RESEARCH ARTICLE OPEN ACCESS

Integrating Deep Learning and Ensemble Techniques for Enhanced Heart Disease Prediction

Shaikh Jamil Ahemad, Pritam Panda, Krishna Patra, Bhagyashree Mohanty

(Department of Computer Science and Engineering, Siksha 'O' Anushandhan (Deemed to be) University, Bhubaneswar, Odisha, India

Email: {skjaah, pritampanda777, krishnamahendra029, bhagyashreemohanty202021}@gmail.com

--*************************----------------------------------

Abstract:

Cardiovascular diseases (CVD) is amongst the leading causes of the increase in mortality rates worldwide. Various heart-related diseases like heart attacks, strokes, arrhythmias, congenital heart defects, ischemic heart disease, and heart failure has been increasing in the population in recent years. This study highlights the need for early and accurate prediction of heart disease by developing a predictive model formed by the integration of machine learning, deep learning and ensemble techniques in order to ensure early treatment and prevent misdiagnosis. A Kaggle dataset including 70,000 data points and 13 features is used for analysis. The approach of this study is to employ base classifiers like K-Nearest Neighbors (KNN), Random Forest (RF), Support Vector Machine (SVM), Multilayer Perceptron (MLP) and XGBoost (XGB) integrated with deep learning through one-dimensional Convolutional Neural Network (1D-CNN). These models are further combined using ensemble techniques like stacking over K-Nearest Neighbors (KNN), Random Forest (RF) Support Vector Machine (SVM) and a voting classifier to give the final prediction with both hard and soft voting methods. The model's performance is evaluated based on standards like accuracy and area under the ROC curve (AUC). SVM performed the best amongst all the base classifiers by attaining a maximum accuracy of 88.02% and AUC 0.95, followed by MLP with an accuracy of 87.85%, XGBoost with 87.83%, RF with 87.77% and lastly KNN with an accuracy of 86.64%. Furthermore, the stacking model gave an accuracy of 87.92%. The voting classifier employed on the model to achieve the best accuracy of 87.98% under hard voting and the highest AUC of 0.96 under soft voting.

Keywords — **Cardiovascular Disease, Deep Learning, Ensemble Techniques, Machine Learning, Prediction**

--*************************----------------------------------

I. INTRODUCTION

A major health disease, Cardiovascular Disease (CVD), has become a remarkable global health concern. This is a prime cause of the increasing mortality rate worldwide. American Heart Association report of 2022 reflects an estimated 19.1 million deaths were due to CVD worldwide. Globally, 244.1 million people suffer from ischemic heart disease [1]. Heart disease poses a major health challenge in the healthcare sector of India, with a significant mortality rate caused due to cardiovascular diseases. From a 2022 report, 'Accidental Deaths Suicides in India,' it is marked that heart failure accounts for a total of 25,000 to 28,000 deaths in India [2]. Certain unique biological channels at play amplify the CVD risk in the Indian population, thereby imposing a complex burden on India. The study suggests that Indians have a higher inclination for Coronary Heart Disease

(CHD) and CHD-related mortality rates in the young population, thereby specifying the challenges faced by the Indian population in combating heart disease [3]. Currently, different varieties of heart diseases are seen prevailing. Fatty deposit accumulated in the coronary arteries leads to the frequent form of heart disease called ischemic heart disease, which reduces the flow of blood to the heart muscle. Conditions like heart failure, where the heart is unable to provide blood needed by the body resulting from heart muscle damage or other underlying conditions. Irregular and abnormal rhythms may lead to complications like blood clots, strokes, and heart failure if left untreated; these are a few symptoms of arrhythmias [4]. Congenital heart defects are some of the configurational abnormalities of the heart present at birth, ranging from minor to complex malformations that require surgical interventions [5]. Genetic factors and acquired conditions like alcohol abuse are effective causes of cardiomyopathies, where the heart muscle affects its ability to pump blood [6]. CVD can affect any age group and is broadly categorized into two main risk factors: modifiable and non-modifiable. To elaborate, non-modifiable factors can be classified as risk factors that cannot be altered by age, genetics, gender, or ethnicity [7]. The severity of CVD and its risk increase as an individual ages due to changes in the heart and arteries and the possibility of developing heart disease at an early age is mostly found in men than in women, although women are more prone to heart disease largely due to the decrease in protective estrogen levels that may even exceed men's post-menopause [8]. Genetic factors have a great role, as people with a family background of heart disease, especially those with heart disease from an early age, are more likely to inherit heart disease [9]. Specific countries and certain ethnicities, like African Americans and South Asians, are susceptible to the risks of particular heart conditions [10]. Modifiable risk factors are more prone to lifestyle choices and medical treatment. Smoking for generations has damaged and increased the risk factor of heart muscle deterioration. Thus, quitting smoking can drastically reduce the risk. Lack of physical activities and a healthy diet are prime contributors to heart disease [11]. Regular exercise and the intake of nutritious foods such as fruits and vegetables can lower the possibility of heart disease [12]. Another risk factor is obesity, which has become a prime concern. Obesity causes many diseases like diabetes and hypertension [13]. With regular exercise and a strict healthy diet, the risk of hypertension and diabetes can be reduced effectively. Further modifiable factors like excessive alcohol consumption, which leads to uncontrollable stress, can increase the effects of heart dis eases [14]. With proper regulations in drinking and stress management techniques such as meditation, the risk can be reduced. Effectively controlling higher blood pressure and potent regulation of higher cholesterol can lower the risk of CVD. High blood sugar levels can increase cardiovascular problems. Hence, diabetes management can be an effective way to reduce such risks. Lastly, maintaining a healthy sleeping habit reduces sleep disorders like Sleep apnea, thereby maintaining overall cardiovascular health [15]. Keeping track of chronic diseases is critical, especially in developing countries where their prevalence is escalating. Generally, people consult with a cardiologist only when the symptoms are pre-dominant and they are in the advanced stages of heart conditions. The existing methods used for predicting the possibilities of coronary heart disease are usually expensive, can result in major side effects, and need specialized expertise, which limits their utility [16]. There is a preference for non-invasive methods, and predictive machine learning models show great promise. The pre-existing systems predict CVD with the use of traditional machine learning methods, ensemble methods, or solely deep learning approaches. The integration of conventional machine learning techniques, deep learning methods, and their combination using ensemble techniques in a cohesive model remains underexplored. This gap presents an opportunity to develop a more powerful system that increases the strengths of each method to improve early diagnosis accuracy. This research is dedicated to creating a clinically effective system to classify heart disease using state-of-the-art predictive techniques. Enhancing the monitoring of heart patients could drastically lower mortality rates [17]. This architecture creates a cost-effective and accessible tool as it helps doctors quickly analyze the patient problem using health data at the base level without needing the help of specialists [18]. Its distinct advantage is its reliance solely on clinical records, which does not necessitate a cardiovascular disease specialist. A more accurate model minimizes the risk of faulty diagnosis, avoiding wrong treatment. The contribution of this research is the creation of a model that has traditional techniques such as K-Nearest Neighbours (KNN), XGBoost (XGB), Support Vector Machine (SVM), and Random Forest (RF) along with Deep Learning models like Multilayer Perceptron (MLP) which acts independently and One-dimensional Convolutional Neural Network (1D-CNN) that is integrated with ensemble techniques like stacking and followed by a voting classifier, which includes hard voting and soft voting. The best of the two voting approaches is chosen to create a smoother decision boundary and make predictions with higher accuracy. This paper attempts to examine all the risks and factors that impact the heart and may lead to any kind of cardiac disease with the proposed approach. Section 2 contains the literature review, which shows the background study related to previous research on the Prediction of Heart Diseases. Section 3 includes the methodology that describes the datasets used and all the techniques used for the model architecture. The result and analysis of the proposed architecture are in Section 4. The conclusion and future scope of this research work have been included in Section 5.

