Educational Administration: Theory and Practice
2024, 30(6), 2640-2646
ISSN: 2148-2403
https://kuey.net/

Research Article

Cardiac Disease Analysis Using Machine Learning Technique

Dr.T. Sundaravadivel[1*], Dr.M. Arasakumar[2]

[1*]Assistant Professor/ Programmer, Department of Information Technology, Annamalai University, Tamilnadu, India.
[2]Assistant Professor, Department of Information Technology, Annamalai University, Tamilnadu, India.

Citation: Dr.T. Sundaravadivel et al. (2024) Cardiac Disease Analysis Using Machine Learning Technique, *Educational Administration: Theory and Practice*, *30*(6), 2640-2646
Doi: 10.53555/kuey.v30i6.5860

ARTICLE INFO

ABSTRACT

One of the most well-known uses of artificial intelligence, and machine learning (ML), is revolutionizing the field of healthcare. In this work, the use of machine learning to determine a person's risk of heart attack is studied. Cardiovascular diseases (CVDs) are common and can possibly be fatal for people anywhere in the globe. A person's age, cholesterol level, chest discomfort, and other characteristics may all be taken into account using machine learning to determine if they have a cardiovascular disease. Cardiovascular disease diagnosis can be facilitated by machine learning classification algorithms based on supervised learning. This study explores the effectiveness of classification approach developed using Artificial Neural Network (ANN), and compares it performance with Random Forest Classifier in predicting the chance of heart attack for a person. Different biological and biochemistry profile were used as clue for classifying whether a patient will suffer from cardiac arrest or not.

Keywords: *Cardiovascular diseases; Artificial Neural Network; Random Forest Classifier; biochemistry profile*

INTRODUCTION

Because of a protracted period of epidemiological shift, the burden of cardiovascular diseases (CVDs) increased over several decades in industrialized nations. This is a reference to the transition from rural to urban communities. However, CVDs and associated risk factors have increased dramatically in India in a very short period of time as a result of the country's economy developing at a very quick pace after independence. As a result, they are now the primary cause of death in the non-communicable illness group. It is well known that metropolitan regions have a greater frequency of CVDs than rural ones.

Growing numbers of people, aging populations, genetic predispositions, and risk factors connected to behavior are all contributing causes to the rising incidence of heart-related illnesses [1]. These consist of things like a sedentary lifestyle, poor dietary habits, long-term stress, tobacco usage, and binge drinking too much alcohol. The long-term consequences of these risk factors might mainly manifest as hypertension, elevated blood sugar and cholesterol, as well as obesity and overweight. Strokes and cardiac arrests are two prominent consequences that might arise from these risk factors.

One of the nation's main concerns has been the economic burden of mortality from CVDs (increase in the death rate due to CVD presented in Fig. 1). Estimates suggest that eliminating non-communicable illnesses might lead to a 10% increase in GDP.

