Machine Learning Methods for Heart Disease Prediction

Hongyu Zhou
Department of Mathematics, Southern University of Science and Technology, Shenzhen, 518000, China

Keywords: Machine Learning, Heart Disease Prediction, Neural Network, Evaluation.

Abstract: Heart disease prediction and treatment play a crucial role in enhancing human health. Numerous studies have highlighted the effectiveness of machine learning models in predicting heart diseases. However, there still have problems with widely use and the accuracy of the prediction. This paper aims to apply different machine learning models, including Naïve Bayes, Decision Tree, Random Forest, XGBoost, and Neural Network System, to a specific dataset and provide a comprehensive evaluation. After thorough analysis using various metrics, the Random Forest model demonstrated the highest recall and F1-score among all models. Additionally, the shallow neural networks model outperformed traditional neural network structures with fewer parameters in this task. In conclusion, this study emphasizes the significance of machine learning models in improving heart disease prediction and treatment. Further research and development in this area are essential to enhance healthcare outcomes and promote overall well-being.

1 INTRODUCTION

Ischemic heart disease is the world's largest cause of death, responsible for 16% of all deaths, according to the World Health Organization. Meanwhile, since 2000, the situation becomes more and more strictly (Soni et al., 2011; Chitra et al., 2022 & Chen and Guestrin, 2016). In 2019, approximately 8.9 million people was killed by this disease.

Over the past decade, the understanding and the treatment technique on the ischemic heart disease has gained a significant improvenment (Kaggle Heart disease dataset). Due to this progress, it is important to receice a professional medical assistance before the disease becomes aggravated. However, there has another problem. As for the patients, some kind of heart disease is difficult to discover without a specialized hospital diagnostic examination. Then machine learning technique are expected to predict the diease with just some basic presonal information (Tsao et al., 2022). Fortunately, several research has listed some features, which are easy to be detected, are related to the heart disease. Then people with no significant cardiovascular disease symptoms can be detected early and can receive early medical treatment which may reduce their death rate significantly (Chintan et al., 2023). Thus, it is meaningful to apply certain machine learning methods to predict whether a person will have heart disease with some physical indicators.

With the development of the machine learning techiniques, there are applications have approached the success in this field. However, there still have problems with widely use and the accuracy of the prediction. This paper focus on analyze 5 different methods for heart diease prediction in the field of machine learning and heart disease prediction technology. Aim to find the strengths and the weakness of different methods.

2 METHOD

2.1 Data Preprocessing

Before establishing and training models, data preprocessing and analysis are crucial. The project uses a remarkable dataset from Kaggle tha dataset which is neat and well-documented.

The dataset comprises 4238 records sourced from the Centers for Disease Control and Prevention, National Center for Health Statistics. For all subsequent model training, this dataset was randomly splited into two parts: one part for training and another for testing (no validation set was used in this experiment). The training part and test part each represent 80% and 20% of the total dataset, respectively. Within this dataset, there are a total of 15 features related to predicting heart disease,

Zhou, H.
Machine Learning Methods for Heart Disease Prediction.
DOI: 10.5220/0012990900004601

In *Proceedings of the 1st International Conference on Innovations in Applied Mathematics, Physics and Astronomy (IAMPA 2024)*, pages 46-52
ISBN: 978-989-758-722-1

including but not limited to physical examination data and various personal demographics. These 15 features are categorized into two groups, starting with a collection of information pertaining to individuals' physical characteristics. Here are the specific meanings of these numerical values:

These features offer crucial insights into cardiovascular health, pivotal for predictive modeling in assessing heart disease risk. Additionally, a set of features encompassing Gender, age, education, currentSmoker, and cigsPerDay reflect individuals' fundamental information and lifestyle habits. Heart disease etiology is shaped not only by congenital factors but also by lifestyle choices and surroundings. Consequently, incomplete data entries were excluded, yielding a final dataset of 3656 entries for model training and prediction.

