EAI Endorsed Transactions
on Pervasive Health and Technology
Research Article EAI.EU

Automatic Data-Driven Classification Systems for Cardiovascular Disease

Muralidharan Jayaraman[1], Shanmugavadivu Pichai[2, *]

[1, 2]Gandhigram Rural Institute, Tamil Nadu, India

Abstract

Cardiovascular disease (CVD) continues to contribute significantly to preventable deaths and avoidable disability worldwide. Prediction and prevention are of utmost importance in the support of public health. Machine learning and deep learning algorithms have emerged as powerful tools to improve the accuracy of diagnosis, prognosis, and treatment of cardiovascular disease. By employing these technologies, medical professionals can gain valuable insights into the risk factors associated with CVD. The focus of this research is to classify and predict cardiovascular diseases using techniques such as support vector machines, ensemble methods, decision trees, random forests, and neural networks. The effectiveness of these algorithms is evaluated based on metrics including accuracy, sensitivity, specificity, area under the curve (AUC), and F1 score. Results show that support vector machines and ensemble methods offer superior accuracy, while neural networks exhibit higher sensitivity and specificity in predicting cardiovascular diseases.

Keywords: Cardiovascular disease, neural networks, machine learning algorithms, classification

Received on 11 02 2024, accepted on 30 05 2024, published on 25 06 2024

doi: 10.4108/eetpht.10.6430

1. Introduction

The world is facing a significant number of preventable deaths and avoidable disabilities due to cardiovascular disease (CVD), a trend that is increasing due to various factors, including lifestyle choices. CVD encompasses several conditions affecting the heart and blood vessels, such as stroke, peripheral artery disease, and coronary artery disease. The causes of CVD can be multifaceted, often involving genetic factors, lifestyle habits, tobacco and alcohol consumption, physical inactivity, poor diet, obesity, and stress-induced hypertension. These risk factors can be managed by adopting lifestyle changes such as quitting smoking, refraining from alcohol consumption, eatig a healthy diet, maintaining healthy blood pressure levels, and engaging in regular physical exercise. [1].

CVD is a major health concern in India. According to the World Health Organization (WHO), CVD accounted for 31% of deaths in India in 2019, a significant increase from 25% in 2015. The prevalence of CVD in India is further highlighted by the fact that the disease is responsible for more than 40% of all hospital admissions in the country and nearly 20% of all outpatient visits. The Indian Heart Association estimates that 80 million Indians are currently living with CVD, which is likely to increase to 135 million by 2030. It indicates that CVD is a major public health concern. These statistics highlight the urgent need for effective prevention and treatment strategies to combat this public health crisis. [2-4].

Artificial intelligence-based prediction models can help healthcare professionals identify high-risk patients and take preventive measures in a timely manner. AI models, trained on large datasets of CVD information, can accurately predict the likelihood of CVD in patients. These models assist doctors in making decisions about treatments, medications and lifestyle changes for their patients. Moreover, AI models can be used to detect early signs of CVD which may not have been detected by traditional methods. The use of AI in predicting CVD is a promising area that can prevent and/or delay heart-related

*Corresponding author. Email: psvadivu67@gmail.com

incidents, reducing healthcare costs and improving patient outcomes [5,6].

Development of Machine Learning (ML) and Deep Learning (DL) algorithms can help to identify, diagnose, and treat CVD more effectively. This can be done by analyzing patient data such as age, medical history, and lifestyle. With the help of these technologies, doctors can make more accurate predictions about a patient's health and provide better treatment options. Deep learning algorithms can identify patterns in large datasets that traditional methods cannot, allowing for more precise predictions about a patient's risk. ML algorithms are critical in detection of anomalies in medical images to provide an indication of CVD symptoms. These algorithms learn complex patterns and can be used to identify biomarkers that are associated with CVD, as well as predict the risk of developing these diseases [7].

This paper contributes several insights into the field of cardiovascular predictions with the help of Cardiovascular Disease Classification (CVDC). Firstly, it presents a comprehensive introduction to the relevant research. Secondly, it provides an in-depth analysis of the significant findings in the literature review. Thirdly, it explores the data pre-processing and classification methods on cardiovascular data. Finally, it provides a detailed analysis to draw meaningful conclusions.

