.20	97.00	97.10	97.05	97.30
	Statlog	97.60	97.40	97.50	97.45	97.70
CT-GAN-RAN	PIMA	97.85	97.12	97.34	97.23	98.00
	Heart	99.75	99.85	99.75	99.80	99.56
	Statlog	99.00	99.20	99.35	99.27	99.00

Tables 4, 5, and 6 evaluate the k-fold values for the developed RAN with the CT-GAN algorithm on the PIMA, heart, and Statlog datasets. The k-fold values of 2, 3, and 5 are evaluated, with k=5 yielding the highest performance compared to other k-fold values. The developed algorithm obtains a t-statistic of 7.14 and a p-value of 0.0002; this value represents a statistically significant difference, indicating high performance compared to LSTM. Furthermore, the F-statistic of 139.62, a p-value of 0.00008, and a statistically significant difference among models.

Table 4: Performance of the K-Fold Values on the PIMA Dataset.

K-fold values	Methods	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
2	FNN	95.12	94.45	94.78	94.61
	RNN	96.34	95.78	95.89	95.83
	LSTM	96.95	96.21	96.47	96.34
	RAN	98.10	97.45	97.68	97.56
3	FNN	95.56	94.89	95.12	95.00
	RNN	96.78	96.12	96.34	96.23
	LSTM	97.34	96.56	96.89	96.72
	RAN	98.25	97.58	97.80	97.69
5	FNN	96.12	95.45	95.78	95.61
	RNN	97.02	96.34	96.67	96.50
	LSTM	97.56	96.89	97.23	97.06
	RAN	97.85	97.12	97.34	97.23

Table 5: Performance of the K-Fold Values on the Heart Dataset.

K-fold values	Methods	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
2	FNN	97.45	97.12	97.30	97.21
	RNN	98.12	97.85	97.90	97.87
	LSTM	98.65	98.34	98.45	98.39
	RAN	99.12	99.25	99.10	99.17
3	FNN	97.85	97.50	97.70	97.60
	RNN	98.45	98.12	98.25	98.18
	LSTM	98.92	98.65	98.78	98.71
	RAN	99.50	99.65	99.55	99.60
5	FNN	98.25	97.85	98.10	97.97
	RNN	98.75	98.45	98.60	98.52
	LSTM	99.10	98.85	98.95	98.90
	RAN	99.75	99.85	99.75	99.80

Table 6: Performance of the K-Fold Values on the Statlog Dataset.

K-fold values	Methods	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
2	FNN	96.80	96.55	96.70	96.62
	RNN	97.50	97.30	97.45	97.37
	LSTM	98.00	97.85	98.10	97.97
	RAN	98.70	98.90	98.95	98.92
3	FNN	97.20	97.00	97.15	97.07
	RNN	97.85	97.65	97.80	97.72
	LSTM	98.40	98.25	98.50	98.37
	RAN	98.95	99.05	99.15	99.10
5	FNN	97.60	97.40	97.65	97.52
	RNN	98.25	98.10	98.30	98.20
	LSTM	98.75	98.60	98.80	98.70
	RAN	99.00	99.20	99.35	99.27

Table 7 represents the statistical performance when compared to different existing methods across three datasets. The results demonstrate that the proposed CT-GAN-RAN method consistently obtains high mean accuracy with lower standard deviation, indicating its superior stability and reliability. Narrow confidence intervals reflect the model's robustness and fewer performance variations across multiple runs. Additionally, t-statistic values are higher for the proposed CT-GAN-RAN with significant p-values, demonstrating statistical improvements over the existing methods. These results illustrate that the integration of CT-GAN for balanced data generation and the RAN for superior classification effectively improves the predictive accuracy and generalization ability across different medical datasets.

Table 7: Statistical Analysis of the Proposed CT-GAN-RAN Model Across all Three Datasets.

