IEC
Institute of Electronics and Computer

Journal of Artificial Intelligence and Systems, 2024, 6, 11-33
https://iecscience.org/journals/AIS
ISSN Online: 2642-2859

Coronary Heart Disease Prediction: A Novel Deep Learning-based Approach

Solomon Kutiame[1, *], Richard Millham[2], Adebayor Felix Adekoya[1], Benjamin Asubam Weyori[3], Mark M. Tettey[4]

[1] Department of Computer Science & Informatics, University of Energy and Natural Resources, Ghana
[2] Department of Information Technology, Durban University of Technology, South Africa
[3] Department of Computer and Electrical Engineering, University of Energy and Natural Resources, Ghana
[4] National Cardiothoracic Center, Korle Bu Teaching Hospital, Accra – Ghana
*Corresponding Author: Solomon Kutiame, Email: solomon.kutiame.stu@uenr.edu.gh

How to cite this paper: Solomon Kutiame, Richard Millham, Adebayor Felix Adekoya, Benjamin Asubam Weyori, Mark M. Tettey (2024). Coronary Heart Disease Prediction: A Novel Deep Learning-based Approach. Journal of Artificial Intelligence and Systems, 6, 11–33. https://doi.org/10.33969/AIS.2024060102.

Received: December 1, 2023
Accepted: February 29, 2024
Published: March 5, 2024

Abstract

Coronary Heart Disease (CHD) is reported to be one of the world's deadliest diseases. Early prediction or detection and diagnosis of the disease can help prevent, cure, and reduce the damage it could cause. Artificial intelligent techniques such as Machine Learning and Deep Learning have proven useful for the early detection and prediction of disease. However, issues of irrelevant and redundant features in open datasets have contributed to low classification accuracy rates and high misclassification rates. This leaves a gap for continuous approaches for smart feature selection and high-performing models in the field for better results. This study investigates the impacts of cardiologist-inspired datasets and PCA dimensionality reduction techniques on the performance of CHD prediction. The compared results show that while PCA improves the CHD prediction accuracy for datasets obtained from cardiologists, there exist no statistically significant differences in the efficacy of the PCA method for CHD classification when applied to open datasets, however, MLP and LSTM offer promising results. The results further indicated that the effectiveness of expert-based feature selection techniques on CHD classification is relatively stable when compared with open-source datasets.

Keywords

Coronary Heart Disease prediction, Dimensionality Reduction, Principal Component Analysis, Expert-based Features, Deep-Learning, Performance

1. Introduction

Coronary heart disease (CHD) is a leading cause of death worldwide, and its early detection and prediction are crucial for the effective treatment and management of the disease [1]. The World Health Organization (2023) reports that in 2016, cardiovascular diseases, including CHD, accounted for 31% of all deaths globally, with an estimated 17.9 million deaths. CHD is a primary cause of loss of lives globally and is more prevalent than other leading causes of death such as cancer, respiratory diseases, lower respiratory infections, stroke, diabetes, kidney disease, and suicide. Cardiovascular diseases remain a key health concern and addressing CHD is critical in reducing the global burden of death and disease. CHD is characterized by the build-up of plaque in the coronary arteries, which can lead to reduced blood flow to the heart and eventually cause a heart attack [3]. The prediction of CHD is challenging due to the complex nature of the disease, as well as the presence of multiple risk factors, such as age, gender, genetics, lifestyle, and comorbidities [4]. Traditionally, CHD risk prediction has been based on classical statistical methods, such as logistic regression and decision trees. However, these methods have limitations when dealing with large and complex data, and they often rely on the assumption of linear relationships between the predictor variables and the outcome. Machine learning (ML) approaches have been advocated in recent years as a solution to overcome these constraints and increase the accuracy of CHD risk prediction [5]. Deep learning is a type of ML that has produced outstanding findings in various fields, such as image and speech recognition, natural language processing, and bioinformatics. Deep learning models, such as Convolutional Neural Networks (CNNs) and Long-Short Term Memory (LSTMs), stand predominantly effective in handling composite and huge data. These models are capable of automatically extracting relevant features from the data, and they can handle non-sequential relationships between the predictor and outcome variables [6].

The problem of high dimensionality in machine learning occurs when datasets contain vast amounts of data that cannot be easily visualized, a phenomenon known as the curse of dimensionality. This can lead to high memory requirements and potential overfitting when processing the data. To address this issue, weighting features can be used to reduce redundancy and processing time. Various feature engineering and selection techniques can also be applied to decrease the dimensionality of the dataset by removing unimportant data. The questions guiding the direction of the current study are as follows: RQ1: How do expert-based features perform when compared to transformed features and existing methods? And RQ2: How do Cardiologists' CHD data compare to the open-source data? Prior studies in disease classification [5], have demonstrated that disease prediction models often exhibit low accuracy and high

misclassification rates in classification due to the large number of features used in the training and classification stages. Similarly, according to Rahim et al. (2021) the success of a prediction model is strongly dependent on how effectively the features are generated and picked for the models. Therefore, one must be aware of the redundant and irrelevant features that result in a higher dimension in datasets (i.e., "curse of dimensionality"). To overcome the challenge, extant research [3], [8] have used different feature-selection strategies prior to training ML models to increase model performance. This study offers the following additions to the CHD and disease prediction literature: (1) Expert-based features for CHD predictions are provided; (2) We provide an improved prediction model features based on selected smart features and primary dataset; (3) We show the effect of dimensionality reduction in the datasets used for CHD prediction models.

This study uses deep learning models for CHD prediction. We compare the performance of different deep learning models, including CNN, MLP, and LSTMs, with statistically dimensionality-improved methods, such as Principal Component Analysis (PCA). The study will be based on a diverse dataset that includes various demographic, lifestyle, and clinical variables collected from cardiologists, as well as laboratory test results and open-source data. We utilize a clinical dataset of which the features are carefully selected based on advice received from cardiologists and compare the results with open datasets. For the open dataset that has redundant and irrelevant features as described in the literature, we use the dimensionality reduction method of feature selection to transform the data for appropriate features for better prediction. The results of this study will provide valuable insights into the potential of deep learning models for CHD risk prediction, and they will help to inform clinical decision-making and improve the early detection and management of CHD. The study will also contribute to the development of more accurate and efficient CHD risk prediction tools, which will ultimately lead to better health outcomes for patients. In the next section, we provide a comprehensive overview of the existing literature on CHD risk prediction. We then describe the methods and materials used in this study, including the data sources, the pre-processing steps, the deep learning models used, and the evaluation metrics. Finally, we will present the results, discussions and the consequences for clinical practice and impending studies.

