Gaussian Process Regressive Accelerated Gradient Convolutional Deep Belief Neural Classification for Healthcare Data Analytics

S.Sathishkumar[1], Research Scholar, Department of Computer Science, Karuppannan Mariappan College, Tirupur, Tamilnadu, India, slsathish@outlook.com

Dr.P.Parameswari[2], Principal, Palanisamy Arts College, Erode, Tamilnadu, India, paramtech20@gmail.com

Abstract

Healthcare data analytics involves collecting, analysing, and interpreting healthcare-related data to improve patient outcomes and decision-making about patient care, treatment plans, and disease management within healthcare organisations. Early detection and effective Healthcare data analytics are crucial for preventing complications and enhancing outcomes in individuals affected by the disease. Numerous machine-learning methods have been developed to address heart problems. However, achieving higher accuracy in Healthcare data analytics with minimal time and space complexity remains challenging. The Gaussian Process Regressive Accelerated Gradient Convolutional Deep Belief Neural Classification (GPRACDBNC) Method was developed to improve the accuracy of healthcare data analytics. The GPRACDBNC Method includes data acquisition, feature selection, and classification. Several features and data samples are collected from the dataset in the acquisition phase. After data acquisition, a Convolutional Deep Belief Neural Network is employed for medical data classification. In the GPRACDBNC Method, the number of medical data samples is inputted in the visible layer. Gaussian Process Regression is then performed to select the relevant features and remove the irrelevant ones. Tversky similarity is applied to healthcare data analytics by analysing the training and testing data samples. Based on this analysis, medical data samples are accurately classified. Finally, the Nesterov Accelerated Gradient method is employed in the fine-tuning phase to minimise error and obtain better classification results. Experimental evaluation is done with the medical dataset for classification accuracy, precision, recall, F-measure, classification time, and space complexity. Quantitative analysis results indicate that the proposed GPRACDBNC Method achieves superior classification accuracy, precision, recall, and F-measure and minimises time consumption compared to existing methods.

Keywords: Healthcare Data Analytics, Convolutional Deep Belief Neural network, Gaussian Process Regression, tversky similarity, Nesterov Accelerated Gradient method

1. Introduction

Healthcare data analytics is pivotal in modern healthcare systems, offering insights that drive improved patient outcomes, operational efficiency, and cost-effectiveness. Healthcare data analytics aims to extract meaningful patterns and correlations from diverse healthcare datasets to support clinical decision-making, enhance patient care, and optimise healthcare delivery processes. Machine learning and deep learning algorithms have been used in healthcare data analytics to build predictive models that forecast patient outcomes, such as disease diagnosis, treatment response, and readmission risk. These models analyse patient data, including electronic health records (EHRs), medical imaging, genetic information, and wearable device data, to identify patterns and predict future events. However, challenges such as accuracy and time minimisation within existing healthcare systems must be addressed to realise the full benefits of this transformative technology in the healthcare domain.

A time series forecasting (TSF) based convolutional neural network (CNN) was developed in [1] to improve the performance of healthcare monitoring for coronary heart disease patients. However, the designed deep learning method was not efficient for optimal analysis of heart diseases and early prediction. A hybrid approach based on deep learning methods called CNN-Bi-LSTM was introduced in [2] to predict whether a person has heart disease and provide a diagnosis based on that prediction using feature selection. However, numerous enhancements were not investigated to increase the accuracy of this prediction system while handling large datasets with a higher number of attributes for the early diagnosis of heart disease.

Deep learning methods integrated with feature augmentation techniques were introduced in [3] to predict whether patients are at risk of cardiovascular disease. However, these methods consumed more time in disease prediction. The Denoising Auto Encoder-based Broad Learning (ABL) system was developed in [4] to improve disease prediction. However, it failed to utilise more explanatory data processing in the disease prediction. An Oversampled Quinary Feed Forward Network (OQFFN) was developed in [5] and provides a less complex framework and a more reliable notification method for accurate disease prediction. However, more data-intensive

layered deep learning-based models were not implemented for further analysis of heart diseases to determine the accurate type and cause of the disease.

An efficient machine learning model was developed in [6] based on relevant feature selection for early and accurate heart disease prediction. However, it failed to generate an effective and efficient diagnosing system. Hybrid Deep Neural Networks (HDNNs) were developed in [7] to extract and learn relevant features from the input data, thereby enhancing prediction accuracy. However, it failed to apply feature selection techniques to minimise the time complexity of disease detection. A deep convolutional neural network (DCNN) was developed in [8] to detect the possible presence of heart disease. However, the accuracy of heart disease prediction was not improved. In [9], a machine learning-based algorithm was designed to predict ischemic disease accurately. However, it failed to incorporate developing deep learning techniques for more diverse and extensive datasets to monitor critical heart patients properly. Big Data analytics in the healthcare domain was developed [10] to analyse structured and unstructured data. Machine learning-based techniques were developed in [11] for detecting early indicators of a disease. However, it failed to identify patients at a higher risk of developing unnecessary chronic diseases such as heart disease.

Machine learning and deep learning techniques were developed in [12] for the healthcare sector with big data analytics to improve patient outcomes and minimise costs. However, the accuracy, precision, and recall performance analysis remained unaddressed. An ensemble approach in machine learning was introduced in [13] for healthcare data analytics, aiming for fewer false negatives and a maximum count of true positives. However, this approach did not incorporate deep feature learning to enhance the accuracy of healthcare data analytics further. A deep convolutional neural network (DCNN) was developed in [14] to classify heart disease at an early stage accurately. However, the issue of time consumption in heart disease diagnosis remained unaddressed. A deep learning-based model was developed in [15] to achieve accurate prediction accuracy and the lowest error rate for the early diagnosis of coronary artery disease.

1.1 Main contribution of the paper

The key contributions of the GPRACDBNC Method are listed as follows,

- To enhance healthcare data analytics, the GPRACDBNC Method is developed, incorporating feature selection and classification.
- To minimise healthcare data analytics time consumption, the GPRACDBNC Method utilises Gaussian Process Regression, which is integrated into the Convolutional Deep Belief Network for selecting the most significant features.
- The GPRACDBNC Method utilises Tversky similarity within the deep learning model to analyse testing and training data samples, thus enhancing the accuracy of cardiovascular disease prediction. The Nesterov Accelerated Gradient method is also applied during the fine-tuning phase to minimise errors.
- Finally, an experimental evaluation is conducted to assess the performance of the GPRACDBNC method using various metrics and comparing it to other deep learning approaches.

1.2 organisation of paper

The paper is structured into five sections as outlined below: Section 2 reviews the related works. Section 3 elaborates on the GPRACDBNC Method with a neat diagram. Section 4 outlines the experimental setup and describes the dataset. Comparative analyses of various performance metrics are presented in Section 5. Finally, Section 6 provides the conclusion.

2. Related works

Generative Adversarial Network (GAN) and Long Short-Term Memory (LSTM) models were developed [16] to classify heart disease using ECG samples from both normal subjects and patients with heart disease as input information. However, they failed to improve the detection performance. In [17], various machine learning (ML) approaches were introduced to predict, classify, and enhance the diagnostic accuracy of cardiovascular disease prediction. However, these approaches failed to minimise the computational complexity associated with cardiovascular disease prediction.

An artificial neural network methodology was developed in [18] to identify potential cardiovascular disease risk factors based on measuring the correlations among risk attributes.

However, it failed to increase the classification accuracy for predicting heart disease. A hybrid smart medical architecture called Edge Convolutional Neural Networks (EdgeCNN) was developed in [19] to address the issue of healthcare data analytics. However, it failed to improve diagnostic accuracy effectively. The Honey Badger Optimization with Modified Deep Learning (ACVD-HBOMDL) method was developed in [20] for accurate cardiovascular disease diagnosis based on feature selection. However, it failed to improve the classification performance of cardiovascular disease prediction. The Louvain Mani-Hierarchical Fold Learning method was developed in [21] to achieve better accuracy for healthcare analytics.

An ensemble deep-learning approach was introduced in [22] to diagnose cardiovascular disease. However, it failed to consider a model incorporating various deep-learning algorithms to accurately predict cardiovascular disease more precisely. Bidirectional Long Short-Term Memory (Bi-LSTM) was developed in [23] for a smart healthcare system to monitor and accurately predict the risk of heart disease. However, precise and timely disease prediction remained a major challenging issue. An ensemble classification model was developed in [24] based on a feature selection approach for accurate heart disease diagnosis.

