HIGH SCHOOL EDITION
Journal of Student Research
Volume 11 Issue 3 (2022)

An Intelligent System for Early Prediction of Cardiovascular Disease using Machine Learning

Aarush Kachhawa[1] and Jeremy Hitt[#]

[#]Saint Francis High School, Mountain View, CA, USA
[#]Advisor

ABSTRACT

Cardiovascular disease (CVD) remains the leading cause of death, responsible for 18.6 million deaths globally in 2019. Given the wide availability of several effective therapeutic treatment options, early diagnosis of CVD is critical for timely intervention and slowing down the progression of the disease. CVD is associated with a multitude of risk markers with non-linear interactions among them, making accurate diagnosis of CVD quite challenging, especially for non-specialized clinicians and under-resourced facilities in developing countries. In recent years, machine learning based computational techniques have shown great promise in becoming a great diagnostic tool. The goal of this research is to leverage multiple machine learning methods such as random forest, gradient boosting, logistic regression and artificial neural network and evaluate their prediction efficacy. This study also evaluates the feasibility of combining multiple UCI datasets in order to improve the prediction accuracy of the models. On a merged dataset of over 700 patients from the UCI machine learning repository, the most accurate model was found to be the random forest classifier, showing an accuracy and F1 score of 94% and AUC of 0.98. It was found that ensemble learning methodologies along with data optimization and hyperparameter tuning techniques were able to achieve higher accuracy relative to prior published studies on these datasets. Finally, this study also proposes how these machine learning workloads can be incorporated into a distributed cloud connected healthcare system to make them widely accessible to practicing doctors and enable them to assess CVD risk of their patients.

Introduction

Cardiovascular disease (CVD) represents a broad set of disorders affecting the heart and blood vessels, and include angina, coronary heart disease, ischemic stroke, heart attack, arrhythmia and other conditions. Even with the advent of effective treatments for CVDs, they were the number one cause of deaths and mortality in the United States (more than Cancers and Chronic Lower Respiratory Disease combined) as per the AHA statistics [1]. Due to the lack of traceable physical symptoms in patients with active CVD, more advanced means of diagnosis are needed. Accurate early classification of patients with higher risk of CVD allows for more intensive procedures to reduce the likelihood of potentially deadly events, like cardiac arrest, and early treatment options or lifestyle changes to slow the progression of the disease.

Data modeling and analysis by supervised machine learning techniques have increasingly become more pivotal in the interpretation and early diagnosis of heart disease in patients. Supervised machine learning is a class of classification algorithms and systems that work by analyzing labeled (training) data and producing a function that maps each training example to its correct label using the input features. This function can be used for predicting the label of a new input data. Each instance of the training data is typically a pair of an input feature vector and a value for the class label [2, 3]. Supervised machine learning algorithms are apt for interpreting and predicting disease onset in individuals in research studies with historical patient data. There have been

several successful research studies for predicting diabetes using machine learning classification techniques like support vector machines (SVM) leading to industry adoption and production of monitoring devices [4].

The goal of this study is to build an early prediction system for a CVD like coronary heart disease in individuals by analyzing historical patient data. The proposed system should have high accuracy and should be a pervasive solution that can be incorporated in connected health care systems via a cloud. Medical practitioners should have a reliable tool to help in early diagnosis of heart disease, which can ultimately help save lives through timely treatment.

In a recent study, the UCI heart disease dataset was used for heart disease prediction using multilayer perceptron [8]. The study achieved an accuracy of 85.71% and used a small neural network architecture with 2 hidden layers and 8 neurons each. The study also proposed a web application tool for heart disease prediction deployed using cloud computing. Another recent study uses random forest classification to predict heart disease [10]. It published an accuracy of 86.9%, sensitivity value of 90.6% and specificity value of 82.7%. Proposed study builds on the prior work and systematically evaluates multiple machine learning methods for their efficacy for prediction. Furthermore, this study also evaluated the strategy of combining multiple datasets from UCI to improve the prediction accuracy of the models.

The rest of the paper is further organized into two sections. The first is a Methods section which is divided into subsections describing the datasets, machine learning methodologies, parameter choices and performance metrics. This is followed by the Results section which has two subsections describing the exploratory data analysis employed and prediction results obtained in this work.

