ORIGINAL ARTICLE

Revitalizing Disease Prediction: Modified Backpropagation and Reformed Feature Extraction Approaches for Classification and Regression of Disease

Jasmine Christabel G [1,2*], A.C. Subhajini [1]

[1] *Department of Computer Application, Noorul Islam Center for Higher Education, Kumaracoil, India*

[2] *Annai Velankanni College, Tholayavattam, India*

*Corresponding Author: Jasmine Christabel G

Email: jasminejcs@gmail.com

Received: 15 June 2024 / Accepted: 06 April 2025

Abstract

Purpose: Diseases are unavoidable because of environmental factors, changes in diet, hereditary issues and many other factors; hence, it is important to detect diseases via various techniques in the healthcare sector to identify and diagnose the disease. Therefore, the proposed model focuses on employing advanced techniques for detecting heart disease, thyroid disease, and hepatitis, as these diseases have become common in recent years, along with the prediction of heart rate.

Materials and Methods: The proposed work employs modified PCA (principal component analysis) for dimensionality reduction to extract appropriate features for the model by utilizing two learning rates (LR1 and L2). Furthermore, the modified back propagation (BP) method is used for effective classification of heart, thyroid, hepatitis, and heart rate prediction by incorporating adaptive Gaussian white noise (AWGN). In the proposed model, three different datasets are utilized: a heart disease dataset, a thyroid dataset, a hepatitis dataset for classification, and a heart rate prediction dataset for regression.

Results: The accuracy, precision, recall, and F1 scores obtained by the proposed model for the heart disease dataset are 97.8%, 98%, 98%, and 98%, respectively. Similarly, 97.2%, 98%, 89%, and 93% for the thyroid dataset, respectively. Finally, the accuracy, precision, recall, and F1 score obtained by the proposed model for hepatitis are 95%, 98%, 88%, and 92%, respectively. Like the classification of diseases, heart rate prediction was also evaluated via different metrics, such as the RMSE, MSE, MAE, and R2. The MAE obtained by the proposed model for the heart rate prediction dataset is 0.112; likewise, the R2 obtained is 0.99, the MSE attained is 0.022, and the RMSE value obtained is 0.1488.

Conclusion: The results of the proposed mechanism reflect its ability to detect different diseases effectively. This is due to the successful implementation of advanced AI approaches in the proposed framework.

Keywords: Disease Prediction; Principal Component Analysis; Backpropagation; Adaptive Gaussian White Noise; Heart Disease; Thyroid Hepatitis.

DOI:

1. Introduction

Disease is inevitable. Therefore, it is important to take care of the body to avoid disease, as it can lead to life-threatening situations. However, humans can't avoid disease completely. Instead, early detection and diagnosis of disease can help patients overcome these problems and lead to a healthy life; hence, analysis of disease has become a vital factor in the healthcare sector. Furthermore, to detect diseases, various manual methods can be employed; however, manual detection of disease can be incorrect [1], inaccurate, time-consuming, and prone to errors due to human intervention. Therefore, in recent years, various AI techniques have been utilized for improved and effective detection of diseases. Some of the common diseases identified in recent years include heart disease, thyroid disease, and hepatitis. Hence, it is extremely important to detect and diagnose these diseases as early as possible. Heart disease, which is also termed Cardiovascular Disease (CVD), involves different conditions that impact the heart and lead to life-threatening situations; therefore, timely and reliable approaches must be taken to accomplish prompt management of the disease. Therefore, the suggested study has employed various ML techniques, such as NB (Naïve Bayes), DT (Decision Tree), K-NN (k-nearest neighbor), and RF (Random Forest). The dataset employed in the existing study consisted of 303 instances and 76 attributes. Among the 76 attributes, only 14 were used for testing. However, the experimental results revealed that the KNN model delivered better outcomes than the existing models did [2]. Similarly, the recommended paper has employed different Machine Learning (ML) and Deep Learning (DL) techniques for the classification of heart diseases, which include algorithms such as the support vector machine (LR), support vector machine (KNN), support vector machine (NB), Multilayer Perceptron (MLP), Artificial Neural Network (ANN) and deep neural network (DNN), as 31% of global deaths are due to heart-associated diseases. The experimental outcome revealed that the RF algorithm delivered better results in terms of accuracy than existing models did, and the evolution was performed via different evaluation metrics, such as the RMSE, precision, accuracy, and recall [3].

Like heart disease, thyroid disease can also lead to life-threatening situations [4] if it is not identified early and diagnosed accordingly [5]. Thyroid disease arises from the abnormal growth of thyroid tissues at the glands of the thyroid gland. Hence, early detection of the thyroid is extremely crucial. Therefore, various ML techniques have been employed for the detection and diagnosis of thyroid disease. Thus, different ML classifiers, such as NB, SVM, LR, DT, and KNN, were used for the identification of TD (thyroid disease) in the body. Furthermore, the dataset used in the suggested study was obtained from DHQ teaching hospitals. The employed dataset is considered unique among the existing datasets because of the incorporation of additional features such as Body Mass Index (BMI), Blood Pressure (BP), and pulse rate. The experimental results revealed that compared with the existing classifiers, the KNN, LR, and NB classifiers achieved satisfactory outcomes [6]. Many people across the world are affected by the thyroid. Hence, efficient prediction and classification techniques should be used for the timely identification of thyroid diseases. Therefore, various ML and DL classifiers, such as the DT, RF, KNN, and ANN models, are used for the timely detection of TD. However, from the experimental outcome, RF delivered better accuracy than the existing classifiers did, and the models were assessed via different metrics [7].

Another lethal and fatal disease that can affect the life of humans is hepatitis. There are various types of hepatitis, such as hepatitis A, B, C, D, and E. However, a previous study focused on hepatitis E, as it can lead to serious diseases, including liver cancer and even death. Hepatitis E, which is an acute liver disease, can lead to painful consequences. Therefore, the authors suggested that the authors employ ML and DL techniques for the detection of hepatitis. The Autoregressive Integrated Moving Average (ARIMA), long short-term memory (LSTM) and Support Vector Machine (SVM) methods were used for the identification of hepatitis, in which the LSTM model outperformed the ARIMA and Support Vector Machine (SVM) models for the prediction of hepatitis by assessing the efficiency of the model via various metrics, such as the root mean square error (RMSE), Mean Absolute Error (MAPE) and Mean Absolute Error (MAE) [8]. Similar to hepatitis E, the recommended method focuses on detecting hepatitis

C. Hepatitis C occurs primarily because the HC virus affects the immune system of the body. HCV spreads primarily via blood contact with infected individuals. Therefore, the detection of hepatitis is crucial for avoiding life-threatening situations. Therefore, an end paper employing an ensemble model, which consists of the MLP, Bayesian, and QUEST models, is recommended. However, the experimental results revealed that the ensemble model delivered better outcomes than the existing models did [9].

Although prevailing studies have aided in delivering reasonable outcomes for the identification and classification of diseases such as heart, thyroid, and hepatitis, they have lagged in delivering accurate outcomes for the effective classification and prediction of heart rates due to the use of ineffective algorithms. Therefore, the proposed study focused on employing modified PCA for dimensionality reduction. The use of conventional methods, such as modified PCA and modified backpropagation, is advantageous for limited datasets. They require less data for effective training, unlike deep learning models, which need large volumes. This enhances interpretability, which is crucial in healthcare for understanding predictions influencing patient care. Additionally, these methods are computationally efficient and simplify feature extraction and classification, thus improving the overall data analysis efficiency in healthcare applications. The normal PCA with BP is a traditional dimensionality reduction method.

In contrast, the proposed Modified Principal Component Analysis (M-PCA) model uses two Learning Rates (LR1 and LR2) to improve feature extraction by allowing different scales of transformation in the dimensionality reduction procedure. The purpose of using dual learning rates is to enable both precise tuning and wide-ranging changes in the feature selection process. This approach enhances feature extraction, improves classification and heart rate prediction, addresses nonlinear relationships, and improves robustness against noise. The M-BP model incorporates AWGN (Adaptive White Gaussian Noise) for both classification and regression, promoting robustness and generalizability. The AWGN acts as a regularization, enabling the model to navigate noise and reducing sensitivity to minor data discrepancies. Thus, M-BP with M-PCA combines neural network training with PCA preprocessing, which reduces input space dimensionality and enhances computational efficiency, assisting in avoiding overfitting and improving generalizability. Therefore, the proposed model is efficient for the classification of different diseases, and regression is employed for heart rate prediction. Therefore, the objectives can be defined as

- M-PCA is employed for dimensionality reduction to extract suitable features for the model.
- To utilize M–BP with the AWGN model for classification of heart, thyroid, and hepatitis disease, and heart rate prediction for regression.
- To assess the model's efficacy, different performance metrics, such as accuracy, recall, F1-score, and precision for classification and RMSE, MAE, R2, and MSE for regression, are employed.

Section 1.1 addresses conventional methods performed on a similar domain with diverse methods, as shown below. Section 2 presents the methodology executed in the projected system. Section 3 shows the results and outcomes accomplished by the proposed method. Finally, Section 4 presents the conclusion and future work of the proposed system.

1.1. Literature Review

Various existing studies have discussed the classification of heart, thyroid, and hepatitis diseases and heart rate prediction. Artificial intelligence plays a crucial role in medicine by addressing various diseases, including heart disease.

Heart disease is considered one of the major issues in human life. Therefore, detecting the presence of heart disease as early as possible is crucial. Several machine learning (ML) and deep learning (DL) algorithms have been used to classify these diseases. The analyzed work applied algorithms such as RF, DT, NB, SVM, and LR, revealing that the RF algorithm showed better accuracy and sensitivity in predicting heart diseases [10]. Similarly, six ML algorithms, namely, NB, gradient boosting, AdaBoost, KNN, LR, and RF, were used for heart disease classification. In addition, Grid SearchCV with 5-fold cross-validation was applied to optimize the hyperparameters. The LR algorithm obtained the best result [11]. Additionally, a previous study developed a

dual-stage stacked machine learning ML model for predicting CVD risk using a dataset of 1,190 patients and eleven significant characteristics. The model has a better range of accuracy, recall, and ROC-AUC score, with a false-negative rate below 1% [12].

Like ML, different DL methods have been used for the prediction and classification of heart disease [13]. The proposed method enhances heart classification quality via a deep learning neural network optimized with Talos, a hyperparameter tuning tool. The model follows the POD process, achieving reasonable accuracy in predicting heart disease [14, 15]. The recommended study implemented a CNN model to predict CVD presence via the Heart Disease Dataset (HDD). The experimental results revealed improved classification accuracy and reduced overfitting compared with existing models [16]. Over 28.1% of deaths are due to heart disease. Therefore, appropriate and timely diagnosis is important for curing patients who suffer from heart disease. Hence, the ANN algorithm can be used for the prediction of [17, 18] heart disease. The existing model utilized the HD dataset with attributes such as BP, age, cholesterol level, and sex. Performance metrics, including precision, F1 score, and recall, were assessed. The ANN model outperformed ML models such as SVM, DT, and KNN, showing superior accuracy for heart disease prediction [19].

Predicting heart disease is a major challenge for medical professionals because of various complications. According to the WHO, heart attack is the leading cause of death worldwide, especially in the West. A study used the DNN method for detection, obtaining satisfactory results in predicting heart disease [20]. The proposed method accurately predicts heart diseases via body indicators. It uses a feature selection method and a DNN (deep neural network) based on the SVC algorithm. The proposed method avoided gradient vanishing, demonstrated superior performance in classification, and was validated on a Kaggle HD dataset [21]. An improved healthcare monitoring framework for CHD utilizing a CNN integrated with an ANN to classify time series data from electrodes was validated through Python simulation on a dataset of 335 records with 36 clinical features. The performance metrics, including accuracies, precisions, recalls, and F-measures, have been assessed to evaluate the effectiveness of the model in disease prediction [22].