Datasets	Methods	Mean Accuracy (%)	Standard Deviation (±)	Confidence Interval (%)	t-Statistic	P-value
PIMA	FNN	94.20	0.90	[93.4–95.0]	6.11	0.015
	RNN	95.50	0.85	[94.8– 96.2]	6.80	0.010
	LSTM	96.80	0.65	[96.2–97.4]	7.00	0.008
	CT-GAN-RAN	97.85	0.55	[97.3–98.4]	7.25	0.003
Heart	FNN	94.80	0.85	[94.1–95.5]	6.72	0.010
	RNN	96.00	0.75	[95.3–96.7]	7.00	0.008
	LSTM	97.20	0.70	[96.5–97.9]	7.15	0.007
	CT-GAN-RAN	99.75	0.40	[99.3–100.0]	7.68	0.004
Statlog	FNN	95.10	0.80	[94.3–95.9]	6.60	0.012
	RNN	96.40	0.70	[95.7–97.1]	6.95	0.008
	LSTM	97.60	0.60	[96.9–98.3]	7.30	0.006
	CT-GAN-RAN	99.00	0.45	[98.5–99.6]	7.84	0.003

Table 8 illustrates the computational analysis of the proposed method with the existing models across all three datasets in terms of training time, inference time, overall execution time, Floating-Point Operations Per second (FLOPs), and parameter count. The results show that a proposed CT-GAN-RAN method introduces slightly higher computation time because of its dual-phase architecture, which includes data balancing by CT-GAN and feature-focused classification through RAN. The proposed CT-GAN-RAN achieves high FLOPs, and parameters show a minimal increase in execution time, and is computationally feasible. This is balanced by the model's significant accuracy gains and stability, as additional complexity ensures superior feature learning, improved attention-based decision-making, and enhanced generalization across all the datasets, making it an effective and high-performing model.

Table 8: Computational Analysis of the Proposed CT-GAN-RAN Model Across all Three Datasets.

Datasets	Methods	Training time (s)	Inference time per sample (s)	Overall execution time(s)	FLOPs($\times 10^6$)	Parameters ($\times 10^6$)
PIMA	FNN	18.40	0.0032	18.893	42.5	1.8
	RNN	24.70	0.0045	25.393	65.2	2.4
	LSTM	32.50	0.0051	33.285	78.9	3.1
	CT-GAN-RAN	39.20	0.0060	40.124	85.7	3.6
Heart	FNN	18.40	0.0032	18.595	42.5	1.8
	RNN	24.70	0.0045	24.975	65.2	2.4
	LSTM	32.50	0.0051	32.811	78.9	3.1
	CT-GAN-RAN	39.20	0.0060	39.566	85.7	3.6
Statlog	FNN	18.40	0.0032	18.746	42.5	1.8
	RNN	24.70	0.0045	25.186	65.2	2.4
	LSTM	32.50	0.0051	33.051	78.9	3.1
	CT-GAN-RAN	39.20	0.0060	39.848	85.7	3.6

Table 9 below represents the stability and robustness of the proposed CT-GAN-RAN model across five independent runs. The performance is presented as the mean ± standard deviation, where lower standard deviation values across all datasets represent consistent convergence and minimal sensitivity to random initialization and GAN sampling variations. Particularly, the heart and Statlog datasets display small performance fluctuations, strong generalization, and reliable minority class learning. The variation in value validates the integration of CT-GAN for balanced synthetic data generation, while RAN with attention-based learning provides statistically stable and reproducible results.

Table 9: Evaluate The Stability and Robustness of the Proposed CT-GAN-RAN Model Across Five Independent Runs.

Dataset	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	AUC (%)
PIMA	97.85 ± 0.52	97.12 ± 0.60	97.34 ± 0.55	97.23 ± 0.57	98.00 ± 0.48
Heart	99.75 ± 0.38	99.85 ± 0.32	99.75 ± 0.35	99.80 ± 0.33	99.56 ± 0.41
Statlog	99.00 ± 0.44	99.20 ± 0.40	99.35 ± 0.37	99.27 ± 0.39	99.00 ± 0.42

Table 10 below represents the class-wise sensitivity and specificity outcomes, demonstrating that the proposed CT-GAN-RAN model maintains a balanced diagnostic ability across the majority and minority classes. The high sensitivity values for minority classes indicate that the proposed model efficiently reduces false negatives, which is significant in medical applications. At the same time, consistent high specificity values represent strong discrimination of non-disease classes, minimizing false positives and avoiding unnecessary clinical interventions. The lower standard deviation across five runs validates the model's stability and reproducibility. Overall, these results indicate that the integration of CT-GAN for the imbalance issue and attention-based RAN for feature prioritization results in reliable and statistically robust performance across all the datasets.

