2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT) | 979-8-3503-3509-5/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICCCNT56998.2023.10306451

Unsupervised Binary Classification of Heart Diseases Using an Autoencoder Model with Boosting Algorithm

Md. Akil Raihan Iftee and Sunanda Das
Department of Computer Science and Engineering
Khulna University of Engineering & Technology
Khulna-9203, Bangladesh
iftee1807002@gmail.com and sunanda@cse.kuet.ac.bd

***Abstract*—Heart diseases are prevalent and encompass a range of cardiovascular disorders with significant health implications and are a leading cause of mortality worldwide. Early detection and effective management are vital in reducing the impact of heart diseases and improving patient outcomes. Traditional approaches in heart disease classification have limitations as they heavily rely on labeled data for training. Obtaining labeled data for training heart disease models is challenging due to privacy concerns and time constraints. This paper presents a new approach to heart disease classification using an unsupervised learning methodology. We employ an autoencoder model with a sigmoid-activated neuron in the last layer of the encoder part and the extracted feature from encoder part is splitted into two clusters based on a threshold value, representing whether or not cardiac disease exists. The unsupervised output of the autoencoder is then fed into the Adaboost algorithm, where misclassified data weights are adjusted iteratively. To ensure generalization across different sections, we train our model using a combined dataset comprising four datasets: Cleveland, Hungarian, Long Beach VA, Switzerland (CHLS dataset) and test on another dataset: Stalog (Heart) Dataset. The outcomes of our experiments validate the efficacy of our approach in obtaining high classification accuracy rates, offering a potential solution for early diagnosis and treatment interventions for heart diseases.**

***Index Terms*—Heart disease classification, Unsupervised learning, Autoencoder, Adaboost algorithm**

I. Introduction

Heart disease is a prevalent and potentially fatal health condition that affects millions of individuals worldwide. Accurate and early detection of heart diseases is crucial for enhancing patient outcomes and minimizing mortality rates.

Traditional approaches to heart disease classification typically rely on supervised learning techniques, which require labeled data for training. However, obtaining labeled data can be challenging and time-consuming, especially in the medical domain. Sometimes, it will arise privacy issues also. Moreover, some traditional methods often face limitations in accurately capturing the complex patterns and relationships present in heart disease data.

Heart disease prediction also uses several common unsupervised machine learning algorithms include k-means clustering, hierarchical clustering, t-distributed stochastic neighbor embedding (t-SNE) [1], principal component analysis (PCA) [2], and outlier detection algorithms such as isolation forest and local outlier factor. These algorithms help in identifying distinct groups or clusters within the data, reducing its dimensionality, and detecting abnormal patterns that may indicate the presence of heart diseases. However, these approaches may face challenges in accurately capturing the intricate relationships and complexities, dealing with overlapping symptoms, and providing interpretability inherent in heart disease data.

We propose an new approach to get beyond these limitations and drawbacks that leverages a unsupervised apporoach using autoencoder model. The autoencoder serves as a powerful unsupervised learning technique capable of learning meaningful representations from unlabeled data. We exploit the last layer of the encoder, consisting of a sigmoid-activated neuron, to divide the data into two clusters representing the presence or absence of cardiac illnesses. Subsequently, we include the AdaBoost algorithm, which utilizes the misclassified instances from the unsupervised output from the encoder to adjust the weights and improve the classification performance. By iteratively boosting the weak learners, our approach achieves enhanced accuracy and robustness in heart disease classification.

We have combined four diverse datasets: Cleveland, Hungarian, Long Beach VA, Switzerland (CHLS dataset) and test on a new dataset, Stalog (Heart) Data Set which is completely unknown to the model. This comprehensive dataset combination enables our approach to learn from a wide range of heart disease cases and effectively handle various scenarios and to assess the generalization performance for all data worldwide, Our model also gets evaluated using the Statlong dataset, where it demonstrates relatively higher accuracy compared to other techniques .

A summary of our contributions is as follows:

- Introducing a new approach that combines unsupervised learning and boosting algorithms for heart disease classification.
- Using autoencoder as an unsupervised learning technique, enabling dimensionality reduction while effectively capturing intricate patterns and relationships within a single-dimensional latent feature vector.

- Demonstrating the effectiveness of the AdaBoost algorithm in improving clustering accuracy.
- Improving the generalization and applicability of the proposed approach by combining diverse heart disease datasets.

In conclusion, our proposed approach presents a promising solution for heart disease classification by utilizing the effectiveness of boosting algorithms and unsupervised learning. The combination of these techniques allows for accurate and efficient detection of heart diseases, contributing to improved patient care and outcomes.