As CVD kills 17.9 million people across the world annually, better classification techniques should be employed for the identification of HD. However, manual detection may lead to inaccurate detection and diagnosis due to human intervention [23]. Similarly, CHD (coronary heart disease) accounts for nearly 13% of US deaths, primarily due to smoking, hypertension, and diabetes. A recommended paper utilized a CNN model to predict CVD incidence via the National Health and Nutrition Examination Survey (NHANES) dataset. Comparisons have shown that the CNN model outperforms existing models, such as the SVM, in accuracy for coronary heart disease prediction [24]. The CART (classification and regression tree) algorithm has been used by the model for the prediction of HDs and for extracting decision rules to clarify the relationships between the input and output variables. Furthermore, the CART model aided healthcare professionals and patients in the treatment of heart disease [25]. The existing work has addressed low data dimensionality, utilizing ML models to reduce human errors. It has employed various regression techniques on the Cleveland Heart dataset, yielding improved outcomes in probability conversion and regression processes [26].

Like heart disease, thyroid disease is considered a complex disease. Thyroid disease is considered to be different in terms of diagnosis. Therefore, the visibility and treatment of thyroid disease differ. Hence, it is important to detect the presence of thyroid problems in the body via AI techniques. Therefore, classification tree algorithms have been used for predicting thyroid conditions in patients [27]. This study focused on determining hyperthyroidism, hypothyroidism, and euthyroidism concerning hormones. Data from 499 patients were collected to predict thyroid parameters, highlighting J48 as the algorithm with the best accuracy in predictions [28]. Similarly, a formal study employed a filter-based feature selection method alongside a stacking-based ensemble ML framework specifically designed for thyroid disease detection. Extensive experiments on a clinical dataset revealed that this approach achieved impressive Receiver Operating Characteristic (ROC)-AUC scores [29]. The recommended methods use the KNN, SVM, XGB, AdaBoost, Gaussian process

classifier, GB classifier, and MLPC to predict thyroid diseases. The UCI dataset was used. The MLPC model has been determined to be the best, as evaluated by F1, accuracy, and AUC [30].

The thyroid is considered a critical medical condition that is caused primarily by Thyroid-Stimulating Hormone (TSH) [31] or by infection of the thyroid organs themselves. Hence, various ML and DL algorithms, such as the DT, RF, SVM, ANN, and LR algorithms, have been used in the suggested methods for the identification of thyroid problems. Various symptoms and reports of the thyroid have been used for the prediction and diagnosis of thyroid problems, and the DT algorithm was found to yield satisfactory outcomes for the prediction of thyroid disease [32, 33]. The research employed various ML algorithms, such as NB, J48, bagging, and boosting, to predict thyroid diseases via data from SMS hospitals in Jaipur. The results revealed that the J48 model was effective, whereas the SVM model was highly accurate. Future research can be improved by integrating feature selection algorithms [34]. Additionally, previous work has introduced a modified XGBoost model that integrates clinical and molecular properties for predicting malignancy in thyroid nodules, utilizing fine-needle aspiration biopsies and gene expression analysis. The model achieved accuracy over other machine learning models, such as C5.0 (92.5%) and CART (88.7%) [35].

Hepatitis is a dangerous condition that can cause severe liver disease, including liver cancer. Protecting the liver is essential, making hepatitis detection crucial to prevent fatal results. A recent study utilized a CNN model to identify hepatitis in liver patients via NHIRD (National Health Insurance Research Database) data from 2002 to 2010, with performance evaluated through accuracy and Area Under the Curve (AUC) analyses [36]. Similarly, various ML and DL techniques, such as NB, SVM, LR, K-NN, MLP, RT, RF, J48, TotF, RepTree, AdaBoost, voting, bagging, and stacking, have been used. However, the experimental results revealed that the voting classifier outperformed the existing model in terms of accuracy, recall, F1, AUC, and precision after the incorporation of SMOTE 10-fold cross-validation [37]. Another approach has utilized classification algorithms, including naïve Bayes, random forest, decision tree, and KNN, to predict diseases on the basis of patient symptoms from a dataset of 4920 records encompassing 41 diseases, demonstrating significant advancements in disease prediction accuracy and healthcare outcomes [38].

1.2. Problem Identification

From the assessment of the above-existing works, core concerns are emphasized as explored below:

- The accuracy obtained by the suggested model is considerably low, which makes it inefficient for predicting heart disease [19].
- Implementing manual techniques for disease detection may lead to inaccurate outcomes. Hence, advanced AI techniques must be incorporated for effective disease prediction [23].
- The reliance on a limited dataset may affect the generalizability of the model predictions, and the complexity of feature selection could lead to challenges in interpretability and clinical applicability [37].
- The dependency on small datasets may hinder model robustness and generalizability, while the complexity of the hybrid approach could complicate interpretability and practical implementation in clinical settings [34]

2. Materials and Methods

Heart, thyroid, and hepatitis diseases are considered fatal diseases worldwide. Hence, it is extremely vital to detect these diseases precisely and correctly to avoid further complications, which may lead to dreadful consequences. Hence, the detection of disease is crucial. However, existing models are quite inefficient for the detection of diseases because of the use of ineffective algorithms. Therefore, the proposed model employs M-PCA and M-BPA with AWGN for the classification of diseases and the prediction of heart rate. Figure 1a and Figure 1b show the process involved in the proposed model for classification and regression.

Figure 1 depicts the classification process for the task of disease prediction, outlining a structured and organized method for analyzing data. The process

starts with loading the dataset and then involves important preprocessing steps that make sure the data is clean and appropriate for additional analysis. Subsequently, an adapted PCA (Principal Component Analysis) method is utilized for dimensionality reduction, which successfully simplifies the dataset while maintaining essential characteristics. Next, a train-test split is performed, enabling an accurate assessment of the model's performance. Ultimately, a revised BP (Backpropagation) algorithm is utilized to execute the classification task, allowing for precise predictions about disease occurrence. This organized process was intended to improve the effectiveness of disease forecasting, also employed to uncover important insights into the fundamental patterns present in the data, thus aiding in a greater comprehension of the elements affecting disease results.

Figure 2 depicts the regression process for the heart rate prediction task, highlighting a methodical and organized approach. The procedure starts with

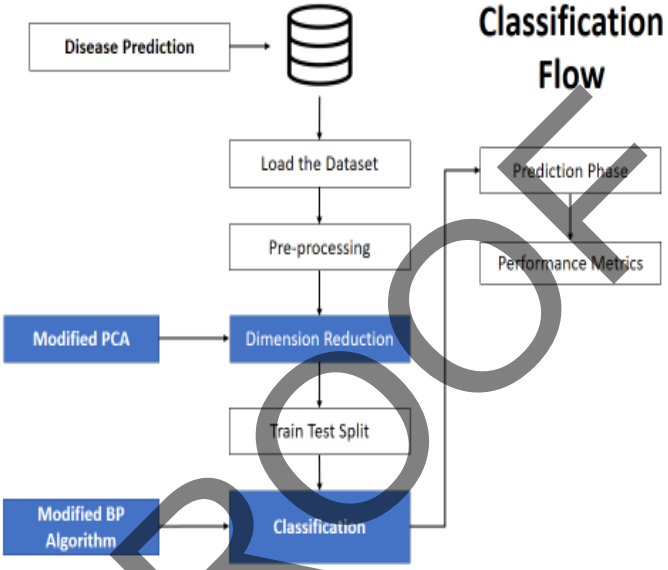

Figure 1. Classification of Proposed Model

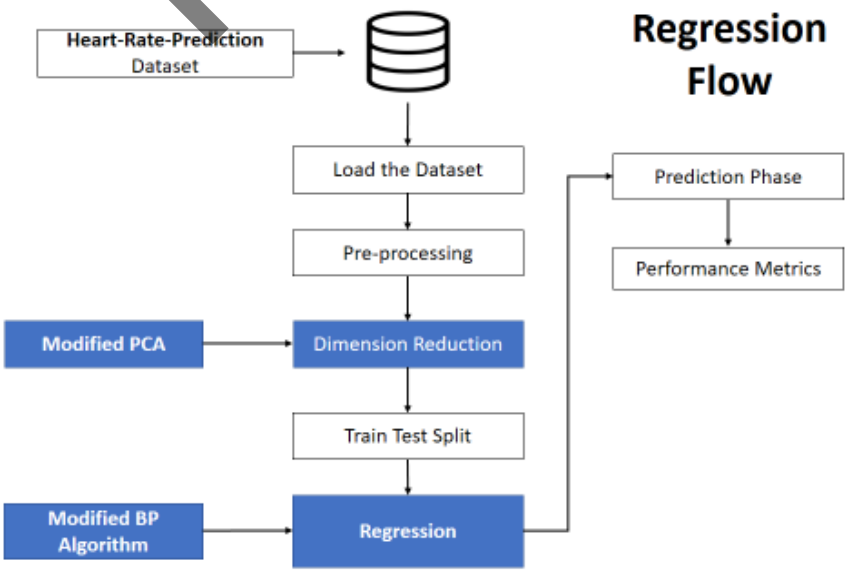

Figure 2. Regression of Proposed Model

importing the dataset and then moves on to crucial preprocessing steps to prepare the data for examination. Following this, a revised PCA (Principal Component Analysis) method is utilized for reducing dimensionality, aiding in simplifying the data by decreasing its complexity while preserving essential information. Subsequently, the dataset is split into training and testing subsets via a train-test division, guaranteeing that the model can be properly assessed. Ultimately, a revised BP (Backpropagation) algorithm is employed for the regression task, enabling precise heart rate predictions. This extensive process, which is proposed to improve the effectiveness of heart rate forecasting, also strives to reveal important insights into the fundamental patterns found within the data.

The overall flow for the classification and regression model is depicted in Figures 1a and 1b. Initially, the heart disease dataset, thyroid dataset, and hepatitis dataset are employed for classification, and the heart rate prediction dataset is implemented for regression. Once the dataset is loaded, the data is preprocessed using various preprocessing techniques in order to remove the noisy data, missing values, thereby improving the quality of the data. After preprocessing, modified PCA is implemented in order to reduce the irrelevant and unwanted features and aid in obtaining the best features suitable for the model. The M-PCA utilizes 2 learning rates (L1 and L2), thereby identifying the best features for the model. After the extraction of the best features, it is split into a train and test split. After training and testing, the trained data enters the process of classification and regression. Classification and regression are carried out by using the proposed Modified backpropagation method (M-BP). M-BP is implemented due to the slow convergence rate of conventional BP. Therefore, M-BP is used, due to its capability to alleviate the problem of local minima that has been provoked by conventional BPA by using the Adaptive White Gaussian Noise Algorithm (AWGN) for enhanced classification and regression process. The Adaptive White Gaussian Noise (AWGN) method aids in improving the robustness of the model by learning to maneuver noise and becomes less sensitive to negligible discrepancies in the data. During the process of iteration, AWGN is fed as the input to the NN, thereby increasing the reliability and proficiency of the model for effective classification of heart, thyroid, and hepatitis disease, and prediction of heart rate. Finally, the efficacy of the model was evaluated by using various metrics such as accuracy, recall, precision, F1 score for classification, and MSE, RMSE, MAE, R2 score for regression.

2.1. Preprocessing

To effectively preprocess the heart disease, thyroid, hepatitis, and heart rate prediction datasets for the proposed model employing modified PCA and modified BP, several key techniques are implemented. Initially, the datasets undergo standardization to ensure that each feature has a mean of zero and a standard deviation of one, which is crucial for PCA, as it relies on variance to identify principal components. Following this, missing values are addressed through imputation or exclusion based on the specific dataset characteristics, ensuring that the integrity of the data is maintained. Noise removal techniques are also applied to enhance data quality by filtering out irrelevant information that could adversely affect model performance. The absence of data augmentation in this context allows for a more straightforward analysis of the original data characteristics. By focusing on these preprocessing steps, the datasets are prepared to maximize the efficacy of dimensionality reduction and classification algorithms, ultimately leading to improved model accuracy and generalization capabilities

2.2. Dimensionality Reduction-Modified PCA

Dimensionality reduction is the process of decreasing the number of variables considered. It is employed primarily to mine the underlying features from the raw datasets or to reduce the amount of data while maintaining the structure. This research proposed PCA for dimensionality reduction, which aided in feature extraction.