Methods

Heart Disease Datasets

This study analyzed 5 datasets from the UCI machine learning repository [11, 14] - Cleveland (303 observations), Hungarian (294 observations), Switzerland (123 observations), Long Beach VA (200 observations) and Statlog (Heart) Data Set (270 observations). Most existing research studies have only used the Cleveland dataset that contains 75 attributes and 303 instances of patient data [12]. There are 12 attributes common to all 5 datasets, hence there are a total of 918 unique observations and 12 common attributes available for this research as are listed in Table 1.

Table 1. Dataset attributes.

Attribute	**Description**	**Value Set**
Age	age of patient	[years]
Sex	sex of patient	[M: male, F: female]
ChestPainType	chest pain type	[TA: Typical Angina, ATA: Atypical Angina, NAP: Non-Anginal Pain, ASY: Asymptomatic]
RestingBP	resting blood pressure	[mm Hg]
Cholesterol	serum cholesterol	[mm/dl]
FastingBS	fasting blood sugar	[1: if FastingBS > 120 mg/dl, 0: otherwise]

Attribute	Description	Value Set
RestingECG	resting electrocardiogram results	[Normal: Normal, ST: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV), LVH: showing probable or definite left ventricular hypertrophy by Estes' criteria]
Thalach	maximum heart rate achieved	[Numeric value between 60 and 202]
ExerciseAngina	exercise induced angina	[Y: Yes, N: No]
OldPeak	ST depression induced by exercise relative to rest	[Numeric value measured in depression]
ST_Slope	slope of the peak ST exercise segment	[Up: upsloping, Flat: flat, Down: downsloping]
HeartDisease	output class	[1: heart disease, 0: Normal]

Classification Methods

Data preprocessing was pivotal in maximizing performance accuracy of the models. The input data was normalized using a MinMaxScaler from the Scikit-learn Python library which normalizes the data between 0 and 1, while preserving the shape of the distribution and maintaining the importance of outliers. The dataset was randomly shuffled and split into 2 subsets. The first subset, the training set, contained 80% of the patient data and was used to train the models. The second subset, the test set, contains the remaining 20% of the patient data and was used to evaluate the performance of models.

The software development and execution for this research study was done on a GPU runtime environment on Google Colab. The ML models were developed on Jupyter Notebook using Numpy, Pandas, Scikit-learn and Keras Python libraries. Four classification models were selected for this study, which are artificial neural network (ANN), gradient boosting (GB), random forest (RF) and logistic regression (LR). GB and LR classifiers were tuned for hyperparameters using the GridSearchCV library from Scikit-learn which further boosted accuracy by taking the most optimal parameters in tuning the models.

LR is used in prediction systems to forecast a categorical variable from a set of independent predictor factors or features. In binary logistic regression, the prediction is a probability value between 0 and 1. LR is well suited for medical research use cases like disease prediction based on statistical patient samples of dependent health stats [5].

The ANN implementation was done with the MLPClassifier from Scikit-Learn. Multilayer perceptron (MLP) is a fully connected multi-layer neural network with 1 or more hidden layers. The MLP model architecture used in this study was one with 4 densely connected hidden layers with output dimensions 256, 128, 64 and 32 respectively. The Relu activation function was used in the hidden layers.

Performance Evaluation of Classification Models

In this study, performance evaluation of the classification models was done using the following measures: classification accuracy, precision, recall, F1 score, receiver operating characteristic (ROC) and area under the ROC

curve (AUC). The underlying metrics used to compute these values are true positive (TP), false positive (FP), true negative (TN) and false negative (FN).

TP: cases where heart disease is TRUE and the system predicted TRUE

TN: cases where heart disease is FALSE and the system predicted FALSE

FP: cases where heart disease is FALSE and the system predicted TRUE

FN: cases where heart disease is TRUE and the system predicted FALSE

Classification accuracy is the ratio of correct predictions to the sum of all predictions made. It works best when the dataset is balanced in terms of the number of samples of each class.

Equation 1. Accuracy = (TP + TN) / (TP + TN + FP + FN)

Precision is the ratio of correct positives predicted to the sum of all positives predicted in one class.

Equation 2. Precision = TP / (TP + FP)

Recall is the ratio of correct positives predicted to the sum of all samples that should have been positive in one class.

Equation 3. Recall = TP / (TP + FN)

F1 score is the harmonic mean of precision and recall. In cases where the dataset is imbalanced with respect to the number of samples of each case or where false negatives can have serious repercussions, for example in disease prediction, F1 score is a better measure of performance of the classification model.