II. LITERATURE REVIEW

In various medical fields, machine learning techniques and many modern methods are employed in medical diagnostics. These methods include deep learning, ensemble learning, and the use of several other machine learning models, which have been made possible by technological advancement and the availability of large amounts of data. One major area that has been studied using these state-of-the-art technologies is heart disease, the leading cause of the increase in mortality rates worldwide. Early prediction and accurate diagnosis of heart diseases are important because they improve treatment outcomes and reduce healthcare costs. These papers propose combining machine learning using ensemble methods with deep learning. Also, these sections showcase what has been done so far to predict heart disease using this approach. Some specific models do not just contribute to this field due to their strengths but also show how ML in the medical context evolves over time.

A. *Heart Disease Detection Models Based on Machine Learning*

The research work by Biswas et al. [19] improves human and artificial intelligence efficiency based on given principles. Various feature selection methodologies are used with different machine learning techniques, and it depicts that Random Forest has an accuracy of 94.51%. Further down the line, from ideas that build on the advantage of optimization of machine learning for clinically accurate work, Pathan et al. [20] report that with carefully selected feature sets, average diagnosis accuracy increases up to 75% from below 73% of all data used and reducing feature size increases the model accuracy within less time. A new decision support system that Rani et al. [21] developed can be used for the early prediction of heart diseases in areas with limited healthcare using different machine learning algorithms. Their system used algorithms such as SVM, Naive Bayes, Logistic Regression, Random Forest, and AdaBoost, with the highest accuracy achieved by Random Forest– 86.6% on the Cleveland Heart disease dataset. This affirms that Random Forest is efficient according to other research and shows that ML can be a part of the clinical workflow for better patient outcomes.

B. *Utilization of Deep Learning for Heart Disease Prediction*

The goal of the research conducted by Arroyo et al. [22] seeks to raise the Artificial Neural Networks model's prediction accuracy. This uses an algorithm to enhance the ANN, a 70,000-instance dataset with 12 variables. however, this is a challenge as determining the best networks, including the number of layers and the neurons in an ANN model, has remained hard to do. The use of GA in a hybrid GA-ANN model provides potential process automation.

The claim is that it could replace the inefficient or slow human-produced results with an efficient mechanism of automated evolution of these parameters. More specifically, the researchers prepared the data and then used a genetic algorithm to evolve neural networks over 25 generations. The networks are judged based on their qualities, and then researchers have moved forward with those that are most successful in terms of accuracy. The results are impressive i.e. a performance boost with the new model, up to 73.43% high, over the standard ANN and other conventional models like Decision Tree, Logistic Regression, Random Forest, KNN, and Support Vector Machines. The researchers conclude that the process used is effective and should be used in predicting cardiovascular disease and explored for similar tasks. Another research by Ramdass and Ganesan shows how large amounts of complex information could be handled through Neighbourhood Com ponent Analysis Optimized Multilayer Feed-Forward Neural Network, which optimized features by performing additional computation steps on existing architecture [23].The accuracy of such a neural network achieved after incorporating the abovementioned factor is determined at 96.03%. Another study by Mohamed Djerioui et al. [24] uses the Heart UCI dataset and MLP and LSTM models to predict CVD. The dataset contains a range of patient information, but the study utilises the most crucial 14 attributes, such as age and cholesterol. LTSM, usually associated with data that show sequencing and possess long-term patterns, outperformed MLP for testing, with 96.5% accuracy and 89.18%, respectively. However, while results show promise, LTC may be limited due to dataset peculiarities. Furthermore, the limited generalizability of results is likely tied to data specificity, while LTC's demands are heightened by computational nature. As the authors suggest, this will enable further use of LTSM for diagnosis. The research by Raniya R. Sarra et al. [25] reflects, an architecture for heart disease prediction which makes use of deep learning models. The main focus is to overcome the challenges of small and imbalanced datasets. In addition to this, One-Dimensional Convolutional Neural Networks (1D-CNN) and Bidirectional Long Short-Term Memory (Bi-LSTM) have been utilised for model creation. Also, Generative Adversarial Networks (GANs) are used for enhancing the data. The data that is generated through GANs helps to improve the efficiency of the datasets that are already in existence. The performance of the models is quite impressive as the outputs reflect that GAN 1D-CNN model achieves an accuracy, specificity, sensitivity, and F1-score of 99.1%, and an area under the curve (AUC) of 100%. Similarly, it can be seen that the GAN-Bi-LSTM model achieves a 99.3% accuracy, 99.2% F1-score, 99.2% specificity and 99.3% sensitivity, and 100% AUC. This indicates that the model has greater predictive abilities. Several machine learning and deep learning techniques are thoroughly analysed, as demonstrated by Tulasi Krishna Sajja et al. [26]. This makes use of the Cleveland dataset to predict heart problems related to the cardiovascular system. The methods include traditional machine learning algorithms such as Naive Bayes, Support Vector Machine (SVM) and Logistic Regression with both linear and radial basis function (RBF) kernels, along with a more complex Neural Network, specifically Convolutional Neural Network (CNN) that is the 'Proposed Network' of the researchers. All the performance of models are measured, which shows that the CNN algorithm performs better than others. Logistic Regression shows a training accuracy of 89.91% and testing accuracy of 86.83%, while 80.62% and 77.04% in training and testing is shown by Naive Bayes, respectively. SVM with a linear kernel achieves 90.61% training and 85.29% testing accuracy. An RBF kernel reaches 85.43% training and 81.96% testing accuracy and is the least useful, with 79.76% training and 68.86% testing accuracy. The proposed network's CNN algorithm outshines all with 95.04% and 94.78% training and testing accuracy, respectively. This results in superiority in diagnosing cardiovascular diseases. Dengqing Zhang et al. [27] reflect various applications of machine learning and deep learning techniques used to predict heart disease effectively. The study uses a new method combining an embedded feature selection technique based on the LinearSVC algorithm, utilizing the L1 norm for penalty and a sophisticated deep neural network (DNN) architecture. The dataset in the study is from Kaggle and includes various physiological and medical features. The accuracy, precision, recall, F1-score, and area under the curve (AUC) for ROC analysis are among the several metrics used to properly assess the network's performance. The model performs far better than all other models, as shown by its 98.56% accuracy, 97.84% precision, 99.35% recall, and 0.983 f1-score. These numbers provided a highly reliable method for early heart disease detection, leading to better clinical results.