2. Literature Review

CVD is a common yet predictable and preventable causes of death. As such, there is a definite requirement for more accurate and reliable methods of detecting CVD. Machine learning and neural network (NN) classification approaches have been identified as a potential solution to this issue. This literature review explores the current state of the classification of CVD.

The advent of machine learning (ML) has revolutionized cardiovascular prediction, allowing for more accurate and reliable predictions. In recent years, many authors have contributed significantly to the field of ML cardiovascular prediction. One notable contribution is from Princy, R. J. P. *et al*. [8], who used ML to develop a predictive model for heart failure. Their work revealed that by training the model using clinical data and supervised machine learning algorithms, it was able to accurately predict heart failure. This work was demonstrated with Decision Tree (DT), Naive Bayes (NB), Logistic Regression (LR), Random Forest (RF), Support Vector Machine (SVM), and KNN-based approaches. The results indicated the potential of ML in predicting CVD.

Jinjri W. M. *et al*. [9] proposed an approach using ML to predict the risk of stroke. A combination of features from electronic medical records and patient-level data was utilized to develop a predictive model that could accurately predict the risk of stroke. ML approaches were used to construct a model that determines CVD, predicting the potential of CVD in individuals. Cardiovascular neural networks (CvNNs) are a powerful tool for predicting CVD outcomes. Naturally, the quality of data is critical in determining the accuracy and reliability of these predictive models, by ensuring the models have a solid training foundation. In recent years, several authors have made significant contributions to the development and optimization of CNNs for predicting cardiovascular outcomes: Sumit Sharma and Mahesh Parmar [10] developed a DL-based NN framework using Talos optimization to predict heart attack risk, whilst Ramprakash *et al*. [11] developed a deep neural networks-based model to predict the onset of congestive heart-related illness, which was compared with artificial NN using evaluation metrics such as accuracy, sensitivity, and specificity.

3. Materials and Methods

3.1. Experimental Dataset

This article used the cardiovascular disease dataset with a collection of 70,000 records from the Kaggle repository [12]. It includes patients' medical history, laboratory test results, and other clinical features. Each record contains demographic information of age, gender, height, and weight, and features like systolic blood pressure, diastolic blood pressure, cholesterol, and other vital signs. Each record is associated with a patient and contains a binary variable, indicating CVD in an individual with a binary output. The results of dataset analysis at a high level are given in Table 1. Table 2 consists of the clinical characteristics of the cardiovascular disease dataset. O, E, and S represent the Objective, Examination, and Subjective features respectively.

Table 1: Overall Analysis of the cardiovascular disease dataset

Dataset	CVD
Patients	70000
Cardiac Absence	34979
Cardiac Presence	35021
Median Age	54
Age Range	30-65
Male	24470
Female	45530

Datasets are usually represented by a distribution of values that help in understanding the data. Data distribution in a dataset can describe the underlying data structure, such as its range, outliers etc. Figures 1 (a) and (b) show the frequency of data distributed in the CVD dataset concerning Age and Body Mass Index respectively.

Table 2: Clinical characteristics of cardiovascular disease clinical Dataset

Attribute	Description
Age (O)	Age in number of days
Height (O)	Height as measured in Centimeter
Weight (O)	Weight as taken in Kilogram
Gender (O)	Gender Code (1: Women, and 2: Men)
ap_hi (E)	Top Number of Blood Pressure
ap_lo (E)	Bottom Number of Blood Pressure
Cholesterol (E)	Intensity of Cholesterol (1: Normal, 2: Above Normal, and 3: Well Above Normal)
Glucose (E)	Intensity of Glucose (1: Normal, 2: Above Normal, and 3: Well Above Normal)
Smoking (S)	Binary values (0: No, and 1: Yes) indicate Smoker or Non-Smoker
Alcohol intake (S)	Binary values indicate Alcoholic or Non-Alcoholic (0: No, and 1: Yes)
Physical activity (S)	Binary values indicate whether the person is involved in physical activities or not (0: Active, and 1: Inactive)
Cardio (Target)	Binary values indicate presence of CVD or Not (0: Absence, and 1: Presence)

