Harnessing Optimized Progressive Dropout Dense Networks for Accurate Heart Disease Prediction in Healthcare Systems

Lakkala Jayasree, Dr. D. Usha

[1]Research Scholar, Department of C.S.E., Dr. M.G.R Educational & Research Institute, Chennai, India.

[2]Associate Professor, Department of C.S.E, Dr. M.G.R Educational & Research Institute, Chennai, India.

Abstract: The field of medical sciences has seen substantial diversification due to developments in computational capabilities and methodologies, namely in diagnosing cardiovascular disorders in humans. Cardiovascular disease (CVD) is a highly challenging condition with significant adverse effects on the global population, greatly impacting human well-being. The prompt and precise detection of cardiovascular diseases in persons can provide substantial advantages in the management of the progression of heart failure during its initial phases, hence enhancing the likelihood of the patient's life. The identification of heart illness by manual processes is prone to prejudice and vulnerable to variations among examiners. Machine learning algorithms have demonstrated efficacy and reliability in identifying and classifying persons with heart illness and those in a healthy state. A novel approach, the Optimised Progressing Dropout Dense Network (OPDDN), has been devised to identify heart illness accurately. By meticulously adjusting and refining particular parameters, we achieved an exceptional accuracy level, reaching an impressive rate of 99%. This accomplishment is deserving of recognition. In this study, a comparison was conducted between the proposed strategy and an alternative method, which led to the observation that our method demonstrated a higher level of effectiveness. This innovative methodology has the potential to function as a more efficient tool for medical practitioners in the diagnosis of cardiovascular conditions. The findings of our study suggest that implementing this innovative approach has the potential to augment the advancement of more efficacious treatment methodologies for persons diagnosed with heart disease.

Keywords: *cardiovascular disease, machine learning, neural networks, evaluation metrics.*

1. Introduction

Cardiovascular diseases (CVDs) were responsible for 17.9 million deaths, or about 32% of all deaths worldwide in 2019, according to data compiled by the World Health Organisation [1]. Furthermore, CVD accounted for almost 17.7 million deaths annually [2]. In 2018, CVD surpassed all other causes of death in Australia, accounting for 42% of all deaths [3], according to the Australian Institute of Health and Welfare research. The current study aims to explore and create a more efficient approach for promptly detecting heart disorders. This is motivated by the limitations of existing diagnostic approaches, which have proven inadequate to identify heart ailments early due to a lack of precision and lengthy computational processes. Diagnosing and managing cardiac disease poses significant challenges when advanced technology and healthcare expertise are not readily accessible [4], [5]. A substantial number of individuals can see a significant increase in their life expectancy through the provision of an accurate diagnosis and effective therapy [6]. Cardiac diseases are diagnosed after thoroughly evaluating the patient's medical history and findings from a physical examination and considering any concerning symptoms.

Nevertheless, the results of this diagnostic approach need to be revised in identifying individuals with cardiac disease. Examining further is a task that incurs significant costs and presents computing challenges [7]. We have developed a noninvasive prediction method to address these challenges utilizing neural network-based machine learning classifiers. Accurately identifying cardiac conditions is efficiently achieved by utilizing a sophisticated decision-making framework that incorporates machine learning classifiers and artificial fuzzy logic. As a result, there has been a decrease in the mortality rate [8, 9]. Numerous studies have employed the Cleveland heart disease dataset. To effectively train and assess prediction models within machine learning, it is imperative to possess appropriate and relevant data. The utilization of a refined or standardized dataset during both the training and testing phases can significantly improve the accuracy of machine learning classifiers.

Moreover, integrating pertinent and interconnected data attributes might augment the model's predictive capabilities. Hence, the accuracy of machine learning classifiers is significantly influenced by data standardization and feature selection processes. Several researchers have employed various predictive techniques in the existing literature, yet these methodologies have proven ineffective in accurately predicting cardiac illnesses. The implementation of data standardisation is necessary in order to enhance the precision of machine learning classifiers. Several standardisation procedures, such as standard scalar (SS), min-max scalar, and others, exclude instances with missing feature values from the dataset.

Multiple tests are necessary for heart disease prediction. Cardiovascular disease prognosis is notoriously challenging, especially in nations with limited access to medical professionals and diagnostic tools [10]. When given enough data, computational classifiers can make accurate medical diagnoses. There are several methods for predicting CVD risks using machine learning. Most of these strategies use publicly available datasets for model training and assessment. These data sets have helped researchers develop state-of-the-art CVD risk prediction systems based on machine learning-based predictive models. Both risk factors and diagnoses are recorded in these databases. Preprocessing is necessary for CVD prediction systems [11] because of the unreliability and redundancy of clinical datasets. In addition, risk variables (features) can be chosen based on criteria including universal occurrence, independent impact on cardiovascular disease, and modifiability. Researchers have used a variety of risk factors and characteristics to develop CVD predictions. Relevant datasets are essential for machine learning algorithms [12]. Lack of sufficient medical data, feature selection, ML algorithm implementations, and in-depth analysis may make heart disease prediction challenging. Our research aims to better forecast CVD by filling in these gaps.

Furthermore, there are limitations to existing study datasets. These databases need to contain more risk factors and clinical data characteristics. The clinical severity of a certain case may influence the precision of a prediction [13]. Previous research should have addressed these limitations. Modern studies have yet to standardize data sets or fine-tuned algorithms. Researchers have created multiple machine learning models, including Support Vector Machines (SVM), K-Nearest Neighbours (KNN), Random Forest (RF), Decision Trees (DT), Logistic Regression (LR), Naive Bayes (NB), and others, to improve the prediction of cardiac diseases [14]. However, the precision of predicting heart disease continues to pose a challenge. The prioritisation of creating a distinctive and economically efficient methodology for precisely predicting the likelihood of cardiovascular illness is of paramount significance. The extent of intricacy surrounding NB, BN, RF, and MLP yet to be clearly defined. The age risk factor, which is not included in NB, BN, RF, and MLP, is denoted as the age element inside the dataset. The examination of the system was conducted using the StatLog datasets. The model utilised in the Cleveland dataset has removed several noteworthy risk factors, such as age, RestECG, ST Depression (Slope), and others [15]. Significance tests were not performed during the standardisation process of the proposed technique. The datasets utilised in this investigation are the StatLog dataset [16] and the Z-Alizadeh Sani dataset. The dataset exhibits a diminished magnitude. The obtained outcome was not subjected to a comparative analysis with other datasets in order to establish standardisation. The Cleveland dataset was employed instead.