II. Related Work

The prediction of heart disease has drawn a lot of attention in the field of machine learning, and several methods have been put out for accurate prediction. In recent years, ensemble learning techniques have gained popularity due to their ability to improve prediction performance by combining multiple base models. The article [3] describes a study using data mining and neural network to predict cardiac illnesses. Dangare et al. evaluated on the Cleveland heart disease dataset and achieved an accuracy of 85.48% with a neural network model. The study being investigated highlights how data mining and machine learning approaches might be utilized to predict heart disease.

Mathan et al. [4] proposes Gini index based decision tree data mining method combined with neural networks for predicting cardiac illnesses. The article [5] proposes a more reliable sparse autoencoder-based artificial neural network (ANN) approach for the prediction of heart disease. As compared to previous approaches, Mienye et al. and his co-authors claim that their proposed approach achieves higher prediction results.

Miao et al. [6] proposes an approach based on deep neural network (DNN) for diagnosing coronary cardiac disease. The study shows promising results in identifying significant features and achieving high classification accuracy. The article [7] presents a diagnostic system based on optimized XGBoost for accurate prediction of cardiac diseases. The system achieved a high accuracy rate of 94.66%, making it a promising tool for diagnosing heart disease. The paper [8] describes the usage and utilization of multiple machine learning methods, including Decision Tree (DT), Support Vector Machine (SVM), and K-Nearest Neighbors (KNN), to predict heart disease. The study achieved its best accuracy by using the SVM algorithm.

Samhitha et al. [9] presents an approach to improving the accuracy in predicting heart disease by using machine learning algorithms. The authors use different models, including logistic regression, DT, random forest (RF), and KNN. The authors [10] of this paper used multiple machine learning algorithms including DT, RF, and neural networks to develop a universal cardiovascular disease prediction system. The efficiency of each algorithm was assessed using metrics such as accuracy, precision, and recall. Nazri et al. [11] proposed a voting based approach for predicting heart disease and they concluded that voting based method performed better than individual model such as SVM, KNN, etc.

Lakshmanarao et al. [12] used an ensemble of multiple machine learning models including RF, XGBoost, and SVM to predict heart disease. They also used feature selection techniques to identify the most important features for the prediction task. Latha et al. [13] employed a method to ensemble categorized to enhance the result of possibility of heart diseases. They utilized a combination of multiple classification models to create an ensemble, leveraging their collective decision-making capabilities to improvement of the performance of predictions.

In the study of Hassan et al. [14] proposed a cardiac illness prediction model that utilizes pre-trained DNNs in combination with principal component analysis. The authors leverage the power of deep learning and dimensionality reduction techniques to enhance the accuracy of heart disease prediction. P. Dileep et al. [15] proposed a novel approach for heart disease prediction. The methodology involved utilizing a cluster-based bi-directional LSTM (C-BiLSTM) algorithm to effectively analyze and predict heart disease.

Liu et al. [16] proposed a hybrid RFRS-based (Random Forest and Rough Set) categorization method for the diagnosis of cardiac disease. The study utilized computational and mathematical methods to improve the performance of cardiac disease detection.

III. Methodology

This section describes the detailed methodology used in this study for predicting heart diseases using AdaBoost with L1-regularized neural networks as base estimators.

A. Data Preprocessing

1) Handling Missing values: We used two techniques to handle missing values. One of them is MICE (Multivariate Imputation by Chained Equations) fills in the missing values in a dataset using repeated imputations of the missing values based on the correlations between the features. Regression models are used to predict the missing values for each variable while taking into consideration the connections between the other variables in the dataset. Another technique that is used is the KNN imputer which fills in the values of the K nearest nearby data points. It selects the K nearest neighbors to fill in the missing values based on the distance between the available features.

2) Feature Encoding: One hot encoding is used in heart diseases prediction because some features, such as chest pain type and resting electrocardiogram results, are categorical in nature and cannot be directly inputted into a neural network. One hot encoding transforms these categorical features into binary vectors of 0s and 1s, where each feature category is represented by a unique vector. This allows the neural network to capture the relationship between the feature categories and the target variable. Without one hot encoding, the neural network may interpret the categorical features as continuous variables, which could lead to incorrect predictions.

3) Data Normalization: Data normalization is here to ensure that all input features are on the same scale, preventing certain features from dominating the model simply because

they have larger values. Normalizing the data also helps our model to converge faster and achieve better accuracy.

B. model build up

1) Base Estimator: An autoencoder network was used as the base estimator in the AdaBoost algorithm. Certainly! The encoder and the decoder are the two fundamental parts of the autoencoder model. The encoder part takes the feature portion of the training dataset as input and consists of several layers with dimensions of 128, 64, 32, and 16. These layers progressively reduce the dimensionality of the input data, extracting and compressing important features. The final layer of the encoder is a single neuron with a sigmoid activation function.