In [25], a decision tree-based random forest (DTRF) classifier algorithm was designed for accurate heart disease prediction. However, deep learning-based classification with machine learning feature analysis remained unaddressed. A hybrid deep learning-based approach for heart disease detection and classification was developed in [26] to achieve the maximum level of accuracy. In [27], a multilayer perceptron-based deep learning model was introduced to enhance coronary heart disease prediction at an early stage. However, the relationships between variables proved challenging for accurate heart disease prediction. Statistical and machine learning techniques were developed in [28] to automate medical data analysis. In [29], Patient Forest was developed, a machine-learned approach for predicting patient outcomes that integrates statistical features. However, it failed to determine the optimal hyperparameters of Patient Forest to enhance accuracy and performance. In [30], a modified ID3 decision-support framework was developed for healthcare disease prediction. However, achieving higher accuracy in healthcare disease prediction remained a challenging issue.

3. Methodology

Healthcare data analytics is applied across various healthcare domains, including clinical care, population health management, disease surveillance, personalised medicine, and healthcare administration. By utilising data analytics, healthcare organisations restructure workflows, reduce costs, enhance patient security, and improve patient care quality. In addition, healthcare professionals identify high-risk individuals through Healthcare data analytics to improve patient outcomes, resource allocation, and operational efficiency. A method called GPRACDBNC has been developed for healthcare data analytics concerning cardiovascular disease prediction. An elaborate description of the GPRACDBNC method is provided in the following sections.

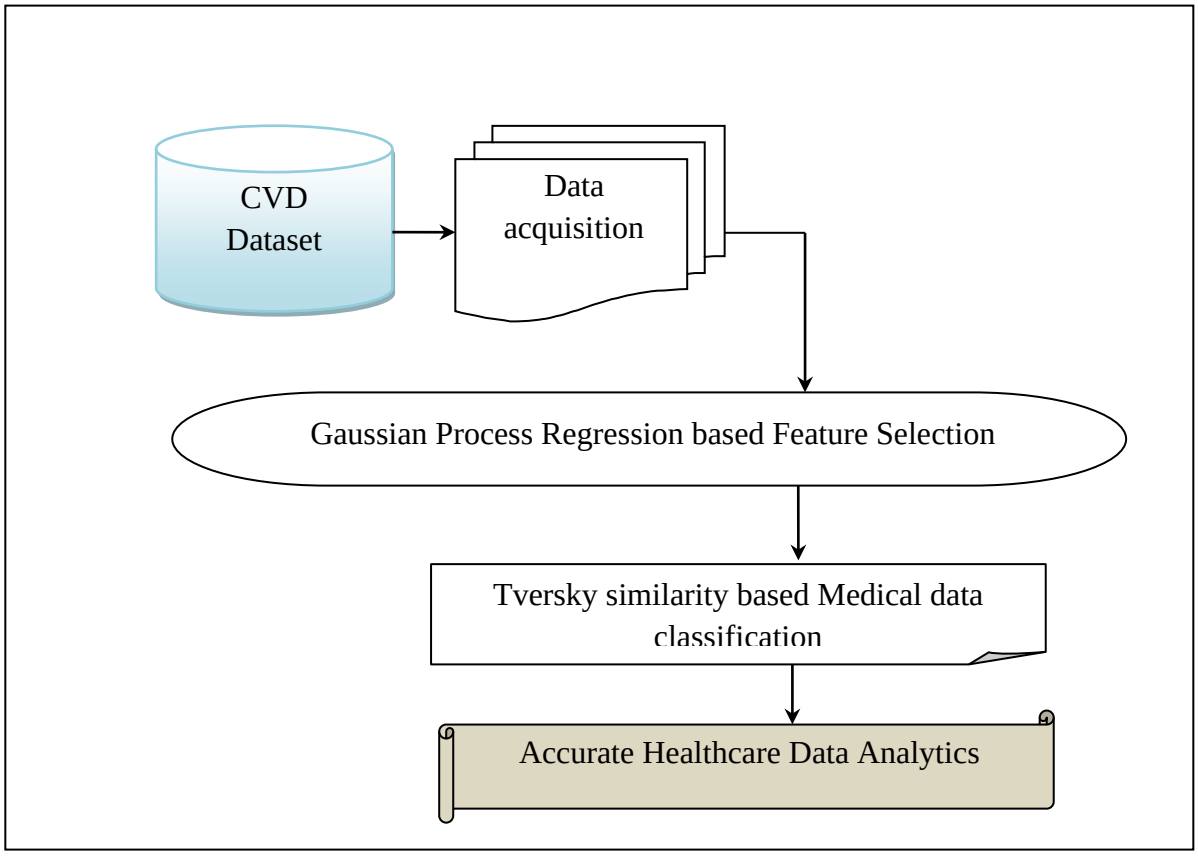

Figure 1 Architecture of proposed GPRACDBNC Method

Figure 1 above illustrates the architecture diagram of the proposed GPRACDBNC Method for Accurate Healthcare Data Analytics in predicting cardiovascular disease. This method involves three fundamental steps: data acquisition, feature selection and classification.

The following subsections briefly explain these essential processes of the GPRACDBNC Method.

3.1 Data acquisition

Data acquisition is an essential step in the proposed GPRACDBNC Method. It involves collecting numerous patient data samples from the cardiovascular disease dataset available at https://www.kaggle.com/sulianova/cardiovascular-disease-dataset. These steps effectively acquire high-quality data for developing and validating predictive cardiovascular disease risk assessment models. The primary aim of this dataset is to identify the presence or absence of cardiovascular disease. The dataset consists of 13 attributes and 70,000 instances. Table 1 below lists the feature descriptions.

Table 1 attribute description

S.no	Attributes	Description
1.	ID	Patient ID
2.	Age	Patient age in days
3.	Height	Patient height in cm
4.	Weight	Patient weight in kg
5.	Gender	1-women, 2-men
6.	ap_hi	Systolic blood pressure
7.	ap_lo	Diastolic blood pressure
8.	Cholesterol	Cholesterol 1: normal 2: above normal 3: well above normal
9.	gluc	Glucose 1: normal, 2: above normal 3: well above normal
10.	Smoke	Smoking 1: Yes 0:no
11.	alco	Alcohol intake 1: Yes 0:no
12.	active	Physical activity
13.	cardio	1 presence

		0 absence

3.2 Convolutional Deep Belief Neural Network

The proposed GPRACDBNC Method uses a Convolutional Deep Belief Neural Network (CDBN) to predict cardiovascular disease accurately. A Convolutional Deep Belief Network (CDBN) is a deep artificial neural network comprising multiple layers of convolutional restricted Boltzmann machines (CRBMs) stacked together. It is a hierarchical generative model for deep learning, making it highly effective in healthcare data analytics. The main advantage of CDBN is that it reduces computational complexity while having many units. Additionally, a CDBN effectively minimises errors in the learning process.