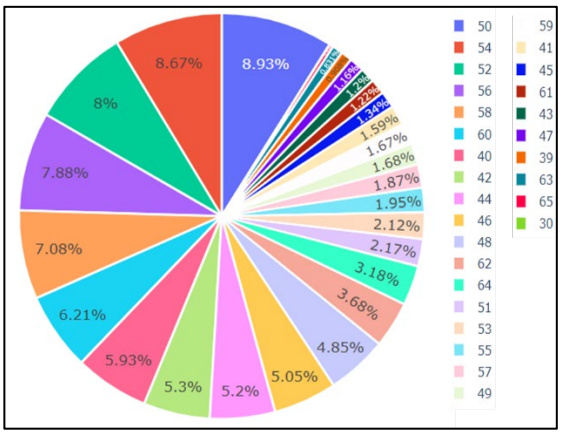

Figure 1 (a): Distribution of Age Group for Non-Cardiac Disease

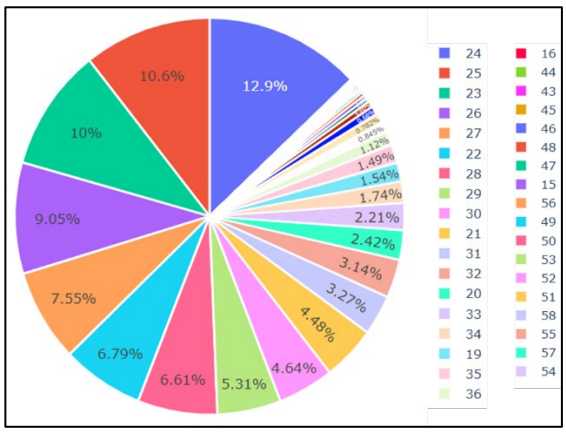

Figure 1 (b): Distribution of BMI for Non-Cardiac Disease

3.2. Data Preprocessing

The initial pre-processing step was to find the null values present in the dataset. Since the CVD dataset has a total of 70000 records, it does not contain null values. An outlier is one which has a value far from the normal pattern. Reasons for a value being an outlier may be due to variability in measurement, or by errors in experiments or data collection. This can generate significant issues in the analysis of data and prediction. Figure 2 represents outliers found in the CVD dataset relating to blood pressure [13].

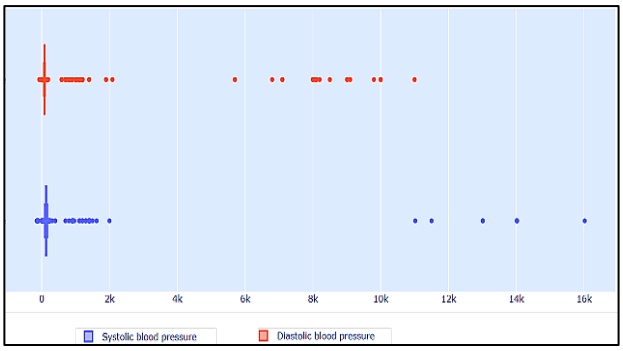

Figure 2: Box Plot for blood pressure with outliers

Correlation reflects the level and direction of the linear relationship between a pair of quantitative variables ranging between -1 and +1. A positive value indicates a strong association, while a negative value indicates otherwise. This phase aims to determine the connection between the attributes [14]. As shown in Figure 3, the correlation heatmap visually represents the relationships within the resulting correlation matrix.

Figure 3: Correlation Matrix Heatmap for Pre-processed CVD Data

In this dataset attribute Cardio is a Target Variable, Systolic Blood Pressure has a high correlation, and alcho_intake and smoking have a low correlation with the target attribute Cardio.

After removing outliers, the CVD dataset which has 69,967 records, was split into two sets with an 80:20 ratio for training and testing. The training dataset consists pf 55,973 records and the test dataset 13,994 records, all balanced, with an approximately equal number of positive and negative cases.

3.3. Machine Learning Algorithms

Machine Learning (ML) techniques are employed for the classification of clinical datasets, with the accuracy of classification being significantly influenced by the distribution and diversity of data [15]. ML can be divided into two main categories based on its learning process: supervised learning and unsupervised learning. The former works on labeled data while the latter uses unlabeled data [16]. All ML algorithms undergo training with both training and test datasets. The effectiveness of ML algorithms is ideally influenced by the data split, which is explained in detail in further sections. At first, ML algorithms are trained with default parameters to determine their accuracies. Then, classifiers are trained with tuned hyperparameters, and a comparison is made based on both sets of accuracy values. Machine Learning methods [17-20] are described in this section, and Figure 4 lists the ML techniques used in CVDC.