Equation 4. F1 score = 2 * (Precision * Recall) / (Precision + Recall)

Receiver Operating Characteristics (ROC) curve is a graph of two parameters, true positive rate (TPR) and false positive rate (FPR). The entire two-dimensional area under the ROC curve is referred to as AUC. It is a value between 0 and 1, with a value of 0 indicating that the model predictions are completely wrong and 1 indicating that the predictions are 100% correct. AUC can be interpreted as the probability that the classification model ranks a random positive sample more highly than a random negative sample in the dataset [13].

Results

Exploratory Data Analysis

The data pre-processing stage included steps to clean the data, normalize the attributes, handle missing values and visualize the data for analysis. All the ML models deployed in the study required numerical data values, hence the first step was to convert any string values to numerical values. Data quality is an important factor in getting accurate results. Hence data cleaning was done by removing unnecessary or irrelevant attributes from the dataset and duplicate rows. This step of the procedure makes the dataset more precise and exact. Additionally, data was verified against the allowed value set for the attributes and any erroneous values were set to NaN (Not a Number). Finally, any NaNs, missing or null values were updated with median values for that attribute from the dataset as these values decrease the productivity of the algorithms.

By studying the data distribution via box and whisker plots, some anomalies were found in Switzerland, VA and Hungarian datasets [9]. A value of 0 for attributes such as *cholesterol* and *restingECG* readout is highly suspicious and can point to data entry or measurement error. The Switzerland dataset had either missing or 0 values for *cholesterol* and the VA dataset had some rows with 0 value for *cholesterol*. However, the VA dataset was otherwise uniform, and could be used after disregarding the rows with 0 value for *cholesterol*. The Hungarian dataset also has an abnormal distribution of values for *restingECG* compared to the other datasets. Most values are 0 with some outliers at 1.0 and 2.0, as shown in Figure 1 below.

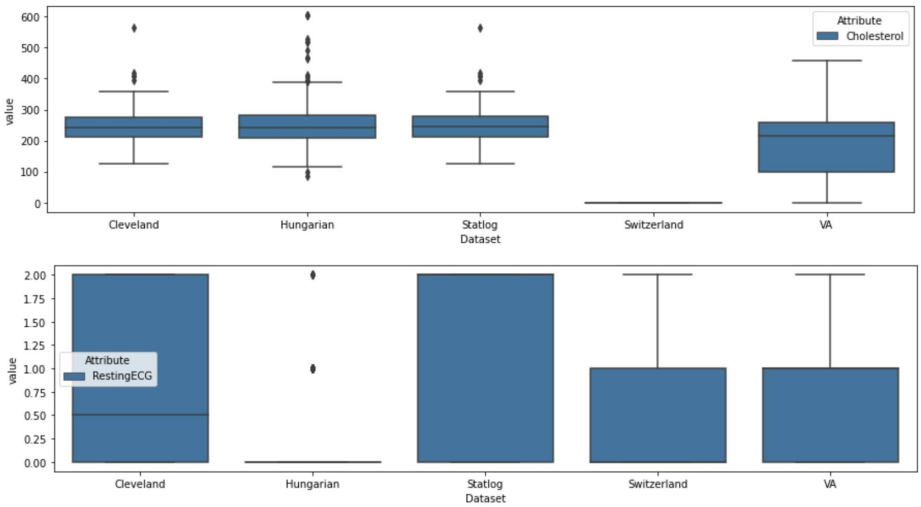

Figure 1. Box plots showing distributions of *cholesterol* and *restingECG* values in all 5 datasets.

Given these findings, this research study used the Cleveland, Statlog, and cleaned VA datasets only since their data distribution is consistent and has comparable ranges for minimum, lower quartile, median, upper quartile and maximum values. The box plots showing distributions for all attributes in these datasets, is shown below in Figure 2.