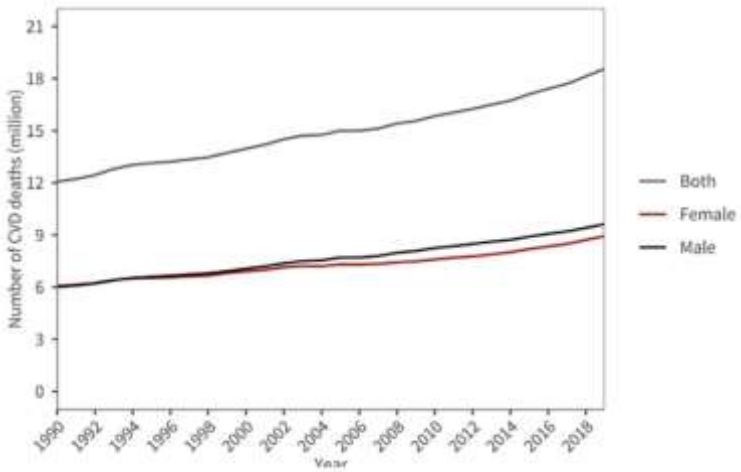

Fig. 1 Mortality due to Cardio Vascular Disease

Most CVDs are recognized to be preventable by addressing risk factors. Preventive steps to lessen them include changing to a better lifestyle, taking the right medicine, and getting surgery or medical attention. The cardiac market may be essentially classified into three categories: thrombosis, hyperlipidemia, and hypertension therapy. The most often recommended medications for cardiac patients receiving palliative, preventative, or post-operative therapy are anti-thrombotic, anti-dyslipidemia, and anti-hypertensive drugs. Anti-dyslipidemia drugs assist in regulating lipid levels, whereas anti-hypertensives refer to drugs that reduce blood pressure. Conversely, anti-thrombotic medications function to stop blood clots from forming. Heart disease prediction can be aided by data mining techniques. By utilizing the information gleaned from databases to identify hitherto unidentified patterns and trends, predictive models may be created [2]. To derive intelligence from vast volumes of data is to engage in data mining [3]. One of the technological advancement that can assist in diagnosing cardiac disease early on before significant harm is done to an individual is machine learning. Machine learning is a rapidly developing subject in science and technology that has the ability to diagnose and categorize cardiac disease in individuals. Conventional machine learning techniques have poor model prediction accuracy and are unable to account for variations in the data. A variety of machine learning methods that may be applied to solve this issue are shown in this study. These models consider various algorithms' training processes as well as ways for observing data. The world's biggest cause of mortality is still cardiovascular illness, which includes conditions including coronary artery disease (CAD), atrial fibrillation (AF), and other cardiac or vascular disorders (10). People with cardiovascular disease (CVD) are becoming more and more common as living standards increase and stress levels rise.

If the prediction of an early diagnosis based on symptoms and indicators is accurate enough, it can help patients avoid heart attacks. For a general decrease in the death rate, early detection of cardiac disease with improved diagnostics and high-risk patients using a model for prediction can be advised. This also improves decision-making for subsequent treatment and prevention. A prediction model is used in the Clinical Decision Support System to assist doctors in determining the risk of heart disease and to recommend suitable therapies to manage subsequent risk. Furthermore, a large body of research has demonstrated that the application of CDSS can enhance clinical decision making, preventative care, and decision quality, in that order [4, 5]. In several nations, coronary artery disease (CAD), also referred to as ischemic heart disease (IHD), is the primary cause of mortality for persons over 35. It rose to the top cause of mortality in China over the same time period. IHD happens when coronary artery stenosis reduces the amount of blood that reaches the heart. Serious outcomes from myocardial injury might include ventricular arrhythmia or even myocardial infarction, which can cause abrupt cardiac death.