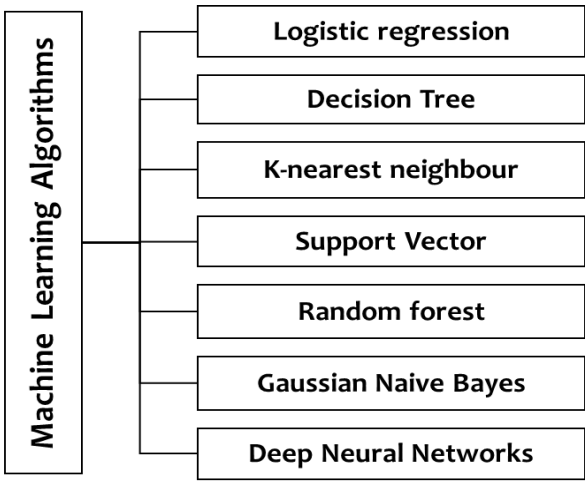

Figure 4: Machine Learning Techniques used in CVDC

Logistic Regression (LR): As a Supervised Learning algorithm to predict discrete outcomes, LR can be used to identify the most important independent variables that have a predictive effect on the outcome of interest. It can also be used to identify interactions between independent variables

and is useful for identifying which variables are associated with higher or lower probabilities of the outcome of interest.

Decision Tree (DT): For Classification, and Regression, DT is used as a Supervisory Learning algorithm, having the primary objective of predicting target values from a smaller data pool and lesser number of iterations. The technique divides a dataset into progressively smaller sections while simultaneously creating an associated decision tree. Each of these subsets is then used to make a decision based on the chosen criteria. Decision Tree algorithms can identify important factors in the data, helping in making better decisions. This can be done by analysing the branches of the tree and the paths that are being taken.

K-Nearest Neighbour (KNN): This classification method classifies data points according to how closely they are related to other data points in a labelled training set. It is a non-parametric, supervised machine learning technique which works by using a distance metric to measure the similarity between data points and then determining the class of the data point from those distances. Euclidean Distance, Manhattan Distance and Cosine Similarity are the most commonly used distance metrics. In this algorithm, data points are classified according to most of their closest neighbors.

Support Vector Machine (SVM): As a sophisticated supervised ML technique, this is used to tackle difficult issues by determining the optimum hyperplane that separates the input points. The goal of SVM is finding a decision boundary known as the “Maximum Margin Hyperlane” which maximizes the margin between the data points of two different classes. This technique is for solving non-linear problems by using a kernel trick. This involves transforming the data into another dimensional space via a kernel function to find a linear boundary. The most commonly used kernels include linear, polynomial, radial basis function, and sigmoid.

Random Forest (RF): As an ensemble method working by building multiple decision trees using a different set of training data, each decision tree is allowed to make its prediction. By combining the predictions of individual trees, the final prediction is made. This reduces the variance in the predictions and makes them more accurate.

Gaussian Naive Bayes (GNB): In this technique, Bayes' Theorem is used. The probability of an event occurring is calculated based on a prior understanding of circumstances which could be relevant to the event. The algorithm works by calculating the probability of a given feature belonging to a category. It then multiplies the probabilities of all features belonging to the category and assigns the category with the highest probability to the given instance. The algorithm requires only minimal training data to produce accurate results, which can be beneficial in cases where data is limited.

Deep Neural Networks (DNN): DNNs are a branch of Artificial Intelligence technology that have been in use for decades in various healthcare industries solving complex problems. DNNs are composed of multiple layers of neurons, each layer working in tandem to process data and information. The layers are connected through weights and biases, which help to determine how the data is processed. DNNs work by connecting input nodes to output nodes through several hidden layers of neurons. Each neuron in the network is capable of storing information and computing outputs from inputs. Data is passed through the NN in the form of numerical values, and the neurons in each layer process the values before passing them to the next layer. The neurons in the hidden layers are trained to detect patterns and make predictions. The output layer of neurons then generates a result based on all the information that was processed in the hidden layers. This result is then compared to the desired output for possible fine-tuning of weights and biases of the neurons in the hidden layers [21]. This process is repeated until the desired output is achieved.