2. Related Work

In this section, we survey the state-of-the-art in heart disease prediction, emphasizing the significance of this problem. Accurate prediction of heart disease risk is vital, not only for the medical profession but also for individuals. Early detection of risk factors through prediction can raise awareness and promote preventative measures,

ultimately leading to better health outcomes. We, therefore, review the various techniques and models that researchers have employed to predict heart disease and present a comprehensive impression of the present status of the field. Katarya and Kumar (2020) conducted a comparative study and analysis of machine learning techniques for predicting heart diseases. The authors evaluated algorithms such as Naïve Bayes, Neural Networks, and Decision Trees and found that the model performance can be impacted by the number of features used. The researchers also developed a prototype system which leveraged data mining methods to make predictions. The prototype considered various heart disease factors and used classification matrices to assess the accuracy of the predictions. The model has the potential to provide cost-effective training and learning opportunities for medical students. Another study used Artificial Neural Network (ANN) as a classification algorithm. The authors utilized Principal Component Analysis (PCA) and chi-square as feature subset selection methods for classifying heart disease. The findings indicated a 96.2% accuracy rate when ANN with PCA and chi-square were utilized, compared to an accuracy of 87.3% using J48, and 77.4% using Naive Bayes. Using a Convolutional Neural Network (CNN) technique, Dutta et al. (2019) evaluated the efficiency of the CNN model compared to traditional machine learning algorithms and found that it outperformed other methods in terms of accuracy and computation time. The proposed CNN model effectively predicted CHD and could potentially be used for early detection and diagnosis. A most recent review by Kutiame et al. (2022) focused on the application of ML algorithms in the prediction of CHD. Deep neural networks were poorly represented in the literature. The best-performing algorithms identified in the study were Deep Neural Network (DNN), Logistic Regression (LR), K-Nearest Neighbors (KNN), Support Vector Machine (SVM), boosting algorithms, and K-Means. It is worth noting that most of the empirical articles reviewed did not mention the use of dimensionality reduction techniques which is a promising aspect of machine learning.

Although CHD prediction using ML systems is studied extensively, several challenges persist unresolved. The use of the expert-based feature and clinical dataset is limited [5], [11], [12]. Such data should be employed in future studies to strengthen the approach to predicting real-world CHDs. Data-based features for CHD prediction have limitations such as missing data, lack of relevant information, and varying quality of data in EHRs. These limitations can affect the accuracy of CHD predictions. Cardiologists' assessments are a valuable source of information for predicting CHD due to their expertise and personalized approach. Cardiologists have extensive training in diagnosing and treating heart disease and take into account the patient's complete medical history, overall health, and specific symptoms when making a

prediction. The use of dimensionality reduction techniques on CHD datasets is an area that has received relatively limited attention in the literature. Dimensionality reduction techniques, such as principal component analysis (PCA) and linear discriminant analysis (LDA), can be used to reduce the complexity of CHD datasets by transforming high-dimensional data into a lower-dimensional representation. This can have several benefits for CHD prediction, including improving the interpretability and visualization of the data, reducing noise and outliers, and improving the performance of predictive models. The gap in the field is that there is limited research [13], [14] that has specifically investigated the use of dimensionality reduction techniques on CHD datasets. There is a need for more systematic evaluations of the potential benefits of using these techniques on CHD datasets, and a need to determine the optimal methods for using these techniques to improve CHD prediction accuracy. Furthermore, there is a need to consider the potential limitations and drawbacks of using these techniques, such as the loss of information, and to determine the trade-offs between the benefits and limitations.

3. Materials and Methods

3.1. Data Description

As mentioned earlier, this study employs datasets from primary (cardiologists) and secondary sources to draw comparative conclusions between experts and academia. The primary dataset consisted of Clinical information on coronary heart disease patients obtained from the c located in Accra, Ghana. The dataset contained 9090 samples of which 8171 are CHD patients and 919 were without the disease. The obtained information includes the following attributes; blood pressure, chest pain type, age, fasting blood sugar, sex, maximum heart rate achieved, age, serum cholesterol in mg/dl, smoking status, and resting blood pressure. According to the experts, the clinical variables were assessed as predictors of CHD. The dataset given had a target column that categorized into two classes: 1 meant the presence of heart disease and 0 indicated the absence of heart disease. The important risk factors that were analyzed in this dataset were displayed in Table 1, which included various risk factors and their respective values along with their encoded values enclosed in brackets. These encoded values were used as input to the proposed framework, and the risk factors were determined based on expert opinions. Some of these risk factors were also commonly known by the general public.

The secondary dataset used for our investigation is the “Heart Disease Health Indicators Dataset (HDI)” which was obtained from Kaggle (a machine learning dataset repository). The dataset is a cleaned version made available from the Behavioural Risk Factor Surveillance System 2015 (BRFSS 2015) dataset. The

Behavioral Risk Factor Surveillance System (BRFSS) is a telephone survey system in the United States that collects data on health-related risk behaviors, chronic health conditions, and use of preventive services among U.S. residents. It was established in 1984 with 15 states and now collects data in all 50 states, the District of Columbia, and three U.S. territories. With over 400,000 adult interviews conducted annually, it is the largest continuously conducted health survey system globally. The extensive data covers various factors, including Age, Environment and Occupation, Family History and Genetics, Lifestyle Habits, Other Medical Conditions, Race or Ethnicity, and Sex, providing clinicians with information to diagnose coronary heart disease. The version of BRFSS adopted for this study includes CHD predictive variables, which aids in diagnosing suspected patients before carrying out bloodwork for detection. The study's sample size comprises 22 attributes and 253,680 instances.

Table 1 Datasets and corresponding encodings for the primary and secondary dataset

	Clinical Dataset		**Open Dataset**	
SNo.	**Risk Factors**	**Values**	**Risk Factors**	**Values**
1	Sex	Male (1), Female (0)	Sex	Male (1), Female (0)
2	Age (years)	20-34 (−2), 35–50 (−1), 51–60 (0), 61-79 (1), >79 (2)	Age	20-34 (−2), 35–50 (−1), 51–60 (0), 61-79 (1), >79 (2)
3	Blood Cholesterol	Below 200 mg/dL - Low (−1) 200–239 mg/dL - Normal (0) 240 mg/dL and above - High (1)	CholCheck	"Below 200 mg/dL - Low (−1) 200–239 mg/dL - Normal (0) 240 mg/dL and above - High (1)"
4	Blood Pressure	"Below 120 mm Hg- Low (−1) 120–139 mm Hg- Normal (0) Above 139 mm Hg- High (1)"	HighBP	"Below 120 mm Hg- Low (−1) 120–139 mm Hg- Normal (0) Above 139 mm Hg- High (1)"
5	Hereditary	Family Member diagnosed with HD - Yes (1) Otherwise –No (0)	Income	Family Member diagnosed with HD -Yes (1) Otherwise –No (0)
6	Smoking	Yes (1) or No (0)	Smoker	Yes (1) or No (0)
7	Alcohol Intake	Yes (1) or No (0)	HvyAlcoholConsump	Yes (1) or No (0)
8	Physical Activity	Low (−1), Normal (0) or High (1)	PhysActivity	Low (−1), Normal (0) or High (1)
9	Diabetes	Yes (1) or No (0)	Diabetes	Yes (1) or No (0)
10	Diet	Poor (−1), Normal (0) or Good (1)	Fruits	Poor (−1), Normal (0) or Good (1)
11	Obesity	Yes (1) or No (0)	High BMI	Yes (1) or No (0)
12	Stress	Yes (1) or No (0)	HighChol	Yes (1) or No (0)
13			AnyHealthcare	Yes (1) or No (0)
14			NoDocbcCost	Yes (1) or No (0)
15			GenHlth	Yes (1) or No (0)
16			MentHlth	Yes (1) or No (0)