2. Related Works

Auscultation was crucial to doctors' diagnosis of irregular heart sounds. Doctors used stethoscopes to diagnose all cardiac diseases. Medical practitioners' auscultation approach for heart issues has limits. Doctors' ability and expertise, gained from rigorous examinations, determine heart sound clarity and classification. Machine learning approaches have been presented as an alternative to manual CVD detection.

To determine the most accurate predictors of cardiovascular disease, Amin et al. [17] conducted a study. Some of the seven classification techniques used include NB, KNN, LR, DT, NN, SVM, and Vote. The 303 records and 76 attributes that make up the Cleveland datasets were obtained from the UCI machine learning library. The 10-fold cross-validation method is used by the authors to train and assess their models. We chose 10-fold cross-validation in place of a data split since it would have resulted in an underestimation of the model's predicted performance because there were fewer samples in the training set. The model can access 90% of the data, though, if you use 10-fold validation. Using the Vote Classifier increased accuracy to 87.4%.

In a study by Anitha et al. [18], learning vector quantization methods were used to forecast heart disease. An accuracy rate of 85.55% was attained by the algorithm in question. The University of California, Irvine's (UCI) machine learning collection provided the datasets used in this work. These datasets contain 76 attributes and 303 records in total. To deal with the issue of missing values, preprocessing processes were used to the data. As a result, only 14 features were included in the sample for the heart disease study, which resulted in a smaller sample size of 302 records. 30% of the dataset is used for model testing, while 70% is used for training.

In a work by Rajdhan et al. [19], they developed a web-based tool for forecasting heart illness using machine learning methods. LR, NB, and SVM are some of the classification techniques used for model training and testing. The Cleveland datasets were split into two sets: a training set containing 75% of the data and a testing set containing 25% using the UCI machine learning repository. Preprocessing the data to remove errors and missing values led to an increase in accuracy of 64.4% when the SVM technique was used. The study's failure to pinpoint the early-stage risk factors in people with heart disease was one of its flaws.

Devansh Shah et al. [20] employed supervised learning techniques, such as decision trees, Naive Bayes, K-NN, and random forests, to analyze a dataset of 303 samples and 76 distinct attributes. The findings of this study indicate that the K-NN algorithm is achieving its maximum level of accuracy. Singh et al. (year) employed machine learning classifiers to develop a predictive model for heart disease [21]. The UCI Cleveland dataset utilizes 14 attributes to train and evaluate various models. The efficiency of the classifiers was as follows: 78% for linear regression, 79% for decision trees, 83% for support vector machines, and 87% for K-NN. The findings indicated that the K-NN algorithm had the highest level of accuracy among the tested methods. In the study by Khan et al. [22], various classification approaches, namely SVM, LR, ANN NN, K-NN, etc., were employed. When comparing the new model to the previous one, it exhibited an accuracy rate of 92.37 percent.

Linda et al. [23] suggested a novel health information system to provide exercise recommendations tailored to individuals with heart disease. The initial findings indicate that medical practitioners require assistance in determining the appropriate level of physical activity to prescribe for patients presenting several risk factors for cardiovascular disease. The provided system is an efficient, time-saving, and evidence-based option for patients. Risk factors for disease can be evaluated using the three-step PB-FARM method presented by Ali et al. [24]. The Z-Alizadeh Sani dataset was also utilized to examine the factors contributing to this disease's spread. The results directly correlate with normal chest discomfort, advanced age, and CAD likelihood. Prediction models for cardiovascular disease were proposed by Rubini et al. [25]. LR, NB, and SVM classifiers contrasted with the suggested technique and achieved the maximum accuracy (84.81%) obtained using random forest in their work.

The main objective of the study undertaken by Drod et al. [26] was to utilize machine learning techniques to identify the most significant risk variables associated with CVD among persons diagnosed with metabolic-associated fatty liver disease (MAFLD). A group of 191 individuals diagnosed with MAFLD underwent a blood chemistry analysis and an evaluation of subclinical atherosclerosis. A machine learning model was developed, incorporating various techniques, such as multiple logistic regression classifier, univariate feature ranking, and PCA, to identify individuals at the greatest risk for CVD. The research found hypercholesterolemia, plaque scores, and duration of diabetes as the most prominent clinical characteristics. The machine learning technique exhibited robust performance, achieving an accuracy rate of 85.11% in correctly classifying 40 out of 47 high-risk patients and 79.17% in reliably identifying 114 out of 144 low-risk patients. This yielded an area under the curve (AUC) value of 0.87. The study's findings suggest that using a machine learning methodology offers benefits in detecting persons with MAFLD who also have existing cardiovascular illness based on patient criteria.

3. Experimental Analysis

Initially, an exposition was provided on the widely recognized machine learning models. Subsequently, the models were trained using the dataset at our disposal. Subsequently, a comparative analysis was conducted to evaluate the performance of each model. By employing this approach, we may gain comprehension and evaluate the efficacy of each model inside our project. Seven algorithms were employed to evaluate machine learning models for predicting cardiovascular disease (CVD) datasets.

Logistic Regression:

Logistic regression is a widely employed statistical model utilized for binary classification. The algorithm calculates the likelihood that a specific occurrence is a member of a specific category. Logistic regression provides a linear boundary that separates the classes by utilizing a logistic function to model the probability. It is simple, fast, and interpretable, making it a good baseline model for binary classification tasks.