3.4. Proposed Methodology

Initially, in this CVD dataset, normalization was used in order to measure the samples individually. In this study, "MinMaxScaler" was used directly as feature_range = (min, max). This was done to normalize the data so that the samples would fall between 0 and 1, as some attributes in the dataset had values close to 0 or 1 after pre-processing. The proposed CVDC was then constructed for the cardiovascular classification of data using an 80:20 training and testing set split. The trained features were input into the ML algorithms and the optimal NN structure. The effectiveness of the NN system and other ML models was evaluated for heart disease classification, with hyperparameter tuning enhancing their performance.

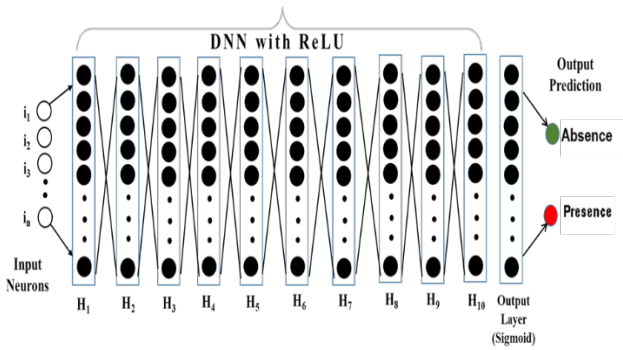

Figure 5: DNN structure used in CVDC

Initially, the DNN architecture was designed with a sequential model with an input layer, multiple hidden layers and an output layer. The CVD dataset used in this research included 10 input features. The dataset was then split into training and testing sets in an 80:20 ratio.

Subsequently, 10 hidden layers were used with ReLu as the activation function. Given that this work addresses a binary classification problem, the sigmoid activation function was employed at the output node. Following the layer formation, the NN model was formed using the binary cross-entropy loss function, which measures the disrepancy between the actual and predicted outputs. Adam was used as the optimizer, and accuracy was the metric for assessing the model's performance. To train the proposed model, two regularizers - L2 regularizer and dropout - were used. The structure of the DNN is shown in Figure 5 and the workflow of CVDC is depicted in Figure 6.

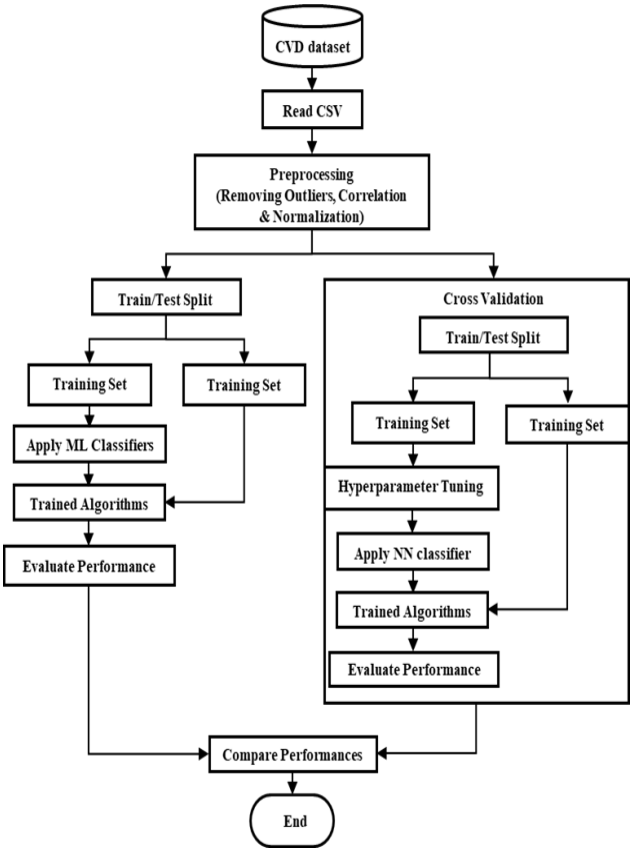

Figure 6: The overall workflow of CVDC

4. Results and Discussions

The results of this study were observed by comparing the performance of ML classifiers against the DNN model. The performance of CVDC on the CVD dataset was measured using accuracy, precision, recall, F1-Score and specificity. The metrics were calculated for classification using the formulas shown in Table 7.

