Heart Disease Prediction Using Machine Learning

Chaimaa Boukhatem
Department of Electrical Engineering
University of Sharjah
U18102614@sharjah.ac.ae

Heba Yahia Youssef
Department of Electrical Engineering
University of Sharjah
U18105651@sharjah.ac.ae

Ali Bou Nassif
Department of Computer Engineering
University of Sharjah
anassif@sharjah.ac.ae

2022 Advances in Science and Engineering Technology International Conferences (ASET) | 978-1-6654-1801-0/22/$31.00 ©2022 IEEE | DOI: 10.1109/ASET53988.2022.9734880

***Abstract*—Cardiovascular disease refers to any critical condition that impacts the heart. Because heart diseases can be life-threatening, researchers are focusing on designing smart systems to accurately diagnose them based on electronic health data, with the aid of machine learning algorithms. This work presents several machine learning approaches for predicting heart diseases, using data of major health factors from patients. The paper demonstrated four classification methods: Multilayer Perceptron (MLP), Support Vector Machine (SVM), Random Forest (RF), and Naïve Bayes (NB), to build the prediction models. Data preprocessing and feature selection steps were done before building the models. The models were evaluated based on the accuracy, precision, recall, and F1-score. The SVM model performed best with 91.67% accuracy.**

Keywords—heart disease prediction, machine learning, support vector machine, multilayer perceptron, naïve bayes, random forest.

I. Introduction

Cardiovascular Disease (CVD), commonly referred to as heart disease, encompasses a wide range of conditions that affect the heart, with the two most common conditions being ischemic heart diseases and strokes. The World Health Organization lists the most significant behavioural risk factors for CVD as maintaining an unhealthy diet, a sedentary lifestyle, tobacco use, and excessive consumption of alcohol. Prolonged exposure to these risk factors can present itself as an initial sign of CVD, which include elevated blood pressure, elevated blood glucose, raised blood lipids, and obesity. Warning signs listed by the American Heart Association include having one or more of the following: shortness of breath, persistent coughing or wheezing, swelling of the ankles and feet, constant fatigue, lack of appetite, and impaired thinking [1]. Moreover, Coronavirus may cause heart disease [2]–[4]. Efficient early diagnosis can substantially reduce the risk and global burden of CVD by initiating treatment rapidly to prevent further health deterioration. Thus, there is an urgent need to develop machine learning models that can predict the probability of developing CVD depending on the risk factors present.

Recently, machine learning models have successfully lent a hand in diverse cases in the medical field [5]. They have been effective in analyzing, evaluating, and predicting different medical conditions [6]. In this paper, we are proposing a machine learning approach to predict the presence of cardiovascular diseases in patients based on major health data.

This paper is organized as follow: Section II covers the related works where machine learning was used for heart disease prediction. Section III explains the methodology, where the dataset is described, preprocessed, and split. As well as the applied algorithms and the corresponding model design parameters, the evaluation metrics selected to evaluate the performance of the model are described. Section IV discusses the experimental results. Finally, in Section V, the remarks and conclusions about this work are presented.

II. Related Work

Heart disease prediction was addressed in the literature using several methods. In [7], Naïve Bayes, SVM, and Functional Trees were used to predict the possibility of heart diseases with an accuracy of 84.5%, using measurements from wearable mobile technologies with the same inputs used in our work. Furthermore, Naïve Bayes was solely used in [8] with a slightly better accuracy of 86.4%, using the same dataset.

Another work [9] used several algorithms; Logistic Regression, KNN, NN, SVM, NB, Decision Tree, and RF, with three feature selection algorithms: Relief, mRMR, and LASSO to predict the existence of heart disease with the same dataset used in this work. The Logistic Regression algorithm had the best performance and yielded predictions with an accuracy as high as 89%.

Moreover, a work done in 2020 [10] applied 4 algorithms with a very high accuracy of 90.8% for the KNN model, and minimum accuracy of 80.3% for the other models.

In [11], a hybrid Random Forest and Naïve Bayes model achieved an accuracy of 84.16% using 10 features, which were selected using Recursive Feature Elimination and Gain Ratio algorithms.

In a recent work done in 2021 [12], Logistic Regression, Random Forest, and KNN were used for the prediction. The maximum accuracy was 87.5%.

All the previous is very promising for the future of heart diseases and failure prediction, especially with the current advances in portable electronic measurement devices.

III. Methodology

A. Data Collection

The dataset was collected from Kaggle [13]. The dataset contains a total of 303 instances with 13 attributes as described in Table I.