C. *Machine Learning and Ensemble approaches-based models*

Machine Learning and Ensemble approaches-based models

The accuracy of machine learning models is increased by using ensemble techniques. This is achieved by combining several models so that the algorithm can learn pattern recognition through voting, averaging, and the like. The significance of Asif et al.'s [28] work lies in its prediction accuracy

for heart disease, which is based on ensemble methods. In this study, predictive accuracy at about 98.15% is reached through grid search CV with Extra Tree classifier during ensemble learning, and hyperparameter tuning over multiple models has been performed. It is clear from this demonstration alone that the application of ensemble approaches has taken heart disease detection beyond imagination in terms of its precision levels. Yewale et al. [29] show the implementation of general ensemble techniques to come up with a holistic heart disease predictor system that has 98% specificity, 100% sensitivity, and an overall accuracy rate equalling 98.73%. This only shows that errors in predicting cardiac problems can be minimized when combinations used are designed well enough to cover different aspects or angles of such issues as they may affect an individual's health state (or outcome). Shorewala [30] also shows how these methods could work even better if applied, more specifically, coronary artery blockage predictions based on botanical boosting stacking and bagging method; here, each patient record had been matched against over 70000 other records from a similar patient's database, thus giving 75.1%. Those numbers could have been bigger if laboratory data utilization during computations had not incurred some penalties leading to incorrect conclusions at the level of optimization of recall scores, as per executor remarks. Another research by Ghasemieh et al. [31] on ensemble methods shows Stacking Ensemble Learner (SEL) for the purpose of predicting urgent care re-admissions among patients with cardiovascular diseases who are hospitalized due to different types of acute myocardial infarction episodes. In SEL, different classes are set based on behavior-based characteristics, and the accuracy is 88.23% compared to standard algorithms like logistic regression or random forest. This particular research is a demonstration of how ensembles can be used to handle emergency health conditions.

Kavitha et al. [32] demonstrate the use of a new hybrid machine-learning model that uses Random Forest and Decision Tree algorithms in combination to detect heart diseases at early stages. That is the reason why the researchers have embraced mixture models. The second model receives probabilities from the first one, thereby improving overall accuracy. In this dataset, the authors show that the method performs best at 88.7% which surpasses individual Decision Trees (79%) and Random Forests (81%). This results in the development of an interface that allows users to enter parameters for predicting heart disease with the hybrid system. So far, no other researchers have proposed such methods of detection. According to these findings, there might be a possibility of helping diagnose and treat this condition at its earliest stages by improving predictive accuracy rates. Chandrasekhar et al. [33] have conducted a lot of research on how to make heart disease predictions more efficient using different machine-learning algorithms and optimization methods. The use of Random Forest, K-nearest neighbour (KNN), Logistic Regression (LR), Naïve Bayes (NB), Gradient Boosting (GB), and AdaBoost as the machine learning algorithms is done. All the datasets in this study, i.e., the Cleveland dataset and the IEEE Dataport datasets, are fully utilized. This aims to optimize accuracy through GridsearchCV, which is one of the optimization techniques used in this research. The further performance evaluation of these algorithms, along with those optimization techniques, is done by model accuracy optimization technique, also known as a fivefold system, where splitting data into five parts is done randomly but stratified by class labels such that each part contains about the same proportion of examples from different classes, then train a classifier on four parts, validated on the fifth part, and this process is repeated for all combinations parts. The logistic regression classifier achieved 90.16% accuracy in Cleveland dataset and AdaBoost outperformed others achieving an accuracy of 90% accuracy. At last, all the base classifiers were combined through soft voting to achieve an accuracy of 93.44% in Cleveland dataset and 95% in IEEE Dataport dataset.

D. Deep Learning and Ensemble approaches-based models.

The study by Baccouche et al. [34] presents an advanced ensemble-learning framework that integrates various neural network architectures for classifying types of heart disease. A dataset from Medica Norte Hospital in Mexico is used, which needs to be more balanced and more challenging. This dataset includes 800 patient records with 141 medical indicators. The preprocessing steps include feature selection, normalization, and dataset balancing through random under- sampling. These are crucial for managing data imbalance. The study employs a combination of CNN, LSTM, GRU, and their bidirectional variants. The use of such methods helps in proper predictions of the disease. The ensemble approach significantly enhances classification accuracy, achieving between 91% and 96% across various conditions. Another study by Almulihi et al. [35] explores an innovative ensemble learning approach to improve heart disease prediction, i.e. integrating two pre- trained hybrid deep learning models, i.e. CNN with LSTM and CNN with GRU. Support Vector Machine (SVM) is a meta- learner for the stack ensemble technique. Recursive Feature Elimination (RFE) selects optimal features from two heart disease datasets. This technique enhances model performance by focusing only on the important features. The proposed ensemble model presents impressive performance over traditional machine learning and individual deep learning models, achieving high metrics such as a 97.17% accuracy and a 97.15% F1-score on the Cleveland dataset.