	Clinical Dataset		Open Dataset	
SNo.	Risk Factors	Values	Risk Factors	Values
17			PhysHlth	Yes (1) or No (0)
18			DiffWalk	Yes (1) or No (0)
19			Veggies	Yes (1) or No (0)
20			Stroke	Yes (1) or No (0)
21			Education	Yes (1) or No (0)
Output	Heart Disease	Yes (1) or No (0)	Output (heart disease)	Yes (1) or No (0)

3.2. Data Pre-processing

Initially, the medical data is preprocessed in two ways, namely missing value removal and data normalisation. The data value, most of which were continuous variables (e.g., age, income, BP, etc.,) were further transformed for the classification task. The tabular data assisted in the efficient detection of patterns related to heart diseases. The research paper introduced a successful approach for recognizing and classifying CHD in clinical and open datasets. The technique involves several preprocessing steps, including eliminating missing values and normalizing the clinical dataset. The next step is to reduce the dataset's dimensionality by applying the PCA method to select the best features. To achieve better improve classification accuracy and reduce processing complexity, we utilized the ranker method technique in conjunction with PCA to produce the most significant set of features. In the end, four processed datasets (datasets transformed with PCA) were used for this study. Data used for the analysis consisted of Original CHD data collected from the Korle Bu Teaching Hospital based on features provided by cardiologists; a transformed version of the Korle Bu data using PCA; an HDI dataset from Kaggle; and a transformed version of the HDI dataset from Kaggle.

The main focus of this study was to investigate the impact smart feature selection techniques have on the performance of CHD prediction. Dimensionality reduction is chosen for this objective. Dimension reduction transforms the data from a high-dimensional state to a low one without compromising the originality of the original dataset [15]. Dimensionality reduction is used for boosting prediction performance, and reducing computational resource requirements, and time. In this study, we chose the principal component analysis method of dimensionality reduction [16].

3.3. Principal Component Analysis

The principal component analysis is a popular dimensionality reduction technique. Using linear procedures, PCA decreases the scale of raw data by putting it into a

smaller space [17]. To evaluate a data set, PCA requires multiple processes, including computing the linear combination, determining the eigenvectors and eigenvalues from the covariance matrix, and selecting eigenvectors according to the size of their associated eigenvalues [16]. Once the eigenvectors have been selected and ranked based on their eigenvalues, the top eigenvectors are considered the principal components of the dataset. The initial dimensionality of the dataset limits the number of principle components that may be produced by PCA, which means that the maximum number of principal components cannot be greater than the amount of variables or features in the dataset. Ideally, increasing the number of features should improve or lead to better predictive performance [18]. However, in practice, a large number of features affect the performance of machine learning algorithms [15]. PCA is used to increase the performance of the models by getting rid of correlated variables that do not affect the models [3].

4. Prediction Technique

CHD prediction is a classification problem. The objective of our predictive model is to classify patients according to whether they may have CHD or not. In this study, three deep learning algorithms namely; Convolutional Neural Network (CNN), Long-Short-Term (LSTM), and Multilayer Perceptron (MLP) are employed for this task. These algorithms are further discussed in the next subsection.

5. Deep Learning Algorithms Applied

5.1. Multilayer Perceptron (MLP)

Multilayer Perceptron (MLP) is considered a basic form of deep learning. Deep learning refers to a subfield of machine learning that utilizes deep neural networks, typically with many hidden layers, to model complex patterns in data. A Multilayer Perceptron (MLP) is a type of feedforward artificial neural network. It is comprised of one or more hidden layers of neurons connected to an input layer and an output layer, where the data flows in only one direction, from input to output, without looping back. The MLP's hyperparameters, namely the number of hidden layers and hidden neurons, play a crucial role in the model's performance and must be selected with care [19]. Cross-validation techniques are commonly employed to determine the optimal values for these hyperparameters. Additionally, activation functions (f) are used for the hidden and output neurons in MLP networks. For a binary classification problem, the sigmoid activation function is often used, as it maps the output to a probability between 0 and 1, representing the confidence of the positive class. In the hidden layers, ReLU is often a preferred activation function due to its simplicity, efficiency, and ease of implementation [20]. MLPs can be used for various tasks such

as classification, regression, and clustering. They are trained using a supervised learning approach and their weights are adjusted during the training process to minimize the prediction error. The basic building block of an MLP is the artificial neuron or perceptron, which receives inputs, weights them, and applies an activation function to produce an output. The formula for a single neuron is given by: y = activation (Σw_i * x_i + b) where x_i are the inputs, w_i are the weights, b is the bias, Σ denotes the sum over all inputs, and activation is the activation function [19]. MLP can handle large datasets with a high number of input features, making it suitable for problems with high-dimensional data, such as CAD prediction, where numerous medical parameters can be used as input features. MLP can learn complex representations of the data through multiple hidden layers, allowing it to capture intricate patterns in the data that are relevant to CAD prediction. However, the performance of MLP for any prediction is also dependent on the quality and representativeness of the data, the choice of hyperparameters, and the architecture of the network. Therefore, careful consideration of these factors is important for achieving good performance with MLP in CAD prediction.

5.2. Convolutional Neural Network (CNN)

Convolutional Neural Networks (CNNs) are a deep learning neural network category that is frequently utilized in computer vision jobs, like image categorization, identifying objects, and semantic segmentation [21]. They are designed to automatically and adaptively learn hierarchical representations of image features, from edges and textures to object parts and entire objects. The fundamental framework of CNN includes several layers, which comprise convolutional layers, activation layers like ReLU, pooling layers, and fully connected layers. In the convolutional layer, a set of filters, also known as kernels, are used to detect specific features in the input image [22]. The filters are slid across the image in a process known as convolution, producing a set of feature maps that capture the presence of different features in the image. The function of the pooling layer is to decrease the spatial dimensions of the feature maps while still preserving the crucial information. This helps to reduce the computational cost of processing the data and also reduces overfitting by removing some of the less important information. The fully connected layer, also known as the dense layer, is used to make the final prediction based on the feature maps produced by the previous layers [23]. In image classification tasks, this layer outputs a probability distribution over the possible classes, and the class with the highest probability is the final prediction.

5.3. Long-Short-Term-Memory (LSTM)

LSTM make predictions based on sequential data, such as time series data, natural

language processing, and speech recognition. The primary advantage of LSTMs over traditional RNNs is their ability to effectively learn and retain long-term dependencies in sequential data [24]. LSTMs have a memory unit that can retain information over a longer period, which helps to prevent the vanishing gradient problem faced by traditional RNNs. LSTMs achieve this by using gates, which control the flow of information into and out of the memory unit. The input gate regulates the information that is added to the memory unit, the forget gate determines which information to discard, and the output gate decides which information to output. A typical LSTM network consists of a series of LSTM cells, each of which contains an input gate, a forget gate, a memory cell, and an output gate. The input to the LSTM network is fed through these cells, and the outputs from one cell are used as inputs to the next. The hidden state and cell state are updated based on the input, forget, and output gate values, and are passed on to the next cell. The finalized network is generated based on the hidden state of the last cell. This structure allows the LSTM network to maintain a memory of past inputs, which is crucial for processing sequential data. The gates in the cells allow the network to regulate the flow of information, effectively deciding what information to retain and what information to discard over time [24].