PCA is considered to be an unsupervised algorithm that is employed primarily for dimensionality reduction. The conversion of observations of correlated features to sets of linearly uncorrelated features with the help of an orthogonal transformation is achieved via a statistical approach called PCA, and the new transformed feature that is obtained is called the PC (principal component). The term PCA is referred to as a technique of DA for constructing linear multivariate models of complex data. Furthermore, PC

is developed by employing orthogonal basis vectors, which include eigenvectors. Although conventional PCA aids in delivering effective dimensionality reduction and other various advantages, it still has drawbacks, such as a loss of information during the process of dimensionality reduction and difficulty in recovering the original parameters; these drawbacks make the model inefficient for the prediction and classification of disease. Therefore, the proposed model incorporates modified PCA (M-PCA) by using a dot product mechanism of the learning rate (LR1 and LR2) to reduce the dimensions of the input. The modified PCA algorithm employs two learning rates (LR1 and LR2) to enhance feature extraction by enabling various scales of transformation in the dimensionality reduction process. Besides, LR1 could be used for updating the principal components more stably, while LR2 could be used for adjusting the weights more dynamically. This dual learning rate approach can help the model converge faster and potentially achieve better performance by balancing stability and adaptability in the learning process. The reason for employing dual learning rates is to facilitate both fine-tuning and broad adjustments in the feature selection procedure. Figure 3 shows the process involved in M-PCA.

Figure 3 shows the process of M-CA, in which the input data are fetched initially; then, the mean and covariance matrix are calculated by using deflated vectors of learning rates LR1 and LR2, in which the previously obtained variance is eliminated to obtain the best features for the model. The iteration is repeated until the best features are attained. Furthermore, the eigenvalue is computed, and the eigenvectors are sorted on the basis of decreasing eigenvalue values, resulting in the best features being obtained for the model. The algorithm for the modified PCA is depicted in Algorithm I.

In the context of dimensionality reduction, the entire dataset, denoted X, comprises multiple features characterized by their dimensionality. To prepare the data for analysis, the mean vector μ is used to center the dataset, followed by the computation of the covariance matrix C. This matrix illustrates both the variance and correlation among the features. During the dimensionality reduction process, we employ two learning rates, LR1 and LR2, to optimize convergence. The analysis reveals eigenvectors, which represent the directions of maximum variance within the data, accompanied by their corresponding eigenvalues, indicating the amount of variance each eigenvector captures. Then, a specific number of eigenvectors denoted as k is selected to form a matrix Z that facilitates data transformation into a new subspace S. This subspace retains the most significant features of the original dataset while reducing its dimensionality, thus enhancing interpretability and efficiency in subsequent analyses. Mathematical equations for the modified PCA are given as follows:

As the modification of PCA involves employing two learning rates, the first update can be aimed at fine-tuning the dimensions as demonstrated in Equation 1,

$$W^{(1)} = W^{0} + LR1.\nabla L \quad (1)$$

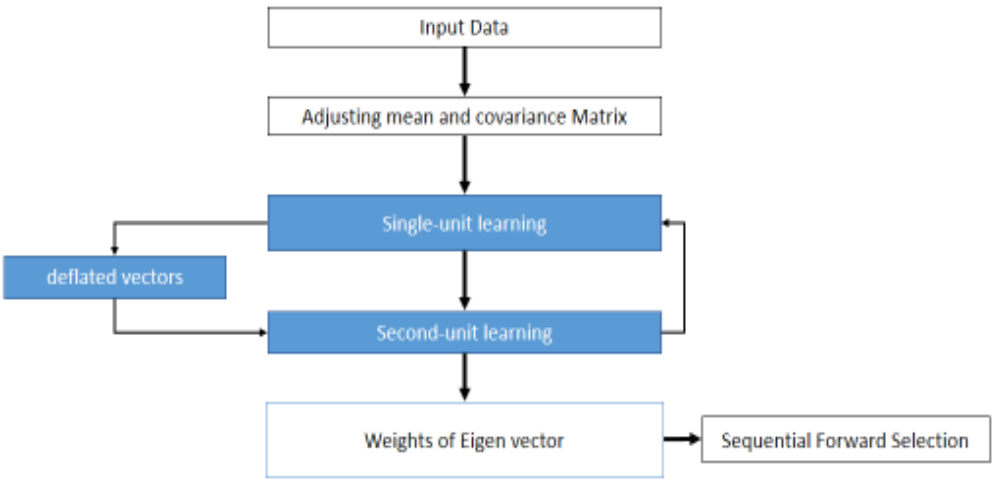

Figure 3. Modified PCA

Algorithm I: Modified PCA
Output: The features with dim − dimensional samples except for the class labels.
Step 1: Take the whole dataset X, including dim − dimensional
PCA($x_1, \ldots\ldots, x_n$ with dim
Step 2: Center the data: $x_t = x_t - \mu$ with μ the mean vector.
Step 3: Compute the covariance matrix using deflated vectors(LR1 and LR2) proposed
$C = \frac{1}{n}\sum_t x_t x_t^T$.
Step 4: Compute eigenvectors and corresponding eigenvalues
Step 5: Sort the eigenvectors based on decreasing eigenvalues. Then choose k eigenvectora based
on the largest eigenvalues to form a dim × e dimensional matrix Z.
Step 6: Use this dimensional dim × eigenvector matrix and then dot the product with the LR1 and LR2 to transform the samples into the subspace S.
Step 7: return new subspace S

In Equation 1, $W^{(0)}$ is represented as the current weight, and $W^{(1)}$ is denoted as the fine-tuned weight.

Likewise, for wider adjustments to the weight, a second learning rate is used, as demonstrated in Equation 2,

$$W^{(2)} = W^0 + LR2.\nabla L \quad (2)$$

This allows different learning rates to be applied depending on the component adjustments needed.

To improve the robustness of M-PCA, the learning rates can be adaptively adjusted based on the iteration number.

$$LR1_t = \frac{LR1_0}{1+\beta_1 t} \quad (3)$$

$$LR2_t = \frac{LR2_0}{1+\beta_2 t} \quad (4)$$

In Equations 3 and 4, $LR1_0$ and $LR2_0$ are denoted as initial learning rates, β_1, and β_2 are identified as the decay rates, and t is the current iteration number.

Therefore, by fine-tuning the principal components, the model can become less sensitive to noise or less important variations in the dataset, potentially leading to more robust classifiers. Additionally, M-PCA can capture and represent varying levels of importance among features, which is crucial in medical datasets where some features may be more indicative of a disease than others.

Table 1 Presents information about the various datasets used, highlighting their components and corresponding learning rates for two different scenarios. Each dataset is associated with a specific medical condition, such as heart disease, thyroid, hepatitis, or heart rate, with the number of components indicating the features or variables included in each dataset. For the Heart Disease and Heart Rate datasets, a higher learning rate of 0.5 is employed, suggesting a more aggressive approach to updating model parameters during training. In contrast, the Thyroid and Hepatitis datasets utilize lower learning rates of 0.05 and 0.007, respectively, indicating a more cautious adjustment to prevent overshooting optimal solutions. This variation in learning rates reflects differing complexities and characteristics of each dataset and influences effective model training.

2.3. Classification and Regression-Modified Backpropagation Method with AWGN

Back propagation is primarily used for training feed-forward NNs. BP is used for updating the weights of parameters and bias to attain proper classification and regression. BP possesses various advantages, such as low computational complexity, simplicity, and adaptability. However, the convergence rate of the model is slow, and it can be trapped in local minima, especially for nonlinearly separable and nonstationary issues. The convergence rate of the BP is slow because the steepest descent technique is implemented to adjust the weights of the networks, resulting in suboptimal solutions. These drawbacks have the potential to result in effective classification and regression processes. Therefore, the proposed model implements modified backpropagation, which can alleviate the problem of local minima provoked by conventional BPA by using AWGN for enhanced

Table 1. Modified PCA parameters

Dataset	Components	Learning Rate 1	Learning Rate 2
Heart Disease	13	0.5	0.2
Thyroid	24	0.05	0.007
Hepatitis	19	0.05	0.007
Heart Rate	15	0.5	0.2

classification and prediction of the heart rate of an individual. The modified back propagation (BP) method is an improved version of the traditional backpropagation technique utilized for training neural networks by using controlled noise. Although the classic BP algorithm adjusts weights and biases by reducing error via gradient descent, the adjusted version incorporates particular modifications to enhance learning efficacy, convergence rate, or precision. Hence, the M-BP model is used for both classification and regression because of the incorporation of AWGN, adaptive white Gaussian noise. This form of regularization forces the model to learn more robustly, and representations with generalizability lead to better performance on unseen test data. Integration of the AWGN model assists in refining the robustness of the model by learning to maneuver noise and becomes less sensitive to negligible discrepancies in the data. Furthermore, a larger space of solutions can be formed by using M-BP.

Thus, incorporating AWGN helps increase the model's resilience by adding controlled noise during training, aiding in preventing overfitting and enhancing generalizability. Crucial hyperparameters, such as learning rates, are determined via empirical testing. LR1 is set at 0.01 for detailed refinements, whereas LR2 is set at 0.1 for wider feature space exploration. The AWGN noise level is set to a standard deviation of 0.05 to ensure adequate regularization with no interference in the learning process. Figure 4 shows the process of M-BP involved in classification and regression.

The reduced features obtained via M-PCA are fed to the ANN. ANNs can be predominantly used for both classification and regression, as ANNs can handle more than one task at the same time. Furthermore, the ANN provides a high-speed classification process, improves the visualization of data, minimizes the training time of the learning algorithm, and aids in enhancing the prediction performance. Initially, the reduced features are fed to the input layer, as this layer accepts the data and passes it to the remaining layers. From the input layers, features are passed to the hidden layer, and these hidden layers are either one or more in number. HL is primarily responsible for the complexity of NNs, as they can perform multiple functions at the same time, which include both classification and regression. The interconnection between the forward layer is passed as input, and the HL assigns weights to each input randomly at the initial point.

Furthermore, bias is added to each input neuron. After this, the amalgamation of bias and weights, which is the weighted sum, is passed through an activation function. The role of the activation function is to consider the nodes for feature extraction from which the output is generated, which is known as forward propagation. Then, the generated outputs are compared with the original output, and the error is known. Finally, the weights are updated in BP to decrease the error, and the process is iterated for a specific number of epochs.

However, the conventional BP model has a slow convergence rate, which can make it inefficient for classification and regression. Therefore, the proposed model uses an additive white Gaussian noise algorithm, which can make it efficient and effective for disease classification and regression. The algorithm for modified backpropagation is depicted in Algorithm II.

The AWGN method aids in improving the robustness of the model by learning to handle noise and becomes less sensitive to negligible discrepancies in the data. During the process of iteration, the AWGN is fed as the input to the NN, thereby increasing the reliability of the model for effective classification

Algorithm II: Modified -BP

Step 1: x → Input Dataset
Step 2: y → labels
Step 3: wt → weight
Step 4: l → numbers of layers in neural network 1 … L
Step 5: D − data
Step 6: a − neuron
Step 7: L − layer
Step 8: $data_{i,j}^{(l)}$ − The error for all l, i, j
Step 9: $t_{i,j}^{(l)}$ − for all l, i, j
Step 10: For i = 1 to m
Step 11: $neuron^{l} < -\text{Backward}\ (x^{(i)}, wt)$
$d^{l} < -\text{neuron(layer)} - y(i)$
$t_{i,j}^{(l)} < - \ t_{i,j}^{(l)} + \ neuron_{j}^{(l)} \ . \ t_{i}^{l+1}$
Step 12: if j ≠ 0 then
$data_{i,j}^{(l)} < - \frac{1}{m} t_{i,j}^{(l)} + \lambda wt_{i,j}^{(l)}$
Step 13: else
$data_{i,j}^{(l)} < - \frac{1}{m} t_{i,j}^{(l)}$

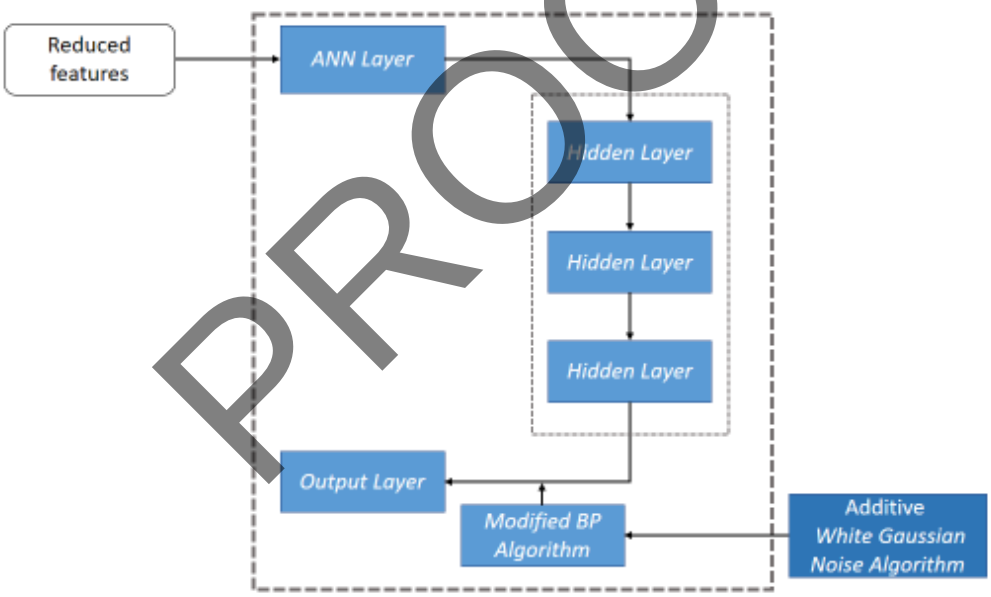

Figure 4. Modified BP with AWGN

and learning process and enhances the generalization learning process and enhances the generalization ability of the proposed model. The algorithm for the adaptive white noise algorithm is depicted in Algorithm III.