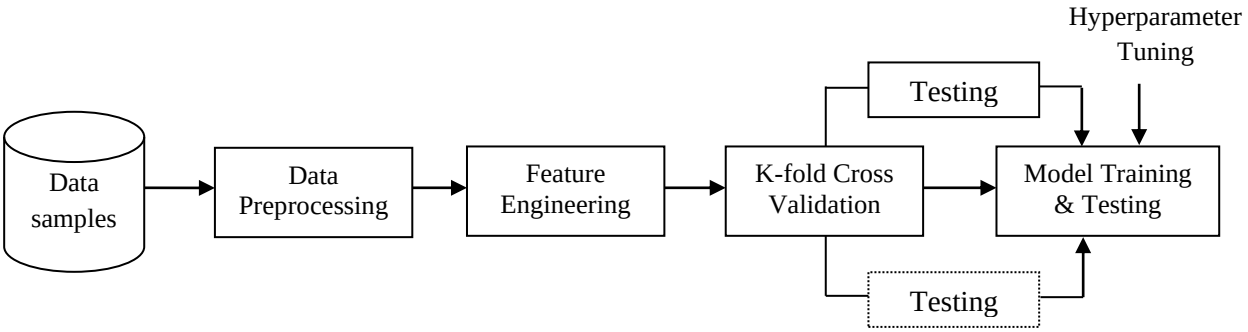

Fig. 2 Flow of Process in the Cardiac Disease Prediction

Thirteen numerically valued features/attributes are fed into the work that is being displayed. According to earlier research, using appropriate engineering and feature selection might enhance prediction. In the medical field, machine learning may be used to identify, anticipate, and diagnose a wide range of illnesses. Giving physicians a tool to identify cardiac issues early on is the main goal of this research. Consequently, it will be simpler to treat patients appropriately without having negative side effects. In order to maximize accuracy and increase performance, this study combines hyperparameter tuning techniques with an experiment using several machine learning models and methodologies. In comparison to other machine learning classifiers like Naïve Bayes, Random Forest, and Support Vector Machine, neural networks demonstrated superior performance. The general flow of process involved in the analytics process has been presented in Fig. 2.

LITERATURE SURVEY

The likelihood of developing heart disease has been predicted using a number of techniques. For instance, genetic algorithms are applied in many different contexts. Weng et al. [6] used clinical data from more than 300,000 UK households to evaluate four alternative models. The more data that were analyzed, the more accurate CVD prediction results were obtained using the NN technique, according to the findings. K-Nearest Neighbor (KNN), Random Forest (RF), and Decision Tree were the three conventional machine learning models that Dimopoulos et al. [7] examined and assessed based on ATTICA data with 2020 observations for the Little CVD dataset. In comparison, it was found that RF's use of the HellenicSCORE tool—a calibration of the ESC Score—produced the greatest results.

The study did not include neural networks with deep learning [8], despite their superior performance in predictions, due to the complexity of the knowledge they acquire. Deep neural nets also need a lot of data for teaching the learning algorithms because learning is continuous [9]. The suggested boosting SVM technique used in this paper outperformed random forests, logistic regression, Naive Bayes, neural networks, and decision trees. These solution strategies rank one of the best-performing algorithms on short datasets and are much simpler to understand.

There is widespread use of neural networks [10, 12]. Swarm-artificial neural network was used by Nandy et al. [11] to predict cardiac heart disease. Increasing accuracy was the research's main objective. Although the research's results were encouraging, there was room for improvement—particularly when compared to the study we suggested—as the accuracy of 95.78% needed to be raised. A data mining and ANN-based strategy for heart illness identification was put out by Sayad and Halkarnikar [13]. This study made use of a backpropagation technique and a multilayer perceptron neural network (MLPNN). Following preprocessing, the residual dataset was divided into two halves. The accuracy of the MLPNN using the backpropagation strategy was 92%, below average. Using data from the Korea National Health and Nutritional Examination Survey (KNHANES-VI), Kim and Kang [14] created a neural network-based method for estimating the risk of heart disease [15]. There are two phases in this procedure. The first step is a feature sensitivity-based selection of features, and the second is a prediction model based on neural networks. Of the 4146 individuals, 3031 were determined to be at minimal risk and 1115 to be with significant risk. A convolutional neural network was proposed by Dutta et al. [16] to predict heart disease by categorizing clinical data that was severely class-imbalanced. However, the study's conclusions were not positive.

Neural networks suffer from temporal complexity and data overfitting, despite their growing popularity and seeming realism. Neural networks also do not converge when dimensionality is low. Moreover, a Support Vector Machine (SVM) has grown in popularity. 2015 saw an investigation of the SVM using sequential minimum optimization procedures; prediction accuracies ranged from 82% to 90%, which was not encouraging. Better results are being obtained, though, with fresh research on SVM algorithms. For instance, Harimoorthy and Thangavelu [17] recently predicted heart disease with 98.7% accuracy using SVM-radial bias kernel technique.