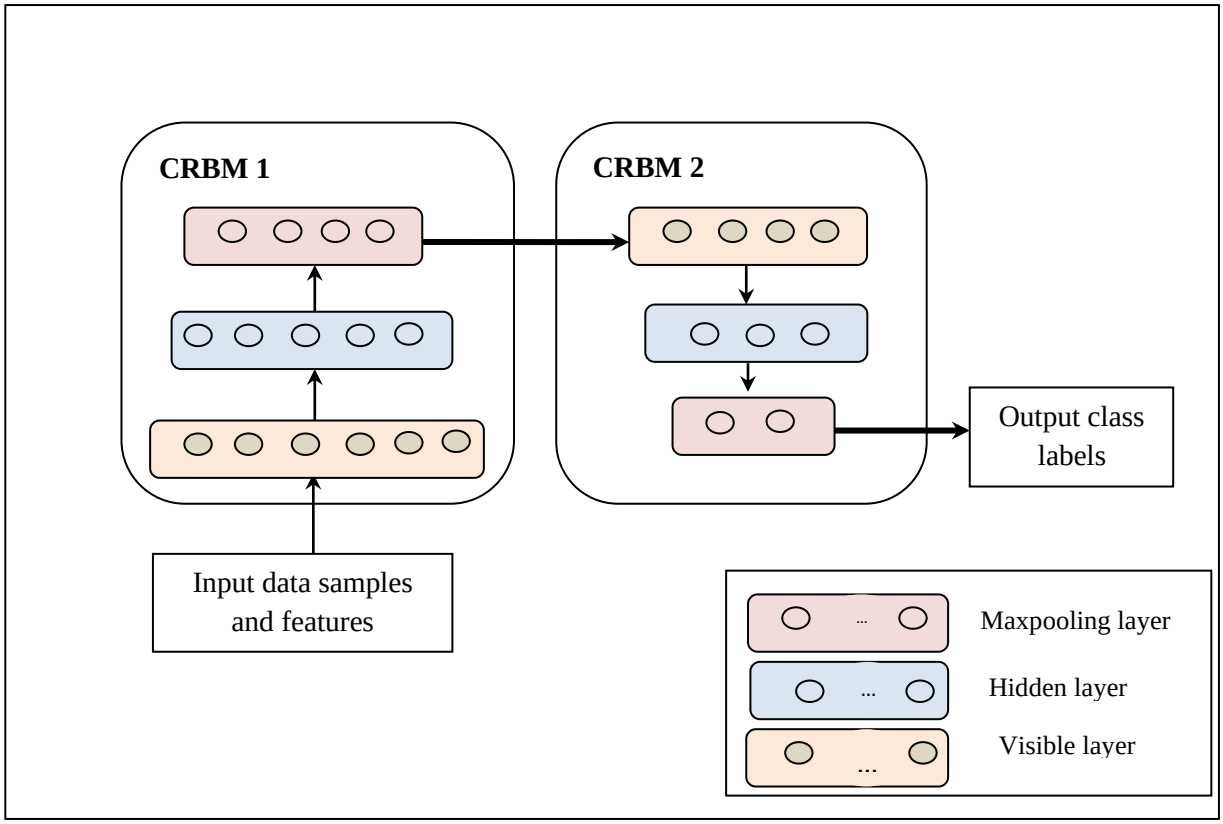

Figure 2 Structural of convolutional deep belief network

Figure 2 depicts the structure of a convolutional deep belief network for classification and deep feature learning. Two steps are involved in training a Convolutional Deep Belief Network (CDBN): layer-by-layer training and fine-tuning. The layer-by-layer training process is a fundamental component in deep learning, where each layer receives weighted input, applies a set of processes to transform it, and passes the output to the next layer. Fine-tuning, on the other hand, involves using error back-propagation algorithms to adjust the hyperparameters of the CDBN by using a Nesterov Accelerated Gradient approach after the initial training is completed.

CDBN uses Convolutional Restricted Boltzmann Machines (CRBMs), stochastic three-layered neural networks in the layer-by-layer training method. This network analyses patient data automatically by reconstructing the input. CRBMs consist of three layers: a visible layer, a hidden layer, and a max-pooling layer, represented in pink, blue, and orange. The visible layer contains neurons (nodes) that receive the input patient data with several features. Each circle represents a neuron, known as a node. The nodes are connected across layers but not within the same layer. The output of one CRBM is fed into the visible layer of the next CRBM, as shown in Figure 2.

As shown in the above figure 2, the visible layer considers that the training set $\{D, Y\}$ where D denotes training data, which includes the number of features and samples $F = \{F_1, F_2\ F_3, \dots, F_n\}$', Samples' $S = \{S_1, S_2, S_3\ , \dots, S_m\}$' collected from the dataset to be learned, and a label or output 'Y' representing its category, which belongs to the different classes such as presence or absence of cardiovascular disease.

The input matrix contains the number of features and the sample values below.

$$IM = \begin{bmatrix} F_1S_1 & F_2S_1 & \dots & F_mS_1 \\ F_1S_1 & F_2S_2 & \dots & F_mS_2 \\ \dots & \dots & \dots & \dots \\ F_mS_1 & F_mS_2 & \dots & F_mS_n \end{bmatrix} \quad m = rows, n = columns \qquad (1)$$

From the above input matrix 'IM' formulation as given in (1), 'n' column features $F = \{F_1, F_2\ F_3, \dots, F_n\}$ are present with overall sample instances $S = \{S_1, S_2, S_3\ , \dots, S_m\}$ of 'm row respectively. Where 'F' represents the features, 'S' denotes the corresponding data samples. The neuron activation probability of input visible layer '$P\ (V)$' is expressed as follows,

$$P\ (V) = \sigma(\sum IM\ * \varphi_v) + b_v \qquad (2)$$

Where, $P\ (V)$ denotes a neuron activation probability of input visible layer in CRBM, σ denotes a sigmoid activation function, 'IM' denotes an input matrix which includes features 'F' and samples 'S', φ_v weights in the visible layer, ' $*$' denotes a convolution operator, b_v denotes a bias of the visible layer. If the neuron activation probability $P\ (V) = 1$, the input is transferred into the hidden layer.

The probability that neurons in the hidden layer get activated in the hidden layer depends on the input from the visible layer and the weights connecting them.

$$P\ (H) = \sigma(\sum v\ * \varphi_{vh}) + b_h \quad (3)$$

Where, $P\ (H)$ denotes a neuron activation probability of the hidden layer in CRBM, σ denotes a sigmoid activation function, 'v' denotes an input from the visible layer, φ_{vh}weights between the visible layer and hidden layer, ' $*$' denotes a convolution operator, b_h denotes a bias of the hidden layer. If neuron activation probability $P\ (H) = 1$, then the input is transferred into the maxpooling layer. The activation probability of each neuron in the max-pooling layer is obtained by calculating the maximum activation probability '$P\ (H) = 1$' in the hidden layer.

The next layer is the max pooling layer, where the feature process is executed to reduce the input dataset's spatial dimensions through the feature selection process. It helps reduce computational complexity and memory usage. Gaussian Process Regression is applied to select the relevant features in that layer, reducing the input dataset's spatial dimensions. Gaussian Process regression is a machine learning method used to measure the relationship between the features and reduce the disease prediction's computational burden.

Gaussian process regression is a machine learning technique used to analyse the extracted feature with the testing features related to the brain tumour.

$$Q = \exp\left(\frac{|F_i - F_j|}{2\ v^2}\right) \quad (4)$$

Where Q denotes a regression output, F_i and F_j denotes a feature vector, $|F_i - F_j|$ Indicates an absolute difference between the values of two feature vectors, 'v' indicates a deviation. The regression provides the output ranges from 0 to 1.

$$Z = \begin{cases} RF, & Q > T \\ IF, & Q < T \end{cases} \quad (5)$$

Where Z denotes a max pooling output, here, the regression output Q is more significant than threshold 'T', then the feature is selected as relevant 'RF'. The regression output Q is lesser than the threshold 'T', then the feature is selected as irrelevant 'IF'. Based on a regression output, the max pooling output minimises the dimensionality of the dataset by removing the irrelevant features with their data samples and selecting the more relevant features with their data samples for accurate cardiovascular disease detection.

Then, the reduced dimensionality output is given to the next CRBM for classification. In CRBM 2, the visible layer receives the selected features set for healthcare data analytics concerning disease prediction.

The Tversky similarity measure is used to perform healthcare data analytics with training samples and testing data samples. It is formulated as follows,

$$TS = \frac{S_r \cap S_t}{\alpha\,(S_r \,\Delta\, S_t) + \beta\,(S_r \cap S_t)} \quad (6)$$

Where TS indicates a Tversky similarity coefficient, S_r denotes training data samples, S_t indicates testing data samples, $S_r \cap S_t$ indicates a mutual dependence between the training and testing data samples, $S_r \cap S_t$ indicates a variance between the training and testing data samples. From (6), α and β represent parameters of the Tversky index ($\alpha, \beta \geq 0$). The coefficient (TS) provides the resultant value between [0, 1].

$$TS = \begin{cases} 1\,; & Disease\ presence \\ 0; & Disease\ absence \end{cases} \quad (7)$$

Where, if the coefficient TS returns '1', the data samples are classified as disease presence. Otherwise, the data samples are classified as disease absence. In this way, accurate healthcare data analytics is performed.