Table 10: Class-Wise Sensitivity and Specificity of the Proposed CT-GAN-RAN Model.

Dataset	Classes	Sensitivity (%)	Specificity (%)
PIMA	Non-Diabetic	97.34 ± 0.52	98.22 ± 0.46
	Diabetic	97.34 ± 0.52	98.22 ± 0.46
Heart	No Disease	99.82 ± 0.35	99.68 ± 0.40
	Disease	99.75 ± 0.38	99.85 ± 0.32
Statlog	No Disease	99.20 ± 0.42	98.95 ± 0.45
	Disease	99.35 ± 0.37	99.10 ± 0.41

3.1 Analysis on Pima Dataset

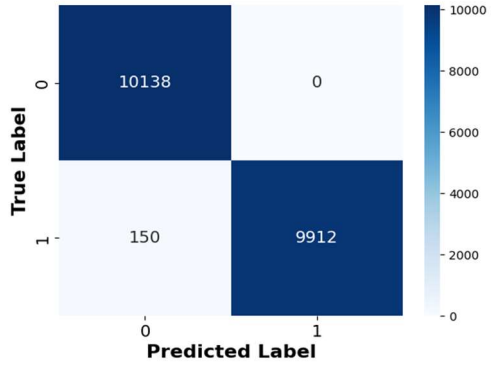

Figure 9: Confusion Matrix on the PIMA Dataset.

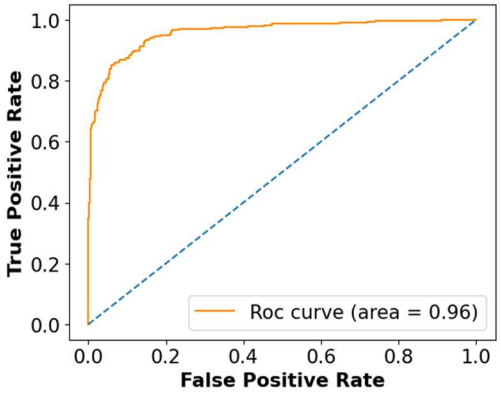

Figure 10: ROC curve on the PIMA dataset.

Figure 9 depicts confusion matrix for PIMA dataset, and Figure 10 displays Receiver Operating Characteristic (ROC) curve. The accuracy and loss curves on the PIMA dataset indicate stable convergence and efficient learning behavior of the proposed CT-GAN-RAN model. The consistent improvement in training and validation accuracy reflects the integration of CT-GAN-generated balanced data, enabling the model to learn representative feature distributions without bias towards the majority class. The lesser gap between training and validation signifies strong generalization and minimized overfitting, which is supported by the confusion matrix indicating reliable classification across classes. Moreover, the ROC curve indicates higher discriminative ability and validates that the attention mechanism in RAN efficiently focuses on the clinically relevant features. These results depict that the integration of data balancing and attention-based recurrent learning improves stability, minority class recognition, and overall predictive robustness on the PIMA dataset.

3.2 Analysis on Heart Dataset

Figure 11 demonstrates the confusion matrix for the heart dataset, and Figure 12 presents the ROC curve. The accuracy and loss curve on the heart dataset show the stable and effective learning behavior of the proposed CT-GAN-RAN model. The

Journal of Theoretical and Applied Information Technology
15st September 2026. Vol.104. No.17

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

consistent improvement in training and validation accuracy through consistent minimization of loss indicates efficient convergence and shows that CT-GAN generated balanced data, ensuring unbiased learning across classes. The alignment between training and validation demonstrates higher generalization ability and lower overfitting. The confusion matrix supports robustness of the model through reliable classification performance for positive and negative classes, enhancing the minority class recognition. Moreover, the ROC curve indicates higher discriminative ability and validates the attention mechanism in RAN, displaying clinically significant features.