A few research projects that used data mining and machine learning techniques to forecast the course of cardiac disease are included in this part. The explanation above makes it rather evident that the precision attained in individual research projects is currently insufficient. It is possible to get superior performance using certain algorithms over others. Through 10-fold cross-validation, the research study has effectively identified three algorithms that provide 100% accuracy. Therefore, the goal of the project is to identify classifiers that can accurately predict cardiac disease in a way that is relevant for clinical settings.

OVERVIEW OF ARTIFICIAL NEURAL NETWORK

An artificial neural network, often known as a neural net or ANN for short, is a model that is based on the architecture and operation of biological neural networks found in the brains of animals. Artificial neurons, which are networked units or nodes that resemble coupled brain neurons. These are linked together by edges that resemble brain synapses. After processing inputs from other linked neurons, each artificial neuron transmits a signal to another connected neuron. Each neuron's output is determined by a non-linear function

known as the activation function, which takes the total of its inputs into account. The "signal" is a real value. A weight that varies throughout learning determines the intensity of the signal at every connection. It is possible for various layers to alter their inputs in different ways. From the first layer, known as the input layer, to the last layer, known as the output layer, signals may pass through a number of intermediate levels, also known as hidden layers.

The usual method for training neural networks is empirical risk minimization. The principle behind this approach is to reduce the empirical risk—the difference between the goal values in a given dataset and the anticipated output—by optimizing the network's parameters.[4] Typically, gradient-based techniques like backpropagation are employed to estimate network parameters.[4] In order to minimize a predetermined loss function, ANNs iteratively update their parameters while learning from labeled training data during the training phase.[5] The network may generalize to previously unknown data using this way.

A hyperparameter is a constant parameter, meaning that its value is predetermined before learning takes place. Parameter values are obtained by learning. The number of hidden layers, batch size, and learning rate are a few examples of hyperparameters.[20] Certain hyperparameters' values may be influenced by those of other hyperparameters. For instance, the total number of layers may affect the size of certain of the layers.

The number of corrective actions the model does to account for mistakes in each observation is determined by the learning rate. [21] While a lower learning rate requires more time, it has the potential to provide higher accuracy, a higher learning rate reduces training time but results in poorer final accuracy. While some optimizations, like Quickprop, focus largely on increasing reliability, others strive to speed up error minimization. Few modifications employ an adjustable learning rate that rises or reduces as necessary to optimize the pace of convergence and prevent oscillation within the network caused by, for example, alternating link weights. [22] The idea of momentum makes it possible to weight the equilibrium between the gradient and the prior modification so that the weight adjustment is dependent on some

Even if a cost function can be defined on the fly, most of the time the decision is made based on the function's desirable characteristics (like convexity) or the fact that it comes from the model (for example, in a probabilistic model, the posterior probability of the model might be utilized as an inverse cost). Backpropagation is a technique that modifies the connection weights to account for every mistake discovered throughout the learning process. In effect, the mistake amount is distributed among the connections. In technical terms, backprop determines the derivative, or gradient, of the cost function linked to a certain state in relation to the weights. Stochastic gradient descent is one way to update the weights; additional approaches include non-connectionist neural networks, extreme learning machines, "no-prop" networks, training without backtracking, "weightless" networks, and training without backtracking.

First of all, depending on the model and the cost function, local minima may exist, which might explain why models don't always converge to a single solution. Second, when the optimization process starts distant from any local minimum, there's a chance that it won't converge. Thirdly, certain strategies become unfeasible for sufficiently huge data or parameters. It's also important to note that training may cross certain saddle points, which might cause the convergence to go in the incorrect direction. More research has been done on the convergence behavior of some ANN design types than others. The ANN inherits the convergence characteristic of affine models when the network width approaches infinity since it is adequately defined by its first order Taylor expansion during training.[214][215] Another illustration is the observation that ANNs frequently match target functions from low to high frequencies when parameters are minimal. The term "spectral bias," sometimes known as the "frequency principle," describes this phenomenon in neural networks. The behavior of several well-studied iterative numerical systems, such the Jacobi technique, is opposed to this phenomena.