In the fine-tuning phase, the error is measured for each classification output as the squared difference between the actual and predicted classification output as follows,

$$e = [Actual_{CR} - Predicted_{CR}]^2 \quad (8)$$

Where 'e' represents the error after the classification, $Actual_{CR}$represents the actual classification output,$Predicted_{CR}$ represents the predicted classification output. In order to minimise the error, the proposed technique utilises the Nesterov Accelerated Gradient method in the proposed technique to update the weight.

$$\varphi_n = \varphi_{old} - \eta\, b_t \quad (9)$$

$$b_t = \delta b_{t-1} + (1-\delta)\left[\frac{\partial e}{\partial \varphi_{old}}\right] \quad (10)$$

Where, φ_n denotes an updated weight, φ_{old} denotes a current weight, η denotes a learning rate($\eta < 1$). A higher learning rate allows the classifier to learn faster than the lesser value, '$\left[\frac{\partial e}{\partial \varphi_{old}}\right]$' denotes a partial derivative of the error 'e' relating to current weight 'φ_{old}', b_t Initialised to 0, common default value δ = 0.9. This process is repeated until the classification output reaches minimal error and higher accuracy. Below are the algorithmic steps for the Gaussian Process Regressive accelerated gradient Convolutional Deep Belief Neural Classification (GPRACDBNC).

//Algorithm 1: Gaussian Process Regressive accelerated gradient Convolutional Deep Belief Neural Classification
Input: Dataset 'DS', features $F = \{F_1, F_2\, F_3, \dots, F_n\}$, Samples' $S = \{S_1, S_2, S_3\, , \dots, S_m\}$' **Output:** Improve healthcare data analysis
Begin **Step 1: Collect the** features $F = \{F_1, F_2\, F_3, \dots, F_n\}$ and Samples' $S = \{S_1, S_2, S_3\, , \dots, S_m\}$' from dataset **Step 2: Input** features 'F' and Samples' S' given to visible layer **Step 3: For each** features F and Samples' S' **Step 4**: Formulate the neuron activation probability using (2) (3) **Step 5: End For** **Step 6: For each** features **Step 7:** Compute relationship using (4) **Step 8:** $\boldsymbol{If}\ (\boldsymbol{Q} > T)$ **then**

Step 9: Feature is said to be relevant

Step 10: $\boldsymbol{Else\ if}\ (\boldsymbol{Q} < T)$ **then**

Step 11: Feature is said to be irrelevant

Step 12: **End if**

Step 13: **For each** relevant feature with training samples

Step 14: **For each** testing samples

Step 15: Measure the tversky similarity using (6)

Step 16: $\boldsymbol{If}\ (\boldsymbol{TS} = \mathbf{1})$ **then**

Step 17: Samples is classified as 'disease presence'

Step 18: **else** $If\ (TS = 0)$ **then**

Step 19: Samples is classified as 'disease absence'

Step 20: **End if**

Step 21: **End for**

Step 22: **End for**

Step 23: **For each classification results**

Step 24: Compute the error rate using (8)

Step 25: Update weight based on error rate using (9) (10)

Step 26: Repeat process until find minimum error

Step 27: Obtain the final classification results **at the output layer**

Step 28: **End for**

End

Algorithm 1 describes the step-by-step process of healthcare data analytics by applying a Gaussian Process Regressive Accelerated Gradient Convolutional Deep Belief Neural Network Classification. First, the number of features and data samples are given as input to the visible layer of the GRBM. Gaussian process regression is employed to select the more relevant features and remove the irrelevant ones. The relevant features are then given as input to the visible layer of the next GRBM. The Tversky similarity is measured between the training and testing samples. Based on the similarity value, the output is either '1' or '0'. If the output is '1', the samples are correctly diagnosed as disease present. Otherwise, samples are classified as disease-absent. After classification, the error rate is calculated based on the actual and predicted output results. To

minimise this error, the Nesterov Accelerated Gradient method is applied to update the weights. This process is repeated until the error in the disease prediction process is minimised. Finally, accurate disease prediction results are obtained.

4. Experimental Setup

This section implements an experimental assessment of the proposed GPRACDBNC method and existing TSF-CNN [1] and CNN–Bi-LSTM [2] using Python, a high-level general-purpose programming language. The cardiovascular disease dataset is utilised to experiment and taken from https://www.kaggle.com/sulianova/cardiovascular-disease-dataset. This dataset includes 13 attributes or features and 70000 instances. This dataset aims to perform healthcare data analytics for cardiovascular disease presence or absence. For the experimental consideration, the number of instances or patient data ranges from 5000, 10000, 15000 ….50000.

5. Performance comparison analysis

In this section, the performance of the GPRACDBNC method and existing TSF-CNN [1] and CNN–Bi-LSTM [2] are evaluated with various metrics, classification accuracy, precision, recall, and F-measure, classification time with different numbers of patient data.

Classification accuracy: it is a metric used for medical data analytics. It is measured as the ratio of the number of samples classified accurately to the total number of samples. The formula for computing the classification accuracy is given below.

$$CA = \sum_{i=1}^{n} \frac{S_{CA}}{S_i} * 100 \qquad (11)$$

Where classification accuracy 'CA' is measured based on the number of correctly predicted.'S_{CA}' and the total number of samplesS_i' respectively. It is measured in terms of percentage (%).

Precision, also known as the positive predictive value, measures the accuracy of the positive predictions made by the classifier. It is the ratio of true positives to the total number of positive predictions (both true and false positives). The precision rate is mathematically computed as given below.

$$Pre = \frac{TP}{TP+FP} \qquad (12)$$

Where precision 'Pre' is obtained based on the true positive rate 'TP' (i.e., diseased diagnosed as it is, diseased as diseased and non-diseased as non-diseased) and the false positive rate 'FP' (i.e., non-diseased as diseased) respectively.

Recall: Recall, also known as sensitivity, measures the proportion of actual positive instances correctly identified by the classifier. It is defined as the ratio of true positives to the total number of actual positives (true positives) and false negatives. The recall rate is measured below.

$$Rec = \frac{TP}{TP+FN} \qquad (13)$$

Where recall rate 'Rec' is measured by taking into consideration true positive 'TP' and false negative 'FN' (i.e., diseased as non-diseased) respectively.

F-measure: it is also called F1-Score and is used to evaluate the performance of a classification model. The harmonic mean of precision and recall provides a single score that balances both metrics.

$$FM = 2 * \frac{Pre*Rec}{Pre+Rec} \quad (14)$$

Where F-measure 'FM' is measured by considering precision 'Pre' and recall 'Rec' respectively. High F1-Score Indicates a good balance between precision and recall, meaning the classifier performs well in identifying positive instances while minimising false positives and false negatives. A low F1 score indicates an imbalance between precision and recall, indicating that the classifier incorrectly classifies many instances as positive (low precision).

Classification time: it is the performance metric for healthcare data analytics. It is the amount of time the algorithm consumes for classifying the data samples. The overall time consumption is mathematically represented below.

$$CT = \sum_{i=1}^{n} S_i * Time\ [Class] \qquad (15)$$

Where 'CT' indicates a classification, 'S_i' indicates a medical data sample, $Time\ [Class]$ denotes an actual time consumed for classifying one sample. It is measured in terms of milliseconds (ms)

Table 2 Comparison of classification accuracy

Samples	Classification accuracy (%)		
	GPRACDBNC	TSF-CNN	CNN–Bi-LSTM
5000	99	97.24	96.22
10000	98.56	93.56	90.56
15000	97.12	92.05	90.05
20000	95.56	91.56	88.56
25000	93.45	89.05	87.05
30000	95.56	92.22	90
35000	96.22	93.5	91.55
40000	97.56	94.5	92.89
45000	97.04	93.56	91.45
50000	94.56	92.05	90.05