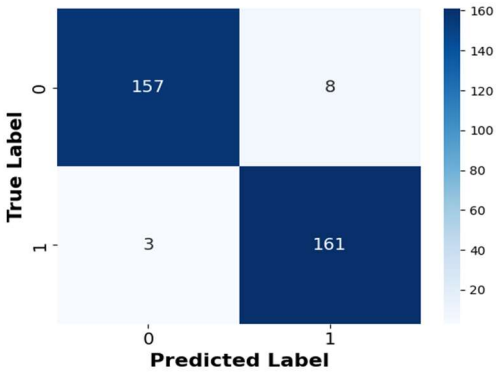

Figure 11: Confusion Matrix on the Heart Dataset.

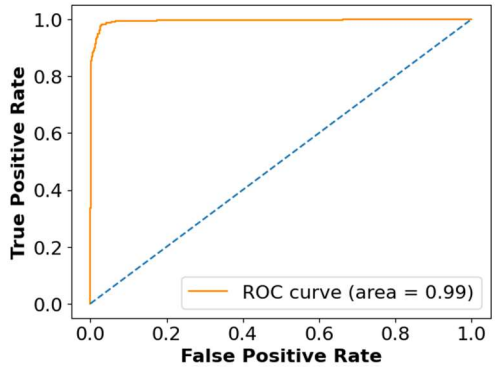

Figure 12: ROC Curve on the Heart Dataset.

3.3 Analysis on Statlog dataset

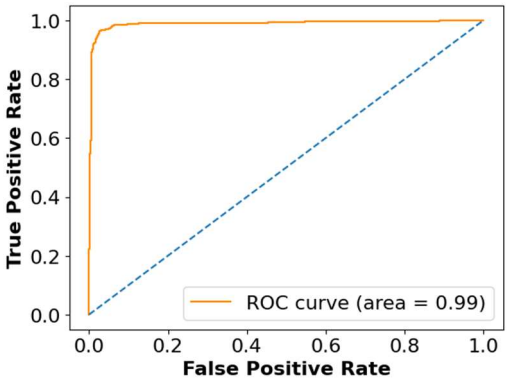

Figure 13: ROC Curve on the Statlog Dataset.

Figure 13 illustrates the ROC curve on the Statlog dataset. The accuracy and loss curves on the Statlog dataset indicate stable convergence of the proposed CT-GAN-RAN model, indicating its capability to learn balanced and discriminative feature representations. The consistent alignment between training and validation accuracy signifies that the CT-GAN-generated samples efficiently overcome class imbalance and prevent biased learning. The minimal divergence between training and validation indicates higher generalization and minimized overfitting. Moreover, the ROC curve depicts higher separability between classes and validates the attention mechanism in RAN, which prioritizes the most informative clinical features.

3.4 Generalization Analysis

The developed CT-GAN+RAN framework demonstrates a stronger generalization ability across multiple medical datasets, such as PIMA, heart, and Statlog, by consistently achieving higher accuracy. The use of CT-GAN effectively balances class distribution and ensures robust learning even in high-imbalanced data. RAN's attention mechanism improves feature relevance and allows the method to generalize well with the unseen samples. The consistent enhancement over baseline methods and conventional GAN-based model indicates minimized overfitting and improved adaptability. Additionally, k-fold cross-validation (k=2,3,5) indicates stability and generalization across different data splits. The method's superior AUC scores suggest reliable performance in different classification scenarios.