On the other hand, the decoder part of the autoencoder is responsible for reconstructing the input data from the encoded representation. It mirrors the structure of the encoder in reverse order, with layers of dimensions 16, 32, 64, and 128. The decoder layers progressively expand the dimensionality of the encoded data to match the original input dimensions. During the training process, The model obtains the ability to most correctly recreate the original input data. The goal is to reduce the discrepancy between the input and output data as much as possible., effectively capturing the underlying patterns and information contained within the training dataset. The L1 regularization applied to each dense layer of the autoencoder brings several benefits, and the L1 regularization strength was chosen by a grid search over a range of values. It helps in reducing overfitting by introducing sparsity in the learned representations.

Forward Propagation:

$$Z^{[l]} = W^{[l]} A^{[l-1]} + b^{[l]}, \tag{1}$$

$$A^{[l]} = \sigma(Z^{[l]}). \tag{2}$$

where $W^{[l]} \in \mathbb{R}^{c \times n^{[l-1]}}$, $A^{[l-1]} \in \mathbb{R}^{n^{[l-1]} \times m}$, $b^{[l]} \in \mathbb{R}^{c \times 1}$ and $A^{[l]} \in \mathbb{R}^{c \times m}$. Here, $W^{[l]}$ is the weight matrix connecting the previous layer $(l-1)$ to the current layer (l). $Z^{[l]}$ represents the weighted sum of the inputs and the bias term $b^{[l]}$. $A^{[l]}$ represents the activation of the current layer, obtained by applying an activation function (σ) to $Z^{[l]}$.

Backward Propagation:

$$dZ^{[l]} = dA^{[l]} \cdot \sigma'(Z^{[l]}) \tag{3}$$

$$dW^{[l]} = \frac{1}{m} dZ^{[l]} A^{[l-1]T} + \frac{\lambda}{m} \text{sign}(W^{[l]}) \tag{4}$$

$$db^{[l]} = \frac{1}{m} \sum_{i=1}^{m} dZ^{[l](i)} \tag{5}$$

$$dA^{[l-1]} = W^{[l]T} dZ^{[l]} \tag{6}$$

where $dZ^{[l]} \in \mathbb{R}^{c \times m}$ $dA^{[l]} \in \mathbb{R}^{c \times m}$, $dW^{[l]} \in \mathbb{R}^{c \times n^{[l-1]}}$, $db^{[l]} \in \mathbb{R}^{c \times 1}$ and $dA^{[l-1]} \in \mathbb{R}^{n^{[l-1]} \times m}$.

Here, $dW^{[l]}$, $dZ^{[l]}$, $dA^{[l]}$ represent the derivative of the weight, cost function, activation function respectably with respect to the weights of layer l.

Upon completion of training, the encoder extracts a single-dimensional latent vector from the final neuron. This process not only reduces the dimensionality of the input data but also captures the intricate patterns and relationships among the features. Subsequently, the latent vector is passed through a sigmoid function, which scales the values between 0 and 1. By comparing these values to a threshold θ, the training features are divided into two distinct clusters, indicating their association with heart diseases or lack thereof. This approach provides a more standardized and gentle method for clustering and identifying heart disease-related patterns. Importantly, this clustering is obtained in an unsupervised manner as it trained only on the training features, without relying on specific labels or information about heart diseases during training.

Here, Figure 1 shows the whole structure of the base estimator.

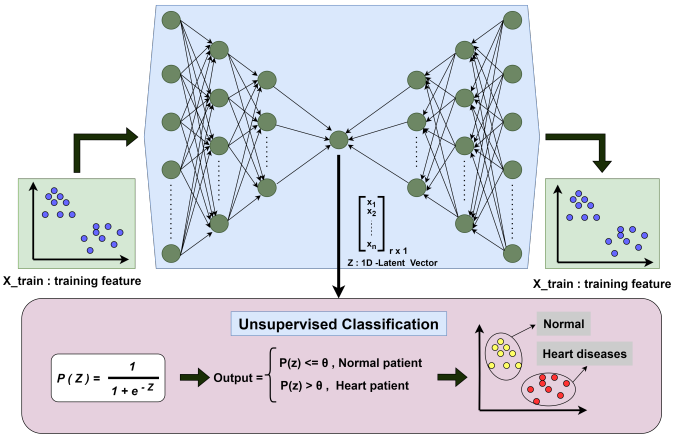


Fig. 1. Proposed weak classifier (Neural Network) for the AdaBoost algorithm.