6. Model Construction

In this section, the model utilized for the experiment is outlined. The design of the model can be visualized in Figure 1. The steps involved in the experiment can be summarized as follows: In this study, we explored the use of a Convolutional Neural Network (CNN) for a classification task involving numerical data. To this end, we first preprocessed the numerical data to ensure it was in the appropriate format for use with a neural network. Specifically, we standardized the input features to have zero mean and unit variance. This step is important as it can improve the convergence and stability of the network during training, as well as its generalization performance. Next, we designed the architecture of CNN for numerical data. Unlike the architecture used for image data, which typically includes convolutional and pooling layers, the architecture for numerical data consisted of a series of fully connected layers. Further information on the hyperparameter tuning is presented in table 2. For MLP the numerical dataset was loaded into Weka and divided into training and test sets. The features were standardized to have zero mean and unit variance.

The architecture of the MLP was set as follows: (a) input layers with neurons equal to the number of features in the dataset, (b) two hidden layers with 128 and 64 neurons, respectively, and ReLU activation, (c) a dropout layer with a rate of 0.3 to reduce overfitting, (d) A layer at the output stage that uses softmax activation and has the same number of neurons as the number of classes present in the dataset. The model

was trained using the Stochastic Gradient Descent optimization algorithm with a learning rate of 0.01 and a batch size of 128 for 200 epochs. The architecture of the LSTM was set as follows: (a) an input layer with the number of neurons equal to the number of features in the dataset, (b) an LSTM layer with 128 neurons, (c) a dropout layer with a rate of 0.2 to reduce overfitting, (d) a fully connected layer with 64 neurons and ReLU activation. The model was trained using the Adam optimization algorithm with a learning rate of 0.005 and a batch size of 64 for 100 epochs. Further details on the parameter tunings of the models are shown in table 2.

Table 2 Models and their Parameter Tuning techniques

Model	Parameters
CNN	hidden_layer_size = a, activation = 'relu', gate activation = 'sigmoid, learning_rate = '0.3 momentum = 0.2, number of epoch= 10, batchsize=100
LSTM	hidden_layer_size = a, activation = 'relu', gate activation = 'sigmoid, learning_rate = '0.3 momentum = 0.2, number of epoch= 10, batchsize=100
MLP	hidden_layer_size = a, activation = 'relu' learning rate = '0.3 momentum = 0.2 optimizer = 'adam', batchsize=100

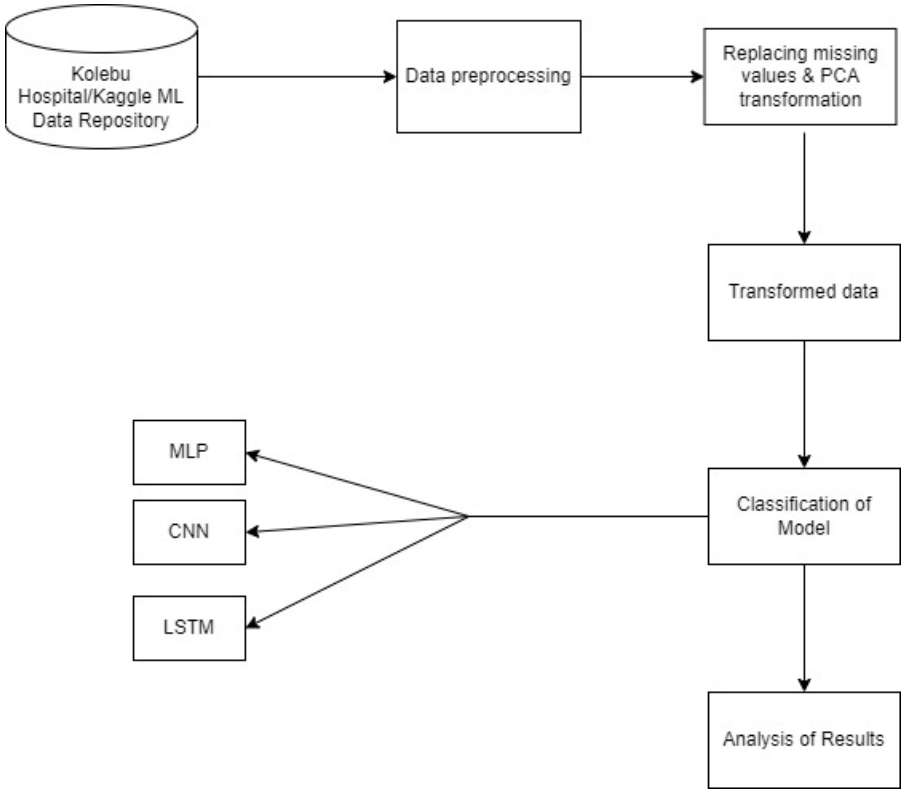

Figure 1. Model construction architecture

7. Performance Measures

Table 2 shows the performance measures utilized to assess the efficacy of the suggested technique. The accuracy of the proposed deep learning models is defined as the correct prediction of the instances. True positives (TP) are positive instances that are accurately classified as positive, while false negatives (FN) are positive instances that are incorrectly labelled as negative. False positive (FP) refers to cases in which there is no disease but the result is projected to be positive. True negative (TN) refers to negative cases in which the disease does not exist in the individual. Accuracy = (TP+TN)/(TP+TN+FP+FN). Precision defines what proportion of positive predictions were correct. The formula TP/(TP+FP) is used to determine precision. Recall is a measure of the proportion of true positive (TP) cases that are correctly identified by the model out of all actual positive (TP + FN) cases. In other words, it is the ratio of the number of correctly identified positive cases to the tot. The formular TP/(TP+FN) is used for computing recall. F1-score also syndicates precision and recalls into a single measure. Mathematically it is computed as F1-score = 2 * (precision * recall) / (precision + recall).

8. Results

The study used two different methods to analyze datasets, involving the application of various deep learning algorithms and PCA to compare differences. The first method directly classified the normal dataset obtained from the Korle-Bu Teaching Hospital and the HDI open dataset, while the second method involved feature selection and did not involve outlier detection. The results were promising, and in the second method, the dataset was normalized to account for outliers and feature selection. The study presents tables and figures that display the various results, which are discussed in the discussion section of the report.

Table 3 KORLEBU-DATA ORIGNAL ANALYSIS WITHOUT PCA

Model	Accuracy	TP	FP	Prec	Rec	F-Mea.	MCC	ROC	PRC
CNN	87.86	0.879	0.769	**0.848**	0.879	**0.86**	0.151	0.728	0.875
LSTM	**89.802**	**0.898**	**0.893**	0.839	**0.898**	0.852	**0.031**	**0.764**	**0.89**
MLP	89.3179	0.893	0.854	0.844	0.893	0.857	0.094	0.732	0.882

Table 4 KORLEBU-DATA ANALYSIS WITH PCA

Model	Accuracy	TP	FP	Prec	Rec	F-Mea.	MCC	ROC	PRC
CNN	89.714	0.897	0.886	0.841	0.897	0.853	**0.05**	0.76	0.889
LSTM	89.802	0.898	0.897	0.825	0.898	0.851	0.008	**0.764**	0.889
MLP	**89.923**	**0.899**	**0.891**	**0.871**	**0.899**	**0.853**	0.065	0.756	0.889