This comprehensive literature review on heart disease pre- diction identifies the lack of integration between machine learning and deep learning using ensemble techniques as a notable research gap. Many researchers have demonstrated the individual efficiency of each model, but there is still much to be explored in terms of ensembling through these methods. Thus, diagnostic accuracy could be improved when they are combined. However, very few studies have used them together within one framework, which represents a missed opportunity for medical advancement. What this means is that if deep learning's intricate feature extraction capabilities are combined with the ensemble's strong generalization abilities while predicting heart disease accurately, it may become more than ever before. Therefore, such integration might create entirely new benchmarks for prognosis statistics, thereby giving doctors better tools to detect diseases earlier so that there are better outcomes among cardiovascular patients.

III. METHODOLOGY

A. Dataset Collection

The dataset named cardiovascular disease dataset is taken from Kaggle [36]. It consists of 70,000 records and 13 attributes (12 attributes + 1 target attribute). The dataset is divided into 80% for training and 20% for testing purposes. The description of the dataset can be found in Table 2.

TABLE I
DATASET DESCRIPTION

Attribute	Description
id	Unique identifier for each patient
age	Age of the patient in days
gender	Gender of the patient (1: female, 2: male)
height	Height of the patient in centimeters
weight	Weight of the patient in kilograms
ap_hi	Systolic blood pressure (the top number)
ap_lo	Diastolic blood pressure (the bottom number)
cholestero l	Cholesterol level (1: normal, 2: above normal, 3: wellabove normal)
gluc	Glucose level (1: normal, 2: above normal, 3: wellabove normal)
smoke	Smoking status (0: non-smoker, 1: smoker)
alco	Alcohol intake (0: non-drinker, 1: drinker)
active	Physical activity status (0: inactive, 1: active)
cardio	Presence of cardiovascular disease (0: no, 1: yes)

The worth of this study is that it is based on the cardio- vascular disease dataset while considering all the major risk factors, making it highly applicable in the clinic. Furthermore, it can be taken as a benchmark when testing other ways of doing medical research. The repository helps to keep relevant data accessible and maintains its quality for other researchers who may want to replicate experiments legally. Moreover, one can easily compare new methods with conventional ones through similar features, and its structure makes it easy to buildpredictive models. The architecture is built by training various independent machine learning, deep learning, and ensemble methods. Combining these models resulted in the hybrid architecture for enhanced heart disease prediction. This proposed methodology section contains the schematic layout of the architecture, the algorithm of the enhanced ensemble learning with Stacking, MLP, XGB and Voting, dataset description and details about how each model works.

B. Data Preprocessing

Different data preprocessing methods have been carried out to increase the quality of the heart disease dataset with 70000 patient records for model training. Initially, missing values have been inspected, but there are no missing values. This ensures the completeness of the heart disease dataset. Further, an inspection of duplicate records has been carried out, as duplicate records can impact the model's performance. Still, the heart disease dataset does not have any duplicate records. The absence of missing values and duplicate records reflects that the dataset is ready to proceed with further preprocessing steps.

1. **Outlier removal:** Outliers are those data points that have extreme values and can impact the model's overall performance. So, to remove these outliers, the percentile method has been used. The points that are absent between the range of 2.5 percentile to 97.5 percentile have been removed from the dataset. The outliers have been removed from the features, i.e., height, weight, ap hi (systolic blood pressure), and ap lo (diastolic blood pressure), as these mainly represent the numeric values and have great chances of having outliers in them. After the removal of outlier points, the dataset is left with 60412 patient records from 70000 patient records.

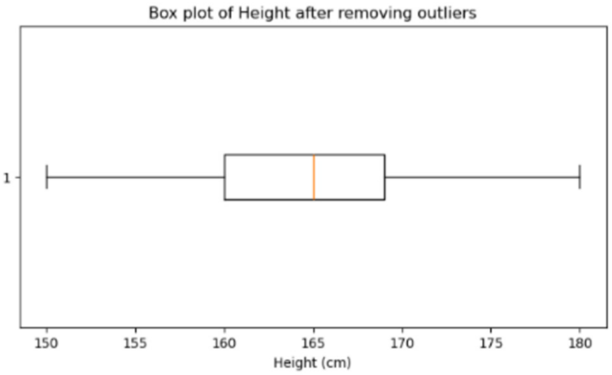

Fig. 1. Height

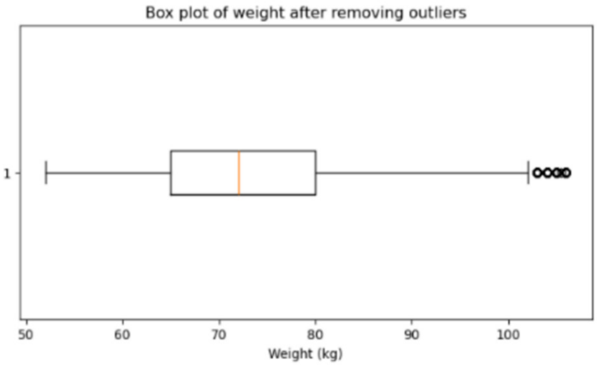

Fig. 2. Weight

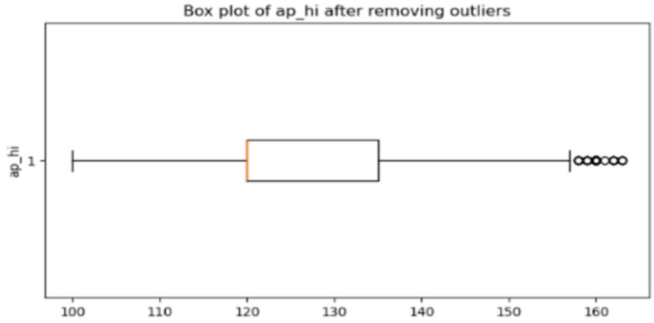

Fig. 3. ap_hi

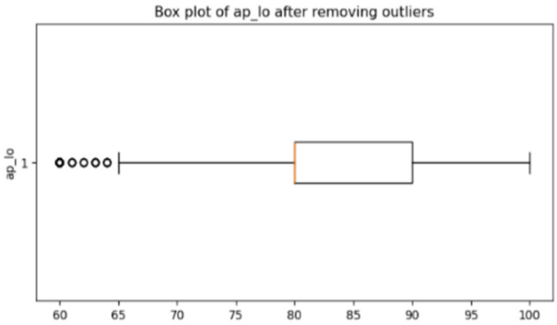

Fig.4. ap_low

2. **Logical consistency in blood pressure measurements:**
   In the measurement of blood pressure, it must be ensured that the systolic blood pressure is greater than the diastolic blood pressure. So, further inspection is done to check whether every record has systolic blood pressure values greater than the diastolic blood pressure. However, no values have been found where diastolic blood pressure is higher than systolic blood pressure, which ensures the correctness of the blood pressure measurements that have been recorded