Table 3: Performance Metrics

Metrics	Precision	Recall	F1-score	Accuracy
Formula	$\frac{TP}{TP+FP}$	$\frac{TP}{TP+FN}$	$\frac{2*recall*precision}{recall+precision}$	$\frac{TP+TN}{TP+TN+FP+FN}$

The performance comparison metrics is presented in Table 4 concerning Precision, Recall, and F1-Score on CVDC. The corresponding illustration is given in Figure 7.

Table 4: Performance Metrics comparison on CVDC concerning Precision, Recall, and F1-Score

Classifiers	Precision	Recall	F1-Score
LR	77.42	69.57	73.29
DT	86.53	79.74	83.00
KNN	74.83	66.47	70.40
SVM	77.50	71.38	74.32
RF	88.00	83.02	85.43
GNB	67.23	62.67	64.87
NN	91.59	87.89	89.70

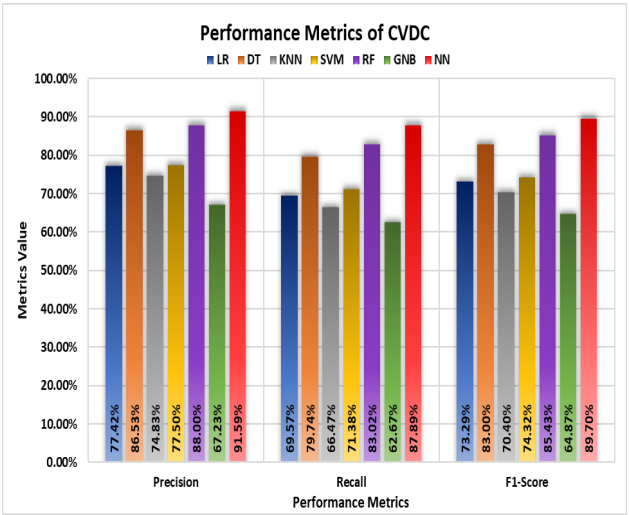

Figure 7: Performance comparison of CVDC over ML and DNN models concerning Precision, Recall, and F1-Score

From Table 4 and Figure 7, it can be deduced that the DNN model performs better than the ML models in terms of Precision, Recall, and F1-Score. The NN model achieved a Precision of 91.59%, Recall of 87.89%, and F1-

Score of 89.70%. Following the NN model, the RF model, an ensemble technique, also demonstrated strong performance with scores of 88%, 83.02%, and 85.43% respectively.

The performance of the CVDC was then further compared by evaluating the training score and testing score, which highlights the importance of using preprocessing techniques according to the nature of their features. This comparison is shown in Table 6, with a pictorial representation of the difference illustrated in Figure 8.

Table 5: Performance of Training and Testing score on CVDC

Classifiers	Score Train	Score Test	Score Difference
LR	71.42	71.38	0.04
DT	66.07	82.16	-16.09
KNN	78.37	67.99	10.38
SVM	72.37	72.82	-0.45
RF	70.27	84.94	-14.67
GNB	61.55	61.83	-0.28
NN	83.19	89.33	-6.14