Based on the figure 1, it can be observed that the original dataset is some imbalanced, with some feature generally not follows a normal distribution. This lends the data is difficult to train and analysis. To makes it suitable for training predictive models, the overSampler method was used in our experienment. Additionally, upon observing the 'age' feature, we roughly categorize the dataset into four major age groups simplify this complex dataset based on age. The similar operations are also be done on ‘BMI’, ‘glucose’ and ‘sysBP’ features. To provide a more intuitive representation of the relationship between the heart disease and the features.

2.2 Data Visualization and Analysis

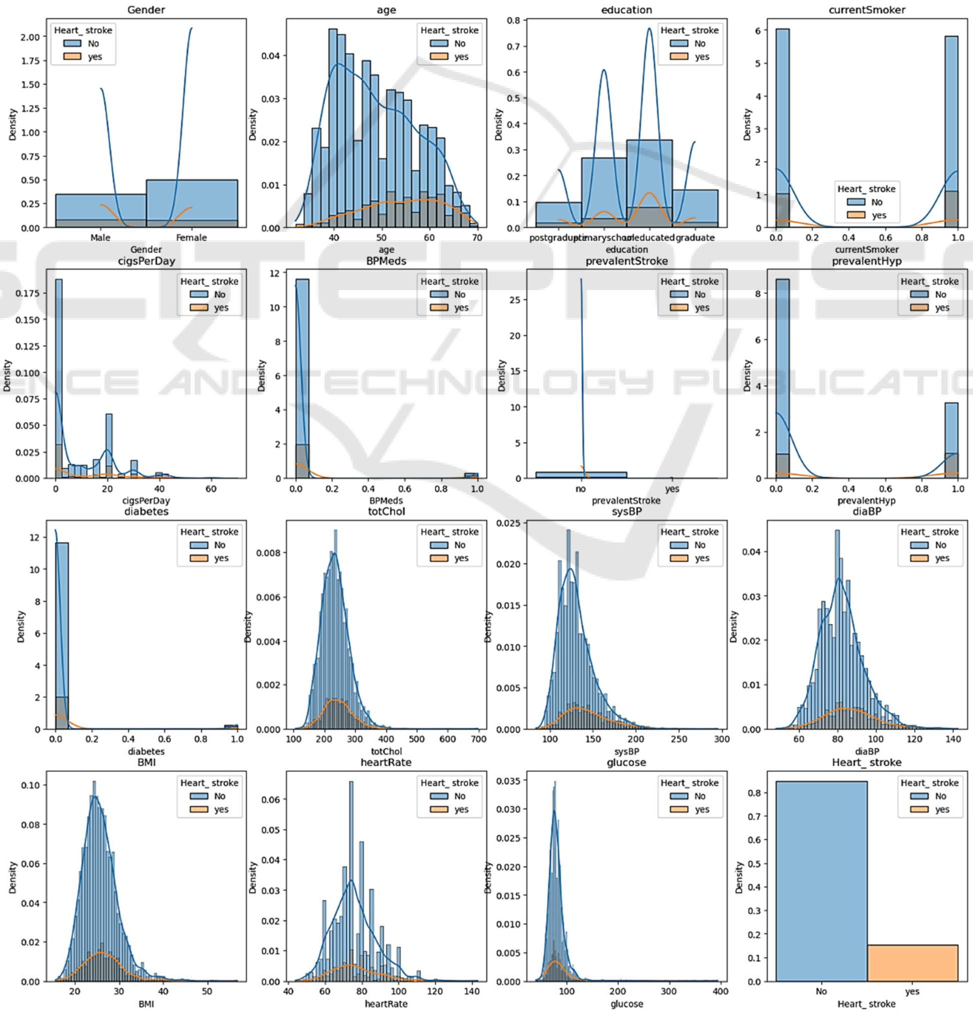

Figure 1: The Preview of the all features (Picture credit: Original).