Random Forest:

The Random Forest algorithm is a popular ensemble learning technique predominantly employed for classification and regression problems. The operational mechanism involves the construction of several decision trees during the training phase, and the resulting output is determined by selecting the mode of the classes from the individual trees for classification purposes. Random Forests are renowned for their user-friendly nature, adeptness in handling a diverse range of feature types, and capacity to manage missing data while upholding high accuracy effectively.

K-Nearest Neighbors:

The K-Nearest Neighbours algorithm is a straightforward approach to machine learning based on instances and is used for classification and regression tasks. By comparing a given test instance with instances within the training set that are nearest in Euclidean space, KNN aims to determine the class of the test instance. The simplicity and effectiveness of KNN make it suitable for baseline modeling in numerous applications.

Decision Tree:

Classification trees, regression trees, and multi-output trees are all tasks that decision trees can handle. Based on the importance of the input features, they divide the data into groups. The ensuing decision tree is the product of this recursive process. Although decision trees are intuitive in their visual presentation and interpretation, they are vulnerable to overfitting when applied to data with many attributes.

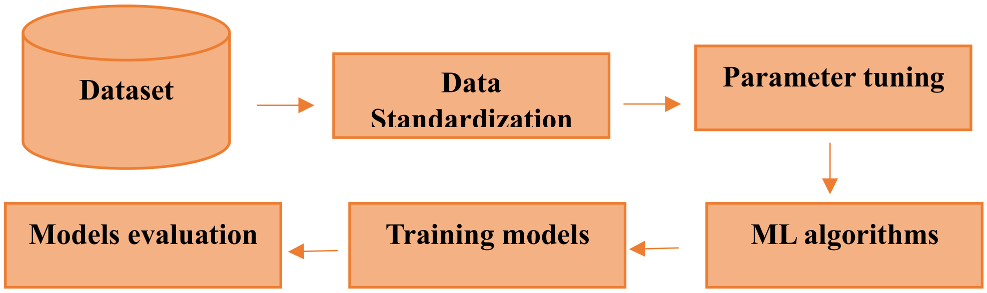

Figure 1: Experimental analysis of numerous models

Support Vector Machine:

Support Vector Machines (SVMs) are highly effective models in supervised learning, specifically designed for binary classification tasks. SVMs function by identifying an optimal hyperplane that effectively partitions the classes within the input feature space. SVMs strive to obtain favorable generalization performance by maximizing the margin between the nearest points of the two classes, referred to as support vectors.

Gradient Boosting (GB):

Gradient Boosting is a widely employed ensemble learning technique utilized for both classification and regression tasks. The construction of robust predictive models is achieved through the amalgamation of predictions made by weak learners that are successively developed, often in decision trees. By correcting the errors of the previous trees, Gradient Boosting improves the model performance iteratively, often resulting in highly accurate models.

AdaBoost:

AdaBoost, short for Adaptive Boosting, is an ensemble learning method that focuses on boosting the performance of simple models, known as weak learners, to form a strong learner. AdaBoost focuses on the misclassified points by re-weighting the instances in the dataset, guiding the model to learn from the mistakes. This iterative correction process often leads to significant improvements in prediction accuracy.

The conventional methodology for the ML pipeline is shown in Figure 1. In this work, we employed a systematic workflow comprising several steps to ensure the robustness and reliability of the predictive models. The process commenced with the acquisition of a well-curated dataset which included numerous features pertinent to cardiovascular health.

Following the data acquisition, we embarked on the data cleaning and exploratory data analysis (EDA) phase, where inconsistencies and missing values within the dataset were addressed. Additionally, we explored the underlying patterns and correlations among different features using various visualization tools, laying the groundwork for the subsequent steps. The standardization step was initiated to ensure a consistent input space for our models. This involved rescaling the features with a mean of 0 and a standard deviation 1. Standardization is crucial for models sensitive to the scale of input features, as it helps speed up the convergence and achieve better performance.

Once the data was standardized, the task of hyperparameter tuning was undertaken. Employing techniques like Grid Search and Random Search, we fine-tuned the configurations of different algorithms to ascertain the optimal set of hyperparameters. This tuning significantly enhanced the performance of the models, ensuring they were well-configured to capture the underlying patterns in the data. With the optimized hyperparameters in place, we applied various machine learning algorithms, including Logistic Regression, Random Forest, K-Nearest Neighbors, Decision Trees, Support Vector Machines, Gradient Boosting, and AdaBoost. Each of these algorithms was trained on a portion of the dataset, setting aside a subset for validation [28], [29].

Post-training, a comprehensive performance evaluation was conducted on each model using the unseen validation data. Metrics such as accuracy, precision, recall, and F1-score were calculated to assess the models' capabilities in predicting cardiovascular diseases. Confusion matrices were generated to provide a clear picture of the true positive, true negative, false positive, and false negative rates. The comparative analysis of the models, based on the computed metrics, revealed Gradient Boosting and AdaBoost as the superior performers with an accuracy of 73%. This rigorous process of evaluation allowed for an in-depth understanding of each model's strengths and weaknesses, thereby guiding the choice of the most suitable model for deployment in predicting cardiovascular diseases [30].

The Logistic Regression (LR) and Random Forest (RF) models both achieved an accuracy of 71%, indicating a fairly good performance in predicting the outcomes. The K-Nearest Neighbors (KNN) model attained an accuracy of 69%, while the Decision Tree (DT) model lagged slightly behind with an accuracy of 63%. On the other hand, the Support Vector Machine (SVM) model demonstrated a slightly better performance with an accuracy of 72%. Noteworthy is the superior performance of the Gradient Boosting (GB) and AdaBoost models, achieving an accuracy of 73%, thus representing the most accurate models for this dataset. This comparative evaluation provides valuable insights into the relative performances of various machine learning models in predicting cardiovascular diseases [31]. Gradient Boosting and AdaBoost models show a slight edge in accuracy over the others. The accuracy of these models is shown in Figure 2.