The dataset input (x) is made up of feature values used for training purposes. Each input is matched with corresponding labels (y), which represent the target outputs. Throughout the training process, the weights (wt) of the neural network are adjusted continuously to reduce errors. The network architecture is determined by the number of layers (l), with each layer containing neurons (a) that process the inputs. As data moves through the network, data (D) is processed during backpropagation, which is crucial for updating weights based on errors. $data_{i,j}^{(1)}$, The error is

Algorithm III: AWGN

Step 1: If ($data_{i,j}^{(l)}$ − white Gaussian noise ≥ T_1) $s_1 = 1$, $s_2 = 1$, $data_{i,j}^{(l)}$ is additive white gaussian noise
Step 2: else if (If ($data_{i,j}^{(l)}$ − white Gaussian noise ≥ T_2) $s_1 = 1$, $s_2 = 0$, $data_{i,j}^{(l)}$ is additive white gaussian noise
Step 3: else
Step 4: $s_1 = 0$, $s_2 = 0$
Step 5: end

computed for all layers (l), instances (i), and neurons (j) in the network, providing a vital metric for performance evaluation. To aid backpropagation, $t_{i,j}^{1}$ a temporary variable is used for intermediate calculations. The training process involves a total of m training instances, each contributing to refining the model. To improve generalization and prevent overfitting, a regularization parameter (λ) is included in the training process.

Additionally, AWGN is applied to increase robustness during training, enabling the model to better adjust to changes in the input data. This process highlights how these variables interact within the context of training a neural network. The mathematical equations are incorporated in the proposed Modified BP.

In standard backpropagation for a neural network, the gradient of the loss function L concerning the weights W is computed using the chain rule (Equation 5),

$$\frac{\partial \mathrm{L}}{\partial \mathrm{W}} = \frac{\partial \mathrm{L}}{\partial \mathrm{A}} . \frac{\partial \mathrm{A}}{\partial \mathrm{Z}} . \frac{\partial \mathrm{Z}}{\partial \mathrm{W}} \tag{5}$$

Here, A is defined as the activation of the neural network, Z is defined as the weighted input to the activations, X is the input data, and b is the bias term.

However, in modified BP with AWGN, noise is added to the gradient during training. Let η_t represent the adaptive noise at time step t.

$$\eta_{t\sim} \mathrm{N}(0, \sigma_t^2) \tag{6}$$

In Equation 6, N indicates a normal distribution, σ_t^2 represents the variance of the noise, which can be expressed as (Equation 7),

$$\widetilde{g_t} = \frac{\partial \mathrm{L}}{\partial \mathrm{W}} + \eta_t \tag{7}$$

Here, $\widetilde{g_{lt}}$ is defined as the modified gradient that includes noise. Hence, incorporating these M-PCA and M-BP aids in a better classification and regression process.

3. Results & Discussion

The proposed design was executed with Python. The first section describes the dataset. The second section describes the performance metrics. The third section presents the proposed model's EDA, and the fourth section describes its performance analysis. Finally, the fifth section presents a comparative analysis to determine the efficacy of the proposed approach over conventional methods.

3.1. Dataset Description

Three different datasets are utilized by the proposed model for the classification of heart, thyroid, and hepatitis. Various attributes are considered for the model, which include dataset characteristics, associated tasks, instances, subject areas, and many more. Tables 2, 3, and 4 show the features of the dataset. The heart disease dataset consists of 303 instances and 13 attributes and includes categorical, real, and integer attribute types. The thyroid dataset consists of 7200 instances and 5 attributes with

Table 2. Heart disease dataset [39]

Characteristic	Description
Source	UCI Machine Learning Repository
Total Instances	303
Number of Attributes	13
Target Variable	Presence of Heart Disease

Table 3. Thyroid dataset [40]

Characteristic	Description
Source	UCI Machine Learning Repository
Total Instances	7200
Number of Attributes	5
Target Variable	Thyroid Condition

Table 4. Hepatitis dataset [41]

Characteristic	Description
Source	UCI Machine Learning Repository
Total Instances	155
Number of Attributes	19
Target Variable	Survival Status

categorical, real values as attribute types, and hepatitis consists of 155 instances and 19 attributes with categorical, real, and integer values as the attribute type.

Similarly, regression is processed by the heart rate prediction dataset of Table 5, implemented for the regression model, which consists of various attributes, which are taken from the signal measures via ECGs recorded for different individuals with different heart rates at the time the measurement was taken. Seven different CSV files were utilized for the dataset. The primary objective of the dataset is to predict the heart rate of an individual.

Table 5. Heart rate prediction dataset [42]

Characteristic	Description
Source	Kaggle
Target Variable	Heart Rate (BPM)
Type	Time Series Data

3.2. Performance Metrics

The model's performance is evaluated in the subsequent section using various performance metrics, such as accuracy, precision, F-measure, and recall for classification and MSE, MAE, RMSE, and R^2 for regression.

A. Accuracy

Accuracy is claimed to be the measure of total accurate classification. The accuracy range is calculated with Equation 8:

$$Acc = \frac{TRN + TRP}{TRN + FLN + TRP + FLP} \tag{8}$$

where TRN signifies a true negative, TRP represents a true positive, FLN represents a false negative, and FLP represents a false positive.

B. Precision

The precision is calculated by measuring the precise classification count. It is measured via improper classification. The precision is calculated via Equation 9:

$$precision = \frac{TRP}{FLP + TRP} \tag{9}$$

C. F-measure

Another term of the F-measure is the F1 score. The F1 score is denoted as the weighted harmonic-mean value of recall and precision, and the F1 score is estimated with Equation 10:

$$F1 - score = 2 \times \frac{P \times R}{P + R} \tag{10}$$

where P denotes precision, and R denotes recall.

D. Recall

Recall is denoted as an isolated part of the generated measure, which calculates the sum of precise positive types conceived of all the optimistic groups and is evaluated by Equation 11:

$$R\_c = \frac{TRP}{FLN + TRP} \tag{11}$$

In Equation 11, the FLN is denoted as a false negative.

3.3. Regression

A. Mean Square Error (MSE)

The MSE is one of the mutual estimation parameters used in measuring quality. If the values are closer to zero, the metric measurement is restored in quality. This is obtained via Equation 12,

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(Y_i - Y_i)^2 \tag{12}$$

Here, n represents the number of data points, Y_i represents the observed values, and $\widehat{Y}_i$ represents the predicted values.

B. Mean absolute error (MAE)

The MAE is used to measure the magnitude, on average, of the errors in a set of forecasts among the paired observations, which are used to express the same phenomenon without considering their direction. This parameter is the variance among the important parameters existing in the data with the projected values in the same dataset. This is obtained via Equation 13,

$$MAE = \sum_{i=1}^{n} |y_i - x_i| / n \quad (13)$$

where y_i represents the prediction rates, x_i represents the true value, and n represents the total number of data points.

C. Root mean square error (RMSE)

R^2 provides information about the goodness of fit of a model. The equation shows the formula for R^2 (Equation 14).

$$RMSE = \sqrt{\sum_{i=1}^{N} (acutal - predicted)^2 / N} \quad (14)$$

where i denotes the variable, and where N is the number of non-missing data points.

3.4. Exploratory Data Analysis

The EDA remains to be considered at the data; previously, it was created with some assumptions. It supports perceptible errors, provides a finer understanding of the data, and aids in detecting deviations or other irregular measures. The EDAs for the proposed model are depicted in the following section.

3.4.1. Classification

The proposed classification process employs three different datasets: a heart disease dataset, a thyroid dataset, and a hepatitis dataset.

Heart Disease Dataset

Figure 5 depicts the target distribution of heart disease. The model shows that the number of disease predictions is greater than that of normal predictions.

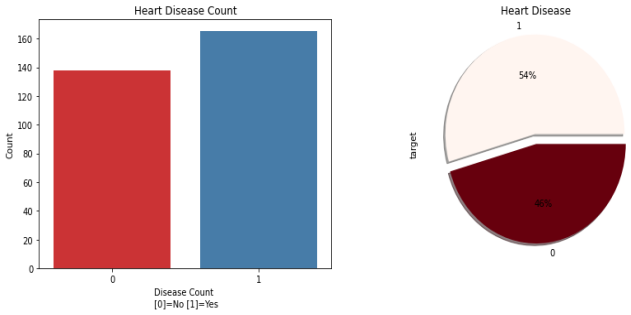

Figure 5. Target distribution

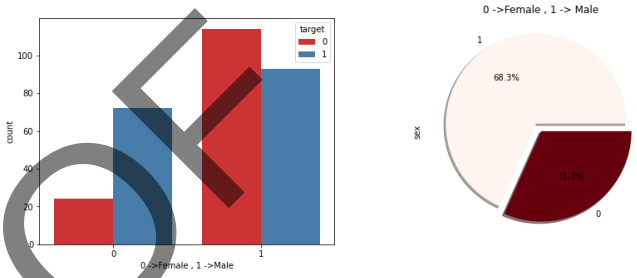

Figure 6. Identification of Disease and Nondisease in Males and Females

Similarly, the model demonstrated no disease or no disease for both males and females.

Figure 6 shows that heart disease is more common in females than in males. The blue color represents no disease, whereas the orange color indicates the presence of heart disease.

Thyroid Dataset

Figure 7 shows the thyroid dataset's heatmap. The heatmap depicts the value for a main variable of interest across the 2-axis variables as a grid of colored squares. A heatmap is primarily used to visualize the strength of correlation among the variables.

Figure 8 shows the count plot of the target variable for the thyroid dataset. The count plot denotes the occurrence of the observation present in the categorical variable. It uses bar charts for visual depiction.

Hepatitis Dataset

Figure 9 depicts the class distribution of the hepatitis dataset employed by the proposed model. Figure 9 shows that approximately 83.7% of people are alive, whereas 16.2% die due to hepatitis.

