Grenze International Journal of Engineering and Technology, January Issue

Deep Learning Approaches for Cardiovascular Disease Prediction

Anushka Sharma[1] and Rizwan Yousuf[2]
[1-2]Department of Mathematics, Chandigarh University, Mohali, Punjab, India, Pin:140413
Email: sharma.anushka.1902@gmail.com, rizwanwar50@gmail.com

***Abstract*— Cardiovascular diseases (CVDs) rank among the world's leading causes of death, and their effective treatments depend on early and precise diagnosis. Conventional diagnostic techniques can be time-consuming and frequently require a high level of medical skill, which limits their effectiveness in real-time applications. Therefore, thorough research is conducted by various researchers and scholars, using machine learning and deep learning algorithms to foretell heart diseases by using datasets and working on attributes related to the same. In our research, we looked over a dataset from Kaggle which contains attributes related to cardiovascular diseases, and explore the potential of deep learning methods in predicting cardiovascular diseases by comparing the accuracy of Deep learning algorithms such as Support vector machine (SVM), Decision tree (DT), K-nearest neighbour (KNN), ADA Boost, and XG Boost with hyperparameter optimization.**

***Index Terms*— deep Learning, KNN, SVM, Decision Tree, ADA Boost, XG Boost, Hyperparameter optimization, and heart disease dataset.**

I. INTRODUCTION

The human heart is the most significant part of the body, which pumps blood to every other organ and every nerve. When the heart does not work properly, it may lead to several serious issues and diseases, which can be fatal. Therefore, all around the world cardiovascular diseases have developed as one of the key reasons for death. As indicated by the World Health Organization (WHO) data, these diseases are accountable for taking 17.7 million lives invariably, and contributing to 31% of every single death across the globe [1]. Because of heart attacks and strokes, nine out of ten cardiovascular disease related deaths occur. CVDs caused 941,652 deaths in the United States alone in 2022, which is more than all types of cancer and chronic lower respiratory disorders combined [2]. Consequently, it has become crucial for the precise prediction of cardiovascular diseases. Several researchers worldwide began attempting to predict heart-related disease through the analysis of datasets related to the heart. Different machine learning methods are used that might potentially analyze large datasets and produce insightful findings. The presence and absence of cardiovascular diseases can be predicted accurately by machine learning models. Therefore, these algorithms have become very crucial in this aspect [3].

II. LITERATURE REVIEW

The incorporation of machine learning and deep learning techniques in the healthcare industry has significantly improved the accuracy of disease prediction models.

Grenze ID: 01.GIJET.12.1.97_1

To increase the predicted accuracy of heart disease diagnosis, V. Chaurasia, and Pal investigated various classification models, such as Naïve Bayes, J48 Decision Tree, and Bagging algorithms. Where bagging obtained a maximum accuracy of 85.03%, the study verified the efficacy of dimensionality reduction without compromising the classification performance [2]. Similarly, Vembandasamy et al. [4] constructed a predictive system for heart disease using the Naïve Bayes classifier. Their predictive system demonstrated that data mining techniques could reduce the need for expensive diagnostic tests without sacrificing high accuracy rates, which yielded an 86.41% classification accuracy. These studies and articles let us know how machine learning models can understand complicated medical data and enhance the accuracy of diagnosis. With ten techniques and four datasets of the UCI heart disease collection, Lakshmi et al. concluded an accuracy of 86.13% using Partial least square discriminant analysis method [5].

Chala Bayen et al. proposed data mining techniques for prediction and analysis of heart disease. J48, Support vector machine (SVM), and Naïve bayes (NB). It provides better outcomes, which save money for individuals and improve service quality [6]. Omankwu et al. have performed a study that utilizes models like KNN, SVM, and Naïve Bayes to compare deep learning methods for the prediction of cardiovascular disease and calculated that Talos hyperparameter tuning yielded improved accuracy up to a maximum of 90.78% [7]. K.C. Tan et al. [8] proposed a method that used combination of two deep learning algorithms at once, which were, Genetic algorithm and Support vector machine. A significant rate of accuracy was observed when the hybrid models were used [9]. R. Sharmila et al. suggested that Support vector machine with parallel fashion gave much improved results than sequential fashion. It was noticed that paraller SVM provided an efficient accuracy of 85% [10].