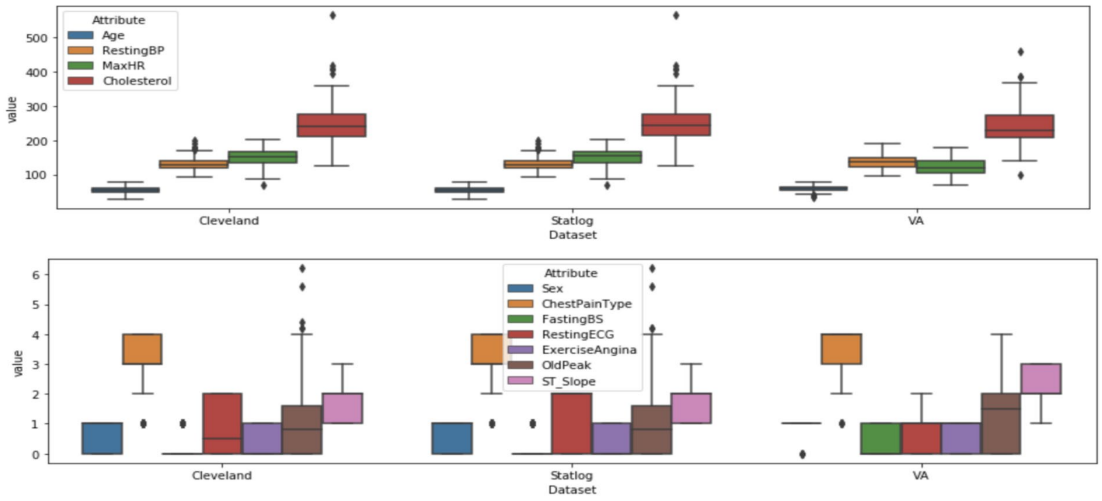

Figure 2. Box plots of all attributes in Cleveland, Statlog and cleaned VA datasets.

The dataset was then balanced for output bias of heart disease presence, by balancing the number of positive and negative data samples. This was done by random up-sampling of the minority classes, which resulted in a boost of 5% in accuracy of the random forest classifier model.

Data visualization using pair plots of bivariate distributions was done to find attributes that show dominant separation of disease and non-disease samples. By doing so, it was found that *oldPeak* and *age* showed stronger separation for heart disease compared to other attributes in pair plots. Both *oldpeak* and *age* were found to have a mild linear separation of disease vs non-disease samples when paired against *maxHR* and *cholesterol*. This is illustrated in Figure 1.

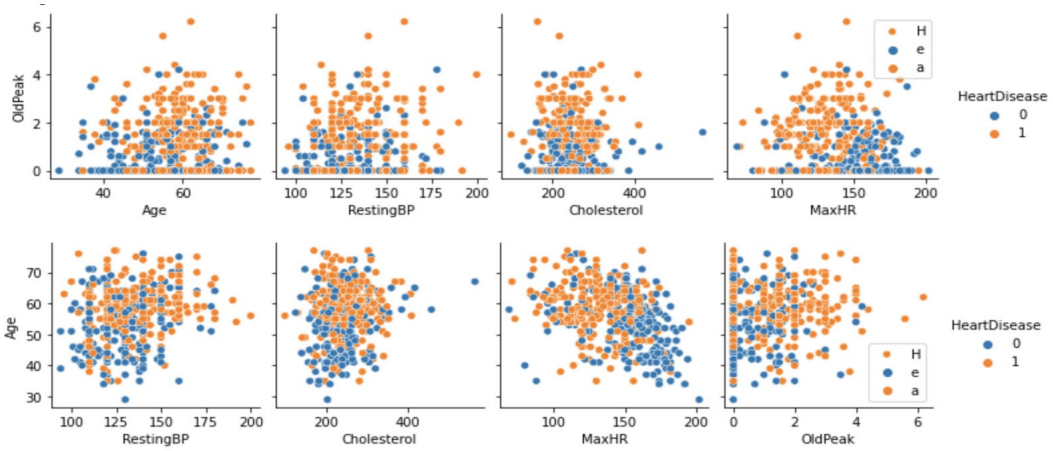

Figure 3. Pair plots of *oldPeak* and *age* against other attributes.

Upon quantifying these results in a correlation coefficient matrix, it was observed that the attributes with highest correlation with *heartDisease* were *oldPeak and age* with coefficients of 0.4 and 0.27 respectively. Additionally, with a coefficient of 0.4, *HeartDisease* and *oldPeak* had the highest overall correlation among continuous attributes. Figure 4 shows the correlation matrix of all numerical attributes made with the Pandas library. Figure 5 helps visualize the values in the correlation matrix with a heatmap.

	OldPeak	Age	RestingBP	Cholesterol	MaxHR	HeartDisease
OldPeak	1.00	0.19	0.21	0.02	-0.27	0.40
Age	0.19	1.00	0.26	0.15	-0.41	0.27
RestingBP	0.21	0.26	1.00	0.12	-0.06	0.18
Cholesterol	0.02	0.15	0.12	1.00	0.02	0.08
MaxHR	-0.27	-0.41	-0.06	0.02	1.00	-0.40
HeartDisease	0.40	0.27	0.18	0.08	-0.40	1.00

Figure 4. Correlation matrix of the numerical attributes.