3.5 Comparative Analysis

The performance of CT-GAN and RAN is evaluated using different algorithms such as EGA-FWSVM [15], HB-SMOTE-ET [16], EHMFFL [17], and DCSGAN [18] on the PIMA, heart, and

Statlog datasets. The proposed CT-GAN and RAN obtain accuracy of 97.85% on the PIMA dataset, 99.75% on the heart dataset, and 99.00% on the Statlog dataset. CT-GAN learns the feature distributions and generates samples that preserve feature correlation, which enhances the model's classification performance. RAN integrates an RNN with an attention mechanism for focusing on the essential feature of the data. This enhances classification performance by allowing the method to dynamically assign attention to relevant attributes. Table 11 presents the comparative analysis of the developed CT-GAN and RAN.

Table 11: Comparative Analysis of the Developed CT-GAN and RAN.

Methods	Dataset	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
EGA-FWSVM [15]	Heart	92.22	93.26	92.16	92.70
HB-SMOTE-ET [16]	Heart	99.20	98.70	99.33	99.00
	Statlog	98.52	98.13	98.09	98.09
EHMFFL [17]	Heart	91.80	91.40	94.10	92.80
	Statlog	88.90	92.60	86.20	89.30
DCSGAN [18]	PIMA	96.27	96.98	96.98	96.98
Proposed CT-GAN and RAN	PIMA	97.85	97.12	97.34	97.23
	Heart	99.75	99.85	99.75	99.80
	Statlog	99.00	99.20	99.35	99.27

The comparative analysis illustrates that the proposed CT-GAN-RAN method outperforms the existing methods. The consistent enhancement in accuracy, stability, and statistical significance shows the method's robustness and reliability. This superior performance is attributed to integration of CT-GAN, which efficiently overcomes data imbalance by generating realistic synthetic samples and RAN that improves temporal feature learning through an attention mechanism. The proposed model obtains a balanced trade-off between computational efficiency and predictive precision, establishing a more efficient and generalizable solution for medical data classification.

3.6 Discussion

The experimental results demonstrate that a proposed CT-GAN-RAN approach handles class imbalance and enhances prediction performance across PIMA, heart, and Statlog datasets. The integration of CT-GAN for synthetic minority sample generation enables the model to preserve complex feature dependencies while avoiding data loss associated with undersampling and overfitting issues caused by conventional oversampling algorithms. This balanced data distribution allows the RAN classifier to learn more representative decision boundaries for the majority and minority classes. The training and validation curves across all datasets show smooth convergence with minimal divergence, representing stable learning, minimized variance, and strong generalization ability. The attention mechanism in RAN dynamically identifies clinically relevant features by enhancing class separability and improving sensitivity toward minority instances. The confusion matrix depicts the reliable classification across both classes, showing the model's capability to minimize false negatives, which is significant in medical scenarios. Additionally, ROC characteristics indicate strong discrimination ability, validating that the proposed model maintains a higher true positive rate across varying decision thresholds. Comparative analysis with the existing models shows consistent performance that is attributed to the integration of generative data balancing and attention-based temporal learning. The model introduces computational overhead because of the two-stage architecture, but the gains in stability, robustness, and minority class recognition justify the overhead. The results reveal that the proposed CT-GAN-RAN model offers scalable, reliable, and clinically relevant classification for imbalanced medical data, by achieving a balanced trade-off between prediction accuracy, sensitivity, and generalization ability.

3.7 Research Implications

The proposed CT-GAN-RAN model introduces several theoretical and practical contributions to field of medical data analysis. The implications of this research are summarized as follows; Improved data quality and balance – The integration of CT-GAN addresses the persistent problem of class imbalance in medical datasets, which ensures fair learning and enhances the model's reliability across minority and majority classes. Enhanced temporal and feature representation – RAN effectively captures long-range dependencies and dynamically focuses on significant features, while also demonstrating how the attention mechanism improves sequential learning. Generalization across medical domains –

The method shows strong adaptability and performance consistency across different datasets, indicating its potential applicability for a wide range of healthcare prediction tasks. Explainable and data-based healthcare – By integrating generative data augmentation with attention-based learning, the proposed model offers improved interpretability and supports much more informed medical decision-making. The proposed model establishes a benchmark for combining generative and attention-based methods, ensuring future exploration of hybrid DL models in clinical and biomedical research.