TABLE I
MODEL SUMMARY OF THE WEAK CLASSIFIER

Layers	Output Shape	Parameters
Input	1x13	-
Dense-1	1x128	1,792
Dense-2	1x64	8,256
Dense-3	1x32	2,080
Dense-4	1x16	528
Dense-5	1x1	17
Dense-6	1x16	32
Dense-7	1x32	544
Dense-8	1x64	2,112
Dense-9	1x128	8,320
Output	1x13	1,677
Total parameters:	-	**25,358**

2) AdaBoost Algorithm: The AdaBoost algorithm (algorithm 1) was used to combine the weak learner into a strong ensemble classifier. It is an iterative algorithm that adjusts the weights of the misclassified samples in each iteration to emphasize the difficult samples in the training set. The final classifier is a weighted sum of the weak classifiers, where the weights depend on their classification accuracy.

In our approach, we employed the unsupervised technique of autoencoder as base estimator of adaboost like Figure 2,

where the encoder predicts the output for the validation data. The misclassified labels are then used to adjust the weights of the subsequent estimators, leading to a gradual improvement in the clustering performance of the autoencoder. This iterative process allows for the refinement and enhancement of the clustering capabilities over time. Such an approach enables the autoencoder to adapt to the underlying patterns and structures within the data, resulting in more accurate and effective clustering outcomes.

Algorithm 1 AdaBoost Training

for $m := 1$ *to* M **do**

(i). Fit the m^{th} autoencoder $h_m(x;\theta_m)$ using the weighted training data (X, y, D_{m-1}).

(ii). Compute the weighted error ϵ_m of $h_m(x;\theta_m)$:

$$\epsilon_m \leftarrow \sum_{i=1}^{n} D_{m-1}(i) \cdot I(y(i) \neq h_m(x(i);\theta_m))$$

(iii). Compute the weight of the m^{th} autoencoder α_m :

$$\alpha_m \leftarrow \frac{1}{2} \cdot \ln\left(\frac{1-\epsilon_m}{\epsilon_m}\right)$$

(iv). Update the weights of the training examples:

$$D_m(i) \leftarrow \frac{D_{m-1}(i) \cdot \exp(-\alpha_m \cdot y(i) \cdot h_m(x(i);\theta_m))}{Z_m}$$

(v). Normalize the weights $D_m(i)$ such that:

$$\sum_{i=1}^{n} D_m(i) \leftarrow 1$$

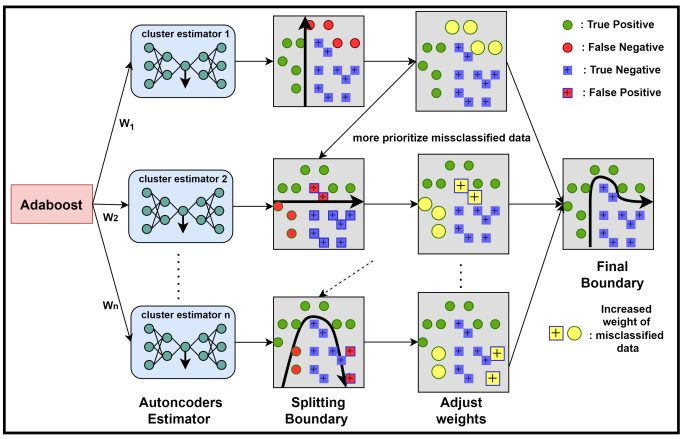


Fig. 2. Working Procedure of the AdaBoost algorithm, where N no of clustering by autoencoder are utilized to create a final strong classifier for the proposed system.

C. Implementation Details

The proposed method was implemented using Python 3.8 and trained on a NVIDIA GTX 1060 and 16 GB RAM. The scikit-learn library was used for machine learning algorithms and cross-validation. The neural network was implemented using the Keras library with TensorFlow backend.

In summary, the proposed methodology consists of pre-processing the dataset, applying L1 regularization for feature selection, using a neural network as the base estimator in AdaBoost, evaluating the performance using cross-validation, and comparing the results with traditional machine learning methods.

IV. Result and Discussion

A. Datasets

We utilized different variants of heart disease datasets to examine the performance of our proposed classification model. The first dataset, known as the Cleveland dataset [17], is widely recognized as one of the most popular datasets for heart disease prediction.

Most of the research work has evaluated on solo dataset. But we used in this study is a combination of four heart disease datasets : Cleveland, Hungary, Long Beach VA, and Switzerland (CHLS dataset) [18]. The dataset includes 14 common features that were used for curation. The Cleveland dataset consists of 303 observations, the Hungarian dataset has 294 observations, the Switzerland dataset has 123 observations, and the Long Beach VA dataset has 200 observations. In total, there are 1025 observations from these four datasets where 272 duplicated observations were removed from the dataset and left 920 observations. This large and curated dataset provides an excellent opportunity for heart disease prediction research.