DATASET DESCRIPTION

In order to create the anticipated model for this investigation, a dataset on heart disease was analyzed. The collection of the dataset came from Kaggle [16]. This dataset has 14 characteristics. All feature information is included in Table 1. 1025 patient records total from 713 males and 312 females of various ages make up the collection. Of them, 499 (48.68%) have normal hearts and 526 (51.32%) have heart disease. Of the patients suffering from heart disease, 226 (52.97%) are female and 300 (57.03%) are male. Four databases—Cleveland, Hungary, Switzerland, and Long Beach V—make up this 1988 data collection. It has 76 properties total, including the anticipated attribute, however only 14 of them are used in the published studies. The patient's cardiac condition is indicated in the "target" field. 0 indicates no illness, whereas 1 indicates disease. Fig. 3 represents the correlation between various attributes present in the dataset.

Table. 1 List of Attributes

Attribute	Description
Age	Age in years
Sex	Gender (Male/ Female)
Chest Pain	Type of chest pain (04 different values)
Resting Blood Pressure	Diastolic blood pressure
Serum Cholesterol	Amount of certain lipids in blood
Glucose (Fasting)	Fasting glucose level in the blood
ECG (results)	Resting 12-lead electrocardiography
Heart Rate (max)	Maximum heart rate
Exercise induced Angina	Stimulation of Angina
1. Oldpeak = ST depression	2. Induced by exercise relative to rest
3. Slope of the peak exercise ST segment	4. ST segment slope at peak
5. Number of major vessels colored by fluoroscopy	6. Examination of blocks
7. Thal	A blood condition known as thalassemia Value Value 1: fixed defect (heart not pumping blood to a certain area) Value 2: regular blood flow Value 3: reversible defect (seen blood flow is abnormal)

This sort of dataset is known as multivariate, which refers to multivariate numerical data analysis that involves or provides a range of distinct mathematical or statistical variables. The 14 attributes that make up this composite are as follows: age, sex, type of chest pain, maximum heart rate achieved, serum cholesterol, blood sugar levels fasted, resting blood pressure, exercise-induced angina, oldpeak, or ST depression induced by exercise relative to rest, number of major vessels, and thalassemia. Although there are 76 characteristics in this database, only 14 of them are used in the published research. So far, ML researchers have only utilized the Cleveland database. One of the main tasks on this dataset is to predict, using the patient's provided attributes, whether or not the individual has heart disease. Another experimental task involves diagnosing the patient and extracting various insights from the dataset that may aid in a deeper understanding of the issue.