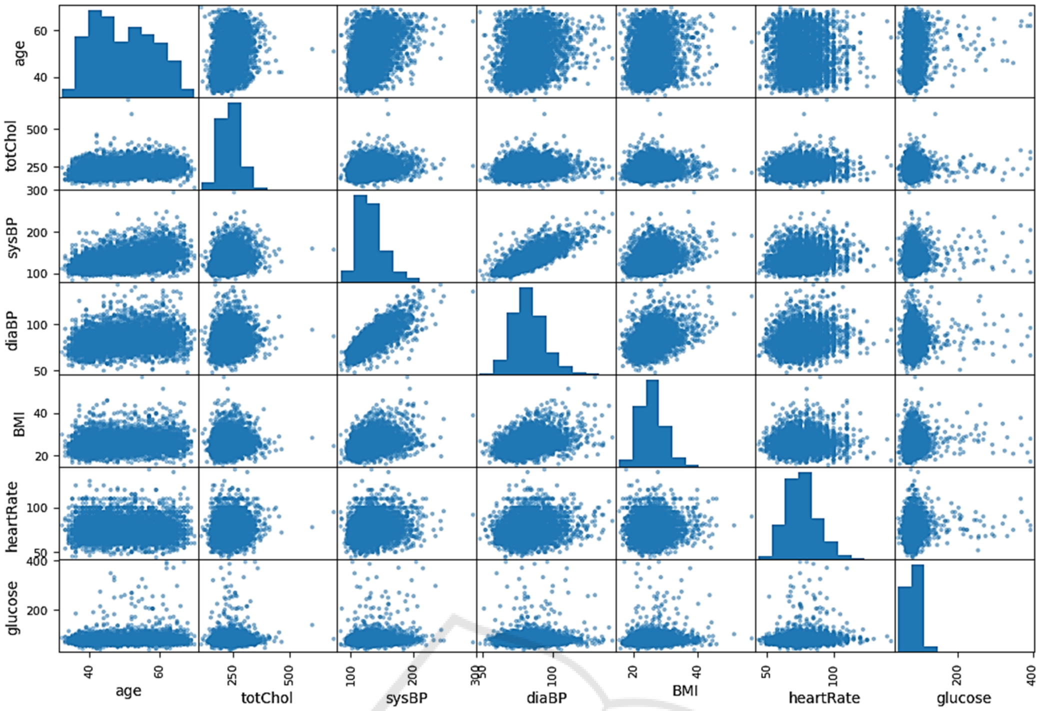

Figure 2: The correlation of features (Picture credit: Original).

The following correlation matrix images provide valuable insights (Figure 2). A noteworthy observation is the strong positive correlation between systolic and diastolic blood pressure class. Additionally, histograms for several features exhibit a Gaussian distribution shaped curve, including diaBP, BMI, heart rate (HR), total cholesterol (totChol), and systolic blood pressure (sysBP).

2.3 Method Choose Comparison

In this study we choose to implement a couple of popular supervised machine learning methods to make predictions on the heart diease task. Accoding to the previous research. Decision Tree and Decision Tree like models, such as Random Forest and XGBoost are widely applied in heart diease prediction (Sonawane and Patil, 2014). Additionally, Naïve Bayes as a classic probablity model, also be implemented in our experiment. In this paper, we do a comparision on neural network systems too.

After constructing the models, to get a proper and better understanding of different methods, the following gives a brief description and the properties of them, especially on prediction tasks.

- Naive Bayes

Naïve Bayes method is a classic and also can be perceived as a one of the most popular supervised learning algorithm. The idea of this model is from the Bayes's theorem. In application, we can simplified the realation to

$$P(y \mid x_1, \dots, x_n) = \frac{P(y)\prod_{i=1}^{n} P(x_i \mid y)}{P(x)}$$

Bayes's theorem is one of the most important theorem in probability theory. Since the equation reveals the relationship between prior probability and posteriol probability. Whence the model has the capability to predict the probability of the heart diease when other features was detected.

The foundation of Naive Bayes analysis is the idea that characteristics, given the class label, are completely independent. When the assumption is broken, less desirable results could result, especially if the features are closely related in complex ways. Therefore, choosing the right characteristics for the Naïve Bayes model's training is important.

- Decision Trees and Random forest

Decision trees are a type of supervised learning method that work well for problems involving regression and classification (Chawla et al., 2002). This model is tree-structured. The tree root and nodes reprsent the selected features and tree branches connect root and nodes or nodes and nodes. Each branches will end by a leaves which means the outcome (Figure 3).