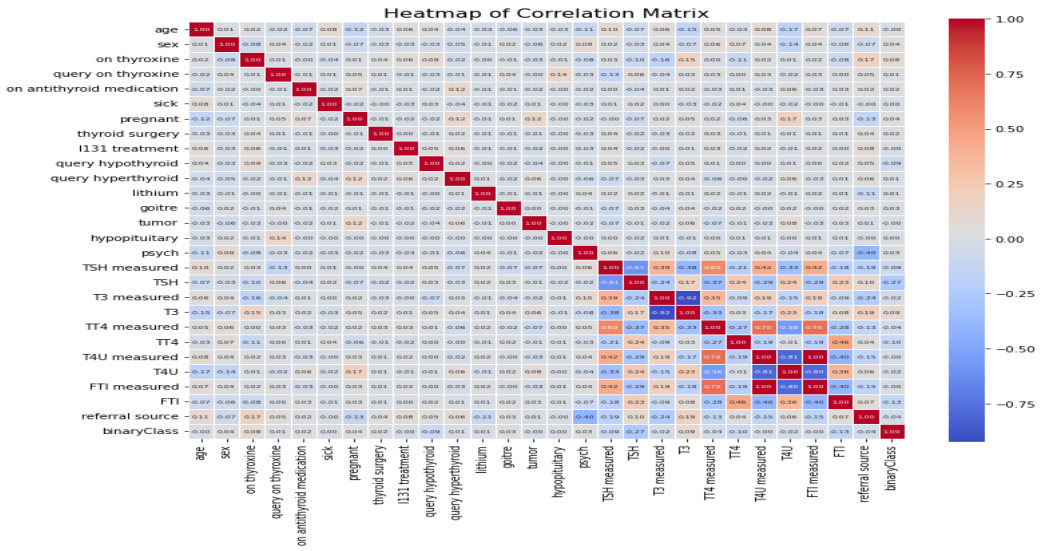

Figure 7. Heatmap for the Thyroid Dataset

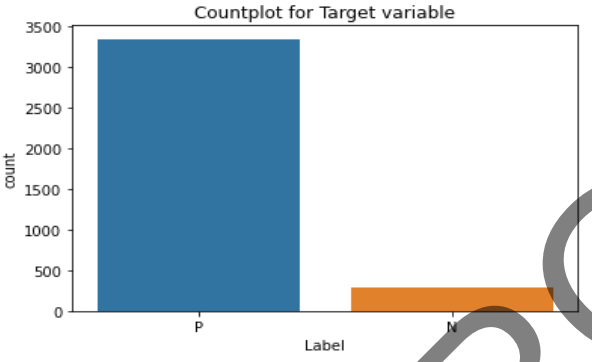

Figure 8. Count plot for the target variable

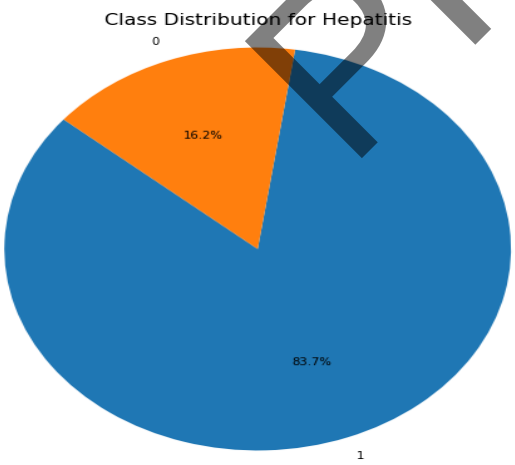

Figure 9. Class distribution for hepatitis

3.4.2. Regression

A regression process is employed by utilizing the proposed model for the heart rate prediction dataset.

Heart Rate Prediction

Figure 10 shows the heatmap for the heart rate prediction dataset. A heatmap aids in visualizing the relationship between variables in the HD space.

Figure 11 shows the signals of the heart rate with time on the X-axis and the heart rate in bpm on the Y-axis. Similarly, Figure 12 shows the scaling of the attributes for heart rate prediction.

Figure 12 depicts the distribution of the participants' resting blood pressure according to density. It shows that it gradually increases from 100 °C, reaches its peak at 130°C, and decreases at 200°C.

Figure 13 depicts the distribution of constrictive pericarditis concerning density. The highest density of 140 is achieved when the CP is 0.0, and the lowest density value of 20 is obtained when the CP is 3.0.

Figure 14 shows the distribution of sex. When the sex value is 1.0, a density value of more than 200 is obtained, and when it is between 0.0 and 0.2, a density value of less than 100 is obtained.

Figure 15 shows the distribution of age, where increasing and low-density values concerning different age counts are demonstrated. The figure shows that the maximum density value is obtained around the age of 60, and the density value increases from the age of 50. However, the bottommost density value is acquired at the age of 30.

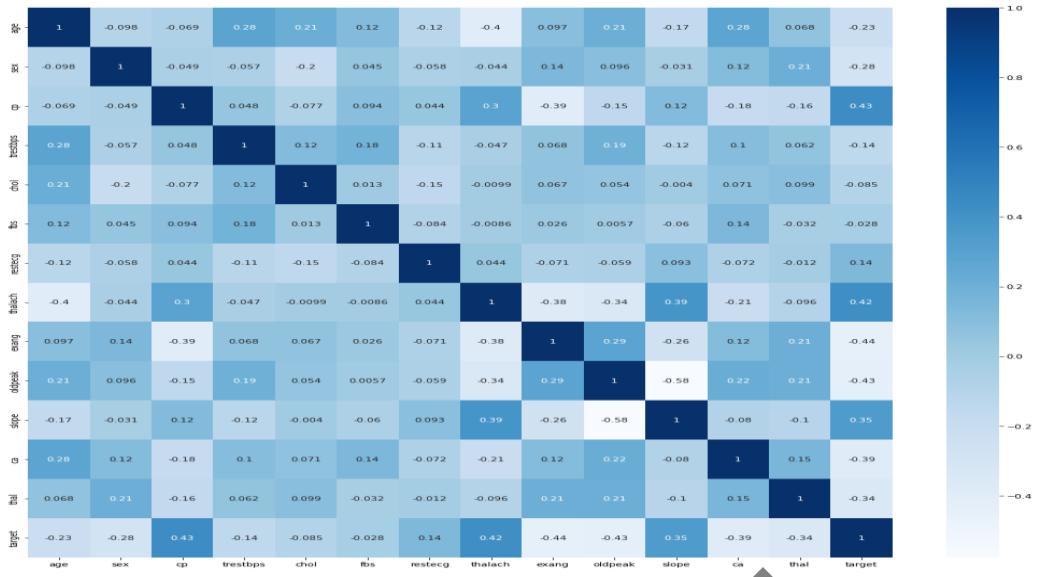

Figure 10. Heatmap of heart rate prediction

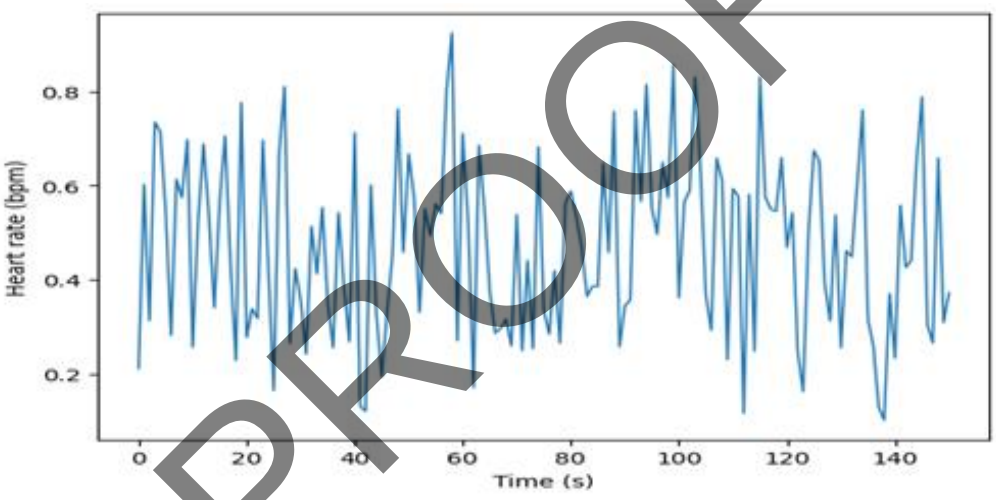

Figure 11. Heart rate

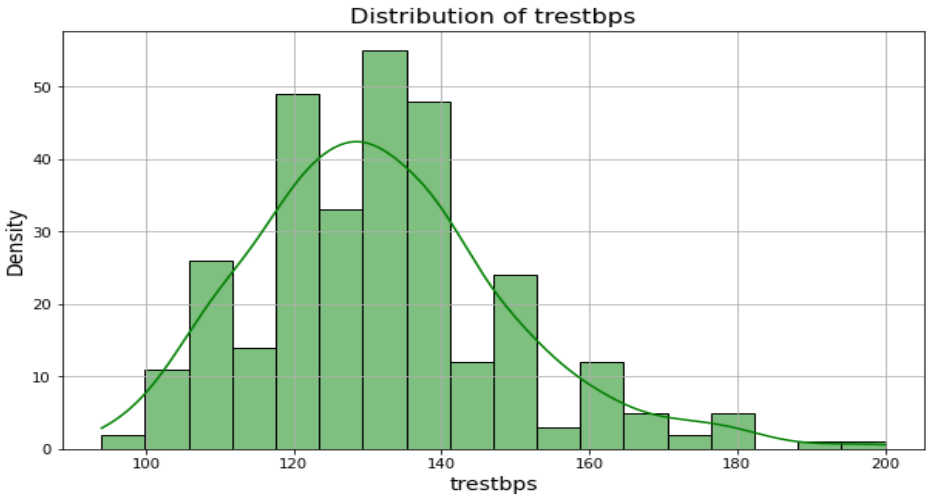

Figure 12. Distribution of Trestbps

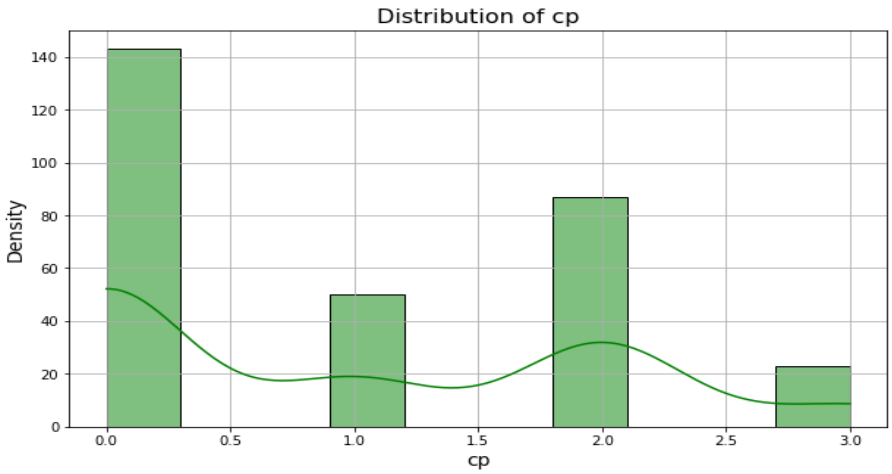

Figure 13. Distribution of CPs

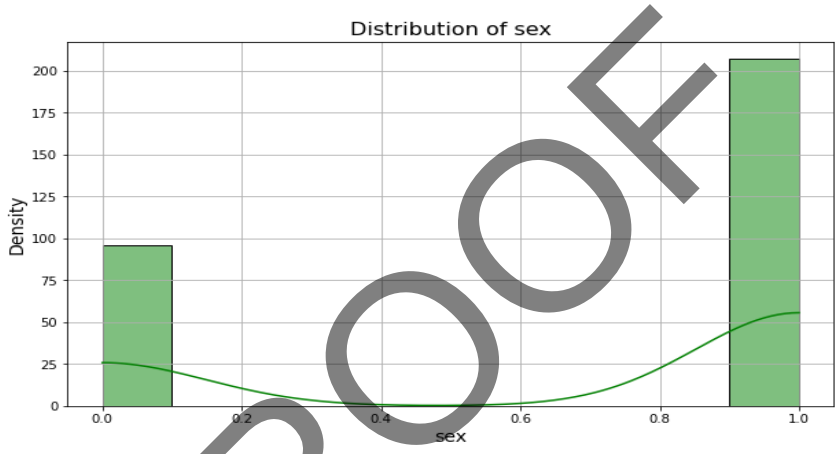

Figure 14. Distribution of sex

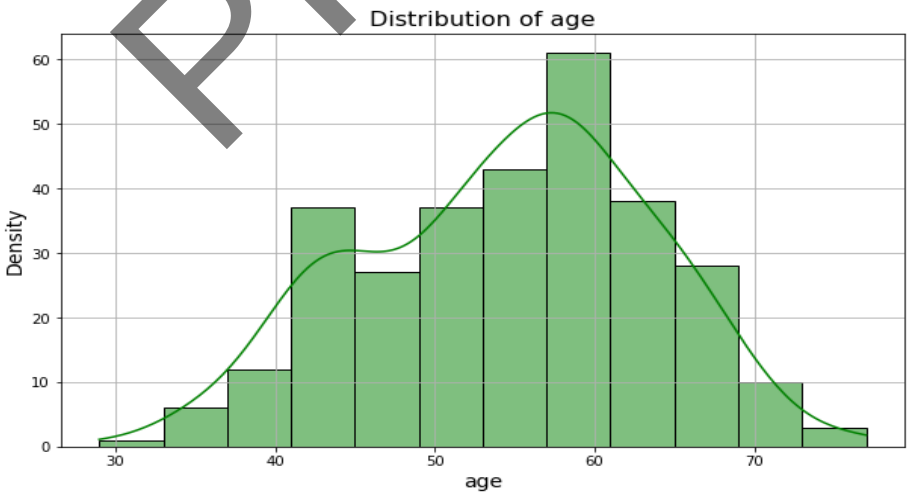

Figure 15. Distribution of Age

Similarly, the CA values for the proposed model are shown in Figure 16, where a density value of 175 is attained when the CA value is 0.0. However, density values of approximately 50–75, 20–50, and 0–25 are attained when the CA is 1.0, 2.0, 3.0, and 4.0, respectively.