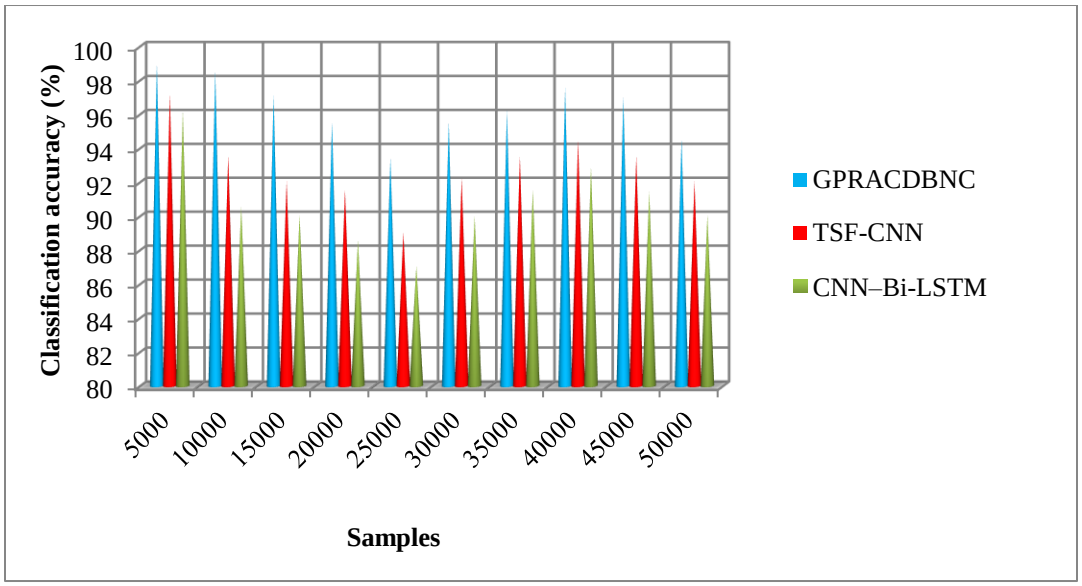

Figure 3 Performance outcomes of classification accuracy

Figure 3 depicts the performance outcomes of classification accuracy using the GPRACDBNC method and existing TSF-CNN [1] and CNN–Bi–LSTM [2]. In this figure, the horizontal axis represents the number of data samples, ranging from 5000 to 50000, while the vertical axis shows the classification accuracy. Among the three methods, the GPRACDBNC demonstrates notably improved performance compared to the existing methods. For example, with 5000 data samples considered in the first iteration, the GPRACDBNC method achieved an

accuracy of 99%. In comparison, the existing models [1] and [2] showed accuracies of 97.24% and 96.22%, respectively. Ten different results were observed for each method with varying data samples. These results were compared to evaluate the performance of the GPRACDBNC method against the existing methods. The overall comparative analysis indicates that the classification accuracy of the GPRACDBNC method improved significantly by 4% and 6% compared to [1] and [2], respectively. This improvement is achieved due to the application of the Convolutional Deep Belief Neural Network, which enhances the accuracy of cardiovascular disease prediction. The Tversky similarity is measured for each training and testing sample in the given dataset. Based on these similarity values, the GPRACDBNC method distinguishes between disease presence and disease absence samples, enhancing accuracy.

Table 3 Comparison of precision

Samples	Precision		
	GPRACDBNC	TSF-CNN	CNN–Bi-LSTM
5000	0.994	0.966	0.945
10000	0.975	0.942	0.922
15000	0.967	0.933	0.915
20000	0.952	0.925	0.905
25000	0.94	0.911	0.89
30000	0.953	0.925	0.887
35000	0.947	0.927	0.905
40000	0.967	0.935	0.911
45000	0.974	0.942	0.918
50000	0.983	0.933	0.905

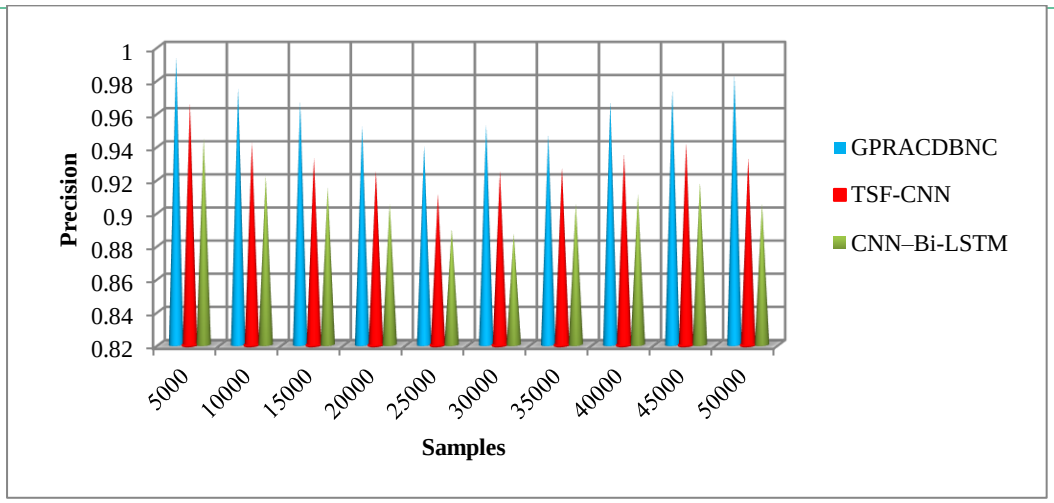

Figure 4 Performance outcomes of precision

In Figure 4, the performance outcomes of precision are depicted against the number of data samples, ranging from 5000 to 50000. Three methods, namely the GPRACDBNC method and existing TSF-CNN [1] and CNN-Bi-LSTM [2], are utilised to evaluate precision during the classification. The horizontal axis represents the number of data samples, while the vertical axis represents precision. The results demonstrate that the GPRACDBNC method achieves higher precision than the other two conventional deep learning methods. It is evident from the experiment where 50,000 samples were taken from the dataset. By applying the GPRACDBNC method, precision performance using the GPRACDBNC method [1], [2] was found to be 0.994, 0.966, and 0.945, respectively. Various results were observed for each method with different counts of input samples. The observed results of the GPRACDBNC method are compared with the existing deep learning methods. The overall comparison reveals that the precision in accurately detecting disease is enhanced by 3% compared to [1] and 6% compared to [2] when applying the GPRACDBNC method. This improved performance is achieved by utilising the Convolutional Deep Belief Neural Network, which utilises the Tversky similarity for analysing the training and testing samples in the given dataset. In addition, the Nesterov Accelerated Gradient method also minimises the error in the output class labels, enhancing accuracy with a better true positive rate and minimising false positives during cardiovascular disease prediction.

Table 4 Comparison of recall

Samples	Recall		
	GPRACDBNC	TSF-CNN	CNN–Bi-LSTM
5000	0.995	0.974	0.954
10000	0.974	0.955	0.925
15000	0.956	0.932	0.911
20000	0.932	0.915	0.895
25000	0.928	0.906	0.887
30000	0.92	0.903	0.896
35000	0.936	0.900	0.874
40000	0.955	0.925	0.893
45000	0.974	0.945	0.911
50000	0.965	0.935	0.907

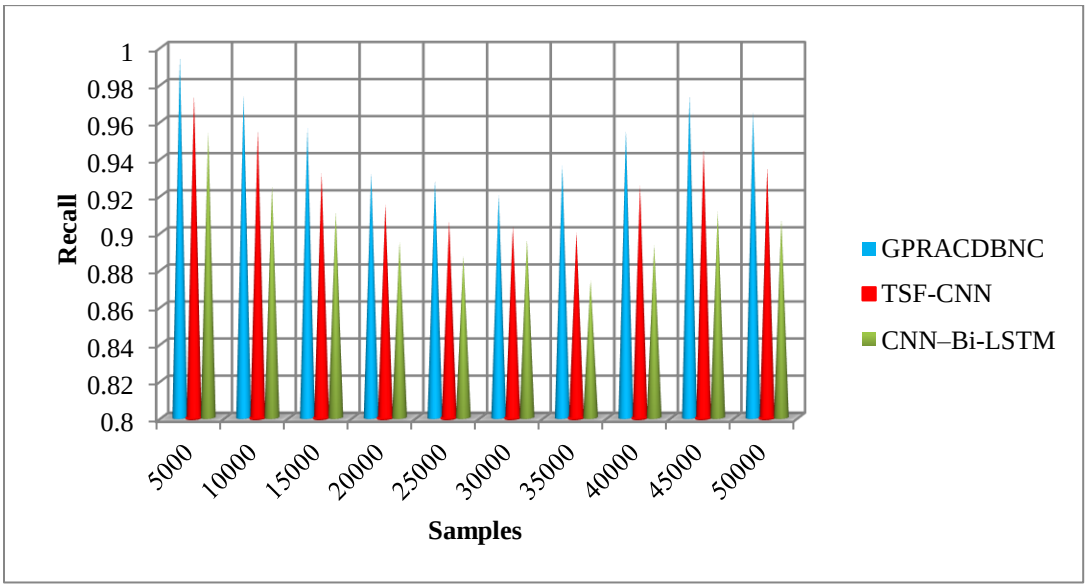

Figure 5 Performance outcomes of recall

Figure 5 presents the performance outcomes of recall versus the number of training samples ranging from 5000 to 10000 for three methods: the GPRACDBNC method existing TSF-CNN [1] and CNN–Bi-LSTM [2]. These methods are used to evaluate recall. The horizontal axis denotes the number of data samples, while the vertical axis represents recall performance. The GPRACDBNC method performs comparatively better in achieving recall than [1] and [2], as shown through statistical analysis. For instance, with 5000 data samples, the

GPRACDBNC method achieved a recall performance of 0.995, while the existing methods [1] and [2] achieved 0.974 and 0.954, respectively. Comparing the GPRACDBNC method with the existing methods, the recall performance is improved by 3% and 5% over [1] and [2], respectively. The proposed deep learning classifier employed in the GPRACDBNC method is the squared difference between the actual and predicted outputs, thereby reducing false negative rates in classifying disease presence or absence in the classification process.