Furthermore, we also trained our model using the combined data from these datasets, and then tested its performance on a separate dataset which is Statlong dataset [19]. This dataset comprises completely unknown data points and serves as a means to assess the generalization capability of our model beyond the training datasets. The Table II shows the details of common features of all mentioned dataset.

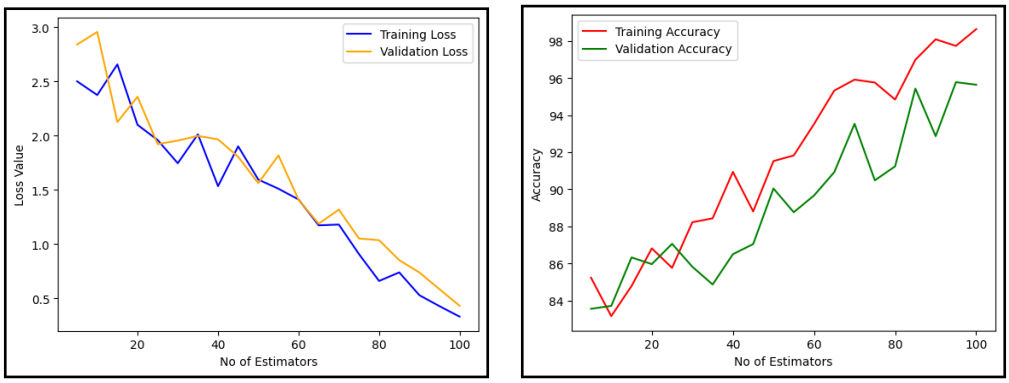
TABLE II
Features Information

Features	Description
Age	patient's age in years (Neumeric)
Sex(/gender)	sex [M: Male, F: Female] of the patient
ChestPain	4 types: typical, atypical, non-anginal, asymptomatic pain
RstingBP	measurement of BP [mm/Hg]
Chol	through BMI sensor, (Cholesterol) serum in mg/dl
FstingBS	(fasting glucose quantity >120 mg/dl) 1: Yes, 0: No
RstingECG	outcomes of a resting ECG (electrocardiogram)
MaxHrtRt	reached maximal heart rate or not
ExAngina	angina brought on by exercise [Y: Yes, N: No]
Ca	color range (0-3) for the quantity of major vessels
Oldpk	number used to quantify depression
ST_Slope	slope of the ST segment's peak exercise
HrtDisease	(heart disease) output class: 1, (normal) output class: 0

The dataset was split into a training set (80%) and a testing set (20%).

B. Model Evaluation

Five-fold cross-validation was used to assess the performance of the proposed strategy. The performance of the model was assessed using the following metrics:
$ND_{correct}$: Number of datas correctly classified as No Diseases, $HD_{correct}$: Number of datas correctly classified as Heart Diseases, $ND_{misclassified}$: Number of datas misclassified of No Diseases, $HD_{misclassified}$: Number of datas misclassified of Heart Diseases, and $Total_{datas}$: Total number of datas in the dataset for evaluation, TPR : True positive rate, FPR: False positive rate. So, here are the formulas of evaluation criteria:



(a) Loss curve of Adaboost (b) Accuracy curve of Adaboost

Fig. 3. The loss and accuracy curve of training and validation data of CHLS dataset

$$Accuracy = \frac{ND_{correct} + HD_{correct}}{Total_{datas}} \quad (7)$$

$$Precision = \frac{HD_{correct}}{HD_{correct} + HD_{misclassified}} \quad (8)$$

$$Recall = \frac{HD_{correct}}{HD_{correct} + ND_{misclassified}} \quad (9)$$

$$F1 - Score = \frac{2 \times (Precision \times Recall)}{(Precision + Recall)} \quad (10)$$

$$TPR = \frac{HD_{correct}}{HD_{correct} + ND_{misclassified}} \quad (11)$$

$$FPR = \frac{ND_{misclassified}}{ND_{correct} + ND_{misclassified}} \quad (12)$$

C. Comparision with other existing work

Initially, our proposed model was trained on the Cleveland dataset. Subsequently, The accuracy, precision, recall, F1 score, and AUC-ROC metrics were used to analyze the classifier's performance. Notably, our model achieved a remarkable accuracy of 97.23%, surpassing the performance of other prominent studies in the field. The detailed accuracy results are shown in Table III.