Correspondingly, Figure 17 shows the distribution of the slope, in which the slope value is 0.00, and the

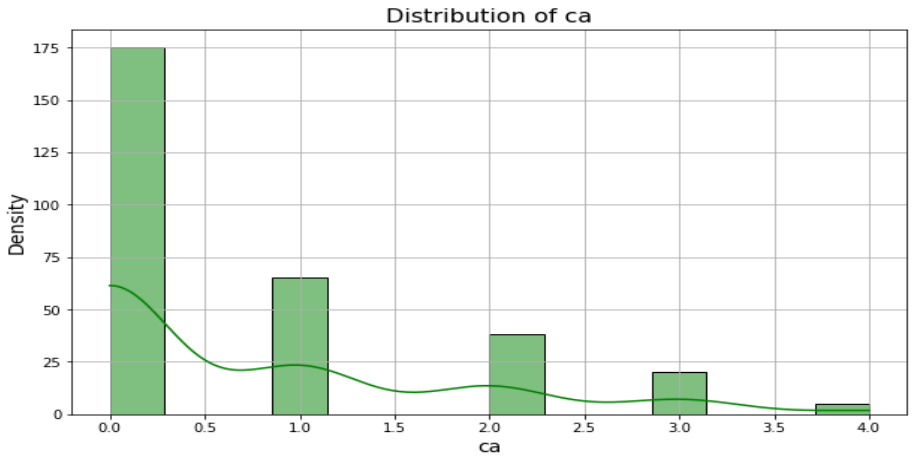

Figure 16. Distribution of CA

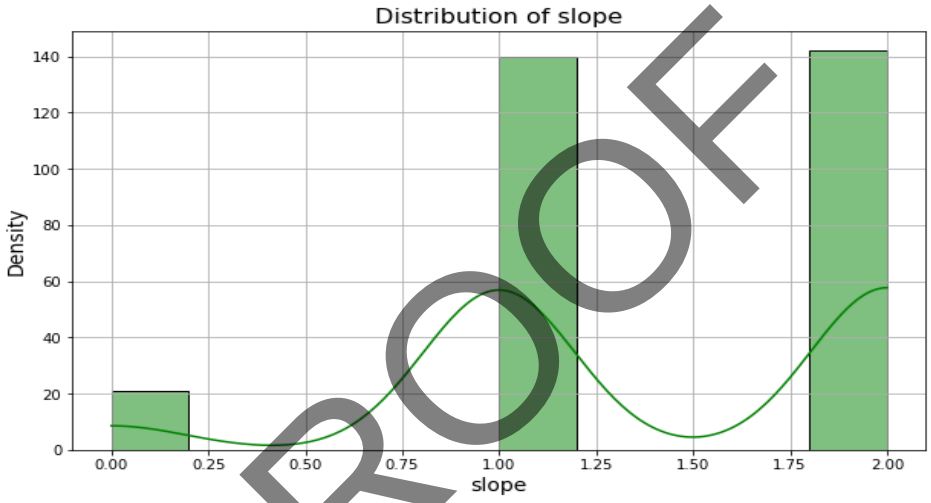

Figure 17. Distribution of slopes

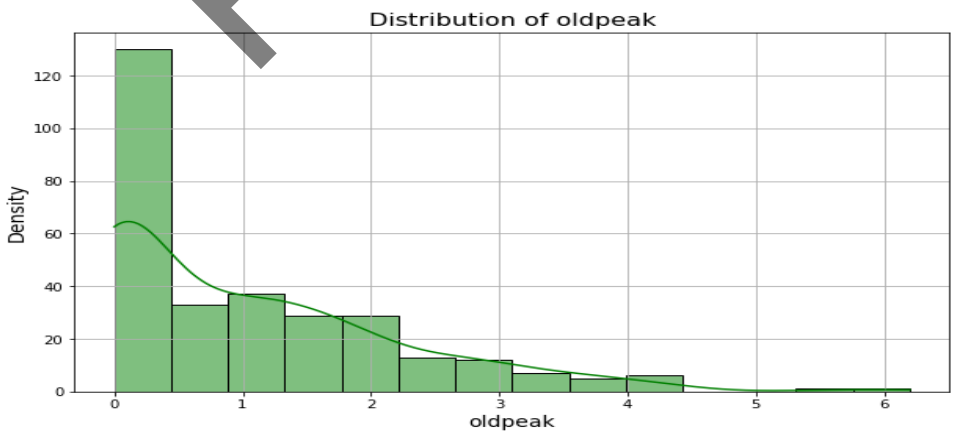

Figure 18. Distribution of OldPeak

density value is 20. However, when the slope value is 1.00, the density value is 140, and when the slope value is 1.75, the density value is greater than 140.

Figure 18 shows the distribution of OldPeak related to the ECG signals, which is high when the depression rate is 0 and gradually decreases when the rate is 6.

Moreover, the density curve is high at peak 0 and shows a decrease in frequency from peaks 1 to 6.

Figure 19 illustrates the distribution between Nonesang and Exang. A value of 0 indicates a greater number of observations, which is approximately 200, and a value of 1 indicates fewer observations, which may be approximately 50. Hence, the density curve nearer to 1 is high when compared with the peak value of 0.

Figure 20 shows that a value of 0 has a high value of approximately 140, and its density curve has a high frequency. A value of 1 has a value above 140, and the density curve shows a high peak similar to peak 0. Moreover, a value of 2 results in fewer observations and is less than both 0 and 1, and its peak is also very low compared with peaks 0 and 1.

Figure 21 depicts the distribution of Fbs. Fbs denotes fasting blood sugar, which is depicted in the figure. Typically, fasting blood sugar levels may not be a good indication of heart illness in the dataset utilized. When FBS is approximately 0.0, the density value exceeds 250; however, if the value of FBS is between 0.8 and 1.0, then it is less than the density value of 50.

3.5. Performance Analysis

The proposed model's performance on three different classification datasets, namely the heart, thyroid, and hepatitis datasets, is analyzed via various metrics, such as accuracy, recall, F1 score, and precision. The heart rate prediction datasets for regression include the MAE, R^2, MSE, and RMSE.

Figure 22 shows the distribution of the cholesterol content, where the value increases and decreases within the range of 0- 400. Therefore, the experimental results indicate that better outcomes are obtained via the proposed model.