Table 5 Comparison of F-Measure

Samples	F-Measure		
	GPRACDBNC	TSF-CNN	CNN–Bi-LSTM
5000	0.994	0.969	0.949
10000	0.974	0.948	0.923
15000	0.961	0.932	0.912
20000	0.941	0.919	0.899
25000	0.933	0.908	0.888
30000	0.936	0.913	0.891
35000	0.941	0.913	0.889
40000	0.960	0.929	0.901
45000	0.974	0.943	0.914
50000	0.973	0.933	0.905

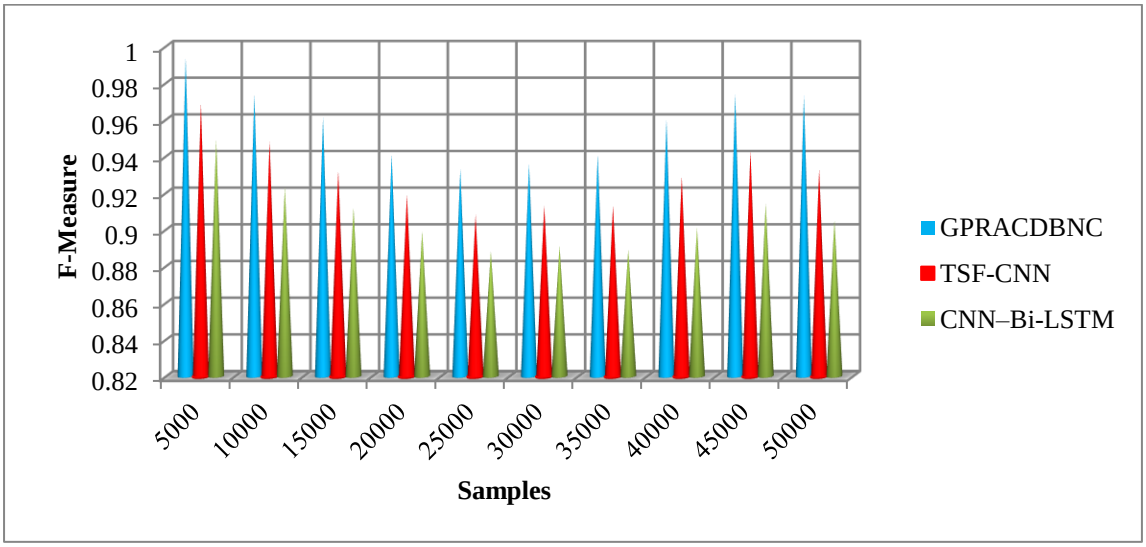

Figure 6 Performance outcomes of F-Measure

Figure 6 illustrates the performance analysis of the F-measure versus the number of data samples taken in the ranges from 5000 to 50000 from the cardiovascular dataset. The F-measure is calculated based on precision and recall performance. The observed results indicate that the GPRACDBNC method provides improved performance in the F-measure analysis compared to conventional deep learning methods. In the first iteration, 5000 data samples were considered to evaluate the performance analysis of the F-measure. The GPRACDBNC method achieved a precision of 0.994 and a recall of 0.995, resulting in an overall F-measure of 0.994. TSF-CNN [1] achieved a precision of 0.966 and a recall of 0.974, leading to an overall F-measure of 0.969. CNN–Bi-LSTM [2] achieved a precision of 0.945 and a recall of 0.954, resulting in an overall F-measure of 0.949. This quantitative analysis indicates that the GPRACDBNC method outperformed the existing methods in achieving a higher F-measure in healthcare data sample classification. On average, across ten comparisons, the GPRACDBNC method's F-measure improved by approximately 3% and 6% compared to [1] and [2], respectively.

Table 6 Comparison of Classification Time

Samples	Classification time (ms)		
	GPRACDBNC	TSF-CNN	CNN–Bi-LSTM
5000	38	47	65
10000	45	59	72
15000	57	67	86
20000	66	79	94
25000	78	91	112
30000	87	107	136
35000	95	121	148
40000	115	145	157
45000	127	155	176
50000	146	178	210

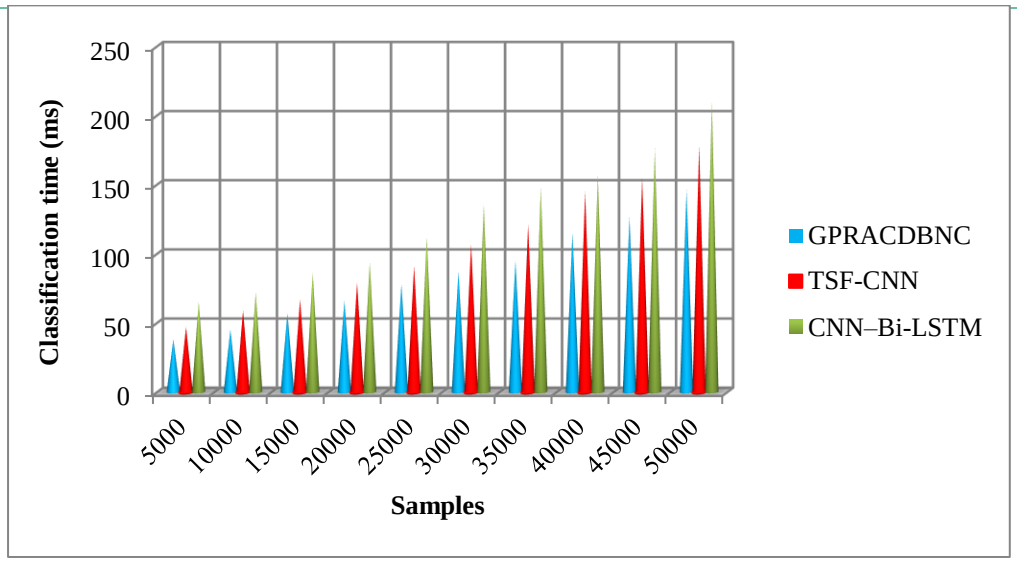

Figure 7 Performance outcomes of Classification time

Figure 7 illustrates the performance outcomes of classification time using three different methods. The figure directly correlates with the number of data samples involved in experiments and the classification—the overall time increases while the number of data samples increases. However, the GPRACDBNC method demonstrates minimised classification time compared to existing deep learning techniques. Consider 5000 data samples for experimentation. The time consumption for computing one data sample is $0.0076ms$ using the GPRACDBNC method, resulting in an overall time consumption of $38ms$. For TSF-CNN [1], the time consumption for computing one data sample is $0.0094\ ms$, with an overall time consumption found to be $47ms$. Using CNN-Bi-LSTM [2], the time consumption for computing one data sample is $0.013ms$, leading to an overall time consumption of $65\ ms$. Different counts of input data samples yield various performance results for all methods. The overall results indicate that the classification time of the GPRACDBNC method is considerably reduced by 19% and 33% compared to [1] and [2], respectively. This improvement is achieved due to the application of significant feature selection processes. The Gaussian Process Regression (GPR) is employed in the GPRACDBNC method to select the most relevant features and remove the irrelevant ones by setting a threshold value. Subsequently, classification is performed with these relevant features and their data samples, reducing the classification time.