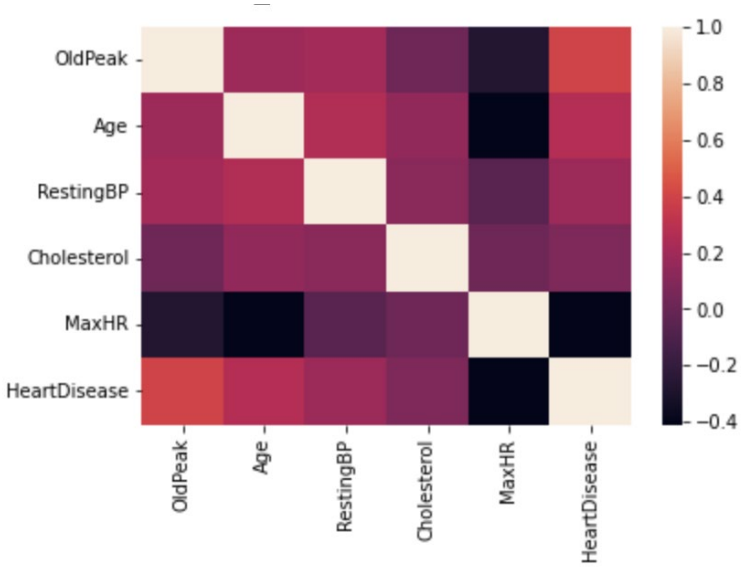

Figure 5. Correlation Heatmap.

Prediction Results

Table 2. Performance evaluation of the models.

Algorithm	Accuracy	F1 score (0, 1)	Precision (0, 1)	Recall (0, 1)	AUC
RF	94%	94%, 94%	93%, 95%	95%, 93%	0.98
GB	94%	94%, 94%	92%, 96%	96%, 92%	0.96
LR	79%	80%, 78%	76%, 82%	84%, 75%	0.87
ANN	91%	91%, 91%	88%, 94%	95%, 88%	0.89

Table 2 provides a detailed performance metrics comparison amongst the models. The best performance was found in the RF model with 94% accuracy, F1 score, precision and recall and 0.98 AUC value. RF is an ensemble classifier that combines bagging and random selection of features in multiple decision trees to achieve high accuracy in prediction and probability estimations [7]. RF can handle hundreds of input variables and can estimate which variables are important for classification. It is less sensitive to outliers in training data and overcomes the problem of overfitting.

For the GB model, the best prediction accuracy of 94% for the test set was found to be with learning rate 0.5, the number of boosting stages of 300 and subsample of 0.5. GB is an ensemble algorithm which in most cases, works by combining multiple decision trees sequentially. The trees are linked in a sequence where each tree attempts to minimize the previous tree's error. As a result of the lengthy connections, boosting algorithms are quite accurate, yet are slow to learn [6]. Hence multiple learning rates were used to tune the GB model to achieve a highest accuracy of 94%, F1 score of 94% for positive and negative cases. In LR, a regularization parameter (penalty) of 12 and C parameter (penalty strength) of 1.0 were found to be most optimal resulting in accuracy of 79%.

The ANN model had a classification accuracy of 91%. Originally, when evaluated on the test data without the feature scaling, the accuracy stabilized around 70%. With data optimization by min-max feature scaling, the accuracy and F1 score increased to 91%. For positive and negative cases of heart disease, the precision and recall scores vary moderately. The precision of patients with heart disease was 94%, while the precision of patients without it was 88%. Similarly recall for positive cases was 88% and 95% for negative cases. Figure 4 further illustrates the ROC curves for all the classification models in this study.