TABLE III
PERFORMANCE REVIEW OF THE PROPOSED METHOD AGAINST OTHER EXISTING WORKS ON CLEVELAND DATASET

Methods	Accuracy (%)
Latha et al. [3]	85.33%
Hybrid approach [13]	85.48%
HRFLM [20]	88.40%
Decision Tree + Random Forest [21]	88.70%
FCMIM-SVM [22]	92.37%
Neural network + PCA [14]	93.33%
C-BiLSTM [15]	94.78%
Gradient Boosting [23]	95.19%
Ours	**97.23%**

In order to enhance our model's overall performance in detecting heart diseases across various scenarios, we further trained it using a hybrid CHLS dataset. The resulting outputs were then evaluated, and the corresponding accuracy values are presented in Table IV. Our model achieved impressive metrics, including a maximum accuracy of 96.65%, precision of 92.40%, recall of 97.33%, and F1-score of 94.80%. Notably, the high recall value demonstrates the significant impact of our model in predicting medical diseases in the given dataset.

TABLE IV
PERFORMANCE COMPARISON OF METHODS TRAINED ON CHLS DATASET AND TEST ON IT WITH OTHER EXISTING MODEL AND EXISTING WORK

Methods	Accuracy	Precision	Recall	F1 Score
Dinesh et al. [24]	0.8651	-	-	-
Kmean clustering	0.8695	0.8695	0.8000	0.8333
Le et al. [25]	0.8993	-	-	-
Logistic Regression	0.9130	0.8470	0.96	0.8999
Xgboost	0.9239	0.8765	0.9466	0.9102
Catboost	0.9402	0.8902	0.9733	0.9299
Ours	**0.9565**	**0.9240**	**0.9733**	**0.9480**

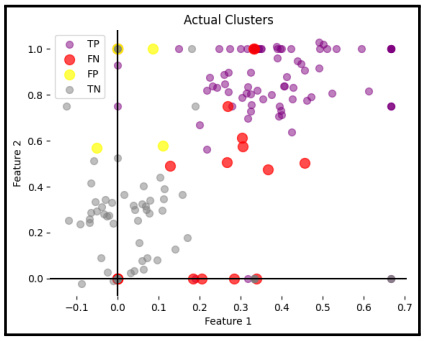
For the final generalized evaluation, the Statlong dataset, which was previously unseen by our model, was used. The performance of our proposed model surpassed that of other methods, indicating its strong capability in heart disease prediction across different datasets. This suggests that our model can be effectively employed for general heart disease classification purposes. The comparison results of the Statlong dataset are presented in Table V.

TABLE V
PERFORMANCE ASSESSMENT OF THE PROPOSED METHOD AGAINST OTHER EXISTING TECHNIQUES ON STATLONG DATASET

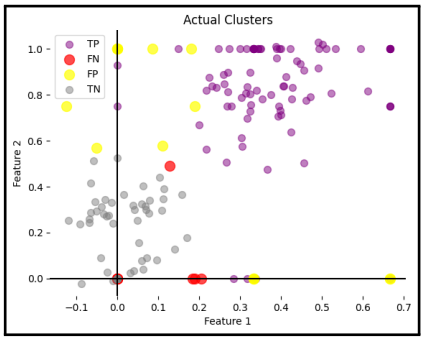
Methods	Accuracy (%)
Dwivedi et al. [26]	85.00%
RFRS classification system [16]	92.59%
Ours (trained with CHLS)	**92.65%**
HPDM [27]	95.90%
Ours (trained with Statlong)	**95.97%**

92.65% accuracy was attained using our proposed technique on the Statlong dataset when trained on the CHLS dataset. When trained on the self Statlong dataset, it achieved a maximum accuracy of 95.97%.

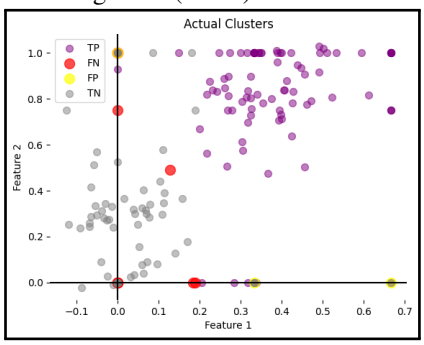
The training and validation losses can be observed in Figure 3, where the loss decreases and accuracy increases as the



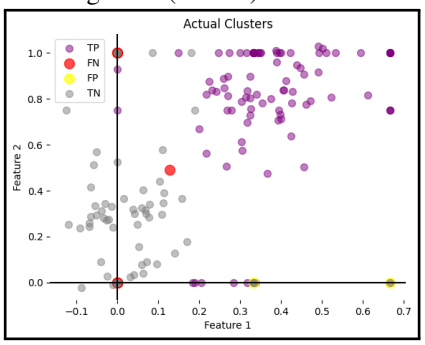
(a) clustering after (n = 1) estimators training



(b) clustering after (n = 20) estimators training



(c) clustering after (n = 50) estimators training



(d) clustering after (n = 100) estimators training

Fig. 4. Enhanced clustering accuracy achieved through incremental training of estimators. Here yellow(false positive) and red(false negative) points are decresing gradually.