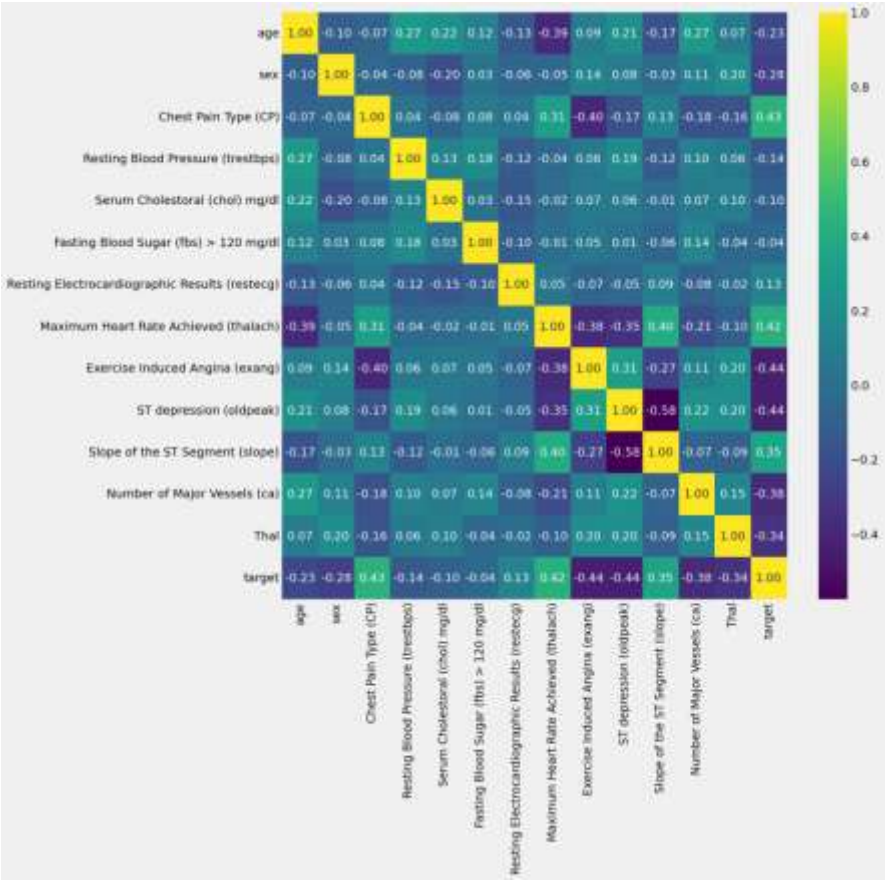

Fig. 3 Plot of Correlation between attributes

EXPERIMENTS & RESULTS

The classifiers have been taught to distinguish between healthy and ill conditions in the medical data set. The precision and recall of a classifier may be used to calculate its accuracy. The characteristics are then ranked in

descending order based on the ANOVA-F value values, which follows the natural rule that the more information an attribute has to contribute about a class, the higher the ANOVA-F value. The most crucial characteristics from a dataset may be chosen using the ANOVA F-test feature selection approach. The F-statistic, which gauges the variation in group means within the dataset, is computed using the ANOVA F-test. Researchers can determine which features most significantly contribute to the classification or prediction job by using the ANOVA F-test. This aids in decreasing the dataset's dimensionality and enhancing the efficacy and efficiency of classification algorithms.
The feature score estimated using ANOVA-F value is presented in the Fig 4.