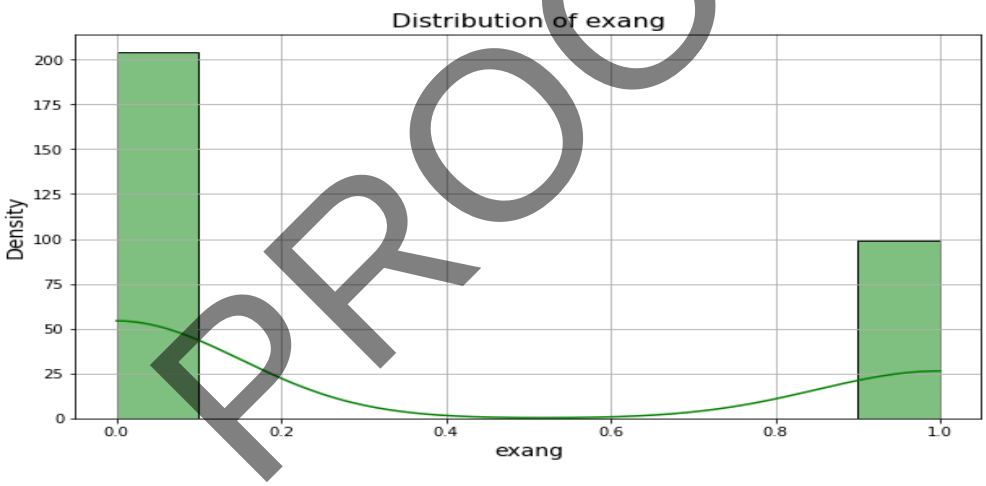

Figure 19. Distribution of Exang

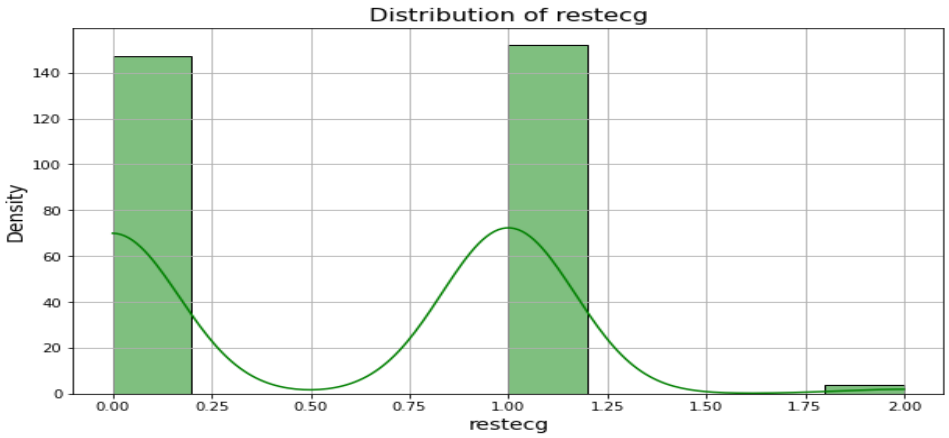

Figure 20. Distribution of Restecg

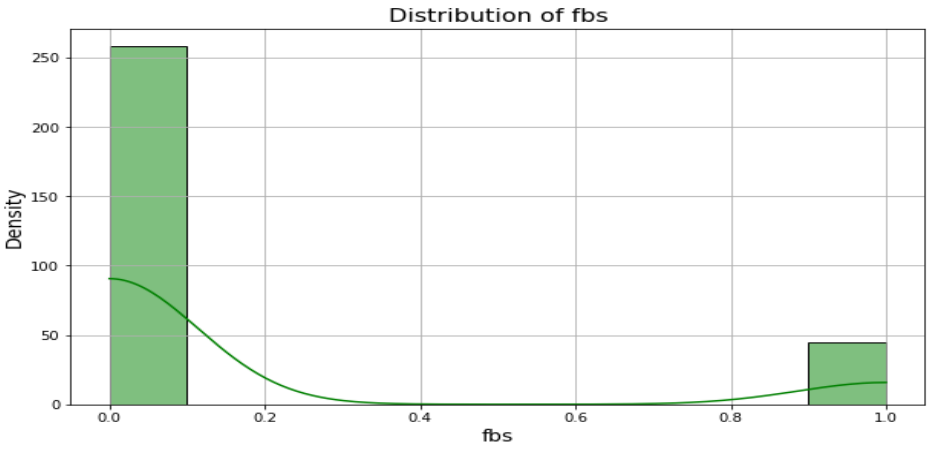

Figure 21. Distribution of Fbs

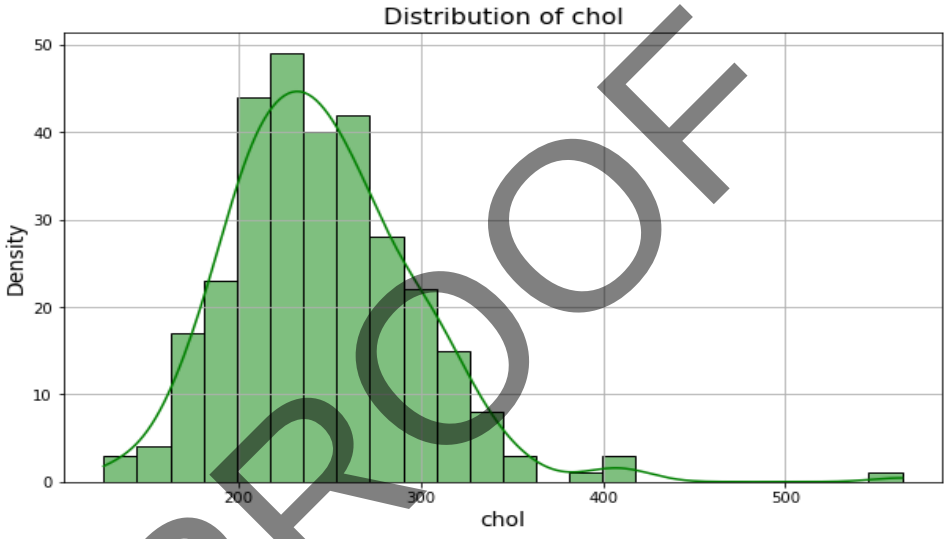

Figure 22. Distribution of Chol

3.5.1. Classification

Various metrics have been considered for the classification of diseases via heart, thyroid, and hepatitis datasets. The accuracy obtained by the proposed model for the heart disease dataset is 97.8%; likewise, the precision, recall, and F1 scores obtained by the proposed model for the heart disease dataset are 98%, 98%, and 98%, respectively. Table 6 shows the performance analysis of the proposed model.

Table 6 shows the performance metrics for three datasets: Heart Disease, Thyroid, and Hepatitis. Heart disease had the highest accuracy at 97.8%, with precision, recall, and F1 Scores all at 98%. The thyroid score was 97.2% accurate, and the hepatitis score reached 95%. This analysis underscores the varied predictive capabilities of models across different health datasets.

Table 6. Performance Analysis of the Proposed Model

Dataset	Accuracy (%)	Precision (%)	Recall (%)	f1-score (%)
Heart Disease	97.8	98	98	98
Thyroid	97.2	89	98	93
Hepatitis	95	98	88	92

3.5.2. Regression

Various metrics, including the MAE, R2, MSE, and RMSE, have been employed to analyze the performance of the proposed model for regression. Table 7 shows the performance analysis of the proposed model, in which the MAE value obtained by the proposed model is 0.112, the R^2 value obtained is 0.99, the MSE value obtained by the proposed model is 0.022, and the RMSE value attained by the proposed model is 0.1488.

Table 7 outlines key performance metrics for evaluating a predictive model. The Mean Absolute Error (MAE) is 0.112, reflecting the average magnitude of the prediction error. The R-squared (R^2) value is 0.99, indicating that a strong model fit explains 99% of the variance. The mean squared error (MSE) is 0.022, indicating a low prediction error, whereas the root mean squared error (RMSE) is 0.1488, indicating deviation from the actual values. Overall, these metrics indicate exceptional accuracy and reliability.

Table 7. Performance analysis of the proposed model

Performance Metrics	Values
MAE	0.112
R^2	0.99
MSE	0.022
RMSE	0.1488

3.6. Comparative Analysis

Comparative analysis is predominantly employed for comparing the existing works with the proposed work to evaluate the efficacy of the proposed model for classification on the basis of the accuracy of the existing and proposed models.

3.6.1. Classification

A comparative analysis of the existing and proposed models for three different datasets is mentioned in the subsequent sections.

Heart Disease Dataset

Table 8 shows the accuracy of the proposed and existing models for the heart disease dataset. The accuracy obtained by the existing model is 90.78, whereas the proposed model attained an accuracy rate of 97.8.

Table 8 illustrates the accuracy percentages of various classification algorithms in a specific analysis. The proposed method achieved the highest accuracy at 97.80%. Hyperparameter optimization via the Talos algorithm followed at 90.78%, whereas the K-NN algorithm reached 90.16%. The logistic regression and naïve Bayes methods both yielded 85.25% accuracy. The random forest results were close to 85.15%, and the SVM results were the lowest at 81.97%.

Table 8. Comparative analysis of the proposed and existing models [14]

Classification Algorithms	Accuracy (%)
Logistic Regression	85.25
K-NN	90.16
SVM	81.97
Naïve Bayes	85.25
Hyperparameter optimization (Talos)	90.78
Random forest	85.15
Proposed Model	**97.80**

Table 9 presents a detailed comparison of the accuracy of various classification algorithms, providing insights into their relative performance. Traditional models like Artificial Neural Networks and Multi-Layer Perceptron with Backpropagation achieve moderate accuracy levels of 80.00% and 80.99%, respectively. XG Boost performs slightly better with an accuracy of 85.68%, while Logistic Regression and a combination of several algorithms, including Logistic Regression, Deep Learning, Random Forest, Decision Trees, SVM, and Gradient Boosting, all reach around 84%. An effective convolutional model also achieves 85% accuracy. However, a stacked model involving KNN, RF, and SVM-RF surprisingly underperforms at 75.10%. Logistic Regression with Boruta Feature Selection improves upon basic logistic regression with an accuracy of 88.52%. Notably, the proposed model stands out with an exceptional accuracy of 97.80% for the heart disease dataset, significantly surpassing all other models.

Table 9. Comparative analysis of the study [43]

Classification Algorithms	Accuracy (%)
ANN	80.00
XG Boost	85.68
Logistic Regression	84.46
MLP with BP	80.99
LR, Deep learning, RF, DT, SVMand Gradient Boosting	84.00
An effective convolutional	85.00
A Stacked model involving KNN, RF, and SVM-RF algorithm	75.10
Logistic Regression with Boruta Feature Selection	88.52
Proposed Model	97.80

Table 10. Comparative analysis of the proposed and existing models [7]

Classifier	Accuracy (%)
KNN	59.00
ANN	94.00
Naïve Bayes	93.00
Random forest	94.80
Proposed	**97.20**

Thyroid Dataset

The accuracies of the proposed and existing models for the thyroid datasets are compared and tabulated in Table 10. The accuracy obtained by the existing model is 94.2, whereas the proposed model attained an accuracy rate of 97.2.

Table 10 shows the Accuracy percentages of classifiers. The proposed classifier performed excellently at 97.20%, followed by random forest at 94.80%. The results of the artificial neural network and naïve Bayes methods were 94.00% and 93.00%, respectively. K-nearest neighbors significantly underperformed at 59.00%, highlighting the effectiveness differences.

Table 11 provides a comprehensive comparison of the accuracy of various classification models, offering insights into their performance. The Convolutional Neural Network achieves an accuracy of 89%, while the SVM slightly outperforms it with 90%. The Random Forest model trails behind with an accuracy of 84%, indicating its limitations in this context. Notably, the NL (SMOTE-NC-LGB) model, which likely incorporates techniques like SMOTE for oversampling and Light Gradient Boosting, reaches a high accuracy of 96%. However, the proposed model surpasses all others with an impressive accuracy of 97.20%, showcasing its superior performance for thyroid disease classification.

Hepatitis Dataset

Table 12 compares and tabulates the accuracies of the proposed and existing models for hepatitis datasets. The accuracy obtained by the existing model is 92%, whereas the proposed model attained an accuracy rate of 95%.

Table 12 shows the accuracy percentages of various algorithms. The proposed algorithm led with 95%, followed by AdaBoost at 92% and XGBoost at 90%. The random forest method achieves 86%, the logistic regression method achieves 82%, and the decision tree

Table 11. Comparative analysis of the proposed and existing models [44]

Classifier	Accuracy (%)
Convolutional Neural Network	89.00
Support Vector Machine	90.00
Random forest	84.00
NL (SMOTE-NC-LGB)	96.00
Proposed	97.20

Table 12. Comparative analysis of the proposed and existing models [45]

Algorithm	Accuracy (%)
Decision Tree	73
Logistic Regression	82
Support Vector Machine	73
Random Forest	86
AdaBoost	92
XGBoost	90
Proposed	**95**

and SVM methods achieve 73%, highlighting the proposed method's superior performance.

Table 13 presents a comparison of the accuracy between an existing model and a proposed model, highlighting a significant improvement in performance. The existing model, which combines Bayesian methods with lion optimization, achieves an accuracy of 77%, indicating a moderate level of effectiveness but with room for improvement. In contrast, the proposed model demonstrates a substantial enhancement, reaching an accuracy of 95%. This 18% increase indicates that the proposed approach is much more effective at making accurate or classifications.

Table 13. Comparative analysis of the proposed and existing models [46]

Algorithm	Accuracy (%)
Bayesian + Lion optimization (Existing Model)	77
Proposed	**95**

This superior classification performance of heart, thyroid, and hepatitis diseases is due to the incorporation of M-PCA and M-BP, which increases the efficacy of the proposed model.

The experimental outcome revealed that the proposed model performed better for both the classification and regression processes than the existing models did because of the incorporation of M-PCA for dimensionality reduction to extract appropriate features for the model by implementing single-unit learning and second-unit learning (L1 and L2). Furthermore, an effective classification and regression process is carried out by employing M-BP with the AWGN algorithm, as the AWGN algorithm makes the model robust for effective classification and regression.

3.7. Statistical Section

Classification Model Cross-Validation Results (k=5)

Table 14 shows the accuracy results for the Heart Disease, Thyroid, and Hepatitis datasets through cross-validation. Heart disease achieves 97.6% to 98.1% accuracy (± 0.4% margin). The accuracy of thyroid parameters ranges from 96.9% to 97.4% (± 0.4% to ± 0.5%). Hepatitis records 94.7% to 95.2% accuracy (± 0.5% to ± 0.7%), indicating model reliability.

Heart rate prediction regression cross-validation results (k=5)

Table 15 presents the performance metrics of the mean absolute error (MAE), R-squared (R^2), mean squared error (MSE), and root mean squared error (RMSE) over five cross-validation folds. The MAEs range from 0.110 to 0.114, the R^2 values range from

Table 14. Classification and cross-validation

k-Fold	Heart Disease Accuracy	Thyroid Accuracy	Hepatitis Accuracy
1	97.6% ± 0.4%	97.0% ± 0.5%	94.8% ± 0.6%
2	98.1% ± 0.3%	97.4% ± 0.4%	95.2% ± 0.5%
3	97.7% ± 0.4%	96.9% ± 0.6%	94.7% ± 0.7%
4	97.9% ± 0.3%	97.3% ± 0.4%	95.1% ± 0.5%
5	97.7% ± 0.4%	97.4% ± 0.5%	95.2% ± 0.6%

Table 15. Regression cross-validation

k-Fold	MAE	R^2	MSE	RMSE
1	0.113 ± 0.008	0.989 ± 0.002	0.023 ± 0.003	0.152 ± 0.010
2	0.110 ± 0.007	0.991 ± 0.001	0.021 ± 0.002	0.145 ± 0.008
3	0.114 ± 0.009	0.988 ± 0.002	0.024 ± 0.003	0.155 ± 0.011
4	0.111 ± 0.007	0.990 ± 0.001	0.022 ± 0.002	0.148 ± 0.009
5	0.112 ± 0.008	0.990 ± 0.001	0.020 ± 0.002	0.141 ± 0.008

0.988 to 0.991, the MSEs range from 0.020 to 0.024, and the RMSEs range from 0.141 to 0.155. These results indicate the model's consistent and reliable predictive performance across different folds.