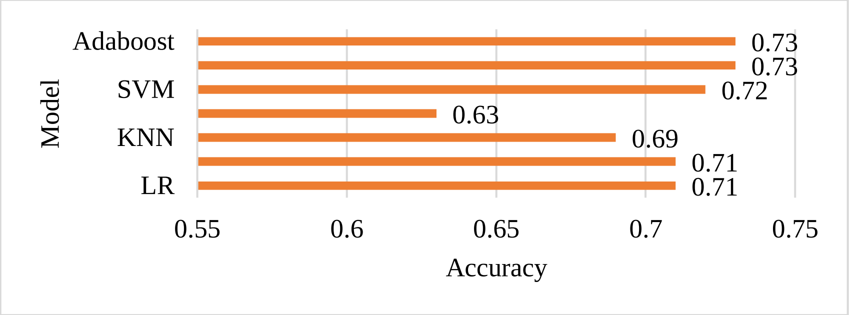

Figure 2: The comparison of accuracy on the CVD dataset using different ML algorithms

4. Proposed Methodology

This section provides a deep neural network pipeline for solving a binary classification problem (i.e., predicting the presence or absence of heart disease). The complete flow of the work is shown in Figure 3. Below are detailed explanations for each step in the pipeline:

Data Collection:

The "CVD dataset" dataset presumably contains various features relevant to cardiovascular diseases. This dataset must be large and diverse enough to provide a broad understanding of the problem.

Data Preprocessing:

Preprocessing involves cleaning and transforming raw data into a format that can be fed into the machine learning algorithms. Common preprocessing steps include handling missing values, encoding categorical variables, scaling numerical variables and sometimes feature engineering to create new informative features or reduce dimensionality.

Train and Test Split:

The dataset is partitioned into separate training and testing sets to assess the subsequent performance of the model. The training process involves utilizing 80% of the available data to train the model, while its performance is evaluated on the remaining 20% of the data.

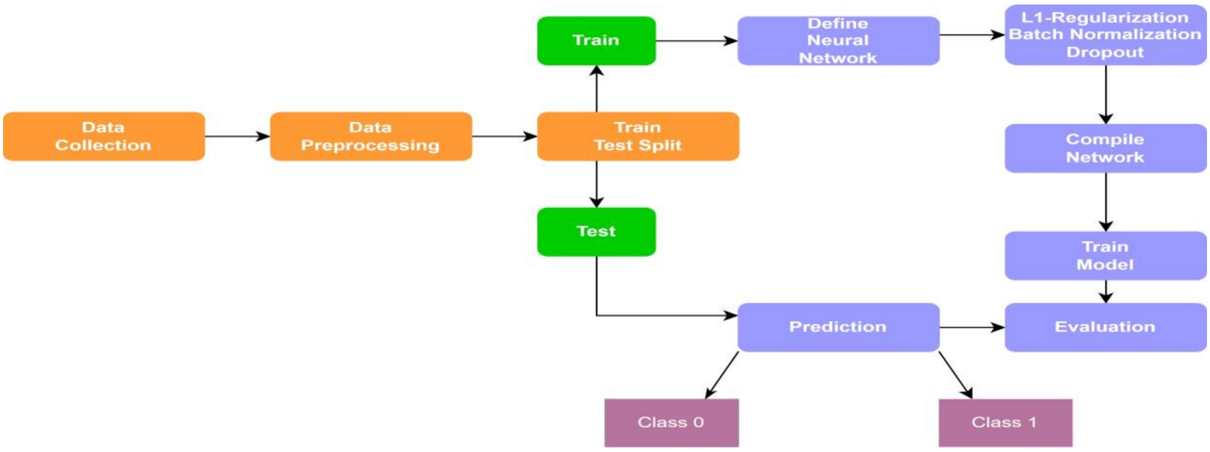

Figure 3: The flow of the proposed method

Defining a Deep Neural Network (Optimized Progressing Dropout Dense Network):

This step involves defining the architecture of the deep neural network. The Optimized Progressing Dropout Dense Network (OPDDN) seems to be a specific type of neural network architecture possibly designed for this problem. It's important to define the architecture properly to capture the underlying patterns in the data.

Incorporating Regularization, Batch Normalization, and Dropouts:

Regularization helps prevent overfitting by adding a penalty on the magnitude of the coefficients. Batch normalization makes the network faster and more stable by normalizing the layers' inputs. Dropout is a regularization method where randomly selected neurons are ignored during training, dropping out a random set of activations in a layer, which forces the network to learn features in a distributed way.

Model Compilation:

Compiling the model involves specifying the loss function, optimizer, and metrics to be monitored. For a binary classification problem like this, a common choice is to use binary cross-entropy as the loss function.

Training the Model:
The model is trained on the training set using the OPDDN. During training, the model learns to map the features to the target variable (heart disease presence) by minimizing the specified loss function.

Model Evaluation on Test Dataset:
The model's performance is evaluated on the unseen test data to gauge how well it generalizes to new, unseen data. Common evaluation metrics for binary classification problems include accuracy, precision, recall, F1 score, and Area Under the ROC Curve (AUC-ROC).

Prediction:
The model is used to make predictions on the test dataset. The binary output indicates the presence (1) or absence (0) of heart disease for each individual in the test dataset.