TABLE I. HEART DISEASE DATASET DESCRIPTION

Data element	Description	Type	Range	Remarks
Age	-	Num[a]	29-77	Average is 54.37
Sex	-	Bi[b]	0: Female 1: Male	32% Female 68% Male
Cp	Chest pain level	Nom[c]	0/1/2/3 0: Asymptotic 2: non-anginal pain 3: Typical angina	Majority have 0 pain
Trestbps	Rest blood pressure	Num	94-200	Average is 131.6
Chol	Cholesterol level	Num	126-564	Average is 246.3
Fbs	Fasting blood sugar level	Bi	0: Level below 120 1: Level above 120	-
Restecg	Resting electrocardiographic results	Nom	0/1/2 0: Showing probable or definite left ventricular hypertrophy. 2: Abnormal	-
Thalach	Maximum heart rate achieved	Num	71-202	-
Exang	Exercise induced angina	Bi	0: None 1: Produced	-
Oldpeak	ST depression induced by exercise relative to rest	Num	0-6.2	Right skewed data, majority of population is between 0 and 0.5
Slope	The slope of the peak exercise ST segment	Nom	0: Unsloping 1: Flat 2: Down-sloping	-
Ca	Number of major vessels	Nom	0/1/2/3/4	-
Thal	Defect type	Nom	1: Fixed defect 2: Normal 3: Reversable defect	There is one outlier of category 0
Target	Diagnosis of heart disease	Bi	0: No disease 1: Disease	-

[a] Numerical, [b] Binary, [c] Nominal.

B. Data Preprocessing

The performance of a machine learning model is greatly determined by the quality of the data used to build it, which makes data preprocessing very important. Data preprocessing includes cleaning the data by removing corrupted or missing data points and outliers, in addition to transforming the data, resampling it, and applying feature selection.

1) Data Visualization and Cleaning

First, we checked for missing values and none were found. Second, we checked for outliers and we found some as reported in Table II.

TABLE II. LIST OF OUTLIERS

Attributes	Outlier values
Age	None
Chol	417, 564, 394, 407, 409
Trestbps	172, 178, 180,180, 200, 174, 192, 178, 180
Thalach	71
Oldpeak	4.2, 6.2, 5.6, 4.2, 4.4

Because the mild outliers contribute to the final diagnosis, only the extreme outliers were removed. The extreme outliers were detected using (1) & (2), where the IQR is the interquartile range, and is a measure for the dispersion of the data, and Q_1, Q_3 are the lower and upper quartiles respectively.

$$(75\% \times Q_3) + 3 \times \text{IQR} \quad (1)$$

$$(25\% \times Q_1) - 3 \times \text{IQR} \quad (2)$$

The data points that are greater than the first expression were removed. Similarly, the data points that are less than the second expression were removed. As a result, 2 out of the 303 instances were removed.

Then, the correlation coefficient matrix was obtained to observe the relation between the different attributes and the output. Fig. 1. illustrates the correlation matrix where the coefficient indicates both the strength of the relationship between the variables as well as the direction (whether it is a positive or negative correlation).

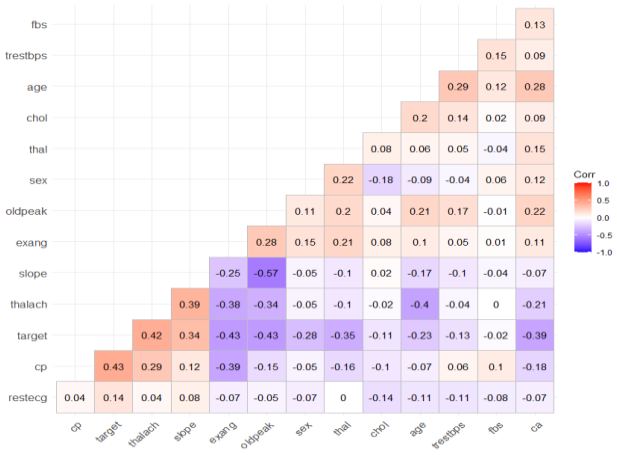


Fig. 1. Correlation coefficient matrix

2) Checking for Imbalances

Imbalance in the output can distort the prediction accuracy. Therefore, the balance of the output "target" was verified as shown in Figure 2. After inspection, the data turned out to be balanced with a 9:11 ratio between the two categories. Thus, there was no need to resample the data.