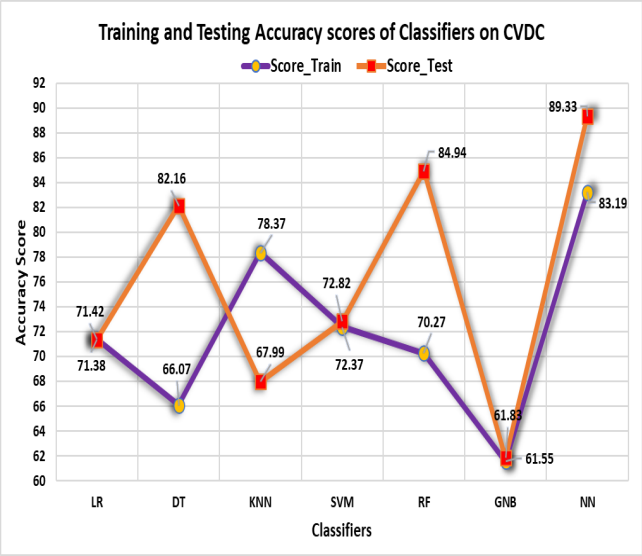

Figure 8: Training and Testing scores of classifiers on CVDC

As per Table 5 and Figure 8, it was found that every classifier has specific differences in training and testing scores. Among all the classifiers, good performance was considered according to the level of accuracy and difference value. If the classifier had high accuracy with less difference, it was considered a good performance classifier. From the above result, the NN classifier had high values of training and testing of 83.19% and 89.33% respectively with a low difference ratio.

The confusion matrix serves to evaluate the performance quality of a classifier. The classifier assigned incorrectly labeled data to off-diagonal elements, with the diagonal elements representing the points where predicted and actual labels were identical. More accurate predictions are shown by a confusion matrix with greater diagonal values [22]. Figure 9 shows the proposed CVDC's confusion matrix, which shows the correlated target properties.

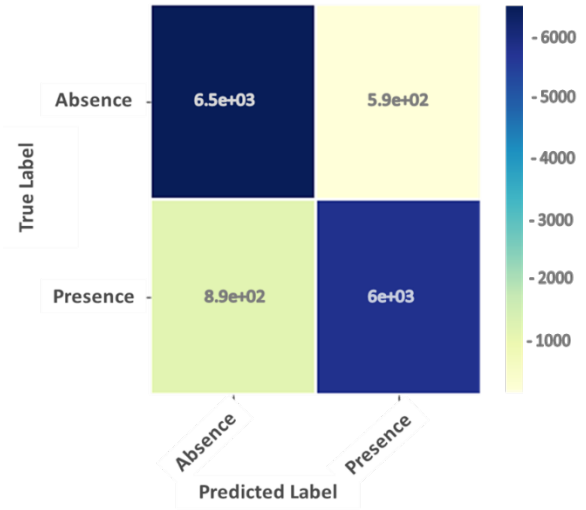

Figure 9: Confusion Matrix on CVDC

The confusion matrix displayed the results of the NN classifier's testing set classification report for the CVD dataset, where NN had a high accuracy rate in terms of performance parameters. Out of 13,994 testing cases, 6,500 instances were detected and estimated as CVD absence cases, which is a True Positive (TP). A total of 6,000 cases were viewed and projected as cases of CVD, which is a True Negative (TN). 895 cases were recognized as absences but estimated as presences, resulting in a False Negative (FN). 597 cases were noted as present but detected as absent, resulting in a False Positive (FP).

5. Conclusion

This study conducted a comparative analysis of cardiovascular disease (CVD) classification and prediction using various ML classifiers and DNNs. The findings underscored that deep neural networks (DNNs) surpassed other classifiers in terms of accuracy and effectiveness. This highlights the DNN classifier's capability to accurately and efficiently predict cardiovascular diseases. As a result, this research significantly enhanced the classification accuracy of classifiers, laying a solid foundation for future research in this field. Employing DNN classifiers in clinical settings could potentially aid in early diagnosis and treatment of cardiovascular disease, and so would be highly beneficial in improving the accuracy of healthcare decision making.

References

[1] Huffman, M. D., Prabhakaran, D., … Sachdev, H. S., & New Delhi Birth Cohort (2011). Incidence of cardiovascular risk factors in an Indian urban cohort results from the New Delhi birth cohort. Journal of the American College of Cardiology, 57(17), 1765–1774.

[2] https://www.who.int/india/health-topics/cardiovascular-diseases

[3] Kumar, A. S., & Sinha, N. (2020). Cardiovascular disease in India: a 360 degree overview. Medical Journal Armed Forces India, 76(1), 1-3.

[4] Wadhawan, S., & Maini, R. (2022). A systematic review on prediction techniques for cardiac disease. International Journal of Information Technologies and Systems Approach (IJITSA), 15(1), 1-33.

[5] Romiti, S., Vinciguerra, M., Saade, W., Anso Cortajarena, I., & Greco, E. (2020). Artificial intelligence (AI) and cardiovascular diseases: an unexpected alliance. Cardiology Research and Practice, 2020.