Table 5 OPEN-DATA ORIGINAL ANALYSIS WITHOUT PCA

Model	Accuracy	TP	FP	Prec	Rec	F-Mea.	MCC	ROC	PRC
CNN	91.7001	0.917	0.869	0.882	0.917	**0.886**	**0.131**	0.56	0.863
LSTM	91.6857	0.917	**0.853**	0.884	0.917	0.888	0.153	**0.787**	**0.916**
MLP	**91.9004**	**0.919**	0.862	**0.891**	**0.919**	0.888	0.162	0.772	0.913

Table 6 OPEN-DATA ANALYSIS WITH PCA

Model	Accuracy	TP	FP	Prec	Rec	F-Mea.	MCC	ROC	PRC
CNN	90.7985	0.908	0.756	0.884	0.908	**0.893**	0.213	0.785	0.914
LSTM	**92.0149**	**0.92**	**0.878**	**0.901**	**0.92**	0.887	0.158	0.793	**0.919**
MLP	91.929	0.919	0.87	0.893	0.919	0.887	**0.155**	**0.783**	0.916

In every classification model, the confusion matrix offers two types of errors, which is false positive and false negatives. A good model should have smaller or zero values for these two types of errors. A higher value for these predicted values leads to misclassification of the predictive model. True positive (TP) indicates the number of samples that were correctly predicted as having CHD. False negative (FN) indicates the number of samples that the model predicted as not having CHD but were actually positive. A high FN count implies that the model is not able to identify CHD patients correctly. False positive (FP) indicates the number of samples that the model predicted as having CHD but were actually negative. A high FP count implies that the model is predicting CHD for samples that do not have CHD. For a CHD prediction model, a high TP count is considered good, as it indicates that the model is correctly identifying patients with CHD. A low FN count is also considered good, as it means that the model is not missing any CHD cases. On the other hand, a low FP count is considered good, as it means that the model is not making false predictions of CHD in healthy individuals. From figure 2 in the Korle Bu dataset, given the sample size of 9090, true positive (TP) count of 8141, false negative (FN) count of 905, false positive (FP) count of 30, and true negative (TN) count of 14. In this case, the model has correctly predicted 8141 samples as having CHD. From figure 3, the LSTM correctly classified 8166 of the instances as healthy and not having CHD and 8 instances as CHD patients. In figure 7, the PCA_CNN model built on the Korlebu hospital data has a higher value for false positives. The PCA_ LSTM MODEL built on the korlebu data in figure 5 achieved the worst false negatives as it incorrectly classified 913 instances of CHD patients as having no CHD. Generally, from figures 4, 6, 8, 9, 10, 11, 12, and 13, the models achieved a relatively good count of FPs, FNs, TN, and TP.