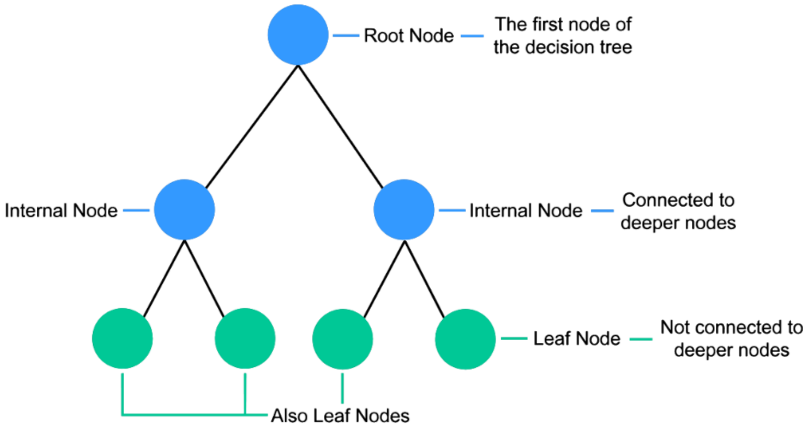
Root Node
The first node of the decision tree
Internal Node
Internal Node
Connected to deeper nodes
Leaf Node
Not connected to deeper nodes
Also Leaf Nodes

Figure 3: Decision Tree (Chawla et al., 2002).

For Decision Tree method applied on heart diease prediction problem is how to select a proper attribute for discrimination among all features in the dataset. There are four popular methods for selecting attribute: information gain, entropy, gain ratio and gini index.

Bagging and boosting is two of the most popular method for ensemble model. After applying bagging on the decision tree model can derive the random forest model,

RF model can be abbreviated as the following steps:

First Step: Select n random subset in the features set.

Second Step: Independently train n decision trees

A single decision tree is trained on a single random subset. Each decision tree's ideal splits are determined by randomly selecting a subset of samples or features.

Third Step: each tree makes the prediction.

Forth Step: Ensemble all the trees and caculate predicted result of each class. Then decide which class is the input belongs to.

In our experiment, the result depends on the majority chosen class of independent trees and the weight is the same for them.

- XGBoost(XGB)

XGBoost model is based on gradient tree boosting algorithm. The objective function is the core problem in this method which can be represented as

$$L(\varphi) = \sum_{i} l(\hat{y}_i, y_i) + \sum_{k} \Omega(f_k) \quad (1)$$

The first term is called loss term. In XGB an additive training idea was used.

$$\begin{gathered}\hat{y}_i^{(0)} = 0 \\ \hat{y}_i^{(1)} = f_1(x_i) = \hat{y}_i^{(0)} + f_1(x_i) \\ \hat{y}_i^{(2)} = f_1(x_i) + f_2(x_i) = \hat{y}_i^{(1)} + f_2(x_i) \\ \cdots \\ \hat{y}_i^{(t)} = \sum_{k=1}^{t} f_k(x_i) = \hat{y}_i^{(t-1)} + f_t(x_i)\end{gathered} \quad (2)$$

This idea enlighten the XGB algorithm to use the gradient to optimize the trees for prediction. The second term is a regularization term which aims to decrease the complexity (eg. Parameters' number, norm of parameters) of this trees model. Specifically, this term in XGB can be written as

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} \omega_j^2 \quad (3)$$

Where T reprensents the number of the selected features(also known as 'leaves'), w is the vectors scores on leaves. XGBoost is always was used for supervised learning problems. That means the dataset includes not only the training data but also the correct label (Bharath et al., 2023).

- Artficial Neural Network

The primary motivation behind the creation of artificial neural networks was to imitate biological neural systems; but, throughout time, these networks have evolved into specialized fields focused on machine learning tasks and engineering. Multi-layered artificial neural networks with fully connected layers are the most common neural networkd in prediction task. A two-hidden-layer neural network system's structure is shown in the figure 4.