3. **Age conversion methodology:** The age column is initially present in the form of days. The conversion from days to years is necessary for better understanding. So, the conversion was carried out by dividing the days by 365 and rounding them further into the nearest integer. This converts the age to years, which provides a better understanding of the dataset.

4. **Binning of age into age groups:** Binning increases the analysis of this dataset by converting the age to different ranges of age groups. This converts the continuous features.
5. **Attribute Combination:** The dataset can be optimized further by reducing the number of features. This can be done by combining some attributes like weight and height to Body Mass Index (BMI) and ap lo and ap hi to Mean Arterial Pressure (MAP). The BMI becomes an important factor for determining if a person is overweight, as overweight people are prone to heart diseases. The BMI is measured using a formula, which is a combination of heights and weights.

$$BMI = \frac{Weight(kg)}{(Height(cm))^2}$$

Similarly, an overall blood pressure can be measured using MAP, which is a combination of systolic and diastolic blood pressures.

$$MAP = \frac{(2 * ap\_lo + ap\_hi)}{3}$$

Further, the BMI and MAP have been categorized into 6 different groups, which simplifies the analysis.

6. **Clustering:** Further, clustering techniques have been used to group the data points in the dataset based on similarities. Commonly, K-means clustering is used for numerical data. It uses distance measures to calculate the distance and then update the clusters. But in this case, the dataset is already converted into a categorical with a set of categorical features. Hence, another famous k-mode clustering technique is used to check similarities in

categorical data and group them based on these categorical similarities. It identifies the frequently occurring values within every cluster for selecting the centroid. The elbow curve method is used to determine the optimal clusters. The costs are plotted against the number of clusters. This method looks for the elbow point in the graph, which represents the optimal number of clusters. In this case, the elbow point is approximately at the value of 2.0. So, the optimal number of clusters is taken as 2.

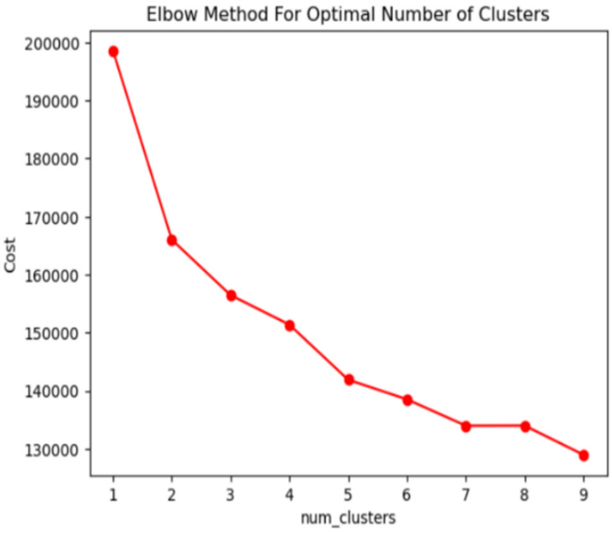

Fig. 5. Elbow Curve

7. **Correlation heatmap:** The correlation heatmap represents the features that have a strong positive correlation, no correlation, or a strong negative correlation. It ranges from -1 to +1, in which the features that correlate closer to -1 are said to have a strong negative correlation. Similarly, +1 is said to have a strong positive correlation. However, in the case of 0, it is said there is no correlation. In this case, we have only removed those points that do not correlate with the target feature 'cardio.' Upon plotting the correlation heatmap, it can be seen that features like gender, smoke, active, and alco are close to having 0 correlation with the target feature. Therefore, these are discarded to reduce the number of features in the dataset.

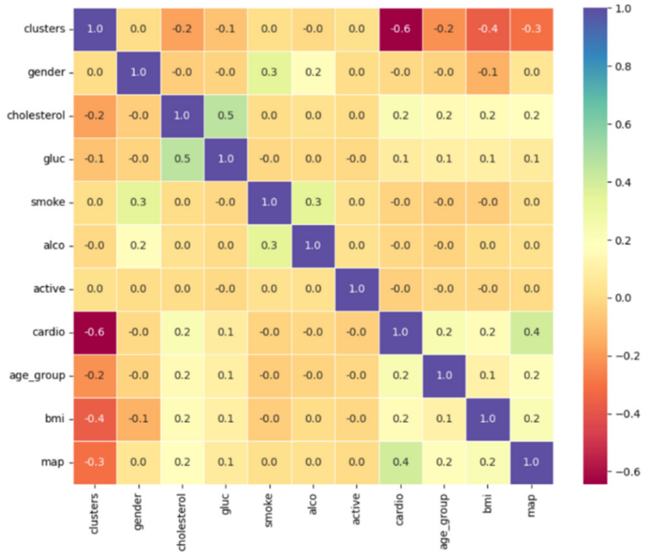

Fig. 6. Heatmap

C. Schematic Layout of Architecture

The dataset undergoes preprocessing before being input into the ensemble learner using the stacking technique. In stacking, traditional machine learning algorithms (KNN, RF, SVM) at Level 0 feed their outputs into a deep learning meta-learner (1D-CNN) at Level 1. Parallelly, the pre-processed data is put in MLP and XGB classifier. The outputs from stacking, MLP, and XGB classifier are then combined through a voting classifier, which applies hard and soft voting to determine the optimal decision boundary. The final model predicts the output (0 for absent, 1 for present).