4. Conclusion

Chronic diseases are fatal and deadly and need to be predicted and diagnosed accordingly. Therefore, effective methods need to be used to identify the presence of disease. However, existing AI methods are feeble for effective classification and prediction of disease due to the incorporation of ineffective algorithms. Additionally, normal PCA with BP serves as a traditional dimensionality reduction method, but the modified PCA model improves feature extraction for heart rate prediction. The proposed M-BP model employs AWGN for increased robustness, helps the model manage noise, and improves generalization, thus enhancing overall data analysis efficiency; hence, to overcome these issues, the proposed model employs M-PCA for dimensionality reduction, which aids in obtaining the best features for the model by considering two learning rates (L1 and L2).

Furthermore, M-BP with AWGN was employed to classify heart, thyroid, and hepatitis diseases and predict patients' heart rates. M-BP was used in the proposed model instead of conventional BP because of the slow convergence rate. Furthermore, the incorporation of the AWGN model made the proposed model robust and efficient for classifying diseases and predicting heart rate. Finally, the performance of the model was evaluated via various performance metrics for both classification and regression.

The accuracy obtained by the proposed model for the heart disease dataset was 97.8%, the precision obtained was 98%, the recall attained was 98%, and the F1 score attained was 98%. Similarly, the accuracy, recall, precision, and f1 score obtained by the proposed model for the thyroid dataset were 97.2%, 98%, 89%, and 93%, respectively. Finally, the values of accuracy, precision, recall, and F1 score obtained by the proposed model for hepatitis were 95%, 98%, 88%, and 92%, respectively. Like the classification of diseases, heart rate prediction was also evaluated via different metrics, such as the RMSE, MSE, MAE, and R2. The MAE obtained by the proposed model for the heart rate prediction dataset was 0.112; likewise, the R2 obtained was 0.99, the MSE attained was 0.022, and the RMSE value obtained was 0.1488. Using conventional methods such as M-PCA and M-BP was beneficial for limited datasets, as they require less data for effective training than deep learning models do. This approach enhances interpretability, which is vital in healthcare for understanding predictions impacting patient care. These methods are computationally efficient, facilitating feature extraction and classification. In the future, various DL algorithms can be employed to obtain enhanced accuracy and to make the existing model more robust for the prediction and classification of diseases, which can aid medical professionals in the early detection and diagnosis of diseases.

References

1- C Kishor Kumar Reddy, "An Efficient Healthcare Monitoring System for Cardiovascular Diseases using Deep Modified Neural Networks." (2023).

2- Devansh Shah, Samir Patel, and Santosh Kumar %J SN Computer Science Bharti, "Heart disease prediction using machine learning techniques." Vol. 1pp. 1-6, (2020).

3- Rahul Katarya, Sunit Kumar %J Health Meena, and Technology, "Machine learning techniques for heart disease prediction: a comparative study and analysis." Vol. 11pp. 87-97, (2021).

4- Yassine Habchi *et al.*, "Machine learning and vision transformers for thyroid carcinoma diagnosis: A review." *arXiv preprint arXiv:2403.13843,* (2024).

5- Yanan Che, Meng Zhao, Yan Gao, Zhibin Zhang, and Xiangyang Zhang, "Application of machine learning for mass spectrometry-based multi-omics in thyroid diseases." *Frontiers in Molecular Biosciences,* Vol. 11p. 1483326, (2024).

6- Hafiz Abbad Ur Rehman, Chyi-Yeu Lin, Zohaib Mushtaq, Shun-Feng %J Arabian Journal for Science Su, and Engineering, "Performance analysis of machine learning algorithms for thyroid disease." pp. 1-13, (2021).

7- Tahir Alyas, Muhammad Hamid, Khalid Alissa, Tauqeer Faiz, Nadia Tabassum, and Aqeel %J BioMed Research International Ahmad, "Empirical method for thyroid disease classification using a machine learning approach." Vol. 2022(2022).

8- Yanhui Guo, Yi Feng, Fuli Qu, Li Zhang, Bingyu Yan, and Jingjing %J Plos one Lv, "Prediction of hepatitis E using machine learning models." Vol. 15 (No. 9), p. e0237750, (2020).

9- Michael Onyema Edeh *et al.*, "Artificial intelligence-based ensemble learning model for prediction of hepatitis C disease." Vol. 10p. 892371, (2022).

10- Arsalan Khan, Moiz Qureshi, Muhammad Daniyal, Kassim %J Health Tawiah, and Social Care in the Community, "A Novel Study on Machine Learning Algorithm-Based Cardiovascular Disease Prediction." Vol. 2023(2023).

11- Nadikatla Chandrasekhar and Samineni %J Processes Peddakrishna, "Enhancing Heart Disease Prediction Accuracy through Machine Learning Techniques and Optimization." Vol. 11 (No. 4), p. 1210, (2023).

12- Subhash Mondal, Ranjan Maity, Yachang Omo, Soumadip Ghosh, and Amitava Nag, "An Efficient Computational Risk Prediction Model of Heart Diseases Based on Dual-Stage Stacked Machine Learning Approaches." *IEEE Access,* (2024).

13- Farida Brahimi, Aicha Aid, Mourad Amad, Abdelghani Mehennaoui, and Abderahmane Baadache, "Enhanced K-Nearest Neighbors for Smart Cardiovascular Disease Prediction in IoT System." *Revue d'Intelligence Artificielle,* Vol. 38 (No. 4), (2024).

14- Sumit Sharma, Mahesh %J International Journal of Innovative Technology Parmar, and Exploring Engineering, "Heart diseases prediction using deep learning neural network model." Vol. 9 (No. 3), pp. 2244-48, (2020).

15- Hosam F El-Sofany, "Predicting heart diseases using machine learning and different data classification techniques." *IEEE Access,* (2024).

16- Awais Mehmood *et al.*, "Prediction of heart disease using deep convolutional neural networks." Vol. 46 (No. 4), pp. 3409-22, (2021).

17- Muhammad Amir Khan *et al.*, "Optimal feature selection for heart disease prediction using modified Artificial Bee colony (M-ABC) and K-nearest neighbors (KNN)." *Scientific Reports,* Vol. 14 (No. 1), p. 26241, (2024).

18- C Ahmed Telmoud, M Mohamed Saleck, and M Cheikh Tourad, "ADVANCING HEART DISEASE DIAGNOSIS AND ECG CLASSIFICATION USING MACHINE LEARNING." *J Theor Appl Inf Technol,* Vol. 102pp. 2608-23, (2024).

19- Syed Nawaz Pasha, Dadi Ramesh, Sallauddin Mohmmad, and A Harshavardhan, "Cardiovascular disease prediction using deep learning techniques."

20- Mohd Ashraf, MA Rizvi, Himanshu %J Asian Journal of Computer Science Sharma, and Technology, "Improved heart disease prediction using deep neural network." Vol. 8 (No. 2), pp. 49-54, (2019).

21- Dengqing Zhang *et al.*, "Heart disease prediction based on the embedded feature selection method and deep neural network." Vol. 2021pp. 1-9, (2021).

22- Shambhu Bhardwaj, Vipul Vekariya, Baldev Singh, Sri Vinay, Alli Arul, and Maria Daya Roopa, "Improved healthcare monitoring of coronary heart disease patients in time-series fashion using deep learning model." *Measurement: Sensors,* Vol. 32p. 101053, (2024).

23- Misha Urooj Khan *et al.*, "Artificial neural network-based cardiovascular disease prediction using spectral features." Vol. 101p. 108094, (2022).

24- Aniruddha Dutta, Tamal Batabyal, Meheli Basu, and Scott T %J Expert Systems with Applications Acton, "An efficient convolutional neural network for coronary heart disease prediction." Vol. 159p. 113408, (2020).

25- Mert Ozcan and Serhat %J Healthcare Analytics Peker, "A classification and regression tree algorithm for heart disease modeling and prediction." Vol. 3p. 100130, (2023).

26- Wasif Akbar *et al.*, "Probability Based Regression Analysis for the Prediction of Cardiovascular Diseases." Vol. 75 (No. 3), (2023).

27- C Uma and P Rathiga, "An Optimized Deep Ensemble Super-Learner Model For Thyroid Disease Classification." *Library of Progress-Library Science, Information Technology & Computer,* Vol. 44 (No. 3), (2024).

28- Dhyan Chandra Yadav, Saurabh %J Indian Journal of Public Health Research Pal, and Development, "Discovery of hidden pattern in thyroid disease by machine learning algorithms." Vol. 11 (No. 1), pp. 61-66, (2020).

29- George Obaido *et al.*, "An improved framework for detecting thyroid disease using filter-based feature selection and stacking ensemble." *IEEE Access,* (2024).

30- Md Asfi-Ar-Raihan Asif *et al.*, "Computer Aided Diagnosis of Thyroid Disease Using Machine Learning Algorithms."

31- Khandaker Mohammad Mohi Uddin, Abdullah Al Mamun, Anamika Chakrabarti, and Rafid Mostafiz, "An ensemble machine learning-based approach to predict thyroid disease using hybrid feature selection." *Biomedical Analysis,* Vol. 1 (No. 3), pp. 229-39, (2024).

32- Mohemmed Sha, "Quantum intelligence in medicine: Empowering thyroid disease prediction through advanced machine learning." *IET Quantum Communication,* Vol. 5 (No. 2), pp. 123-39, (2024).

33- Md Riajuliislam, Khandakar Zahidur Rahim, and Antara Mahmud, "Prediction Of Thyroid Disease (Hypothyroid) In Early Stage Using Feature Selection And Classification Techniques."

34- Yasir Iqbal Mir, Sonu %J International Journal of Scientific Mittal, and Technology Research, "Thyroid disease prediction using hybrid machine learning techniques: An effective framework." Vol. 9 (No. 2), pp. 2868-74, (2020).

35- Surjeet Dalal *et al.*, "Enhancing thyroid disease prediction with improved XGBoost model and bias management techniques." *Multimedia Tools and Applications,* pp. 1-32, (2024).

36- Dinh-Van Phan, Chien-Lung Chan, Ai-Hsien Adams Li, Ting-Ying Chien, and Van-Chuc %J International Journal of Cancer Nguyen, "Liver cancer prediction in a viral hepatitis cohort: A deep learning approach." Vol. 147 (No. 10), pp. 2871-78, (2020).

37- Elias Dritsas and Maria %J Computers Trigka, "Supervised machine learning models for liver disease risk prediction." Vol. 12 (No. 1), p. 19, (2023).

38- Pyla Jyothi, A LOKESH KUMAR, G KAVYA SRI, D DAKSHAYANI, and K KAVYA, "DISEASE PREDICTION USING NAIVE BAYES, RANDOM FOREST, DECISION TREE, KNN ALGORITHMS." *I-Manager's Journal on Computer Science,* Vol. 11 (No. 4), (2024).

39- [Online]. Available: https://archive.ics.uci.edu/dataset/45/heart+disease.

40- Ross Quinlan, "Thyroid disease data set." *Thyroid Disease Data Set. https://archive. ics. uci. edu/ml/datasets/thyroid+ disease (accessed Jul. 03, 2022),* (1987).

41- [Online]. Available: https://archive.ics.uci.edu/dataset/46/hepatitis.

42- Heart Rate Prediction Dataset. [Online]. Available: https://www.kaggle.com/datasets/saurav9786/heart-rate-prediction?select=time_domain_features_train.csv.

43- G Manikandan, B Pragadeesh, V Manojkumar, AL Karthikeyan, R Manikandan, and Amir H %J Informatics in Medicine Unlocked Gandomi, "Classification models combined with Boruta feature selection for heart disease prediction." Vol. 44p. 101442, (2024).

44- Ali Raza, Fatma Eid, Elisabeth Caro Montero, Irene Delgado Noya, Imran %J BMC Medical Informatics Ashraf, and Decision Making, "Enhanced interpretable thyroid disease diagnosis by leveraging synthetic oversampling and machine learning models." Vol. 24 (No. 1), p. 364, (2024).

45- George Obaido *et al.*, "An interpretable machine learning approach for hepatitis b diagnosis." Vol. 12 (No. 21), p. 11127, (2022).

46- C Vijayalakshmi and S Pakkir %J Measurement: Sensors Mohideen, "Survival prediction model with lion optimization for Hepatitis-B patients using supervised and unsupervised learning algorithm." Vol. 32p. 100958, (2024).