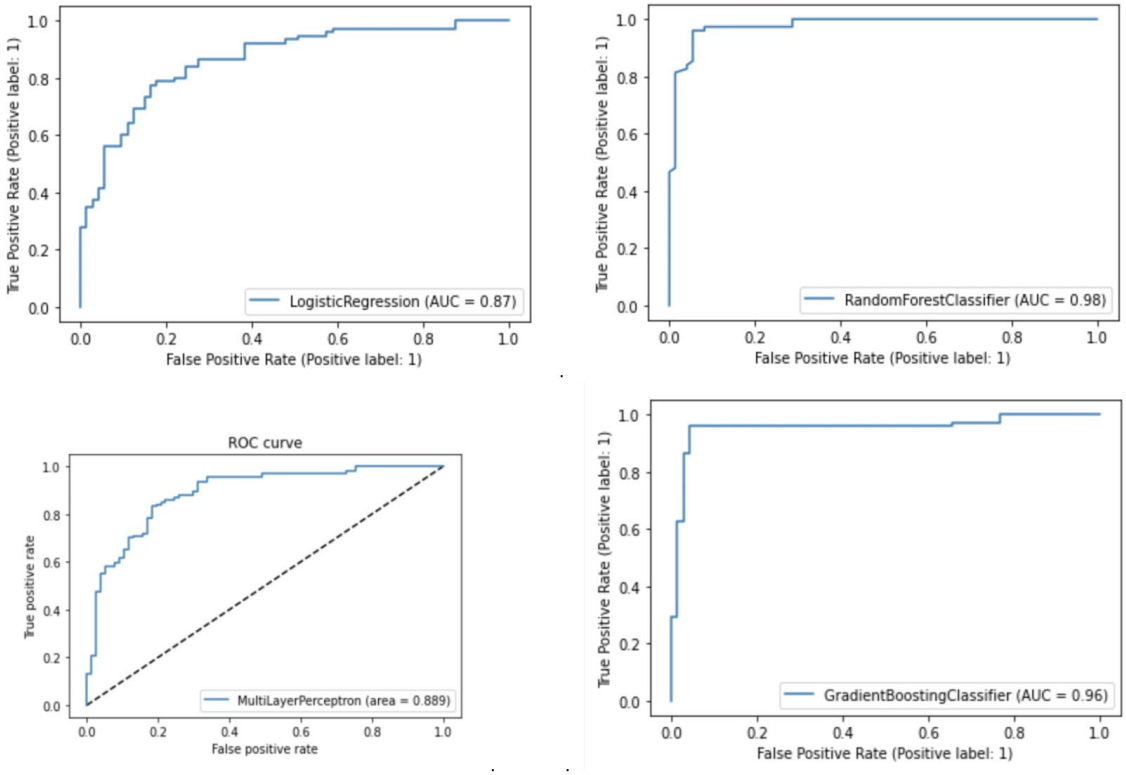

Figure 6. Receiver Operating Characteristic (ROC) graphs of LR, ANN, RF and GB models.

Discussion

This research demonstrates the efficacy of using Machine Learning models for CVD prediction. These models can be deployed as an open source, cloud-based system with a web interface and API for programmatic access to everyone. A more advanced end-to-end system for accurate and real time heart disease prediction would involve continuous monitoring of vital health metrics via wearable sensor technologies, advanced wireless and an AI powered healthcare software system running in the cloud. One such study reviewed some popular sensors on the market for monitoring patients' health parameters [8]. AliveKor is a touchpad based device for ECG monitoring and HealthGear devices can provide monitoring for physical parameters, blood sugar and lipids. In this study we propose a system for collecting an individual's health metrics and connecting to a cloud application that will comprise the ML workloads for heart disease prediction. Using cloud computing we would deploy this prediction system as a web application on the client side (doctors office and hospitals) that would help primary care and cardiovascular specialist physicians in early diagnoses of high risk patients and allow them to prescribe appropriate treatment plans. Another application of this prediction system can be to generate personalized statistics for individuals and monitor their cardiovascular disease risk over time . Subsequently, aggregated predictions across all patients can be useful for demand forecasting in hospitals for medical equipment. Moreover,

as cardiovascular disease is the number one cause of death and mortality in the US [1], the system will reduce the burden of emergency treatment and help reduce healthcare costs.

Conclusion

This study has researched multiple machine learning methods to predict CVD using publicly available heart disease datasets. The proposed models have achieved higher accuracy and predictability compared to the recent benchmarking studies. Ensemble learning models such as random forest and gradient boosting classifiers were found to be most effective and accurate. Given the relatively small size of the dataset, MLP was found to be the most accurate neural network model. As future improvements, convolutional neural networks (CNN) can be utilized to train bigger models with larger datasets and features and more dense hidden layers to achieve higher rate of classification accuracy. CVDs have a slow progression and can largely be controlled with lifestyle changes and timely treatment. Currently CVD diagnosis is mostly done using clinical methods that can require expensive laboratory tests to be performed regularly. Accurate and high precision supervised learning models can help speed up the diagnostic process, allowing for more time to successfully treat patients. In a connected health care system, such ML based prediction systems can be deployed on a cloud application. An end to end system for heart disease prediction, would require smart devices like wearables or mobile phones to continuously monitor patients' health attributes and send data to the cloud for processing and analysis. Additionally, the core functionality of machine learning based classification would run on a cloud computing layer.