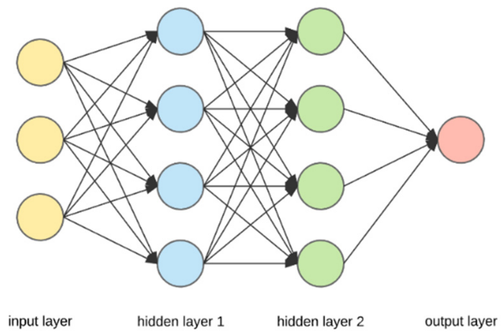

Figure 4: A 2-hidden-layer neural network (Picture credit: Original).

- Shallow Neural Networks

Shallow neural network is a particular kind of neural network architectures specially consisting of one hidden layer. The system with more hidden layers are known as the deep neural network system. So intuitively, shallow neural networks is more simple. However, it is incredibly expressive and also be able to handle the big data modeling and machine learning task with less model compexity.

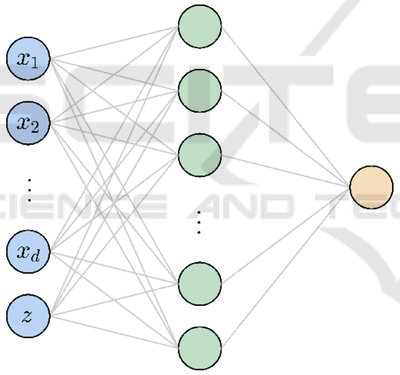

Figure 5: Shallow neural networks architecture (Picture credit: Original).

The structure of shallow neural networks is relatively simple, making them easier to understand and implement (Figure 5). During training, the backpropagation algorithm was used to tune the weights and biases. The process is likely to enable the modle learn patterns and features in the input data, producing corresponding outputs. While they may not capture complex data patterns, shallow neural networks generally perform well on simple to moderately complex problems and train relatively quickly (Karthiga et., 2017).

How is the model work? Firstly, the model receives the feature $\{x_i\}$, which are going to be classified. Then they are passed to a layer of nodes W_i each of which will be activated by some function $\sigma(\cdot)$ acting on the weighted sum of those values. The result of each unit in the hidden layer is then passed to a final, output layer α_i (which may consist of a single unit). At last, the sum of these results will be passed to another activation function as the final output (Fahd Saleh, 2019).

With expressive capabilities afforded by neural networks, we hereby construct a basic feedforward, fully-connected model, comprising a shallow architecture (consisting of just one hidden layer), to approximate the prediction.

$$p = \sigma(\sum_{j=1}^{N} \alpha_j \sigma(W_j \mathrm{x}^T + b_j)) \qquad (4)$$

Here, σ is the sigmoid activation function, N is the number of parameters in output layer. The weights and are formed as learnable parameters. Notice that, the output is after the sigmoid function. So the output is in the interval [0,1], which can be perceived as the possibility of each class.

As for the model, with a given training set , which the input dimension is 32 (32 features), the neural net parameters (weights W 130 *130 and with biases)) are learned via minimizing the cross entropy error. Specifcly, the binary cross entropy error can be represented as the form:

$$H(\mathrm{y}, \hat{y}) = -(y\log(\hat{y}) + (1-y)\log(1-\hat{y})) \qquad (5)$$

Cross entropy is defined in information theory, which is widely used to define the metric of the difference between two probability distributions in information theory.In machine learning and deep learning, cross entropy is commonly used as the model loss, especially in classification tasks. In this experiment, $H(y,\hat{y})$ close to 0 means the model is perfect , otherwise $H(y,\hat{y})$ *closes to* 1 means model is underfitting.