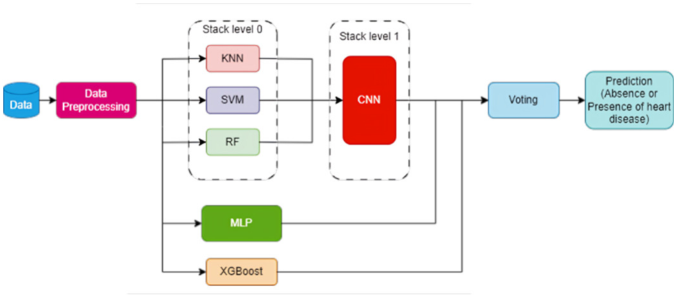

Fig. 7. Schematic layout of architecture

D. Enhanced Ensemble Learning with Stacking, MLP, XGB and Voting

1) Data:

• Training data Dtrain

• Testing data Dtest

2) Components:

- Ensemble classifiers S1 (stacked ensemble) and S2
- Traditional machine learning algorithms c1,c2,...,c3 (KNN, RF, SVM)
- Meta-classifier M (1D-CNN)
- XGBoost Classifier (XGB)
- MLP Classifier (MLP)

3) Output:

Z contains the predictions for each instance in (D test), indicating the presence (1) or absence (0) of the condition.

E. Ensemble models in CVD prediction

To find out the best results, the proposed methodology makes use of several machine learning, deep learning, and ensemble methods-based classification algorithms. In stacking, traditional methods (KNN, SVM, RF) are used as base classifiers at level 0. At level 1, 1D-CNN is used as a Meta-Learner. Simultaneously, the pre-processed data is also fed into the Multilayer Perceptron classifier. The final output of stacking, XGB, and MLP classifier is sent to the voting ensemble classifier. This voting classifier uses hard and soft voting to predict the final output.

1) *Stacking ensemble for enhanced heart disease prediction:*

Stacking is a machine learning algorithm that trains base classifiers in level 0, and another algorithm acts as a meta learner in level 1. The base classifiers are trained using the training data, and then the model gives its predictions on the test data. The meta-classifier uses these predictions as an input to train itself. The result given by the meta-classifier is the final prediction of the stacking model. In this case, the base classifiers are KNN, SVM and RF. The 1-D CNN acts as the meta-learner and gives the final prediction.

• KNN: This classifies the new data points with the help of existing classified data points. It calculates the distance between the new instance and its k nearest points using the distance formula. Euclidean distances are popularly used for this purpose. The majority voting takes place between the nearest points, and the new instance is assigned to the class having a majority vote in the nearest points.

• SVM: In this case, the classification takes place by finding the best hyper-plane that separates the data points in two different classes in binary classification. The algorithm finds different hyperplanes which linearly separate the data points within which that hyperplane is chosen in which the distance between the marginal planes is high. This distance creates nearly a perfect separation between the classes.

• RF: A random forest creates many

decision trees at training the model and outputs the class that has been voted for by most of the individual trees. Each tree in the forest is built from a randomly drawn sample from the training dataset, and at every node in the tree, a random set of features is considered for splitting. The randomness contributes to building diverse trees, which, when combined together, give more robust predictions than using only one model.

- 1D CNN: A 1D CNN is a type of neural network which uses convolutional layers to process 1-dimensional data. The primary mathematical operation in a 1D-CNN is the convolution operation. Given the concatenated outputs from the base classifiers as an input sequence, a convolutional operation involves applying one or more filters (kernels) across this sequence. After the convolution operation, an activation function is applied for non linearity. It is important for learning complex patterns. After convolutional layers, the output is flattened and sent to one or more fully connected layers.

2) *Voting ensemble for enhanced heart disease prediction*: The outputs of the final predictions of all the predictions are sent to the voting ensemble technique for the final output. The types of voting techniques:

- Hard voting: It involves counting the votes from different classes and choosing the class with the highest number of votes.
- Soft voting: It uses the concept of probability for voting and prediction. It calculates the average probabilities of all the classes across all the classifiers and predicts the output by selecting the class having the highest average probability. The base classifiers used in the case of voting are both individual classifiers and the ensemble model. The ensemble model is the stacking model, and the individual classifiers include:

- MLP: Multilayer Perceptron (MLP) is an artificial neural network-based model that is effective in heart disease prediction. A multilayer perceptron is a neural network model that estimates the normal function of a brain with links that strengthen or weaken it. Multiple layers form the network, and each neuron in one layer is connected to all neurons in the next artificial layer. The input layer is a borrower of factors to predict the output. Each neuron receives signals and processes them before passing them to the next one. Hence, the whole network processes data until it reaches the final layer.

- XGB: It is an advanced gradient-boosted decision tree implementation. Unlike other models, XGBoost improves the model’s accuracy by focusing on the errors of the trees built earlier and continuously improving the predictions. It uses L1 and L2 regularization to prevent overfitting and retain more generalized results. The model is computationally efficient in that it can use as many cores as possible and can process massive datasets, typical for medical ones.

IV. RESULTS & ANALYSIS

Jupyter Notebook has been used in this study on an AMD Ryzen 9 machine equipped with an octa-core 6900 HS processor and 16 GB of RAM. Following cleaning and preprocessing, the dataset’s

70,000 rows and 13 features have been reduced to roughly 60,412 rows and 7 features. Outliers have been eliminated to increase the model's efficiency because every attribute is categorical. The Random Forest, K Nearest Neighbors, and Support Vector Machine are trained as base classifiers for stacking. This prediction system employs multilayer perception, XGBoost, and stacking techniques as base classifiers for voting. This study uses a few performance metrics, including area under the ROC curve and accuracy. The dataset has been divided into two parts: 80% of the dataset has been utilized to train the model, and the remaining 20% has been used for testing. The GridSearchCV method has been a part of an automated hyperparameter tuning process. It returns a set of best hyperparameters required for optimal model training by passing a parameter grid. This approach uses the scikit-learn module and checks the effectiveness of various sets of hyperparameters using k-fold cross-validation. The Table shows the results of various algorithms used in the architecture. With cross-validation, the support vector machine surpassed every base classifier, with an accuracy score of 88.02%. Similarly, the highest 0.96 AUC has been achieved by Random Forest, Multi-Layered Perceptron, and XGBoost. Similarly, the Table shows the results of all the ensemble methods. Stacking has an accuracy of 87.92% and 0.90 AUC. The final predictions of the voting classifier reflect that hard voting has better accuracy than soft voting. On the other hand, soft voting has a higher AUC than hard voting.