[6] Mathur, P., Srivastava, S., Xu, X., & Mehta, J. L. (2020). Artificial intelligence, machine learning, and cardiovascular disease. Clinical Medicine Insights: Cardiology, 14, 1179546820927404.

[7] Krittanawong, C., Johnson, K. W., Rosenson, R. S., Wang, Z., Aydar, M., Baber, U., ... & Narayan, S. M. (2019). Deep learning for cardiovascular medicine: a practical primer. European heart journal, 40(25), 2058-2073.

[8] Princy, R. J. P., Parthasarathy, S., Jose, P. S. H., Lakshminarayanan, A. R., & Jeganathan, S. (2020, May). Prediction of cardiac disease using supervised machine learning algorithms. In 2020 4th international conference on intelligent computing and control systems (ICICCS) (pp. 570-575). IEEE.

[9] Jinjri, W. M., Keikhosrokiani, P., & Abdullah, N. L. (2021, July). Machine learning algorithms for the classification of cardiovascular disease-A comparative study. In 2021 International Conference on Information Technology (ICIT) (pp. 132-138). IEEE.

[10] Sharma, S., & Parmar, M. (2020). Heart diseases prediction using deep learning neural network model. International Journal of Innovative Technology and Exploring Engineering (IJITEE), 9(3), 2244-2248.

[11] Ramprakash, P., Sarumathi, R., Mowriya, R., & Nithyavishnupriya, S. (2020, February). Heart disease prediction using deep neural network. In 2020 International Conference on Inventive Computation Technologies (ICICT) (pp. 666-670). IEEE.

[12] https://www.kaggle.com/datasets/sulianova/cardiovascular-disease-dataset

[13] Hasan, N., & Bao, Y. (2021). Comparing different feature selection algorithms for cardiovascular disease prediction. Health and Technology, 11, 49-62.

[14] Kanimozhi, G., Shanmugavadivu, P., & Rani, M. M. S. (2020). Machine Learning-Based Recommender System for Breast Cancer Prognosis. Recommender System with Machine Learning and Artificial Intelligence: Practical Tools and Applications in Medical, Agricultural and Other Industries, 121-140.

[15] Salama GI, Abdelhalim MB, Zeid MA. Experimental comparison of classifiers for breast cancer diagnosis. In2012 Seventh International Conference on Computer Engineering & Systems (ICCES) 2012 Nov 27 (pp. 180-185). IEEE.

[16] Agarwal A, Saxena A. Malignant Tumor Detection Using Machine Learning through Scikit-learn. International Journal of Pure and Applied Mathematics. 2018;119(15):2863-74.

[17] Rana SP, Dey M, Tiberi G, Sani L, Vispa A, Raspa G, Duranti M, Ghavami M, Dudley S. Machine Learning Approaches for Automated Lesion Detection in Microwave Breast Imaging Clinical Data. Scientific reports. 2019 Jul 19;9(1):1-2.

[18] Swathy, M., & Saruladha, K. (2022). A comparative study of classification and prediction of Cardio-vascular diseases (CVD) using Machine Learning and Deep Learning techniques. ICT Express, 8(1), 109-116.

[19] Ganesan, K., Pichai, S., Kavitha, M. S., & Takahashi, M. (2022). Data imputation in deep neural network to enhance breast cancer detection. International Journal of Imaging Systems and Technology, 32(6), 2094-2106.

[20] Naganandhini, Shanmugavadivu, P., Kanimozhi, & Kavitha, M. S. (2021, September). Data imputation of brain MRI features with enhanced multinomial logistic regression for Alzheimer's disease classification. In Proceedings of the 6th International Conference on Sustainable Information Engineering and Technology (pp. 339-347).

[21] Ali, L., Rahman, A., Khan, A., Zhou, M., Javeed, A., & Khan, J. A. (2019). An automated diagnostic system for heart disease prediction based on ${\chi^{2}} $ statistical model and optimally configured deep neural network. Ieee Access, 7, 34938-34945.

[22] Khan, A., Qureshi, M., Daniyal, M., & Tawiah, K. (2023). A Novel Study on Machine Learning Algorithm-Based Cardiovascular Disease Prediction. Health & Social Care in the Community, 2023.