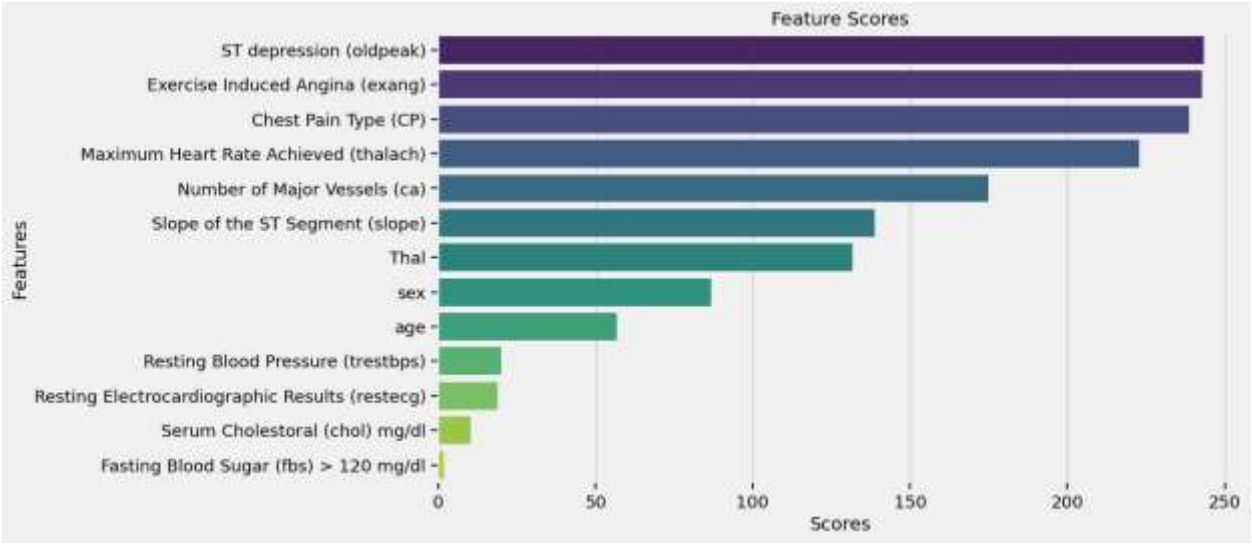

Fig. 4 Feature Importance Score

Table 2. Performance Analysis of NN with all attributes

Optimizer	Loss	Accuracy	F1 Score	Precision	Recall
ADAM	0.4711	0.7866	0.6758	0.6826	0.6782
SGD	0.7872	0.6133	0.6057	0.4942	0.7975
RMSProp	0.4569	0.7866	0.6918	0.6525	0.7443

Based on the estiamted feature importance score the less important attributes were removed from the training and the performance of the model were evaluated using different metric in Table 3. The reduced feature set yielded better results when irrelevant features were included in the training process. In machine learning, overfitting is a prevalent issue that arises when a model learns the train data—including noisy data—too well, leading to subpar generalization performance on test data. Generalization is the capacity to apply information to new contexts, while overfit models are incapable of doing so. One method that penalizes the coefficient is regularization. The coefficients of an overfit model are typically exaggerated. Regularization prevents having the parameters weigh too much by adding penalties to them. The linear equation's cost function is increased by the coefficients. The cost function will therefore rise if the coefficient inflates. In the experiments L2 regularization was used which defines the regularization term as the sum of the squares of all the feature weights.

Table 3. Performance Analysis of NN with selected features

Optimizer	Loss	Accuracy	F1 Score	Precision	Recall
ADAM	0.1853	0.9200	0.8007	0.8072	0.7990
SGD	0.4318	0.8044	0.7052	0.6818	0.7381
RMSProp	0.2399	0.8622	0.7499	0.7662	0.7392

CONCLUSION

In this research, attributes that don't contribute much to a specific high-level are filtered using the ANOVA-F value. The classification algorithm employed was ANN. This study demonstrates how feature selection enhances classification accuracy while boosting computing efficiency. Additionally, they lessen the dimensionality of the dataset, which lowers the system's complexity. It lowers the amount of computing power needed, the amount of storage space needed, the cost of the health checklist, and the quantity of patient characteristics that must be collected. Also the analysis of neural network under different optimization algorithms was tabulated. Among the evaluated optimization algorithms the ADAM- Adaptive Momentum yielded better results.

REFERENCE

[1] Fuchs, Flávio D., and Paul K. Whelton. "High blood pressure and cardiovascular disease." *Hypertension* 75.2 (2020): 285-292.

[2] Bhatla, N., & Jyoti, K. (2012). An analysis of heart disease prediction using different data mining techniques. International Journal of Engineering, 1(8), 1-4.

[3] Patel, J., TejalUpadhyay, D., & Patel, S. (2015). Heart disease prediction using machine learning and data mining technique. Heart Disease, 7(1), 129-137.

[4] Zhenya Q., Zhang Z. A hybrid cost-sensitive ensemble for heart disease prediction. *BMC Medical Informatics and Decision Making* . 2021;21(1):p. 73. doi: 10.1186/s12911-021-01436-7.

[5] Alarsan F. I., Younes M. Analysis and classification of heart diseases using heartbeat features and machine learning algorithms. *Journal of Big Data* . 2019;6(1):p. 81.