TABLE II
BASE CLASSIFIER COMPARISON

Base Classifiers	**Accuracy without CV**	**Accuracy with CV**	**AUC**
Random Forest	87.64	87.77	0.96
Support Vector Machine	85.18	88.02	0.95
K-Nearest Neighbors	85.78	86.64	0.95
Multilayered Perceptron	87.50	87.85	0.96
XGBoost	87.75	87.83	0.96

TABLE III
ENSEMBLE LEARNER COMPARISON

Ensemble Learners	**Accuracy**	**AUC**
Stacking	87.92	0.90
Hard Voting	87.98	0.89
Soft Voting	87.90	0.96

The proposed model surpassed existing systems that built prediction models using only traditional or ensembled approaches. The table shows the comparison of this.

TABLE IV
RESULT COMPARISON

Author	**Models**	**Dataset Used**	**Accuracy**
Jan Carlo T. Arroyo et al. (2022) [22]	GA-ANN	Cardiovascular Disease dataset (70000 records)	73.43%
Muhammad Salman Pathan et al. (2022) [20]	Support Vector classifier	CVD and Framingham dataset	74%
Vardhan Shorewala (2021) [30]	Stacking of KNN, Random Forest, and SVM with logistic regression	Cardiovascular Disease dataset (70000 records)	75.1%
Pooja Rani et al. (2021) [21]	Random Forest	Cleveland dataset (UCI)	86.60%
Proposed	**Stacking (SVM,RF,KNN) and voting with MLP and XGBoost**	**Cardiovascular Disease dataset (70000 records)**	**87.98%**

The above comparison shows the scope for hybrid ensemble models, which have deep learning and machine learning, to have a greater impact on improving the model's overall performance.

V. CONCLUSIONS

Classification of heart disease is the primary objective of this study. The dataset has been pre-processed by converting the features into categorical with the help of binning. Age has been binned with a span of 5 years range.

Similarly, after converting height and weight to BMI and systolic and diastolic blood pressures to MAP, binning was done to convertthem into six bins. Using the k-mode clustering technique has significantly increased the accuracy of this prediction model.It groups the heart patients based on similarities by cluster formation, which helps extract patterns from the dataset. The elbow curve method provides a visualization of the clusters concerning the cost. This visually helps us identify the elbow shape, representing the optimal number of clusters that providebetter accuracy.

Among the base classifiers, SVM has achieved 88.02%accuracy, which outperforms every other base classifier. Asa meta-classifier, CNN takes the predictions from the base classifiers and then combines the projections. It provides maximum weightage to the algorithm that works better to improveoverall accuracy. The final voting predicts the presence of heart disease. The majority of votes are considered in the case of hard voting. In contrast, soft voting adapts the average probability approach to predict the result, which chooses the maximum of all the average probabilities. Hard voting has surpassed soft voting with a remarkable accuracy of 87.98%, whereas soft voting has a higher AUC of 0.96.

The outcome of this research reflects that the K-mode clustering algorithm is a tool that can effectively boost the model's accuracy. The model building reflects the hybrid combination of models that can capture essential patterns fromthe dataset. The performance of base classifiers improves the performance of the hybrid model by giving a solid foundation for the proposed model. This strengthens the hybrid model's base and reduces the number of False Positive and False Negative cases, which is quite impressive.

Even if the model is performing better, it has some limitations. The model has been trained using a single datasetof a specific region, which may not apply to people livingin different areas and lifestyles. Further, the dataset had only binary target attributes, which can only predict the presenceor absence of the disease but do not talk about the severity of the disease. Also, the model architecture is well designed to handle the tabular dataset and won't be effective for the imagedata.

In future, developments can be made by making the model capable of working on datasets of different geographic regions, including the disease's severity. Also, significant developments should be made to be compatible with the image dataset. Ultimately, the primary future developments mainly focus on making the model capable of working with various kinds of datasets, including the disease's severity. This would enhance the model's capacity to work under different conditions.

REFERENCES

[1] American Heart Association, Inc. (2022). 2022 Heart disease & Stroke statistical update fact sheet. https://www.heart.org/-/media/PHD- Files-2/Science-News/2/2022-Heart-and-Stroke-Stat-Update/2022-Stat-

[2] Update-factsheet-Global-Burden-of-Disease.pdf

[3] Online, E. (2023, September 29). World Heart Day: 27% of deaths in India are caused by cardiovascular diseases. The Economic Times. https://economictimes.indiatimes.com/magazines/panache/world-heart-day-27-per-cent-of-deaths-in-india-are-caused-by-cardiovascular-diseases/articleshow/104031929.cms?from=mdr

[4] Ghosh, J. (2023, January 1). A review on understanding the risk factors for coronary heart disease in Indian college students. International Journal of Noncommunicable Diseases. https://doi.org/10.4103/jncd.jncd 68 23.

[5] Non-modifiable Risk Factors - University of Ottawa Heart Institute — Prevention & Wellness Centre. (n.d.).

[6] UCSF Health. (2023, May 8). Understanding your risk for heart disease. ucsfhealth.org. https://www.ucsfhealth.org/education/understanding-your-risk-for-heart-disease

[7] Non-Modifiable Risk Factors — University of Ottawa Heart Institute — Prevention & Wellness Centre. (n.d.). https://pwc.ottawaheart.ca/awareness/heart-health-portal/risk-factors/non-modifiable-risk-factors

[8] Cimmino, G., Natale, F., Alfieri, R., Cante, L. C., Covino, S., Franzese, R., Limatola, M., Marotta, L., Molinari, R., Mollo, N., Loffredo, F., & Golino, P. (2023). Non-conventional risk factors: "Fact" or "Fake" in cardiovascular disease prevention? Biomedicines, 11(9), 2353. https://doi.org/10.3390/biomedicines11092353

[9] Nindrea, R. D., & Hasanuddin, A. (2023). Non-modifiable and modifiable factors contributing to recurrent stroke: A systematic review and meta-analysis. Clinical Epidemiology and Global Health, 20, 101240.

[10] Hart, D. A. (2022). Sex differences in biological systems and the conundrum of menopause: Potential commonalities in post-menopausal disease mechanisms. International Journal of Molecular Sciences, 23(8), 4119.