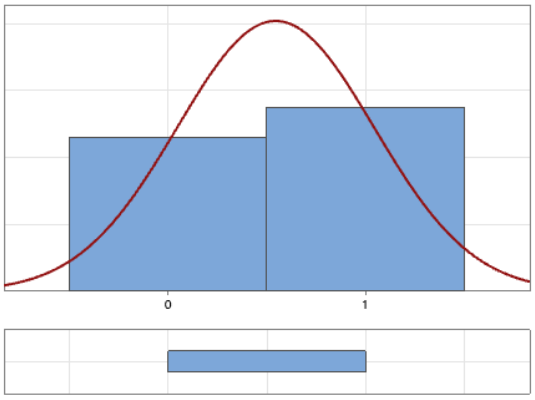


Fig. 2. Histogram and boxplot of the output "target"

3) Data Transformation

Transformation is applied when the dataset includes data of different formats, or when different datasets are combined. In this case, the nominal features were transformed into factors, for them to be used in Rstudio.

4) Dimensionality Reduction

In machine learning, dimensionality reduction refers to the process of reducing the number of features to decrease the complexity and prevent overfitting, by either feature selection or extraction.

Feature selection is done by selecting a subset of features from the original set, and is done by methods such as CFS (Correlation-based Feature Selection), Chi-squared test and ridge regression. In this paper, the feature selection method used was CfsSubsetEval, which evaluates the worth of a subset of the attribute by considering both the individual predictive ability of each feature, and the degree of redundancy between them. Weka software was used for feature selection as it has several options of attributes evaluator to test and use.

Slightly different than feature selection, feature extraction is when a new set of features is generated from the original set. Principal Component Analysis (PCA) is widely used. It calculates the projection of the original data into a smaller dimension space.

5) Data Splitting

In machine learning, the data is usually split into training and testing sets, where the training set is used to train the model, and the testing set is to test it and predict the output. Hold-out was used in this work with 80% of the data used in training and 20% used for testing.

C. Applied Algorithms

1) Naïve Bayes (NB)

Naïve Bayes is a supervised learning algorithm, that is based on the Bayes Theorem, and assumes that all features are independent and have equal contribution to the target class. Bayes' theorem calculates the posterior probability of an event A, given some prior probability of event B, as in (3).

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \quad (3)$$

2) Random Forest (RF)

RF is also a supervised machine learning algorithm, used for both classification and regression. It utilizes ensemble learning, which is a technique that combines several classifiers to make accurate predictions in complex situations. RF algorithms establish the prediction based on the results of multiple decision trees through bagging or bootstrap aggregation as shown in Figure 3.

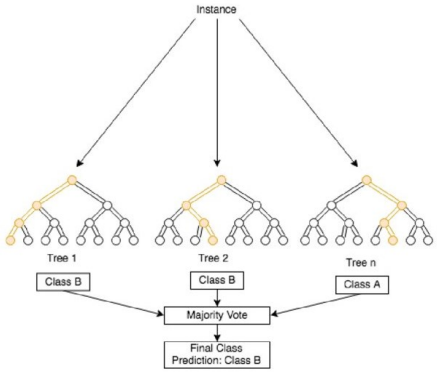


Fig. 3. Random Forest prediction method [14]

When training the model in RStudio, the following parameter were fixed: 500 decision trees, 3 variables tries at each split, in classification method.

3) Neural Networks (NN)

Neural Networks are universal approximators that can map any relation between the inputs and outputs of a system, regardless of its complexity as shown in Figure 4. Their working principle is mimicked from the human brain, where during training, they assign weights (w) to each input to indicate its significance to the output.

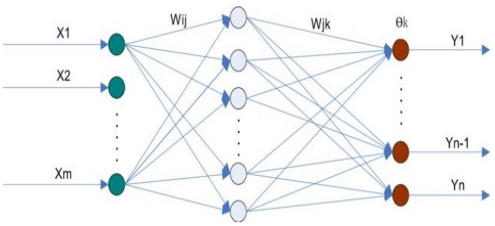


Fig. 4. Neural Network Diagram [15]

Each node is called a neuron and has its own activation function. The number or neurons, network layers, and activation functions, all depend on the application, and influence the performance of the model.

In this work, the MLP network was chosen to have 5 hidden nodes with sigmoid activation function.

4) Support Vector Machine (SVM)

SVM is a supervised learning method used for classification, regression, and outlier detection. It seeks to establish a decision boundary between different classes, to label prediction using one or more feature vectors as shown in Figure 5. This decision boundary, known as the hyperplane, is oriented to be as far away as possible from the nearest data points. Those nearest locations are referred to as support vectors.