Ashwini Shetty et al. [11] used the WEKA tool and MATLAB, which are data mining methods for predicting diseases in the healthcare industry. They calculated an accuracy of 84% through Neural networks and an accuracy of 89% through Hybrid systems. Sahaya et al. did a comparitive analysis on present works of cardiovascular disease prediction that used data mining. They discussed the databases from the UCI repository, and the softwares like WEKA, KEEL, Data Melt, R data mining, Apache Mahout, Rattle, etc. Through their study, they came to the conclusion that prediction of diseases can be improved with better accuracy by using ensemble learning [12]. Maratea et al. [13] proposed a feature selection approach in sequential fashion that worked on Neuro fuzzy classifier and produced Cleveland Set accuracy of 88.2 percent.

III. DATASET

The dataset used in this study from Kaggle consists of 1,025 patient records with 14 attributes, sourced for cardiovascular disease prediction with both categorical and numerical features relevant to heart disease diagnosis.

TABLE I. ATTRIBUTES AND DESCRIPTION OF THE DATASET USED

No.	Attribute	Description
1.	Age	age of patient in years.
2.	Sex	1 = male, 0 = female.
3.	cp	type of chest pain.
4.	Resting BP (trestbps)	BP measured in mm Hg.
5.	Serum (cholestrol)	level of cholesterol in mg/dl.
6.	Blood sugar (fbs)	blood sugar while fasting >120 mg/dl
7.	restecg	ECG readings.
8.	Maxium heart rate attained (thalach)	feature that indicates maximun heart rate.
9.	Exercise instigated angina (exang)	0 = none, 1 = yes.
10.	Depression (oldpeak)	depression induced by exercise.
11.	Slope of peak (slope)	indicator of health of the heart.
12.	ca	Number of vessels colored by fluoroscopy.
13.	Thalassemia (thal)	variable related to blood disorder.
14.	Target variable	1 = heart disease is present, 0 = no heart disease.

IV. METHODOLOGY

A. Data Collection

This research employs a preprocessed cardiovascular disease dataset drawn from publicly available repositories to provide consistent and integrated data. The dataset contains several patient health factors, including age, gender, blood pressure, cholesterol level, and other clinical parameters relevant to cardiovascular health.

B. Data Preprocessing

Data preprocessing is needed to transform the dataset for the training and development of Machine learning models. Below are the steps that have been executed in preprocessing the dataset:
Deleting missing values: Inconsistent information or missing values are deleted.
Encoding: Categorical features are encoded numerically using label encoding.
Feature scaling: Normalized numeric values by StandardScaler for feature scaling.
Train-Test: Divide the data into training (80%) and test (20%) sets by stratified sampling.

C. Model Construction

Below are the five different constructed machine learning algorithms and their optimized versions:
K - Nearest Neighbor (KNN): We trained the KNN model using the Euclidean distance method. For determining the number of optimal neighbours, i.e., k, we conducted hyperparameter tuning using GridSearchCV.
Support Vector Machine (SVM): We used both radial basis function kernels and linear basis kernal when using SVM. Kernel parameters and C (penalty) were tuned with grid search as well as probability estimates for ROC analysis.
Decision Tree (DT): We used a Decision Tree Classifier with maximum depth as the primary hyperparameter. To prevent overfitting and underfitting, we used Grid search optimization for the identification of the best tree depth.
AdaBoost Classifier: Decision stump was utilized in applying AdaBoost, an ensemble method targeted towards boosting weak learners (commonly underfit the data). To optimize the model's learning rate and the number of estimators, we utilized grid search.
XGBoost Classifier: XG Boost is a framework for gradient boosting with an aim, to be faster and more performance-oriented. Some of the configurable parameters during the application of XGBoost classifier were learning rate, no. of estimators, and max depth of the trees. GridSearchCV facilitated the easy selection of the optimal hyperparameters.