In summary, this study makes three key contributions to the existing body of work (i) It proposes several machine learning algorithms that achieve state of the art performance using attributes or test results of patients that are commonly measured or can be easily collected by practicing physicians on the front lines (ii) suggests a guide towards creating acceptance criterion for a prediction for a patient using exploratory data analysis and (iii) proposes an intelligent system and its deployment such that it is accessible to the practicing primary care and cardiovascular specialist physicians.

Limitations

The proposed system relies on patient health attributes to be provided which require several medical examinations to be done and monitored on a regular basis, hence may not always be available.

Acknowledgments

I would also like to thank my mentor Jeremy Hitt for guiding me in this project. I am also grateful for my parents' support and encouragement during this research project.

References

[1] *2021 Heart Disease and Stroke statistics update fact sheet at-a-glance*. (n.d.). Retrieved June 1, 2022, from https://www.heart.org/-/media/phd-files-2/science-news/2/2021-heart-and-stroke-stat-update/2021_heart_disease_and_stroke_statistics_update_fact_sheet_at_a_glance.pdf?la=en

[2] *Machine learning: What it is and why it matters*. SAS. (n.d.). Retrieved May 31, 2022, from https://www.sas.com/en_us/insights/analytics/machine-learning.html

[3] Nasteski, V. (2017). An overview of the supervised machine learning methods. HORIZONS.B, 4, 51-62. https://doi.org/10.20544/horizons.b.04.1.17.p05

[4] *Diabetes prediction using support Vector Machines*. Sisense. (2022, March 18). Retrieved May 31, 2022, from https://www.sisense.com/blog/diabetes-prediction-using-support-vector-machines/

[5] *What is logistic regression?* Master's in Data Science. (n.d.). Retrieved May 31, 2022, from https://www.mastersindatascience.org/learning/introduction-to-machine-learning-algorithms/logistic-regression/

[6] Yıldırım, S. (2020, February 17). *Gradient boosted decision trees-explained*. Medium. Retrieved May 31, 2022, from https://towardsdatascience.com/gradient-boosted-decision-trees-explained-9259bd8205af

[7] Brownlee, J. (2020, December 2). *Bagging and Random Forest Ensemble algorithms for Machine Learning*. Machine Learning Mastery. Retrieved May 31, 2022, from https://machinelearningmastery.com/bagging-and-random-forest-ensemble-algorithms-for-machine-learning/

[8] Bhoyar, S., Wagholikar, N., Bakshi, K., & Chaudhari, S. (2021). Real-time heart disease prediction system using Multilayer Perceptron. *2021 2nd International Conference for Emerging Technology (INCET)*. https://doi.org/10.1109/incet51464.2021.9456389

[9] *Whisker plot*. Whisker Plot - an overview | ScienceDirect Topics. (n.d.). Retrieved May 31, 2022, from https://www.sciencedirect.com/topics/mathematics/whisker-plot

[10] Pal, M., & Parija, S. (2021). Prediction of heart diseases using Random Forest. *Journal of Physics: Conference Series*, *1817*(1), 012009. https://doi.org/10.1088/1742-6596/1817/1/012009

[11] UCI Machine Learning Repository: Heart disease data set. (n.d.). Retrieved May 31, 2022, from https://archive.ics.uci.edu/ml/datasets/heart+disease

[12] Singh, A., & Kumar, R. (2020). Heart disease prediction using machine learning algorithms. *2020 International Conference on Electrical and Electronics Engineering (ICE3)*. https://doi.org/10.1109/ice348803.2020.9122958

[13] Mishra, A. (2020, May 28). *Metrics to evaluate your machine learning algorithm*. Medium. Retrieved May 31, 2022, from https://towardsdatascience.com/metrics-to-evaluate-your-machine-learning-algorithm-f10ba6e38234

[14] UCI Machine Learning Repository: Statlog (heart) data set. (n.d.). Retrieved May 31, 2022, from https://archive.ics.uci.edu/ml/datasets/statlog+(heart)