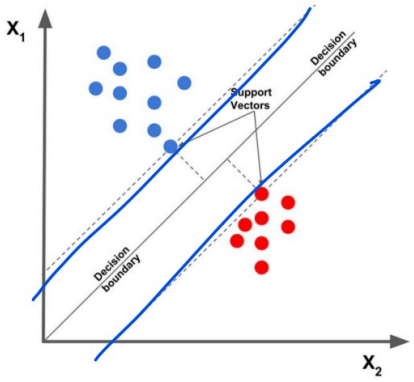


Fig. 5. SVM hyperplane separation of data points [16]

D. Evaluation Metrics

Evaluation metrics are used to test the quality and performance of the machine learning model. In this paper, the best model was chosen based on the following evaluation metrics:

Confusion Matrix: An N*N matrix, where N is the number of classes being predicted. For a prediction problem of two possible outputs, like in this work, the confusion matrix dimension is 2*2.

		Predicted class	
		P	N
Actual class	P	True Positive (TP)	False Negative (FN)
	N	False Positive (FP)	True Negative (TN)

The elements of the matrix are the counts of the correct and incorrect predictions, separated by class. For example, a True Positive is the number of the correctly classified Positive class (in this case, the number of correctly diagnosed heart diseases). Similarly, a True Negative is the number of correctly classified Negative class (In this case, the count of correctly predicted absence of heart disease).

Accuracy: The percentage of the total number of predictions that were classified correctly, and is obtained from the confusion matrix by the following equation:

$$A = \frac{TP + TN}{TP + TN + FP + FN}$$

Precision: The percentage of the positive cases that were classified correctly, and is obtained from the confusion matrix by the following equation:

$$PR = \frac{TP}{TP + FP}$$

Sensitivity or Recall: The percentage of the actual positive cases that were classified correctly, and is obtained from the confusion matrix by the following equation:

$$RE = \frac{TP}{TP + FN}$$

F1 Score: If the target is to get the best precision and recall, F measure would be the best choice as it provides a harmonic mean of the recall and the precision values in classification problem, and is obtained from the confusion matrix by the following equation:

$$F1 = \frac{2TP}{2TP + FP + FN}$$

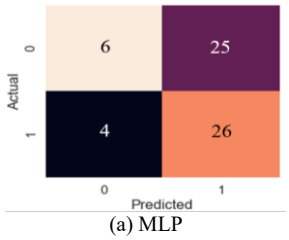
IV. Results and Discussion

The four selected machine learning techniques were used to build the heart disease prediction model, and the results were obtained in three different stages to reach to the best final model. In the first stage, the prediction was done without data cleaning. Whereas in the second stage, the prediction was done after data cleaning. And to enhance the performance further, the last stage predictions were done after applying feature selection.

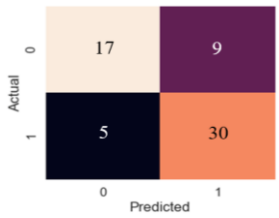
The results of each model before data cleaning, based on the confusion matrix shown in Fig. 6, are listed in Table III.

TABLE III. Results of the Predicted Model Before Removing Extreme Outliers

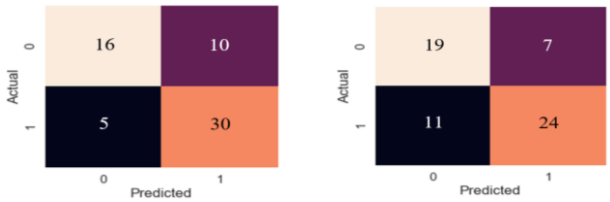
Metric \ Model	MLP	SVM	RF	NB
Accuracy	52.46 %	75.41%	77.05%	70.49%
Precision	60%	76.19%	77.27%	63.33%
Recall	19.36%	61.54%	65.38%	73.08%
F1 Score	29.27%	68.08%	70.82%	67.85%