4.1 Data Source

According to [23], this research used a dataset with 70,000 patient records and 12 different features, and some sample rows (5 records) and columns (6 columns) from the dataset are shown in Table 1 for representation. Age, sex, systolic blood pressure, and diastolic blood pressure are all examples of such factors. Cardiovascular disease is represented by the target class "cardio," while good health is shown by the "cardio-healthy." Figure 4 displays the classwise number of samples. The observation shows that 50.03% of samples belong to class 0, and 49.97% belong to class 1, demonstrating that the dataset contains a balanced distribution of values. The dataset is partitioned into two subsets: a training dataset, which comprises 80% of the total data, and a testing dataset, which constitutes the remaining 20%. The training dataset is utilised to train a model, whereas the testing dataset is employed for evaluating its performance. The efficacy of various classifiers, including the decision tree classifier, random forest classifier, multilayer perceptron, and XGBoost, is evaluated using the clustered dataset. The efficacy of each classifier is afterwards evaluated based on four measures, namely accuracy, precision, recall, and F-measure.

Table 1. Sample records from the CVD dataset (source:kaggle [23])

active	height	smoke	ap_lo	cholesterol	gender	alco
1	156	0	90	2	1	0
0	161	0	90	1	2	0
1	151	0	80	2	1	0
1	165	0	70	1	1	0
1	169	1	80	1	2	0

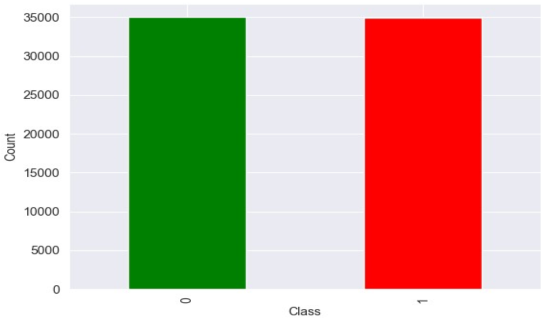

Figure 4: Class Classwise count in the dataset (no-disease:0, disease:1)

5. Results and Discussion

The technique begins with the execution of a Sequential model. This model represents a linear stack of layers, making it appropriate for a simple stack of layers with exactly one input tensor and one output tensor. The first layer is the dense or fully linked input layer. To allow each feature to be processed by a single neuron, the number of neurons in this layer is configured to equal the number of features in the dataset. The Rectified Linear Unit (ReLU) is the activation function employed in this layer. ReLU is popular because it introduces non-linearity into the model without changing the network's receptive field, allowing the model to learn from the error back-propagated through the network.

Following the input layer, numerous hidden layers with variable numbers of neurons are defined. These layers, too, are dense and employ the ReLU activation function. The selection of different numbers of neurons and the ReLU activation function could be driven by empirical findings or the problem and the data being dealt with. Dropout layers with varied dropout rates are added between these dense layers. Dropout is a regularisation technique intended to prevent network overfitting. When a network learns to perform very well on training data but poorly on unseen or new data, this is called overfitting. Dropout layers randomly change a fraction of the inputs to zero during training, preventing any neuron from becoming highly specialized depending on the training data.

The output layer is the final layer in the architecture, and it is another dense layer with two neurons. This layer employs the softmax activation function, widely utilized for multi-class classification tasks since it provides a probability distribution over the classes, implying that the sum of the output values from the neurons in this layer is 1. Each neuron in this layer represents a class; its output is the likelihood that the incoming data belongs to that class. The model is compiled after the architecture has been defined. By establishing the optimizer, loss function, and metrics, this stage prepares the model for training. Because of its efficiency and minimal memory demand, the Adam optimizer is frequently a viable choice for problems of this type. Categorical Cross entropy is the loss function appropriate for multi-class classification applications because it gauges model performance by comparing predicted probability to genuine class labels. The metric system Categorical Accuracy is used to assess the model's performance throughout the training and validation phases.

The model is then trained on a dataset for a fixed number of epochs, each representing one complete run through the training dataset. The model's weights are modified throughout training to minimize the loss function. At the end of each epoch, validation data is delivered to the model to evaluate its performance on unseen data, which aids in understanding how well the model is likely to perform on unseen data. The training procedure is carried out with a set batch size, which determines the amount of samples used to update the model's weights. Compared to full-batch and stochastic gradient descent, this mini-batch gradient descent balances computational efficiency and convergence speed. Finally, the training method generates a history object with records of training and validation loss and accuracy for each epoch. This data can be used to analyze how the model's performance varies over time and to determine whether the model is overfitting or underfitting the training data. Figure 5 illustrates a comparative analysis of various evaluation metrics over 15 epochs utilizing the PDDN model, shedding light on the model's performance and consistency across different training phases.

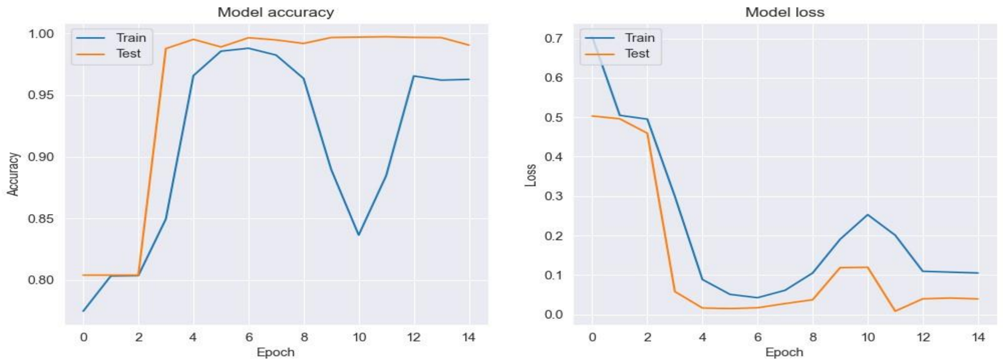

Figure 5: Comparison of evaluation metrics for 15 epochs using PDDN

Table 2: Evaluation measures on two different classes

	Precision	Recall	F1 score
0	0.95	1.00	0.98
1	1.00	0.99	0.99

Table 2 presents the evaluation metrics, such as the trained model's precision, recall, and f1-score. These values indicate that the classification model performed very well on this task. A precision and recall near or at 1.00 indicate that the model has high accuracy and completeness in identifying the positive instances of each class. The F-Score, being a harmonic mean of precision and recall, also reflects a high level of performance by having values close to 1.00. The confusion matrix of this method is shown in Figure 6.