D. Model Evaluation

The trained models were evaluated using multiple performance metrics:
Accuracy: Accuracy is the basic metric to calculate the performance of a Classification model. It can be calculated as the sum of true positive and true negative upon the total number of input samples.

$$Accuracy = \frac{TP + TN}{+++} . \qquad (1)$$

Equation (1) is, TP = True positive, and TN = True negative.
Precision: It can be calculated as the ratio of true positive instances to the sum of true positive and false positive instances.

$$Precision = \frac{TP}{TP+FP} . \qquad (2)$$

Equation (2) is, TP = True positive, and FP = False positive.

Recall: It can be calculated as the ratio of the predicted positive instances to the total no. of actual positive instances.

$$Recall = \frac{TP}{TP+F} . \qquad (3)$$

Equation (3) is, TP = True positive, and FN = False negative.

F1 Score: It is a Harmonic Mean of recall and precision and ranges from [0,1].
ROC and AUC:
ROC (Receiver Operating Characteristic) and Area Under the Curve (AUC) score is a metric that justifies how well a classifier can separate positive and negative instances.
We plot the true positive (TP) rate in comparison with the false positive (FP) rate at certain thresholds and compute the area under the curve.

E. Model Comparison

Bar chart: Displayed accuracy percentages of all optimized models.

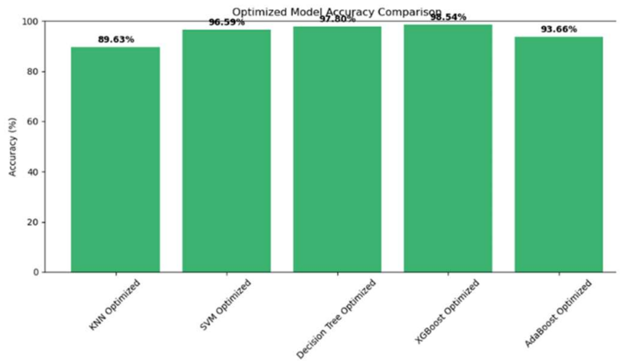

Figure 1. Accuracy Comparison with Bar Chart

ROC Curves: Were plotted to evaluate true positive vs. false positive rates.

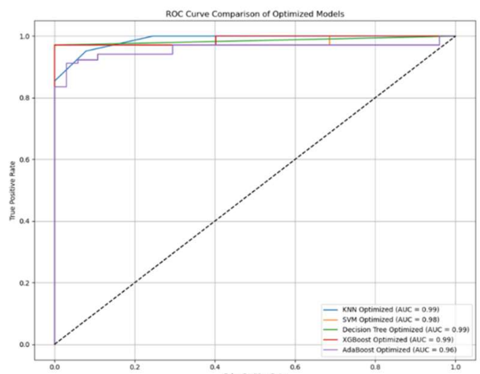

Figure 2. ROC Curve

F. Model Validation

Cross Validation: To measure model's performance, we divided the dataset into several folds. Then, a particular set of fold was used to train the model and the rest half to test it.

Grid SearchCV: It is used for optimal hyperparameter selection. It does cross-validation and picks the best configuration under a certain scoring criterion.