6. Conclusion

In this paper, an improved healthcare data analytics system based on a deep learning method is proposed, aimed at enhancing the medical surveillance of patients. The GPRACDBNC method performs relevant feature selection from the dataset using Gaussian Process Regression, resulting in minimised time consumption for patient data sample analysis. Following this, a Convolutional Deep Belief Neural Network utilises the Tversky similarity for analysing the training and testing samples and classifies the results as either disease presence or absence. Based on this analysis, accurate healthcare data analytics are performed. A comprehensive experimental evaluation uses various performance metrics such as classification accuracy, precision, recall, F-measure, and classification time concerning the number of data samples. The overall performance results illustrate that the proposed GPRACDBNC method achieved higher accuracy, precision, recall, and F-measure with minimal time compared to conventional deep learning methods.

References

[1] Shambhu Bhardwaj, Vipul Vekariya, Baldev Singh, Sri Vinay, Alli Arul, Maria Daya Roopa, "Improved healthcare monitoring of coronary heart disease patients in time-series fashion using deep learning model", Measurement: Sensors, Elsevier, Volume 32, 2024, Pages 1-8. https://doi.org/10.1016/j.measen.2024.101053

[2] Prashant Kumar Shrivastava, Mayank Sharm, Pooja Sharma, Avenash Kumar, "HCBiLSTM: A hybrid model for predicting heart disease using CNN and BiLSTM algorithm", Measurement: Sensors, Elsevier, Volume 25, 2023, Pages 1-7. https://doi.org/10.1016/j.measen.2022.100657

[3] María Teresa García-Ordás, Martín Bayón-Gutiérrez, Carmen Benavides, Jose Aveleira-Mata & José Alberto Benítez-Andrades, "Heart disease risk prediction using deep learning techniques with feature augmentation", Multimedia Tools and Applications, Springer, Volume 82, 2023, Pages 31759–31773. https://doi.org/10.1007/s11042-023-14817-z

[4] Haewon Byeon, Prashant GC, Shaikh Abdul Hannan, Faisal Yousef Alghayadh, Arsalan Muhammad Soomar, Mukesh Soni, Mohammed Wasim Bhatt, "Deep Neural network model for

enhancing disease prediction using auto encoder based broad learning", SLAS Technology, Elsevier, Volume 29, Issue 3, 2024, Pages 1-12. https://doi.org/10.1016/j.slast.2024.100145

[5]Md. Ishan Arefn Hossain, Anika Tabassum, Zia Ush Shamszama, "Deep edge intelligence-based solution for heart failure prediction in ambient assisted living", Discover Internet of Things, Springer, Volume 3, 2023, Pages 1-17. https://doi.org/10.1007/s43926-023-00043-4

[6] Farhat Ullah , Xin Chen , Khairan Rajab , Mana Saleh Al Reshan , Asadullah Shaikh, Muhammad Abul Hassan , Muhammad Rizwan , and Monika Davidekova, "An Efficient Machine Learning Model Based on Improved Features Selections for Early and Accurate Heart Disease Predication", Computational Intelligence and Neuroscience, Hindawi, Volume 2022, July 2022, Pages 1-12. https://doi.org/10.1155/2022/1906466

[7] Mana Saleh A Reshan, Samina Amin, Muhammad Ali Zeb, Adel Sulaiman, Hani Alshahrani, and Asadullah Shaikh, "A Robust Heart Disease Prediction System Using Hybrid Deep Neural Networks", IEEE Access, Volume 11, 2023, Pages 121574 – 121591. **DOI:** 10.1109/ACCESS.2023.3328909

[8] Sadia Arooj, Saif ur Rehman, Azhar Imran, Abdullah Almuhaimeed, A. Khuzaim Alzahrani and Abdulkareem Alzahrani, "A Deep Convolutional Neural Network for the Early Detection of Heart Disease", Biomedicines, Volume 10, Issue 11, 2022, Pages 1-15. https://doi.org/10.3390/biomedicines10112796

[9] Ghulam Muhammad, Saad Naveed, Lubna Nadeem, Tariq Mahmood, Amjad R. Khan, Yasar Amin and Saeed Ali Omer Bahaj, "Enhancing Prognosis Accuracy for Ischemic Cardiovascular Disease Using K Nearest Neighbor Algorithm: A Robust Approach", IEEE Access, Volume 11, 2023, Pages 97879 – 97895. **DOI:** 10.1109/ACCESS.2023.3312046

[10] Kornelia Batko and Andrzej Ślęzak, "The use of Big Data Analytics in Healthcare", Journal of Big Data, Springer, Volume 9, 2022, Pages 1-24. https://doi.org/10.1186/s40537-021-00553-4

[11] Mohd Javaid, Abid Haleem, Ravi Pratap Singh, Rajiv Suman, Shanay Rab, "Significance of machine learning in healthcare: Features, pillars and applications", International Journal of

Intelligent Networks, Elsevier, Volume 3, 2022, Pages 58-73. https://doi.org/10.1016/j.ijin.2022.05.002

[12] Juli Kumari, Ela Kumar & Deepak Kumar, "A Structured Analysis to Study the Role of Machine Learning and Deep Learning in The Healthcare Sector with Big Data Analytics", Archives of Computational Methods in Engineering, Springer, Volume 30, 2023, Pages 3673–3701. https://doi.org/10.1007/s11831-023-09915-y

[13] Deepali Pankaj Javale, Sharmishta Suhas Desai, "Machine learning ensemble approach for healthcare data analytics", Indonesian Journal of Electrical Engineering and Computer Science, Volume 28, Issue 2, 2022, Pages 926-933. DOI: 10.11591/ijeecs.v28.i2.pp926-933

[14] Sadia Arooj, Saif ur Rehman, Azhar Imran, Abdullah Almuhaimeed, A. Khuzaim Alzahrani and Abdulkareem Alzahrani, "A Deep Convolutional Neural Network for the Early Detection of Heart Disease", Biomedicines, Volume 10, Issue 11, 2022, Pages 1-15. https://doi.org/10.3390/biomedicines10112796

[15] Varun Sapra, Luxmi Sapra, Akashdeep Bhardwaj, Salil Bharany, Akash Saxena, Faten Khalid Karim, Sara Ghorashi, Ali Wagdy Mohamed, "Integrated approach using deep neural network and CBR for detecting severity of coronary artery disease", Alexandria Engineering Journal, Elsevier, Volume 68, 2023, Pages 709-720. https://doi.org/10.1016/j.aej.2023.01.029

[16] Adyasha Rath, Debahuti Mishra, Ganapati Panda, Suresh Chandra Satapathy, "Heart disease detection using deep learning methods from imbalanced ECG samples", Biomedical Signal Processing and Control, Elsevier, Volume 68, July 2021, Pages 1-11. https://doi.org/10.1016/j.bspc.2021.102820

[17] Osman Taylan, Abdulaziz S. Alkabaa, Hanan S. Alqabbaa, Esra Pamukçu and Víctor Leiva, "Early Prediction in Classification of Cardiovascular Diseases with Machine Learning, Neuro-Fuzzy and Statistical Methods", Biology, Volume 12, Issue 1, 2023, Pages 1-31. https://doi.org/10.3390/biology12010117

[18] Jyotismita Talukdar and Thipendra P. Singh, "Early prediction of cardiovascular disease using artificial neural network", Paladyn, Journal of Behavioral Robotics, Volume 14, Issue 1, 2023, Pages 1-10. https://doi.org/10.1515/pjbr-2022-0107

[19] Yan He, Bin Fu, Jian Yu, Renfa Li and Rucheng Jiang, "Efficient Learning of Healthcare Data from IoT Devices by Edge Convolution Neural Networks", Applied Sciences, Volume 10, Issue 24, Pages 1-19. https://doi.org/10.3390/app10248934

[20] Marwa Obayya, Jamal M. Alsamri, Mohammed Abdullah Al-Hagery, Abdullah Mohammed, And Manar Ahmed Hamza, "Automated Cardiovascular Disease Diagnosis Using Honey Badger Optimization With Modified Deep Learning Model", IEEE Access, Volume 11, 2023, Pages 64272 – 64281. **DOI:** 10.1109/ACCESS.2023.3286661