(a) MLP



(b) RF



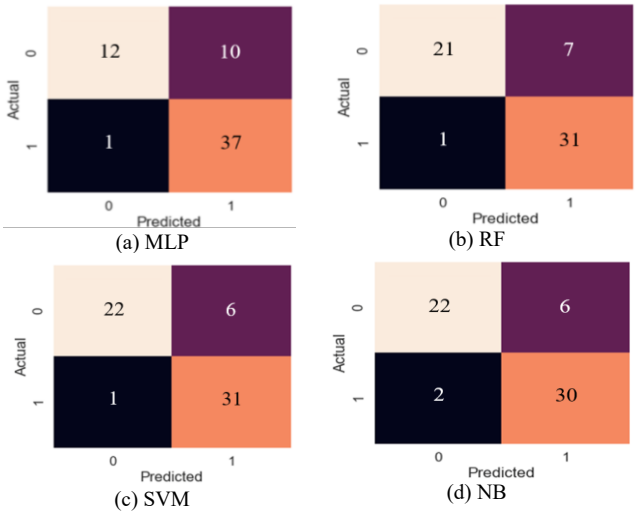
(c) SVM (d) NB

Fig. 6. Confusion matrices before removing the outliers

Table IV shows the results of each model after removing the extreme outliers. The metrics are calculated using RStudio based on the confusion matrix shown in Fig. 7.

TABLE IV. RESULTS OF THE PREDICTED MODEL AFTER REMOVING EXTREME OUTLIERS

Metric \ Model	MLP	SVM	RF	NB
Accuracy	81.67%	88.33%	86.67%	86.67%
Precision	92.31%	95.65%	95.45%	91.67%
Recall	54.55%	78.57%	75%	78.57%
F1 Score	68.67%	86.27%	83.99%	84.61%



(a) MLP (b) RF

(c) SVM (d) NB

Fig. 7. Confusion matrices after removing the outliers

The comparison between Table III and IV clearly shows the improvement in prediction accuracy due to data cleaning. In both cases, the SVM performed best according to all evaluation metrices with an accuracy of 88.33%, precision of 95.65%, recall of 78.57% and F1 score of 86.27%.

The result shown in Fig. 8. were obtained using WEKA, after feature selection was applied to improve the performance of the models.

For feature extraction, PCA was used for the nominal components in the dataset. The maximum proportion of variance between the features was 36%, which means that about 1 third of the information in the variables can be encapsulated by just that one Principal Component. However, since this ratio is not very large, the variables were kept without modification.

```
Selected attributes: 3,8,9,10,11,12,13 : 7
                     cp
                     thalach
                     exang
                     oldpeak
                     slope
                     ca
                     thal
```

Fig. 8. Selected attributes based on WEKA software

According to the previous evaluation, the attributes Cp, Thalach, Exang, Oldpeak, Slope, Ca, Thal, were chosen to be the most important inputs. Hence, Age, Sex, Fbs, Chol, Restecg, Trestbps are the least important attributes. However, the decision was to not remove all the least important, but the three least significant (Fbs, Restecg, Sex), which are indicated in RStudio as shown in Fig. 9.

```
          MeanDecreaseGini
age               9.444118
sex               4.471178
cp               15.819341
trestbps          7.494398
chol              9.047356
fbs               1.136530
restecg           2.039839
thalach          14.055739
exang            10.180642
oldpeak          11.185531
slope             4.612805
ca               16.174579
thal             13.247638
```

Fig. 9. Level of Importance

After removing the unimportant attributes, the SVM model was retrained using the new data. Fig. 10. shows the confusion matrix of the retrained model and Table V illustrates the final results that were obtained from the confusion matrix.

TABLE V. RESULTS OF THE PREDICTED MODEL AFTER APPLYING FEATURE SELECTION

Model \ Metric	Accuracy	Precision	Recall	F1 Score
SVM	91.67%	92.31%	88.89%	90.56%

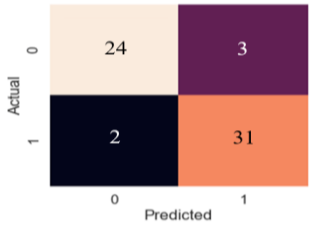


Fig. 10. Confusion matrix of the SVM model after feature selection

V. CONCLUSION AND FUTURE WORK

This work aims to predict the existence of heart disease in patients according to specific health measurements.

The paper demonstrated 4 classification mechanism to build the prediction model. The data was collected and cleaned from any missing values and extreme outliers. In addition, it was preprocessed to fit the model requirements, where it went into different phases of visualizing the imbalances, obtaining the correlation matrix, using

dimensionality reduction techniques, and finally splitting using Hold-out. The model was trained and tested for each machine learning algorithm. SVM algorithm with linear kernel had the best results with a 91.67% accuracy, 92.31% precision, 88.89% recall, and F1 Score of 90.56%. The algorithms used were able to extract the complex relations between the symptoms and the disease. Machine learning algorithms can also be applied to other types of diseases, especially with the generation of more accurate datasets in the medical field in the future.