3 EXPERIMENT AND RESULT

3.1 Evaluation Metrics

- Accuracy

By calculating the proportion of each example's sickness that was correctly detected

$$Accuracy = \frac{properly\ identified\ cases}{num\ of\ all\ cases} \times 100\% \ (6)$$

- Precision

Out of all positively predicted instances, precision is the ratio of positively occurrences that are correctly predicted.

$$Precision = \frac{positively\ properly\ identified\ cases}{num\ of\ all\ predicted\ positively\ cases} \times 100\% \qquad (7)$$

- Recall

The ratio of properly predicted positive cases among all actual positive cases.

$$Recall\ = \frac{positively\ properly\ identified\ cases}{num\ of\ all\ actual\ positive\ \ cases} \times 100\% \qquad (8)$$

- F1-score

The F1 score evaluates the balance of the model. The formulation is as follows:

$$F1 - Score\ = (1 + \beta^2) \times \frac{precision \times recall}{(\beta^2 \times precision) + recall} \qquad (9)$$

- ROC Curves and AUC

The performance of a classifier across a variety of discriminants is visually represented by the ROC curve. For various threshold values, it plots the true positive rate against the false positive rate. Then, the classify effficency of the model is measured by the proportion under the ROC curve; a greater proportion value, closer to 1, denotes better classification performance (Institute of Medicine Committee on Social Security Cardiovascular Disability Criteria 2010 Cardiovascular Disability).

3.2 Experiment Settings

In this research, the experiment is all implemented by python 3.11.5. Additionally, Some libraries such as pandas, scikit-Learn, pytorch are used. The hardware configuration comprise a 3.20GHzAMD Ryzen 7 5800H CPU, a RTX 3060 GPU and 16.0GB RAM.

- Naïve Bayes

Ordinary Naïve Bayes model is used with no special setting.

- Decision Trees

Critirion was chosen as entropy in decision trees model.

- Random Forest

The Random Forest model includes a total of 1000 trees, and the split criterion is entropy.

- XGB

The basic tree model is hist and a total of 8783 gradient boost trees are utilized.

- Artificial Neural Network

In this experiment, we use a deep NN model with 2 hidden layers, and each hidden layer had 40 dimensions. The input is 15 dimensions, which includes all the features just with normalization steps. And the output is 2 dimensions reprensents the possibility of each class. the Adam optimizer was used with default settings and the maximum number of optimization iterations is 10000. Our loss function was categorical cross-entropy.

- Shallow Neural Network

We adjusted parameters multiple times and settled on setting the dimension of the hidden layer weights to 130. Other setting is as same as three layers neural network.

3.3 Results and Evaluation

Because of the heart disease property, this task is more likely to imporve the possibilities of discovering the patients who already has the heart diease. Therefore, recall is the most important metric in these models evaluation. According to the result in the table above, Decision Tree model and Random forest provides the highest recall at the same time compared to the other models. However, Random Forest model get a higer accuracy and also has the highest F1-score. So Random Forest models can be perceived as the most balanced model among other algorithms (Table 1).

Table 1: Results from different methods.

Model	accuracy	recall	precision	F1 score	AUC
Naïve Bayes	0.6203	0.3644	0.6944	0.4780	0.69
Decision Trees	0.9117	0.9970	0.8455	0.9150	0.92
Random Forest	0.9812	0.9970	0.9647	0.9806	1.00
XGB	0.9970	0.9635	0.8755	0.9174	0.98
Two layers NN	0.8407	0.8834	0.8026	0.8411	0.85
Shallow NN	0.9033	0.9869	0.8389	0.9069	0.85

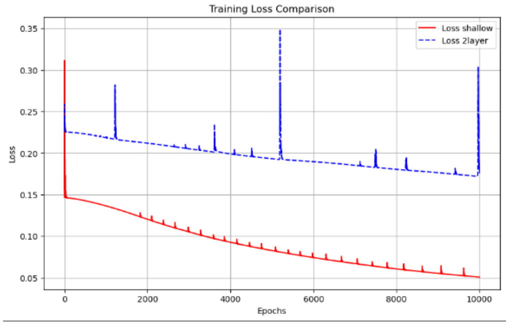

Figure 6: Loss of NN model (Picture credit: Original).

Figure 6 is the loss of two neural networks models. Loss of shallow neural networks can be easier optimized and also get a better results compared to the two layer models.