[21] Sarah Shafqat, Maryyam Fayyaz, Hasan Ali Khattak, Muhammad Bilal, Shahid Khan, Osama Ishtiaq, Almas Abbasi, Farzana Shafqat, Waleed S. Alnumay & Pushpita Chatterjee, "Leveraging Deep Learning for Designing Healthcare Analytics Heuristic for Diagnostics", Neural Processing Letters, Springer, Volume 55, 2023, Pages 53–79. https://doi.org/10.1007/s11063-021-10425-w

[22] David Opeoluwa Oyewola, Emmanuel Gbenga Dada, Sanjay Misra, "Diagnosis of Cardiovascular Diseases by Ensemble Optimisation Deep Learning Techniques", International Journal of Healthcare Information Systems and Informatics, Volume 19, Issue 1, 2024, Pages 1–21. DOI: 10.4018/IJHISI.334021

[23] A Angel Nancy, Dakshanamoorthy Ravindran, P M Durai Raj Vincent, Kathiravan Srinivasan and Daniel Gutierrez Reina, "IoT-Cloud-Based Smart Healthcare Monitoring System for Heart Disease Prediction via Deep Learning", Electronics, Volume 11, Issue 15, 2022, Pages 1–19. https://doi.org/10.3390/electronics11152292

[24] Jafar Abdollahi & Babak Nouri-Moghaddam, "A hybrid method for heart disease diagnosis utilising feature selection based ensemble classifier model generation", Iran Journal of Computer Science, Springer, Volume 5, 2022, Pages 229–246. https://doi.org/10.1007/s42044-022-00104-x

[25] Anil Pandurang Jawalkar, Pandla Swetcha, Nuka Manasvi, Pakki Sreekala, Samudrala Aishwarya, Potru Kanaka Durga Bhavani & Pendem Anjani, "Early prediction of heart disease with data analysis using supervised learning with stochastic gradient boosting", Journal of Engineering and Applied Science, Springer, Volume 70, 2023, Pages 1-18. https://doi.org/10.1186/s44147-023-00280-y

[26] Dwarakanath B., Latha M., Annamalai R., Jagadish S. Kallimani, Ranjan Walia, and Birhanu Belete, "A Novel Feature Selection with Hybrid Deep Learning Based Heart Disease Detection and Classification in the e-Healthcare Environment", Computational Intelligence and Neuroscience, Hindawi, Volume 2022, September 2022, Pages 1-12. https://doi.org/10.1155/2022/1167494

[27] Nancy Masih, Huma Naz, Sachin Ahuja, "Multilayer perceptron based deep neural network for early detection of coronary heart disease", Health and Technology, Springer, Volume 11, 2021, Pages 127–138. https://doi.org/10.1007/s12553-020-00509-3

[28] Sarinder Kaur Dhillon, Mogana Darshini Ganggayah, Siamala Sinnadurai, Pietro Lio and Nur Aishah Taib, “Theory and Practice of Integrating Machine Learning and Conventional Statistics in Medical Data Analysis”, Diagnostics, Volume 12, Issue 10,2022, Pages 1-25. https://doi.org/10.3390/diagnostics12102526

[29] Atieh Khodadadi, Nima Ghanbari Bousejin, Soheila Molaei, Vinod Kumar Chauhan, Tingting Zhu and David A. Clifton, "Improving Diagnostics with Deep Forest Applied to Electronic Health Records", Sensors, Volume 23, Issue 14, 2023, Pages 1-15. https://doi.org/10.3390/s23146571

[30] Arun Agarwal, Khushboo Jain & Rakesh Kumar Yadav, "A mathematical model based on modified ID3 algorithm for healthcare diagnostics model", International Journal of System Assurance Engineering and Management, Springer, Volume 14, 2023, Pages 2376–2386. https://doi.org/10.1007/s13198-023-02086-w

1. Classification accuracy

$$CA = \sum_{i=1}^{n} \frac{S_{CA}}{S_i} * 100$$

$$CA\ (\text{GPRACDBNC}) = \frac{4950}{5000} * 100 = 99\%$$

$$CA\ (\text{TSF} - \text{CNN}) = \frac{4862}{5000} * 100 = 97.24\%$$

$$CA\ (\text{CNN} - \text{Bi} - \text{LSTM}) = \frac{4811}{5000} * 100 = 96.22\%$$

Samples	Classification accuracy (%)		
	GPRACDBNC	**TSF-CNN**	**CNN–Bi-LSTM**
5000	99	97.24	96.22
10000	98.56	93.56	90.56
15000	97.12	92.05	90.05
20000	95.56	91.56	88.56
25000	93.45	89.05	87.05
30000	95.56	92.22	90
35000	96.22	93.5	91.55
40000	97.56	94.5	92.89
45000	97.04	93.56	91.45
50000	94.56	92.05	90.05

Precision

$$Pre = \frac{TP}{TP + FP}$$

$$Pre\ (\text{GPRACDBNC}) = \frac{3200}{3200 + 18} = 0.994$$

$$Pre\ (\text{TSF} - \text{CNN}) = \frac{2300}{2300 + 80} = 0.966$$

$$Pre\ (\text{CNN–Bi}-\text{LSTM}) = \frac{2100}{2100+120} = 0.945$$

Samples	Precision		
	GPRACDBNC	**TSF-CNN**	**CNN–Bi-LSTM**
5000	0.994	0.966	0.945
10000	0.975	0.942	0.922
15000	0.967	0.933	0.915
20000	0.952	0.925	0.905
25000	0.94	0.911	0.89
30000	0.953	0.925	0.887
35000	0.947	0.927	0.905
40000	0.967	0.935	0.911
45000	0.974	0.942	0.918
50000	0.983	0.933	0.905

2. Recall

$$Rec = \frac{TP}{TP+FN}$$

$$Rec\ (\text{GPRACDBNC}) = \frac{3200}{3200+15} = 0.995$$

$$Rec\ (\text{TSF}-\text{CNN}) = \frac{2300}{2300+60} = 0.974$$

$$Rec\ (\text{CNN–Bi}-\text{LSTM}) = \frac{2100}{2100+100} = 0.954$$

Samples	**Recall**		
	GPRACDBNC	**TSF-CNN**	**CNN–Bi-LSTM**
5000	0.995	0.974	0.954
10000	0.974	0.955	0.925

15000	0.956	0.932	0.911
20000	0.932	0.915	0.895
25000	0.928	0.906	0.887
30000	0.92	0.903	0.896
35000	0.936	0.9	0.874
40000	0.955	0.925	0.893
45000	0.974	0.945	0.911
50000	0.965	0.935	0.907

4. F-measure:

$$FM = 2 * \frac{Pre * Rec}{Pre + Rec}$$

$$FM(\text{GPRACDBNC}) = 2 * \frac{0.994 * 0.995}{0.994 + 0.995} = 0.994$$

$$FM(\text{TSF} - \text{CNN}) = 2 * \frac{0.966 * 0.974}{0.966 + 0.974} = 0.969$$

$$FM(\text{CNN} - \text{Bi} - \text{LSTM}) = 2 * \frac{0.945 * 0.954}{0.945 + 0.954} = 0.949$$

Samples	F-Measure		
	GPRACDBNC	**TSF-CNN**	**CNN–Bi-LSTM**
5000	0.994	0.969	0.949
10000	0.974	0.948	0.923
15000	0.961	0.932	0.912
20000	0.941	0.919	0.899
25000	0.933	0.908	0.888
30000	0.936	0.913	0.891
35000	0.941	0.913	0.889
40000	0.960	0.929	0.901
45000	0.974	0.943	0.914
50000	0.973	0.933	0.905

5. Classification time

$$CT = \sum_{i=1}^{n} S_i * Time\ [Class]$$

$$CT\ (\text{GPRACDBNC}) = 5000 * 0.0076ms = 38ms$$

$$CT\ (\text{TSF} - \text{CNN}) = 5000 * 0.0094ms = 47ms$$

$$CT\ (\text{CNN} - \text{Bi} - \text{LSTM}) = 5000 * 0.013ms = 65ms$$

Samples	Classification time (ms)		
	GPRACDBNC	**TSF-CNN**	**CNN–Bi-LSTM**
5000	38	47	65
10000	45	59	72
15000	57	67	86
20000	66	79	94
25000	78	91	112
30000	87	107	136
35000	95	121	148
40000	115	145	157
45000	127	155	176
50000	146	178	210