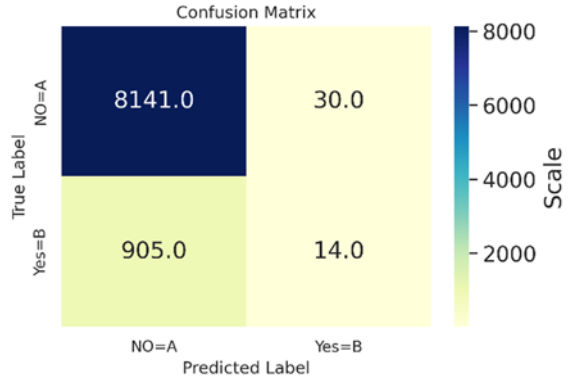

Figure 2. ORIGINAL KORLEBU DATA: CNN MODEL

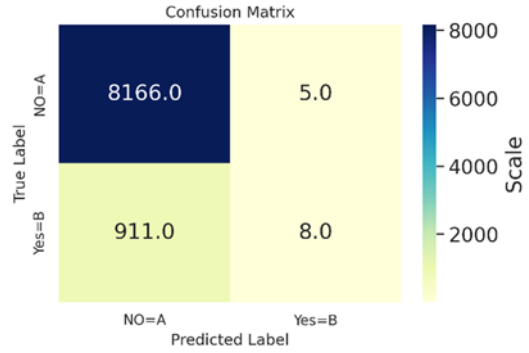

Figure 3. ORIGINAL KORLEBU DATA: LSTM MODEL

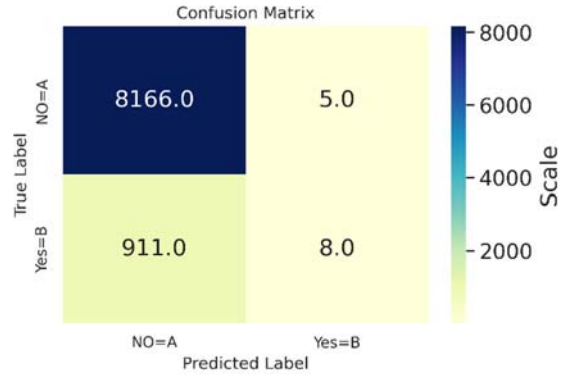

Figure 4. ORIGINAL KORLEBU DATA: MLP MODEL

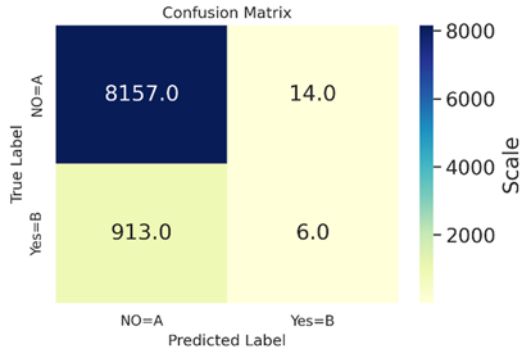

Figure 5. KORLEBU DATA: PCA_ LSTM MODEL

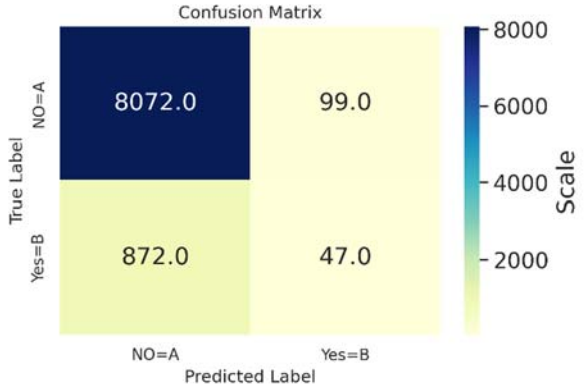

Figure 6. KORLEBU DATA: PCA_ MLP MODEL

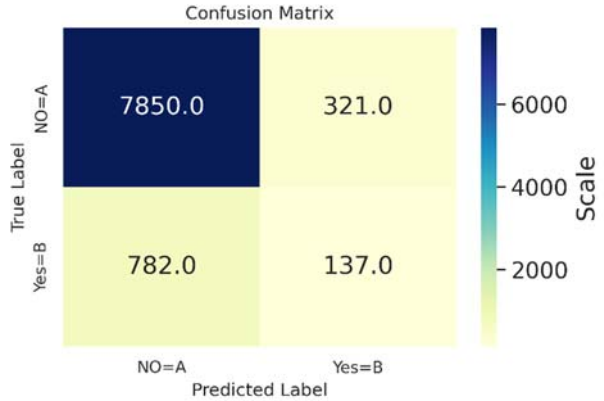

Figure 7. KORLEBU DATA: PCA_ CNN MODEL

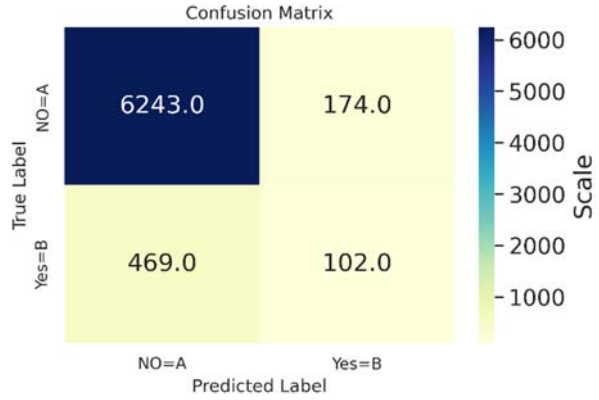

Figure 8. Original Open Data_ CNN MODEL

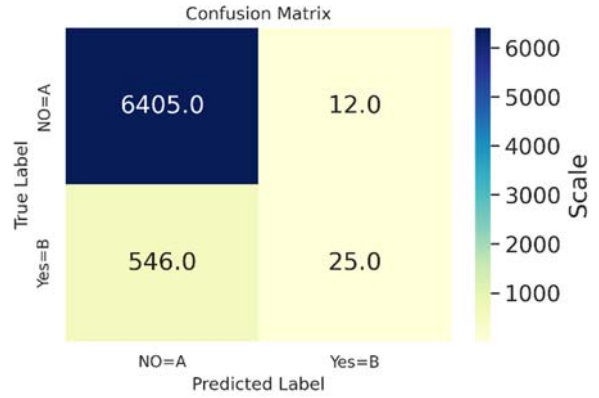

Figure 9. Original Open Data_ LSTM MODE

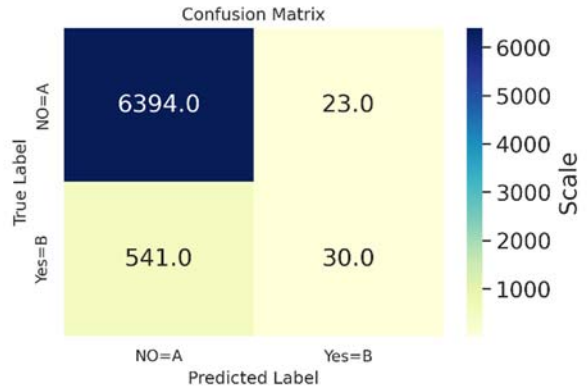

Figure 10. Original Open Data_ MLP MODEL

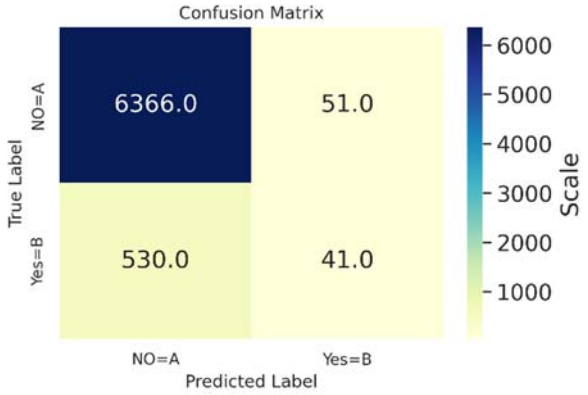

Figure 11. PCA Open Data_ LSTM MODEL

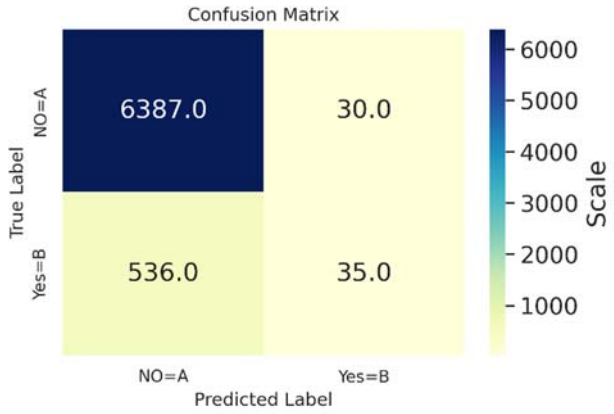

Figure 12. PCA Open Data_ MLP MODEL

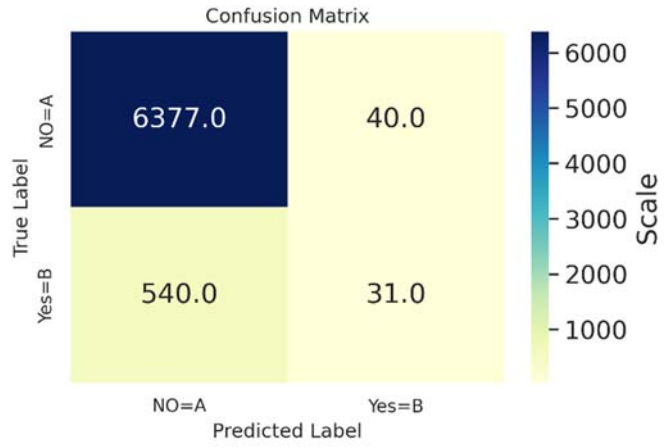

Figure 13. PCA Open Data_ CNN MODEL

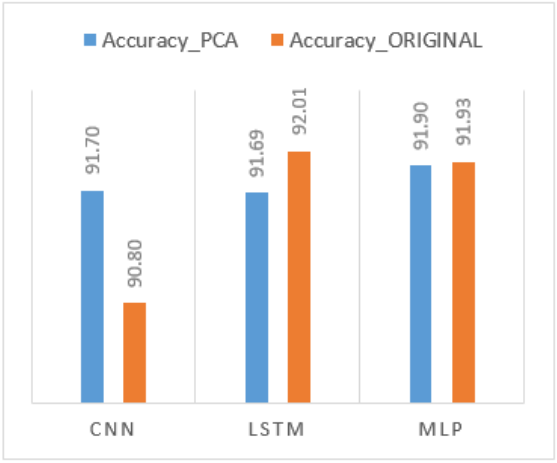

Figure 14. Performance of Original Open data vs PCA version

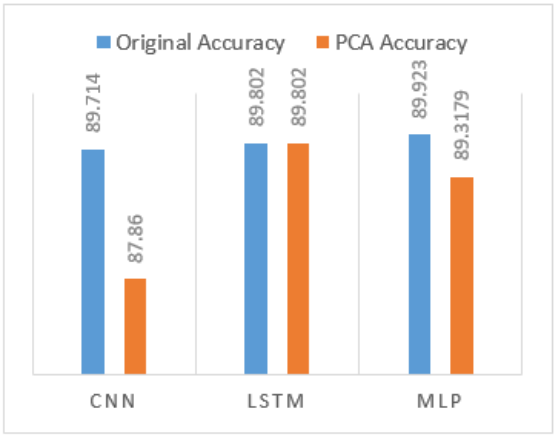

Figure 15. Performance of Original Korlebu data vs PCA version

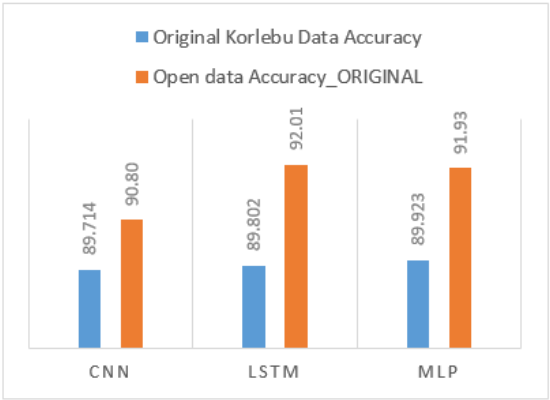

Figure 16. Performance of Original Korlebu vs Open data

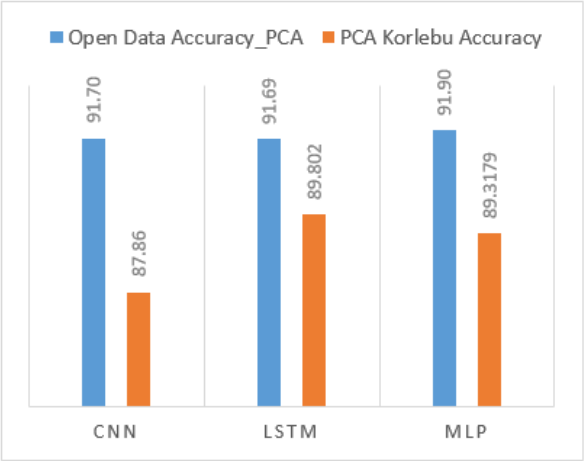

Figure 17. Performance of PCA Version of Korlebu and PCA Open data

9. Discussion

A modern approach to solving classification problems involves the use of artificial intelligence for intelligent data processing. This is achieved by solving optimization tasks. This research suggests that deep learning can be used to create a highly accurate and effective model for predicting CHD which has been seen in existing studies [25], [26]. The study aimed to achieve two objectives. The first objective was to investigate the impact of datasets obtained from cardiologists and open-source datasets on the accuracy of diagnosing coronary heart disease (CHD). The second objective was to compare the effectiveness of using PCA as a method to reduce spatial dimensions to assess the accuracy of CHD diagnosis. The achieved results present detailed insights worthy of discussion. For the dataset obtained from cardiologists, the results show that when PCA was not applied, the LSTM model had the highest accuracy (89.802%), true positive rate (0.898), and recall (0.852) among the three models, followed by the MLP model with an accuracy of 89.3179%, true positive rate of 0.893, and recall of 0.857. The CNN model had the lowest accuracy (87.86%), true positive rate (0.879), and recall (0.86) among the three models. The precision of the models ranged from 0.844 to 0.848, indicating that the models had relatively high precision in their diagnostic predictions. The F-measure scores ranged from 0.86 to 0.852, indicating that the models had reasonably good performance in terms of balancing precision and recall. The MCC scores ranged from 0.031 to 0.151, suggesting that the models had a moderate to weak correlation with the ground truth. The ROC area and PRC area values ranged from 0.728 to 0.764 and 0.875 to 0.882, respectively. These values indicate the overall performance of the models in terms of balancing the true positive rate and false positive rate and how well they rank the positive samples, respectively.

The updated results provided in the second set of data after applying PCA are shown in table 3. The results show that the accuracy of all three models improved compared to the previous results when PCA was not applied. The MLP model had the highest accuracy (89.923%), followed by the LSTM model (89.802%) and the CNN model (89.714%). The true positive rate for all three models remains relatively high, ranging from 0.897 to 0.899. The false positive rate for all models has decreased compared to the previous results, with rates ranging from 0.886 to 0.891. The precision of the models ranges from 0.825 to 0.871, indicating that the models have reasonably high precision in their diagnostic predictions. The F-measure scores range from 0.853 to 0.899, indicating that the models have good performance in terms of balancing precision and recall. The MCC