4. CONCLUSION

This manuscript developed a model integrating CT-GAN and RAN to address the class imbalance issue in medical datasets. CT-GAN was used in the preprocessing phase to generate synthetic minority samples while preserving complex feature correlations. Then the data was normalized by a standard scalar and classified by RAN, which integrated recurrent learning with an attention mechanism to focus on significant features. The proposed model was evaluated on the PIMA, heart, and Statlog datasets, and its effectiveness was validated by the k-fold cross-validation, statistical analysis, and computational evaluation.

The conclusion of this research was presented as follows. The CT-GAN learned joint feature distributions and generated high-quality synthetic samples, also addressing the drawbacks of conventional oversampling and undersampling algorithms.The integration of RAN improved feature representation by attention-based learning, which led to superior accuracy across all three datasets. Statistical analysis determined the significant enhancements across baseline models, with t-statistic and p-values depicting reliability and stability of a proposed model. CT-GAN-RAN introduced high computational complexity; the increase in training and inference time was moderate, but justified by improvement gains in predictive performance and model stability.

As future work, the CT-GAN-RAN model will be extended by combining explainable AI techniques to improve interpretability in clinical settings. Moreover, the model will be adapted for multi-class or multi-label to further assess scalability and deployment feasibility.

Nomenclature

Variables	Description
B	dimensionless heat source length
CP	sepecific heat, J. kg-1. K-1
g	gravitational acceleration, m.s-2
k	thermal conductivity, W.m-1. K-1
Nu	local Nusselt number along the heat source
α	thermal diffusivity, m2. s-1
β	thermal expansion coefficient, K-1
ϕ	solid volume fraction
Θ	dimensionless temperature
μ	dynamic viscosity, kg. m-1.s-1
p	nanoparticle
f	fluid (pure water)
nf	nanofluid

REFERENCES

[1] S.S. Bhat, M. Banu, G.A. Ansari, and V. Selvam, "A Risk Assessment and Prediction Framework for Diabetes Mellitus Using Machine Learning Algorithms," *Healthcare Analytics*, Vol. 4, 2023, p. 100273.

[2] M.M. Chowdhury, R.S. Ayon, and M.S. Hossain, "An Investigation of Machine Learning Algorithms and Data Augmentation Techniques for Diabetes Diagnosis Using Class Imbalanced BRFSS Dataset," *Healthcare Analytics*, Vol. 5, 2024, p. 100297.

[3] E.A. Ogundepo and W.B. Yahya, "Performance Analysis of Supervised Classification Models on Heart Disease Prediction," *Innovations in Systems and Software Engineering*, Vol. 19, 2023, pp. 129-144.

[4] G. Obaido, B. Ogbuokiri, C.W. Chukwu, F.J. Osaye, O.F. Egbelowo, M.I. Uzochukwu, I.D. Mienye, K. Aruleba, M. Primus, and O. Achilonu, "An Improved Ensemble Method for Predicting Hyperchloremia in Adults with Diabetic Ketoacidosis," *IEEE Access*, Vol. 12, 2024, pp. 9536-9549.

[5] V.V. Khanna, K. Chadaga, N. Sampathila, S. Prabhu, P.R. Chadaga, D. Bhat, and K.S Swathi, "Explainable Artificial Intelligence-Driven Gestational Diabetes Mellitus Prediction Using Clinical and Laboratory Markers," *Cogent Engineering*, Vol. 11, No. 1, 2024, p. 2330266.

[6] I. Tasin, T.U. Nabil, S. Islam, and R. Khan, "Diabetes Prediction Using Machine Learning and Explainable AI Techniques," *Healthcare Technology Letters*, Vol. 10, No. 1-2, 2023, pp. 1-10.

[7] I. Shaheen, N. Javaid, N. Alrajeh, Y. Asim, and S. Aslam, "Hi-Le and HiTCLe: Ensemble Learning Approaches for Early Diabetes Detection Using Deep Learning and Explainable Artificial Intelligence," *IEEE Access*, Vol. 12, 2024, pp. 66516-66538.