G. Results and Discussions

The accuracy and performance of each model was analyzed based on multiple Evaluation Metrics. The optimized versions of the algorithms showed varying levels of accuracy:

TABLE II. RESULTS

S. No.	Model	Accuracy
1	KNN Optimized	89.63%
2	SVM Optimized	96.59%
3	Decision Tree Optimized	97.80%
4	XG Boost Optimized	98.54%
5	ADA Boost Optimized	93.66%

V. CONCLUSION

This research was inspired by the growing cardiovascular-related problems. We aimed to study this dataset and determine which classification model would better predict diseases. Our study found that the XG Boost Optimized Model performed better at making cardiovascular disease predictions with this dataset, as indicated by the highest classification accuracy it reached. The ensemble AdaBoost and XGBoost models worked more stably across categorization thresholds, as per ROC curves. Classification reports and bar charts proved that ensemble methods generally achieved a better tradeoff between recall and precision.

FUTURE SCOPE

There is great potential in enhancing the early detection of diseases with the application of deep learning models, Ensemble learning, and CNNs i.e., Convolutional Neural Networks. They might offer better forecast precision rates than traditional machine learning models. With leveraging the integration of electronic health records (EHR) and wearable medical technology, we can enable early diagnosis and preventive treatment. Optimized models like XGBoost can be embedded with real-time cardiovascular risk prediction systems and other personalized systems. Principal Component Analysis (PCA) or autoencoders can be used in future work to reduce and minimize dimensionality and eliminate redundant features to ensure and enhance model efficiency and potential.

REFERENCES

[1] T. Z. Z. Z. D. W. J. H. a. X. D. S. Xu, “Cardiovascular Risk Prediction Method Based on CFS Subset Evaluation and Random Forest Classification Framework,” in IEEE 2nd International Conference on Big Data Analysis ,228-232., Beijing China, 2017.

[2] V. a. S. P. Chaurasia, “Data Mining Approach to Detect Heart Dieses,” in International Journal of Advanced Computer Science and, 2013.

[3] S. N. Pasha, “Cardiovascular disease prediction using deep learning techniques,” in OP Conf. Series: Materials Science and Engineering 981 (2020) 022006, 2020.

[4] K.Vembandasamy, “Heart Diseases Detection Using Naive Bayes Algorithm,” in IJISET - International Journal of Innovative Science, Engineering & Technology, Vol. 2 Issue 9, September 2015. , 2015.

[5] K. Lakshmi, “Performance Comparison of Data Mining Techniques for Predicting of Heart Disease Survivability,” in International Journal of Scientific and Research Publications, Volume 3, Issue 6, June 2013 1, 2013.

[6] M. C. Beyene, “Survey on Prediction and Analysis the Occurrence of Heart Disease Using Data Mining Techniques,” in International Journal of Pure and Applied Mathematics Volume 118 No. 8 2018, 165-174, 2018.

[7] O. Omankwu, “Cardiovascular (Heart) Diseases Prediction using Deep Learning Neural Network Model,” in Journal of Science and Technology Research 6(1) 2024 pp. 28-40 ISSN-2682-5821 , 2024.

[8] E. J. T. Q. Y. a. K. C. G. K. C. Tan, “A hybrid evolutionary algorithm for attribute selection in data mining,” in Expert Systems with Applications, Vol.36, No.4, pp.8616-8630, 2009. , 2009.

[9] M. P. Sumit Sharma, “Heart Diseases Prediction using Deep Learning Neural Network Model,” in International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-9 Issue-3, January 2020.

[10] S. C. R. Sharmila, “A conceptual method to enhance the prediction of heart diseases using the data techniques,” in International Journal of Computer Science and Engineering, May 2018., 2018.

[11] C. N. Ashwini Shetty A, “Different Data Mining Approaches for Predicting Heart Disease,” in International Journal of Innovative in Science Engineering and Technology, Vol.5, May 2016, 2016.

[12] G. M. A.Sahaya Arthy, “A survey on heart disease prediction using data mining techniques,” 2018.

[13] H. R. M. a. S. Goudarzi, “A noninvasive method for coronary artery diseases diagnosis using a clinically interpretable fuzzy rule-based system,” in Journal of Research in Medical Sciences, Vol. 20, Issue 3, pp.214-223, March 2015., 2015.