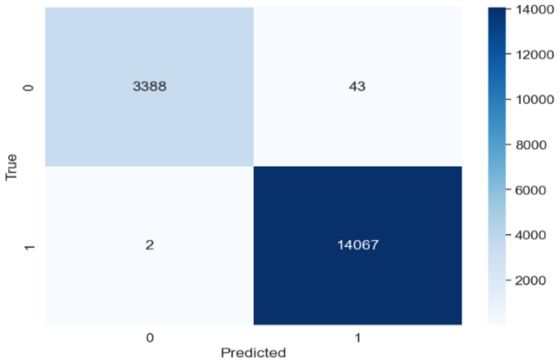

Figure 6: Confusion matrix on the test dataset (PDDN)

The ROC curve, as illustrated in Figure 7, offers valuable insights into the performance of the model by evaluating its sensitivity and specificity across various threshold values.

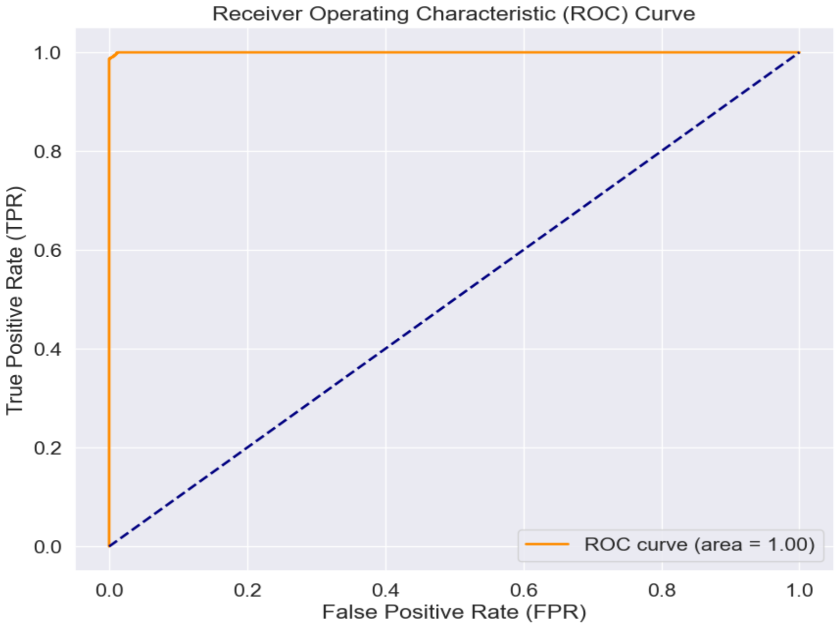

Figure 7: ROC curve using PDDN

We further improved our neural network design to enhance its performance. Several important improvements were made, such as the integration of L2-regularization to address the issue of overfitting, the implementation of batch normalization to stabilize the learning process, and the utilization of learning rate decay to facilitate a more accurate convergence during the training phase. The alterations above had a crucial role in enhancing the network's ability to accurately forecast outcomes and improve its overall effectiveness in learning.

Power-of-two units (like 32 or 64) were chosen on purpose for the neurons in each layer of the refined architecture. This common practice has been seen to improve performance in some cases, possibly because it works better with low-level hardware optimizations. To avoid the risk of overfitting, L2 regularisation was created. This penalizes big weight values, encouraging the model to learn a simpler, more general way to represent the data. Also, after the dense layers, batch normalization was used to deal with the internal covariate shift problem. This is when the distribution of network activations changes as the network learns. Batch normalization helps to speed up and make the training process more stable by adjusting the results of each layer to a standard distribution. Also, a learning rate decline strategy was added to the Adam optimizer. This strategy slows down the rate of learning over time. This method ensures the model gets closer to a local minimum of the loss function towards the end of training, making the convergence more precise. Each of these changes was made to make the network's way of learning stronger and more effective. This improved the network's ability to predict cardiovascular disease and its general performance.

Figure 8 displays a comparative study of the training and testing accuracies and the corresponding losses during the learning epochs. The presented visualization provides a clear representation of the learning trajectory of the model, highlighting the dynamic relationship between the advancement of training and the model's ability to generalize to new, unseen data. The trends illustrated in the data offer valuable information regarding the model's efficacy in the learning process and its susceptibility to overfitting or underfitting. The entire categorization report is presented in Table 3, which thoroughly assesses the model's performance across many classes. The table provides a comprehensive overview of important metrics for each class, including precision, recall, and F1-score. Additionally, it includes an aggregate performance score, which contributes to a more nuanced evaluation of the model's classification capabilities.

Figure 9 presents the confusion matrix, concisely representing the model's rates of accurate and inaccurate classifications. The matrix provides a visual representation of the counts for true positive, true negative, false positive, and false negative, hence serving as a valuable tool for assessing the model's ability to distinguish between classes and identifying areas where enhancements may be made. By utilizing graphical and tabular representations, we aim to assess the model's performance comprehensively, highlighting its strengths and identifying areas that can be further improved.