[8] K. Alnowaiser, "Improving Healthcare Prediction of Diabetic Patients Using KNN Imputed Features and Tri-Ensemble Model," *IEEE Access*, Vol. 12, 2024, pp. 16783-16793.

[9] H. Hairani and D. Priyanto, "A New Approach of Hybrid Sampling SMOTE and ENN to the Accuracy of Machine Learning Methods on Unbalanced Diabetes Disease Data," *International Journal of Advanced Computer Science and Applications*, Vol. 14, No. 8, 2023, pp. 585-590.

[10] K. Abnoosian, R. Farnoosh, and M.H. Behzadi, "Prediction of Diabetes Disease Using an Ensemble of Machine Learning Multi-Classifier Models," *BMC Bioinformatics*, Vol. 24, 2023, p. 337.

[11] G. Mulugeta, T. Zewotir, A.S. Tegegne, L.H. Juhar, and M.B. Muleta, "Classification of Imbalanced Data Using Machine Learning Algorithms to Predict the Risk of Renal Graft Failures in Ethiopia," *BMC Medical Informatics and Decision Making*, Vol. 23, 2023, p. 98.

[12] X. Wang, J. Ren, H. Ren, W. Song, Y. Yiao, Y. Zhao, L. Linghu, Y. Cui, Z. Zhao, L. Chen, and L. Qiu, "Diabetes Mellitus Early Warning and Factor Analysis Using Ensemble Bayesian Networks with SMOTE-ENN and Boruta," *Scientific Reports*, Vol. 13, No. 1, 2023, p. 12718.

[13] O. Olabanjo, A. Wusu, O. Olabanjo, M. Asokere, O. Afisi, and B. Akinnuwesi, "A Novel Deep Learning Model for Early Diabetes Risk Prediction Using Attention-Enhanced Deep Belief Networks with Highly Imbalanced Data," *International Journal of Information Technology*, Vol. 17, 2025, pp. 1933-1955.

[14] M.J. Sai, P. Chettri, R. Panigrahi, A. Garg, A. K. Bhoi, and P. Barsocchi, "An Ensemble of Light Gradient Boosting Machine and Adaptive Boosting for Prediction of Type-2 Diabetes," *International Journal of Computational Intelligence Systems*, Vol. 16, No. 1, 2023, p. 14.

[15] G. Sugendran and S. Sujatha, "Earlier Identification of Heart Disease Using Enhanced Genetic Algorithm and Fuzzy Weight Based Support Vector Machine Algorithm," *Measurement: Sensors*, Vol. 28, 2023, p. 100814.

[16] A. Abdellatif, H. Abdellatef, J. Kanesan, C.O. Chow, J.H. Chuah, and H.M. Gheni, "An Effective Heart Disease Detection and Severity Level Classification Model Using Machine Learning and Hyper Parameter Optimization Methods," *IEEE Access*, Vol. 10, 2022, pp. 79974-79985.

[17] M. Shokouhifar, M. Hasanvand, E. Moharamkhani, and F. Werner, "Ensemble Heuristic–Metaheuristic Feature Fusion Learning for Heart Disease Diagnosis Using Tabular Data," *Algorithms*, Vol. 17, No. 1, 2024, p. 34.

[18] X. Feng, Y. Cai, and R. Xin, "Optimizing Diabetes Classification with a Machine Learning-Based Framework," *BMC Bioinformatics*, Vol. 24, 2023, p. 428.

[19] P. Nandakumar and S. Narayan, "Cardiac Disease Detection Using Cuckoo Search Enabled Deep Belief Network," *Intelligent Systems with Applications*, Vol. 16, 2022, p. 200131.

[20] PIMA dataset: https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database

[21] UCI repository: https://archive.ics.uci.edu/

[22] V.K. Daliya and T.K. Ramesh, "A Cloud Based Optimized Ensemble Model for Risk Prediction of Diabetic Progression—An Azure Machine Learning Perspective," *IEEE Access*, Vol. 13, 2025, pp. 11560-11575.