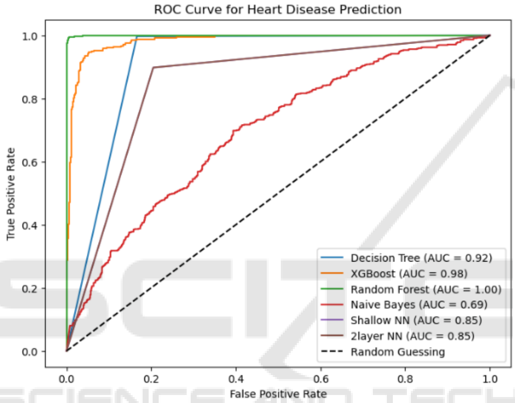

Figure 7: ROC Curve (Picture credit: Original).

According to the experiment results (Figure 7), Random Forest model is the most perfect model on this task. As for the nerual network system, we can conclude that shallow structure gets a higher accuracy and recall scores comparing to a deeper system. At the same, in this sturctures, there are less parameters are used and is convenient to optimize.

4 CONDUCTION

The aim of this research was to evaluate several machine learning prediction techniques and provide an overview of their variations and effectiveness on the heart disease system. These machine learning methods are Naïve Bayes , Decision Tree, Random Forests, XGBoost, Aritificial Neural Network and Shallow Neural Network. During the experimental comparisons, we aimed to evaluate the strengths of each method by conducting specific data preprocessing and analysis tailored to their characteristics. This allowed for a comprehensive comparison of algorithms, from model selection and data analysis to model training and prediction.

In this study, we also proposed using a Shallow Neural Network as an alternative to the traditional ANN model. Through our comparisons, we found that the Shallow Neural Network achieved comparable or even better results while decreasing the number of parameters required. This highlights the effectiveness of the Shallow Neural Network in achieving similar or improved performance with fewer parameters compared to conventional ANN models.

REFERENCES

J. Soni, U. Ansari, D. Sharma and S. Soni 2011 *Inter.J. Comput. Appli.* **17** 8

R. Chitra, K. A. Warsi, S. Muzamil, P. N. Shankar, D. Ghai and B. C. Dharmani 2022 Performance Evaluation of Machine Learning Algorithms in Design and Development of Heart Disease Detection *2022 Inter. Conf. Comput., Communic. Secur. Intelli. Sys.* 1-8

Chen, Tianqi and Carlos Guestrin 2016 XGBoost: A Scalable Tree Boosting System. *22nd Inter. Conf. Knowl. Discov. Data Mini.* 785–794

Kaggle Heart Disease Dataset. Available online: https://www.kaggle.com/datasets/ mirzahasnine /heart-disease-dataset/data

Tsao CW, Aday AW, Almarzooq ZI, et al. 2022 Heart Disease and Stroke Statistics-2022 Update: A Report From the American Heart Association. *Circulation* **145(8)** e153e639

Chintan M. Bhatt and Parth Patel and Ta Ghetia and Pier Luigi Mazzeo 2023 *Algorithms* **16** 88.

Jayshri S. Sonawane and Dharmaraj R. Patil 2014 *Inter.Conf. Inform. Communic. and Embed. Sys.* 1-6

N. Chawla and K. Bowyer and Lawrence O. Hall and W. Philip Kegelmeyer 2002 SMOTE: synthetic minority over-sampling technique *ArXiv* **abs/1106. 1813**

Kakarla Sai Bharath and Anudeep Sanakkayala and Abhishek Kadiyam and Gudapati Pradeep Chandra and Iwin Thanakumar Joseph S and K. B. V. Brahma Rao 2023 *2023 3rd Inter. Conf. Smart Data Intelli.* 383-386.

Karthiga, A. and Mary, Safish and Yogasini, M. 2017 Early Prediction of Heart Disease Using Decision Tree Algorithm *Inter. J. Adv. Res. Basic Engin. Sci. Technol.* **3**

Fahd Saleh S. Alotaibi 2019 Implementation of Machine Learning Model to Predict Heart Failure Disease *Inter. J. Adv. Comput. Sci. Appl.*

Institute of Medicine (US) Committee on Social Security Cardiovascular Disability Criteria 2010 Cardiovascular Disability: Updating the Social Security Listings *Washington (DC): National Acad. Press*