scores range from 0.008 to 0.065, indicating that the models have a low to moderate correlation with the ground truth. The ROC area and PRC area values are similar for all three models, ranging from 0.756 to 0.764 and 0.889 for both measures, respectively. In sum, applying PCA relatively improved the accuracy of all three models, with the MLP model showing the highest accuracy and precision. These findings suggest that using PCA as a dimensionality reduction technique can be an effective way to improve the accuracy of CHD diagnosis using deep learning models. However, it's important to note that the effectiveness of PCA, as with any dimensionality reduction technique, depends on the specific characteristics of the dataset and the machine learning models used. Therefore, it's important to evaluate the impact of dimensionality reduction techniques on model performance on a case-by-case or database. Applying PCA to the open dataset used to predict CHD had a negative impact on the accuracy of the CNN model but had no significant impact on the accuracy of the LSTM and MLP models. The accuracy of the CNN model decreased from 91.7% to 90.8% when PCA was applied, while the accuracy of the LSTM and MLP models remained relatively stable, ranging from 91.9% to 92.0%. Additionally, the other performance metrics were also relatively stable across the two sets of results for the LSTM and MLP models. However, the CNN model had a slightly lower TP Rate and a higher FP Rate when PCA was applied, which may have contributed to the decrease in accuracy.

Conversely, it has been proven in the prior literature [10], [27] that the PCA dimensionality reduction method is efficient in improving the accuracy of diagnosing various types of diseases. In the current study, no statistically significant differences were found in the efficacy of the PCA method for CHD classification. The results also indicate that the selected features/terms by cardiologists (experts) did not outperform the performance demonstrated when the open dataset and the PCA extracted feature model. The evaluation metrics used in the study revealed only minor discrepancies of the models in terms of performance. This further supports the

conclusions drawn in previous studies. Consequently, based on the results presented in this research, it can be inferred that: The effectiveness of expert-based features on CHD prediction is relatively stable when compared with open-source datasets.

10. Conclusion

The study aimed to achieve two objectives. The first objective was to investigate the impact of datasets obtained from cardiologists and open-source datasets on the accuracy of diagnosing coronary heart disease (CHD). The second objective was to compare the effectiveness of using PCA as a dimensionality reduction technique to assess the accuracy of CHD diagnosis. The results show that PCA improves the CHD prediction accuracy for datasets obtained from cardiologists. The effectiveness of expert-based feature selection techniques on CHD classification is relatively stable when compared with open-source datasets. Applying PCA to the open dataset used to predict CHD had a negative impact on the accuracy of the CNN model but had no significant impact on the accuracy of the LSTM and MLP models. The current study suggests there is no statistically significant impact on the efficacy of the PCA method for CHD classification, however, MLP and LSTM offer promising results. The limitation of the findings could sample size of datasets utilized in this study. Consequently, the findings indicated that, given a sufficiently enough population of training data, it is possible to get a very accurate prediction of CHD using PCA with datasets received from cardiologists. Sophisticated machine-learning algorithms were critical in dealing with the noise, redundancy, heterogeneity, and nonlinearity of disease prediction. Furthermore, expert-based CHD data received from cardiologists revealed new features useful for prediction. Our findings emphasize the significance of collecting large amounts of data from cardiologists in order to make reliable CHD forecasts.

Conflicts of Interest

The authors declare no conflict of interest.