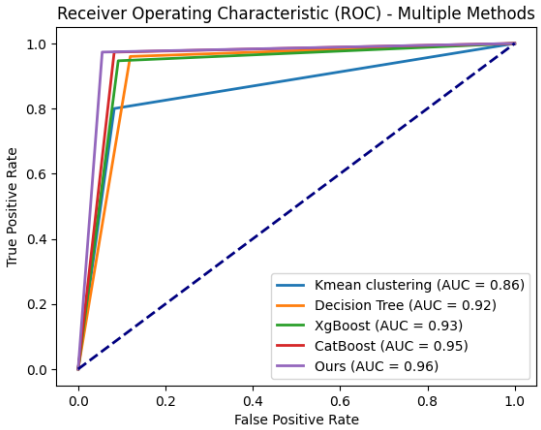


Fig. 5. Receiver Operating Characteristics Curve (ROC) for the proposed model and other existing model on CHLS dataset.

number of trained estimators increases. Additionally, a feature clustering analysis was performed by increasing the number of base estimators in our Adaboost model, as shown in Figure 4. The results demonstrated a gradual improvement in clustering performance with an increase in the number of trained estimators. This was observed through a reduction in the number of misclassified data points as the number of estimators increased.

When comparing the AUC-ROC curve with other existing techniques, our proposed method (In Figure 5) achieved a maximum value of 0.96. What this shows is that the classifier demonstrated a high level of discrimination between patients with and without heart disease. Overall, the results demonstrate the effectiveness of using AdaBoost with unsupervised base estimators such as autoencoder, incorporating L1 regularization on each layer, for predicting heart disease. This approach enables accurate and interpretable predictions of heart disease risk using readily available patient characteristics. The results emphasize the potential of this method in improving patient risk assessment and informing clinical decision-making.

V. Conclusion

This study focuses on a new unsupervised heart disease prediction model using the AdaBoost classifier with an autoencoder as the base estimator with L1 regularization. The model was trained and evaluated on diverse datasets, including a comprehensive dataset consisting of 920 patients with 13 input features. Comparative analysis with other widely used classification algorithms demonstrated the superior performance of our proposed model in terms of accuracy, precision, recall, and F1 score.

One key advantage of Our approach's capability to alleviate the reliance on labeled data, which is often scarce due to privacy concerns and the time-intensive nature of data collection. It presents a generalizable framework for heart disease prediction across various datasets. Future research could explore the applicability of this approach in other medical domains, harnessing its potential to enhance diagnostic capabilities and inform clinical decision-making.

REFERENCES

[1] J. Pirgazi, A. Ghanbari Sorkhi, and M. Iranpour Mobarkeh, "An accurate heart disease prognosis using machine intelligence and iomt," *Wireless Communications and Mobile Computing*, vol. 2022, 2022.

[2] A. K. Gárate-Escamila, A. H. El Hassani, and E. Andrès, "Classification models for heart disease prediction using feature selection and pca," *Informatics in Medicine Unlocked*, vol. 19, p. 100330, 2020.

[3] C. Dangare and S. Apte, "A data mining approach for prediction of heart disease using neural networks," *International Journal of Computer Engineering and Technology (IJCET)*, vol. 3, no. 3, 2012.

[4] K. Mathan, P. M. Kumar, P. Panchatcharam, G. Manogaran, and R. Varadharajan, "A novel gini index decision tree data mining method with neural network classifiers for prediction of heart disease," *Design automation for embedded systems*, vol. 22, pp. 225–242, 2018.

[5] I. D. Mienye, Y. Sun, and Z. Wang, "Improved sparse autoencoder based artificial neural network approach for prediction of heart disease," *Informatics in Medicine Unlocked*, vol. 18, p. 100307, 2020.

[6] K. H. Miao and J. H. Miao, "Coronary heart disease diagnosis using deep neural networks," *international journal of advanced computer science and applications*, vol. 9, no. 10, 2018.

[7] K. Budholiya, S. K. Shrivastava, and V. Sharma, "An optimized xgboost based diagnostic system for effective prediction of heart disease," *Journal of King Saud University-Computer and Information Sciences*, vol. 34, no. 7, pp. 4514–4523, 2022.