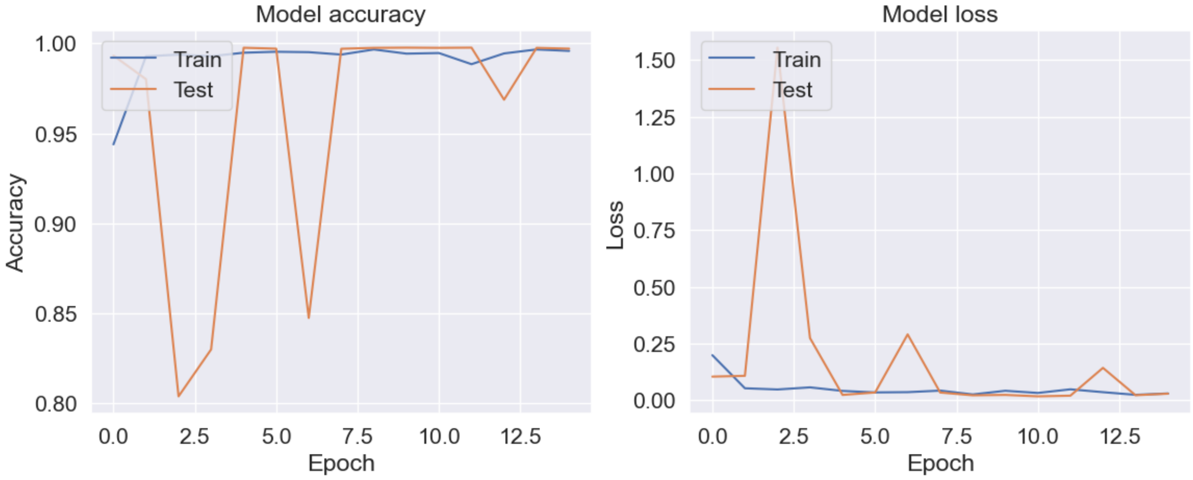

Figure 8: Comparison of evaluation metrics for 15 epochs using OPDDN

Table 2: Evaluation measures on two different classes

	Precision	Recall	F1 score
0	1.00	0.99	0.99
1	1.00	1.00	1.00

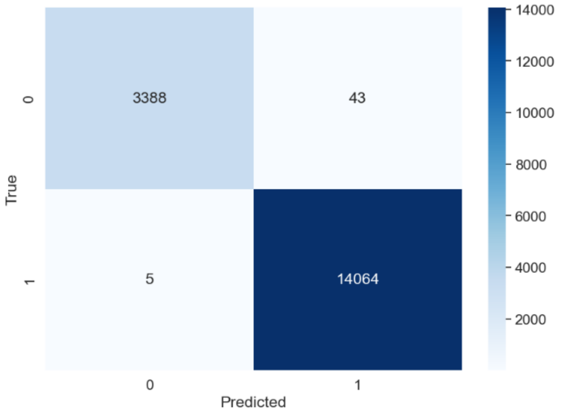

Figure 9: Confusion matrix on the test dataset (PDDN)

6. Conclusion

The major goal of this study was to apply the PDDN technique to cardiac disease categorization on a large, real-world dataset. Before analysis, the dataset was subjected to several successful data transformations and scalings. The PDDN model was the most accurate, with a 99% accuracy rate. These findings indicate that OPDDN clustering has promise. Specific strategies for identifying and treating sickness are refined to accurately predict cardiac abnormalities, highlighting the algorithm's potential utility. The study used the 70,000-case Kaggle cardiovascular disease dataset, and all algorithms are currently being experimented with in Google Colab. The conclusions, however, are tempered by significant cautions. Keep this in mind. The study's conclusions may only apply to some populations because it depended on a single dataset. Patients are one example of a heterogeneous population. Furthermore, the study only examined a subset of the population. However, additional risk variables should have been addressed besides demographic and clinical criteria. Poor habits and a family history of heart disease are also risk factors for the disease.

6.1 Future Directions

The current study had some gaps, and we've outlined future work to address these issues and expand on the initial findings.

- Examine the comparative efficacy of the PDDN approach about other frequently employed machine learning methodologies to determine superiority or ascertain the potential for amalgamation to enhance outcomes.
- The use of machine learning algorithms is widespread, and a crucial aspect of this practice involves comprehending the impact of factors such as missing data or atypical data points on the model's correctness. Therefore, it is advisable to go into this particular component and devise strategies to effectively address these challenges to ensure the model yields dependable outcomes.
- Ensuring the robustness and generalizability of the study's findings necessitates implementing a rigorous validation process, which involves verifying their consistency over time and across various contexts. This approach can contribute to establishing robust findings that can be generalized to a wide range of contexts outside the specific scenario under investigation.
- Through the examination of the clusters generated by the PDDN technique, a deeper comprehension of the implications and practical applications of the findings may be attained. This enhanced understanding may contribute to advancing knowledge and strategies for addressing heart disease, utilizing the insights derived from the present investigation.

References

[1] Cardiovascular Diseases (Cvds), “World health organization,” https://www.who.int/news-room/fact%20sheets/detail/cardio vascular-diseases-(cvds).

[2] A. K. Dwivedi, S. A. Imtiaz, and E. R. Villegas, “Algorithms for automatic analysis and classification of heart sounds - a systematic review,” IEEE Access, vol. 7, 2019.

[3] A Coronary, “Heart disease,” Available from: https://www. aihw.gov.au/reports/australias-health/coronaryheart-disease, 2020

[4] Bakar, W. A. W. A., Josdi, N. L. N. B., Man, M. B., & Zuhairi, M. A. B. (2023, March). A Review: Heart Disease Prediction in Machine Learning & Deep Learning. In *2023 19th IEEE International Colloquium on Signal Processing & Its Applications (CSPA)* (pp. 150-155). IEEE.

[5] Bhatt, C. M., Patel, P., Ghetia, T., & Mazzeo, P. L. (2023). Effective heart disease prediction using machine learning techniques. *Algorithms*, *16*(2), 88.

[6] Dileep, P., Rao, K. N., Bodapati, P., Gokuruboyina, S., Peddi, R., Grover, A., & Sheetal, A. (2023). An automatic heart disease prediction using cluster-based bi-directional LSTM (C-BiLSTM) algorithm. *Neural Computing and Applications*, *35*(10), 7253-7266.