References

[1] Centers for Disease Control and Prevention, “Heart Disease Facts,” CDC, 2022. https://www.cdc.gov/heartdisease/facts.htm (accessed Feb. 12, 2023).
[2] World Health Organization, “Cardiovascular diseases,” Cardiovascular diseases, 2023. https://www.who.int/health-topics/cardiovascular-diseases#tab=tab_1 (accessed Feb. 12, 2023).
[3] R. Rajendran and A. Karthi, “Heart disease prediction using entropy based feature engineering and ensembling of machine learning classifiers,” Expert Syst. Appl., vol. 207, p. 117882, 2022, doi: https://doi.org/10.1016/j.eswa.2022.117882.

[4] J. Jeyaranjani, T. D. Rajkumar, and T. A. Kumar, “Materials Today : Proceedings Coronary heart disease diagnosis using the efficient ANN model,” Mater. Today Proc., no. xxxx, 2021, doi: 10.1016/j.matpr.2021.01.257.
[5] S. Kutiame, R. Millham, A. F. Adekoya, M. Tettey, B. A. Weyori, and P. Appiahene, “Application of Machine Learning Algorithms in Coronary Heart Disease : A Systematic Literature Review and Meta-Analysis,” Int. J. Adv. Comput. Sci. Appl., vol. 13, no. 6, pp. 153–164, 2022, [Online]. Available: https://dx.doi.org/10.14569/IJACSA.2022.0130620
[6] M. Swathy and K. Saruladha, “A comparative study of classification and prediction of Cardio-Vascular Diseases (CVD) using Machine Learning and Deep Learning techniques,” ICT Express, vol. 8, no. 1, pp. 109–116, 2022, doi: https://doi.org/10.1016/j.icte.2021.08.021.
[7] A. Rahim, Y. Rasheed, F. Azam, M. W. Anwar, M. A. Rahim, and A. W. Muzaffar, “An Integrated Machine Learning Framework for Effective Prediction of Cardiovascular Diseases,” IEEE Access, vol. 9, pp. 106575–106588, 2021, doi: 10.1109/ACCESS.2021.3098688.
[8] M. Nilashi, N. Ahmadi, S. Samad, L. Shahmoradi, H. Ahmadi, and O. Ibrahim, “Journal of Soft Computing and Decision Support Systems Disease Diagnosis Using Machine Learning Techniques : A Review and Classification,” J. Soft Comput. Decis. Support Syst., vol. 7, no. 1, pp. 19–30, 2020.
[9] R. Katarya and S. Kumar, “Machine Learning Techniques for Heart Disease Prediction : A Comparative Study and Analysis,” Health Technol. (Berl)., no. 0123456789, 2020, doi: 10.1007/s12553-020-00505-7.
[10] A. Dutta, T. Batabyal, M. Basu, and S. T. Acton, “An Efficient Convolutional Neural Network for Coronary Heart Disease Prediction,” ArXiv, 2019.
[11] C. Tesche and V. Brandt, “Calling for a New Framingham: Machine Learning in Cardiovascular Risk Assessment—The Key for Improved Outcome Prediction?∗,” JACC Cardiovasc. Imaging, vol. 14, no. 3, pp. 626–628, 2021, doi: https://doi.org/10.1016/j.jcmg.2020.12.027.
[12] A. K. Dubey and K. Choudhary, “A systematic review and analysis of the heart disease prediction methodology,” International Journal of Advanced Computer Research, vol. 8, no. 38. pp. 240–256, 2018. Doi: 10.19101/IJACR.2018.837025.
[13] O. Subramaniam and R. Mylswamy, “Prediction of coronary artery disease using core principal component analysis based support vector machine,” Int. J. Sci. Technol. Res., vol. 8, no. 8, pp. 791–798, 2019.
[14] R. Bharti, A. Khamparia, M. Shabaz, G. Dhiman, S. Pande, and P. Singh, “Prediction of Heart Disease Using a Combination of Machine Learning and Deep Learning,” Comput. Intell. Neurosci., vol. 2021, p. 11, 2021.
[15] M. F. Kabir, T. Chen, and S. A. Ludwig, “A performance analysis of dimensionality reduction algorithms in machine learning models for cancer prediction,” Healthc. Anal., vol. 3, p. 100125, 2023, doi: https://doi.org/10.1016/j.health.2022.100125.
[16] K. Keerthi Vasan and B. Surendiran, “Dimensionality reduction using Principal Component Analysis for network intrusion detection,” Perspect. Sci., vol. 8, pp. 510–512, 2016, doi: 10.1016/j.pisc.2016.05.010.
[17] G. Ivosev, L. Burton, and R. Bonner, “Dimensionality reduction and visualization in principal component analysis,” Anal. Chem., vol. 80, no. 13, pp. 4933–4944, 2008, doi: 10.1021/ac800110w.

[18] M. F. Ansari, B. Alankarkaur, and H. Kaur, A Prediction of Heart Disease Using Machine Learning Algorithms. Springer International Publishing, 2021. Doi: 10.1007/978-3-030-51859-2.
[19] J. Yan et al., "A clinical decision support system for predicting coronary artery stenosis in patients with suspected coronary heart disease," Comput. Biol. Med., vol. 151, p. 106300, 2022, doi: https://doi.org/10.1016/j.compbiomed.2022.106300.
[20] Z. Kalinić, V. Marinković, L. Kalinić, and F. Liébana-Cabanillas, "Neural network modeling of consumer satisfaction in mobile commerce: An empirical analysis," Expert Syst. Appl., vol. 175, no. February, pp. 0–3, 2021, doi: 10.1016/j.eswa.2021.114803.
[21] C. Krittanawong et al., "Machine learning prediction in cardiovascular diseases: a meta-analysis," Sci. Rep., vol. 10, no. 1, pp. 1–11, 2020, doi: 10.1038/s41598-020-72685-1.
[22] D. Han, J. Liu, Z. Sun, Y. Cui, Y. He, and Z. Yang, "Deep learning analysis in coronary computed tomographic angiography imaging for the assessment of patients with coronary artery stenosis," Comput. Methods Programs Biomed., vol. 196, p. 105651, 2020, doi: https://doi.org/10.1016/j.cmpb.2020.105651.
[23] N. Yuan et al., "Prediction of Coronary Artery Calcium Using Deep Learning of Echocardiograms," J. Am. Soc. Echocardiogr., 2022, doi: https://doi.org/10.1016/j.echo.2022.12.014.
[24] A. M. Johri et al., "Deep learning artificial intelligence framework for multiclass coronary artery disease prediction using combination of conventional risk factors, carotid ultrasound, and intraplaque neovascularization," Comput. Biol. Med., vol. 150, p. 106018, 2022, doi: https://doi.org/10.1016/j.compbiomed.2022.106018.
[25] S. Rani and S. Masood, "Predicting congenital heart disease using machine learning techniques," J. Discret. Math. Sci. Cryptogr., vol. 23, no. 1, pp. 293–303, 2020, doi: 10.1080/09720529.2020.1721862.
[26] H. Khdair and N. M. Dasari, "Exploring Machine Learning Techniques for Coronary Heart Disease Prediction," vol. 12, no. 5, 2021.
[27] J. Zhang, H. Zhu, Y. Chen, C. Yang, H. Cheng, and Y. Li, "Echocardiography-based screening for coronary heart disease using an ensemble machine learning approach," 2020.