[11] Khandaker, G. M., Zuber, V., Rees, J. M., Carvalho, L., Mason, A. M., Foley, C. N., ... & Burgess, S. (2020). Shared mechanisms between coronary heart disease and depression: findings from a large UK general population-based cohort. Molecular psychiatry, 25(7), 1477-1486.

[12] Patel, A. P., Wang, M., Kartoun, U., Ng, K., & Khera, A. V. (2021). Quantifying and understanding the higher risk of atherosclerotic cardiovascular disease among South Asian individuals: results from the UK Biobank prospective cohort study. Circulation, 144(6), 410-422.

[13] Global Cardiovascular Risk Consortium. (2023). Global effect of modifiable risk factors on cardiovascular disease and mortality. New England Journal of Medicine, 389(14), 1273-1285.

[14] Zurbau, A., Au-Yeung, F., Blanco Mejia, S., Khan, T. A., Vuksan, V., Jovanovski, E., ... & Sievenpiper, J. L. (2020). Relation of different fruit and vegetable sources with incident cardiovascular outcomes: a systematic review and meta-analysis of prospective cohort studies. Journal of the American Heart Association, 9(19), e017728.

[15] Natsis, M., Antza, C., Doundoulakis, I., Stabouli, S., & Kotsis, V. (2020). Hypertension in obesity: novel insights. Current hypertension reviews, 16(1), 30-36.

[16] Ng, R., Sutradhar, R., Yao, Z., Wodchis, W. P., & Rosella, L. C. (2020). Smoking, drinking, diet and physical activity—modifiable lifestyle risk factors and their associations with age to first chronic disease. International journal of epidemiology, 49(1), 113-130.

[17] Manolis, T. A., Manolis, A. A., Apostolopoulos, E. J., Melita, H., & Manolis, A. S. (2021). Cardiovascular complications of sleep disorders: a better night's sleep for a healthier heart/from bench to bedside. Current vascular pharmacology, 19(2), 210-232.

[18] Muhammad, Y., Tahir, M., Hayat, M., & Chong, K. T. (2020). Early and accurate detection and diagnosis of heart disease using intelligent computational model. Scientific reports, 10(1), 19747.

[19] Biswas, N., Ali, M. M., Rahaman, M. A., Islam, M., Mia, M. R., Azam, S., ... & Moni, M. A. (2023). Machine Learning-Based Model to Predict Heart Disease in Early Stage Employing Different Feature Selection Techniques. BioMed Research International, 2023.

[20] Pathan, M. S., Nag, A., Pathan, M. M., & Dev, S. (2022). Analyzing the impact of feature selection on the accuracy of heart disease prediction. Healthcare Analytics, 2, 100060.

[21] Rani, P., Kumar, R., Ahmed, N. M. S., & Jain, A. (2021). A decision support system for heart disease prediction based upon machine learning. Journal of Reliable Intelligent Environments, 7(3), 263-275.

[22] Arroyo, J. C. T., & Delima, A. J. P. (2022). An optimized neural network using genetic algorithm for cardiovascular disease prediction. Journal of Advances in Information Technology, 13(1).

[23] Ramdass, P., & Ganesan, G. (2023). Leveraging Neighbourhood Compo- nent Analysis for Optimizing Multilayer Feed-Forward Neural Networks in Heart Disease Prediction. Mathematical Modelling of Engineering Problems, 10(4).

[24] Djerioui, M., Brik, Y., Ladjal, M., & Attallah, B. (2020, September). Heart Disease prediction using MLP and LSTM models. In 2020 International Conference on Electrical Engineering (ICEE) (pp. 1-5). IEEE.

[25] Sarra, R. R., Dinar, A. M., Mohammed, M. A., Ghani, M. K. A., & Albahar, M. A. (2022). A robust framework for data generative and heart disease prediction based on efficient deep learning models. Diagnostics, 12(12), 2899.

[26] Sajja, T. K., & Kalluri, H. K. (2020). A Deep Learning Method for Prediction of Cardiovascular Disease Using Convolutional Neural Network. Rev. d'Intelligence Artif., 34(5), 601-606.

[27] Zhang, D., Chen, Y., Chen, Y., Ye, S., Cai, W., Jiang, J., ... & Chen, M. (2021). Heart disease prediction based on the embedded feature selection method and deep neural network. Journal of healthcare engineering, 2021, 1-9.

[28] Asif, D., Bibi, M., Arif, M. S., & Mukheimer, A. (2023). Enhancing heart disease prediction through ensemble learning techniques with hyperparameter optimization. Algorithms, 16(6), 308.

[29] Yewale, D., Vijayaragavan, S. P., & Bairagi, V. K. (2023). An Effective Heart Disease Prediction Framework based on Ensemble Techniques in Machine Learning. International Journal of Advanced Computer Science and Applications, 14(2).

[30] Shorewala, V. (2021). Early detection of coronary heart disease using ensemble techniques. Informatics in Medicine Unlocked, 26, 100655.

[31] Ghasemieh, A., Lloyed, A., Bahrami, P., Vajar, P., & Kashef, R. (2023). A novel machine learning model with Stacking Ensemble Learner for predicting emergency readmission of heart-disease patients. Decision Analytics Journal, 7, 100242.

[32] Kavitha, M., Gnaneswar, G., Dinesh, R., Sai, Y. R., & Suraj, R. S. (2021, January). Heart disease prediction using hybrid machine learning model. In 2021 6th international conference on inventive computation technologies (ICICT) (pp. 1329-1333). IEEE.

[33] Chandrasekhar, N., & Peddakrishna, S. (2023). Enhancing heart disease prediction accuracy through machine learning techniques and optimization. Processes, 11(4), 1210.

[34] Baccouche, A., Garcia-Zapirain, B., Castillo Olea, C., & Elmaghraby, A. (2020). Ensemble deep learning models for heart disease classification: A case study from Mexico. Information, 11(4), 207.

[35] Almulihi, A., Saleh, H., Hussien, A. M., Mostafa, S., El-Sappagh, S., Alnowaiser, K., ... & Refaat Hassan, M. (2022). Ensemble learning based on hybrid deep learning model for heart disease early prediction. Diagnostics, 12(12), 3215.

[36] Cardiovascular Disease dataset. (2019, January 20). Kaggle. https://www.kaggle.com/datasets/sulianova/cardiovascular-disease-dataset/data