[8] G. Choudhary and S. N. Singh, "Prediction of heart disease using machine learning algorithms," in *2020 International Conference on Smart Technologies in Computing, Electrical and Electronics (ICSTCEE)*, pp. 197–202, IEEE, 2020.

[9] B. K. Samhitha, M. S. Priya, C. Sanjana, S. C. Mana, and J. Jose, "Improving the accuracy in prediction of heart disease using machine learning algorithms," in *2020 International Conference on Communication and Signal Processing (ICCSP)*, pp. 1326–1330, IEEE, 2020.

[10] E. Maini, B. Venkateswarlu, and A. Gupta, "Applying machine learning algorithms to develop a universal cardiovascular disease prediction system," in *International Conference on Intelligent Data Communication Technologies and Internet of Things (ICICI) 2018*, pp. 627–632, Springer, 2019.

[11] R. A. Nazri, S. Das, and R. T. H. Promi, "Heart disease prediction using synthetic minority oversampling technique and soft voting," in *2021 International Conference on Automation, Control and Mechatronics for Industry 4.0 (ACMI)*, pp. 1–6, IEEE, 2021.

[12] A. Lakshmanarao, A. Srisaila, and T. S. R. Kiran, "Heart disease prediction using feature selection and ensemble learning techniques," in *2021 Third International Conference on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV)*, pp. 994–998, IEEE, 2021.

[13] C. B. C. Latha and S. C. Jeeva, "Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques," *Informatics in Medicine Unlocked*, vol. 16, p. 100203, 2019.

[14] D. Hassan, H. I. Hussein, and M. M. Hassan, "Heart disease prediction based on pre-trained deep neural networks combined with principal component analysis," *Biomedical Signal Processing and Control*, vol. 79, p. 104019, 2023.

[15] P. Dileep, K. N. Rao, P. Bodapati, S. Gokuruboyina, R. Peddi, A. Grover, and A. Sheetal, "An automatic heart disease prediction using cluster-based bi-directional lstm (c-bilstm) algorithm," *Neural Computing and Applications*, vol. 35, no. 10, pp. 7253–7266, 2023.

[16] X. Liu, X. Wang, Q. Su, M. Zhang, Y. Zhu, Q. Wang, and Q. Wang, "A hybrid classification system for heart disease diagnosis based on the rfrs method," *Computational and mathematical methods in medicine*, vol. 2017, 2017.

[17] M. Lichman *et al.*, "Uci machine learning repository," 2013.

[18] "Heart disease dataset." Available: https:https://www.kaggle.com/datasets/johnsmith88/heart-disease-dataset. [Accessed Jan. 11, 2023].

[19] R. D. King, "Statlog databases," *Department of Statistics and Modelling Science, University of Strathclyde, Glasgow, UK*, vol. 535, 1992.

[20] S. Mohan, C. Thirumalai, and G. Srivastava, "Effective heart disease prediction using hybrid machine learning techniques," *IEEE access*, vol. 7, pp. 81542–81554, 2019.

[21] M. Kavitha, G. Gnaneswar, R. Dinesh, Y. R. Sai, and R. S. Suraj, "Heart disease prediction using hybrid machine learning model," in *2021 6th international conference on inventive computation technologies (ICICT)*, pp. 1329–1333, IEEE, 2021.

[22] J. P. Li, A. U. Haq, S. U. Din, J. Khan, A. Khan, and A. Saboor, "Heart disease identification method using machine learning classification in e-healthcare," *IEEE Access*, vol. 8, pp. 107562–107582, 2020.

[23] S. I. Sherly *et al.*, "An ensemble basedheart disease predictionusing gradient boosting decision tree," *Turkish Journal of Computer and Mathematics Education (TURCOMAT)*, vol. 12, no. 10, pp. 3648–3660, 2021.

[24] K. G. Dinesh, K. Arumugaraj, K. D. Santhosh, and V. Mareeswari, "Prediction of cardiovascular disease using machine learning algorithms," in *2018 International Conference on Current Trends towards Converging Technologies (ICCTCT)*, pp. 1–7, IEEE, 2018.

[25] H. M. Le, T. D. Tran, and L. Van Tran, "Automatic heart disease prediction using feature selection and data mining technique," *Journal of Computer Science and Cybernetics*, vol. 34, no. 1, pp. 33–48, 2018.

[26] A. K. Dwivedi, "Performance evaluation of different machine learning techniques for prediction of heart disease," *Neural Computing and Applications*, vol. 29, pp. 685–693, 2018.

[27] N. L. Fitriyani, M. Syafrudin, G. Alfian, and J. Rhee, "Hdpm: an effective heart disease prediction model for a clinical decision support system," *IEEE Access*, vol. 8, pp. 133034–133050, 2020.