[7] Nandy, S., Adhikari, M., Balasubramanian, V., Menon, V. G., Li, X., & Zakarya, M. (2023). An intelligent heart disease prediction system based on swarm-artificial neural network. *Neural Computing and Applications*, *35*(20), 14723-14737.

[8] Kumar, D. V. S., Chaurasia, R., Misra, A., Misra, P. K., & Khang, A. (2023). Heart disease and liver disease prediction using machine learning. In *Data-Centric AI Solutions and Emerging Technologies in the Healthcare Ecosystem* (pp. 205-214). CRC Press.

[9] Hassan, D., Hussein, H. I., & Hassan, M. M. (2023). Heart disease prediction based on pre-trained deep neural networks combined with principal component analysis. *Biomedical Signal Processing and Control*, *79*, 104019.

[10] Ozcan, M., & Peker, S. (2023). A classification and regression tree algorithm for heart disease modeling and prediction. *Healthcare Analytics*, *3*, 100130.

[11] Nouman, A., & Muneer, S. (2022). A systematic literature review on heart disease prediction using blockchain and machine learning techniques. *International Journal of Computational and Innovative Sciences*, *1*(4), 1-6.

[12] Sekar, J., Aruchamy, P., Sulaima Lebbe Abdul, H., Mohammed, A. S., & Khamuruddeen, S. (2022). An efficient clinical support system for heart disease prediction using TANFIS classifier. *Computational Intelligence*, *38*(2), 610-640.

[13] Reddy, D. J., & Kumar, M. R. (2021, May). Crop yield prediction using machine learning algorithm. In *2021 5th International Conference on Intelligent Computing and Control Systems (ICICCS)* (pp. 1466-1470). IEEE.

[14] Kumar, M. R., & Gunjan, V. K. (2020). Review of machine learning models for credit scoring analysis. *Ingeniería Solidaria*, *16*(1).

[15] Swetha, A., Lakshmi, M. S., & Kumar, M. R. (2022). Chronic Kidney Disease Diagnostic Approaches using Efficient Artificial Intelligence methods. *International Journal of Intelligent Systems and Applications in Engineering*, *10*(1s), 254-261.

[16] Abdar, M., Książek, W., Acharya, U. R., Tan, R. S., Makarenkov, V., & Pławiak, P. (2019). A new machine learning technique for an accurate diagnosis of coronary artery disease. *Computer methods and programs in biomedicine*, *179*, 104992.

[17] Amin, M. S., Chiam, Y. K., & Varathan, K. D. (2019). Identification of significant features and data mining techniques in predicting heart disease. *Telematics and Informatics*, *36*, 82-93.

[18] Anitha, S., & Sridevi, N. (2019). Heart disease prediction using data mining techniques. *Journal of analysis and Computation*.

[19] Rajdhan, A., Agarwal, A., Sai, M., Ravi, D., & Ghuli, P. (2020). Heart disease prediction using machine learning. *INTERNATIONAL JOURNAL OF ENGINEERINGRESEARCH & TECHNOLOGY (IJERT)*, *9*(O4).

[20] D. Shah, S. Patel, and S. Kumar Bharti, Heart Disease Prediction Using Machine Learning Techniques, Springer Nature Singapore Pte Ltd, Berlin, Germany, 2020

[21] Singh, A., & Kumar, R. (2020, February). Heart disease prediction using machine learning algorithms. In *2020 international conference on electrical and electronics engineering (ICE3)* (pp. 452-457). IEEE.

[22] Li, J. P., Haq, A. U., Din, S. U., Khan, J., Khan, A., & Saboor, A. (2020). Heart disease identification method using machine learning classification in e-healthcare. *IEEE access*, *8*, 107562-107582.

[23] Pescatello, L. S., Wu, Y., Panza, G. A., Zaleski, A., & Guidry, M. (2021). Development of a novel clinical decision support system for exercise prescription among patients with multiple cardiovascular disease risk factors. *Mayo Clinic Proceedings: Innovations, Quality & Outcomes*, *5*(1), 193-203.

[24] Yavari, A., Rajabzadeh, A., & Abdali-Mohammadi, F. (2021). Profile-based assessment of diseases affective factors using fuzzy association rule mining approach: A case study in heart diseases. *Journal of Biomedical Informatics*, *116*, 103695.

[25] Rubini, P. E., Subasini, C. A., Katharine, A. V., Kumaresan, V., Kumar, S. G., & Nithya, T. M. (2021). A cardiovascular disease prediction using machine learning algorithms. *Annals of the Romanian Society for Cell Biology*, 904-912.

[26] Drożdż, K., Nabrdalik, K., Kwiendacz, H., Hendel, M., Olejarz, A., Tomasik, A., ... & Lip, G. Y. (2022). Risk factors for cardiovascular disease in patients with metabolic-associated fatty liver disease: a machine learning approach. *Cardiovascular Diabetology*, *21*(1), 240.

[27] Kaggle Cardiovascular Disease Dataset. Available online: https://www.kaggle.com/datasets/sulianova/cardiovascular-diseasedataset (accessed on 1 November 2022)

[28] Kumar, M. R., & Gunjan, V. K. (2020). Review of machine learning models for credit scoring analysis. *Ingeniería Solidaria*, *16*(1).

[29] Ramana, K., Kumar, M. R., Sreenivasulu, K., Gadekallu, T. R., Bhatia, S., Agarwal, P., & Idrees, S. M. (2022). Early prediction of lung cancers using deep saliency capsule and pre-trained deep learning frameworks. *Frontiers in oncology*, *12*, 886739.

[30] Reddy, K. U. K., Shabbiha, S., & Kumar, M. R. (2020). Design of high security smart health care monitoring system using IoT. *Int. J*, *8*.

[31] Swetha, A., Lakshmi, M. S., & Kumar, M. R. (2022). Chronic Kidney Disease Diagnostic Approaches using Efficient Artificial Intelligence methods. *International Journal of Intelligent Systems and Applications in Engineering*, *10*(1s), 254-261.