This work can be enhanced by applying more extensive data analysis and trying additional algorithms to reach the maximum possible accuracy.

REFERENCES

[1] S. Rehman, E. Rehman, M. Ikram, and Z. Jianglin, "Cardiovascular disease (CVD): assessment, prediction and policy implications," *BMC Public Health*, vol. 21, no. 1, p. 1299, 2021, doi: 10.1186/s12889-021-11334-2.

[2] O. Atef, A. B. Nassif, M. A. Talib, and Q. Nassir, "Death/Recovery Prediction for Covid-19 Patients using Machine Learning," 2020.

[3] A. B. Nassif, I. Shahin, M. Bader, A. Hassan, and N. Werghi, "COVID-19 Detection Systems Using Deep-Learning Algorithms Based on Speech and Image Data," *Mathematics*, 2022.

[4] H. Hijazi, M. Abu Talib, A. Hasasneh, A. Bou Nassif, N. Ahmed, and Q. Nasir, "Wearable Devices, Smartphones, and Interpretable Artificial Intelligence in Combating COVID-19," *Sensors*, vol. 21, no. 24, 2021, doi: 10.3390/s21248424.

[5] O. T. Ali, A. B. Nassif, and L. F. Capretz, "Business intelligence solutions in healthcare a case study: Transforming OLTP system to BI solution," in *2013 3rd International Conference on Communications and Information Technology, ICCIT 2013*, 2013, pp. 209–214, doi: 10.1109/ICCITechnology.2013.6579551.

[6] A. Nassif, O. Mahdi, Q. Nasir, M. Abu Talib, and M. Azzeh, "Machine Learning Classifications of Coronary Artery Disease." Jan. 2018.

[7] A. F. Otoom, E. E. Abdallah, Y. Kilani, A. Kefaye, and M. Ashour, "Effective diagnosis and monitoring of heart disease," *Int. J. Softw. Eng. its Appl.*, vol. 9, no. 1, pp. 143–156, 2015, doi: 10.14257/IJSEIA.2015.9.1.12.

[8] K. Vembandasamyp, R. R. Sasipriyap, and E. Deepap, "Heart Diseases Detection Using Naive Bayes Algorithm," *IJISET-International J. Innov. Sci. Eng. Technol.*, vol. 2, no. 9, 2015, Accessed: Dec. 11, 2021. [Online]. Available: www.ijiset.com.

[9] A. U. Haq, J. P. Li, M. H. Memon, S. Nazir, R. Sun, and I. Garciá-Magarinõ, "A hybrid intelligent system framework for the prediction of heart disease using machine learning algorithms," *Mob. Inf. Syst.*, vol. 2018, 2018, doi: 10.1155/2018/3860146.

[10] D. Shah, S. Patel, · Santosh, and K. Bharti, "Heart Disease Prediction using Machine Learning Techniques," vol. 1, p. 345, 2020, doi: 10.1007/s42979-020-00365-y.

[11] K. Pahwa and R. Kumar, "Prediction of heart disease using hybrid technique for selecting features," *2017 4th IEEE Uttar Pradesh Sect. Int. Conf. Electr. Comput. Electron. UPCON 2017*, vol. 2018-January, pp. 500–504, Jun. 2017, doi: 10.1109/UPCON.2017.8251100.

[12] H. Jindal, S. Agrawal, R. Khera, R. Jain, and P. Nagrath, "Heart disease prediction using machine learning algorithms," doi: 10.1088/1757-899X/1022/1/012072.

[13] "Heart Disease UCI | Kaggle." https://www.kaggle.com/ronitf/heart-disease-uci (accessed Jan. 10, 2022).

[14] D. Murphy, "Using Random Forest Machine Learning Methods to Identify Spatiotemporal Patterns of Cheatgrass Invasion through Landsat Land Cover Classification in the Great Basin from 1984 - 2011," 2019.

[15] S. Liu, Z. Fang, and L. Zhang, "Research on Urban Short-term Traffic Flow Forecasting Model," *J. Phys. Conf. Ser.*, vol. 1237, no. 5, Jul. 2019, doi: 10.1088/1742-6596/1237/5/052026.

[16] "Support Vector Machines (SVM) | LearnOpenCV #." https://learnopencv.com/support-vector-machines-svm/ (accessed Jan. 10, 2022).