[6] Weng, SF, Reps, J, Kai, J, Garibaldi, JM, and Qureshi, N. Can machine-learning improve cardiovascular risk prediction using routine clinical data? *PLoS One*. (2017)

[7] Dimopoulos, AC, Nikolaidou, M, Caballero, FF, Engchuan, W, Sanchez-Niubo, A, Arndt, H, et al. Machine learning methodologies versus cardiovascular risk scores, in predicting disease risk. *BMC Med Res Methodol*. (2018) 18:1–11.

[8] Shin H. C., Roth H. R., Gao M., et al. Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning. *IEEE Transactions on Medical Imaging* . 2016;**35**(5):1285–1298.

[9] Onan A. Sentiment analysis on product reviews based on weighted word embeddings and deep neural networks. *Concurrency and Computation: Practice and Experience* . 2020;**33**(23)

[10] 13. Karayılan T., Kılıç Ö. Prediction of heart disease using neural network. Proceedings of the 2017 International Conference on Computer Science and Engineering (UBMK); October 2017; Antalya, Turkey. pp. 719–723.

[11] 14. Nandy S., Adhikari M., Balasubramanian V., Menon V. G., Li X., Zakarya M. An intelligent heart disease prediction system based on swarm-artificial neural network. *Neural Computing and Applications* . 2021:1–15.

[12] 16. Awan S. M., Riaz M. U., Khan A. G. Prediction of heart disease using artificial neural network. *VFAST Transactions on Software Engineering* . 2018;**13**(3):102–112.

[13] 17. Sayad A. T., Halkarnikar P. P. Diagnosis of heart disease using neural network approach. *International Journal of Advances in Science Engineering and Technology* . 2014;**2**(3):88–92.

[14] 18. Kim J. K., Kang S. Neural network-based coronary heart disease risk prediction using feature correlation analysis. *Journal of Healthcare Engineering* . 2017;**2017**:13.

[15] 19. Kweon S., Kim Y., Jang M. J., et al. Data resource profile: the Korea national health and nutrition examination survey (KNHANES) *International Journal of Epidemiology* . 2014;**43**(1):69–77.

[16] 20. Dutta A., Batabyal T., Basu M., Acton S. T. An efficient convolutional neural network for coronary heart disease prediction. *Expert Systems with Applications* . 2020;**159**

[17] Harimoorthy K., Thangavelu M. Multi-disease prediction model using improved SVM-radial bias technique in healthcare monitoring system. *Journal of Ambient Intelligence and Humanized Computing* . 2021;**12**(3):3715–3723.

[18] Vapnik VN, Vapnik VN (1998). *The nature of statistical learning theory* (Corrected 2nd print. ed.). New York Berlin Heidelberg: Springer. ISBN 978-0-387-94559-0.

[19] Ian Goodfellow and Yoshua Bengio and Aaron Courville (2016). *Deep Learning*. MIT Press.

[20] Lau S (10 July 2017). "A Walkthrough of Convolutional Neural Network – Hyperparameter Tuning". *Medium*.

[21] Wei J (26 April 2019). "Forget the Learning Rate, Decay Loss". arXiv:1905.00094.

[22] *Li Y, Fu Y, Li H, Zhang SW (1 June 2009). "The Improved Training Algorithm of Back Propagation Neural Network with Self-adaptive Learning Rate". 2009 International Conference on Computational Intelligence and Natural Computing. Vol. 1. pp. 73–76.*

[23] Lee J, Xiao L, Schoenholz SS, Bahri Y, Novak R, Sohl-Dickstein J, et al. (2020). "Wide neural networks of any depth evolve as linear models under gradient descent". *Journal of Statistical Mechanics: Theory and Experiment*. **2020** (12): 124002.

[24] Arthur Jacot, Franck Gabriel, Clement Hongler (2018). *Neural Tangent Kernel: Convergence and Generalization in Neural Networks* (PDF). 32nd Conference on Neural Information Processing Systems (NeurIPS 2018), Montreal, Canada.