Prediction and Prevention of Heart Disease Using Interpretable Machine Learning

Pranit Deshpande [1] **Siddhesh Bhande**[2] **Harika Tamma**[3] **Aksheetha Muthunooru**[4]
University of Maryland Baltimore County
Maryland
{pranitd1,sbhande1,ci75173,me52139}@umbc.edu

Abstract

The greatest cause of death globally is coronary heart disease, which is a kind of cardiovascular disease (CVD). If it is identified or diagnosed early on, the prognosis is favorable. The classification of data sets related to coronary heart disease using machine learning (ML) techniques is covered in the current work. The current work developed and examined a number of categorization models based on machine learning. To address unbalanced classes, the dataset was manually resampled, and feature selection techniques were used to evaluate the influence on two different performance metrics. The findings demonstrate that, when compared to the other methods used, the gradient boosting classifier obtained the highest performance score on the original dataset. In the second part i.e., of prevention, LIME (local interpretable model-agnostic explanations) is used to find the most important feature for that instance. In conclusion, Cleveland clinic website is used to retrieve the precautionary steps by using the most important feature.

1 Introduction

The World Health Organization estimates that 17.9 million deaths worldwide in 2019 were caused by cardiovascular the diseases (CVDs), or 32% of all fatalities. Cardiovascular diseases (CVDs) include coronary heart disease, cerebrovascular disease, peripheral arterial disease, congenital heart disease, and other conditions that affect the heart and blood vessels. The prevalence of coronary heart disease (CHD) has been rising over time despite breakthroughs in healthcare and medical research, and scientists from all over the world are striving to determine the variables that are linked to increased risk of developing coronary heart disease in the future (CHD).

The current methods used by doctors to anticipate and diagnose heart disease are largely based on an evaluation of the patient's medical history, symptoms, and the patient's physical reports. While they may sometimes forecast a patient's heart problems with up to 67% accuracy, medical professionals typically struggle to do so. In order to aid in the decision-making process, the medical sector needs an automated intelligence system that can reliably forecast cardiac disease. The vast amount of patient data that is currently available in the medical field, along with machine learning or deep learning algorithms, and intelligent decision-making systems, can be used to achieve this. If appropriately handled, the big data contained in healthcare database repositories will aid in lowering the prevalence of certain of these disorders. This will make it simpler and faster for doctors to diagnose patients and predict diseases.

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

2 Related Work

In the last few years, there has been a significant amount of study into machine learning and deep learning algorithms in a number of industries, including transportation (12), medicine (6), and other physical sciences (13). Different modeling approaches have been used by certain researchers to develop classification models for the forecasting of coronary heart disorders. A Naive Bayes, Neural Network, and Decision Trees model was proposed by Sellappan Palaniappan et al. to develop the Intelligent Heart Disease Prediction System (IHDPS) (14). To examine the Framingham dataset, other researchers utilized the Random Forest (RF), K-Nearest Neighbor (KNN), Support Vector Machine (SVM), and Logistic Regression (LR) methods (15). The RF approach outperformed all of these strategies when they were put side by side. Some writers employed machine learning (ML) techniques on the well-known Cleveland heart dataset and Statlog data sets to forecast heart illnesses based on some key variables. Results showed that the LR and NB classifiers performed better on the Statlog dataset, whereas the RF and SVM with grid search algorithms performed better on the Cleveland dataset (4).

3 Method

3.1 Data set

The coronary heart disease dataset (Framingham dataset), which is available for download from University of Washington's Bio-Statistics Class, was used in this study. This data collection is derived from current cardiovascular research of people living in Framingham, Massachusetts. To determine whether the patient will experience coronary heart disease during the following ten years is the classification goal (CHD). Information about the patients is included in the dataset. There are a total of 16 qualities and approximately 4,000 records. Each trait has the potential to be risky. Risk factors include those related to demographics, behavior, and health.

Category	Description
Demographic	*Sex:* male or female (Nominal) *Age:* Age of the patient (Continuous -Although the recorded ages have been truncated to whole numbers, the concept of age is continuous) *Education:* Educational background of the patient ranked from 1 to 4 (continuous)
Behavioral	*Current Smoker:* whether or not the patient is a current smoker (Nominal) *Cigs Per Day:* the number of cigarettes that the person smoked on average in one day. (can be considered continuous as one can have any number of cigarettes, even half a cigarette.)
Medical History	*BP Meds::* whether or not the patient was on blood pressure medication (Nominal) *Prevalent Stroke:*whether or not the patient had previously had a stroke (Nominal) *Prevalent Hyp:*whether or not the patient was hypertensive (Nominal) *Diabetes:*whether or not the patient had diabetes (Nominal)
Medical Current	*Tot Chol:*total cholesterol level (Continuous) *Sys BP:*systolic blood pressure (Continuous) *BMI:*Body Mass Index (Continuous) *Heart Rate:*heart rate (Continuous - In medical research, variables such as heart rate though in fact discrete, yet are considered continuous because of large number of possible values.) *Glucose:*glucose level (Continuous)
Predict Variable	10-year risk of coronary heart disease CHD (binary: "1", means "Yes", "0" means "No")

The distribution of the target variable which is shown in Figure 1.

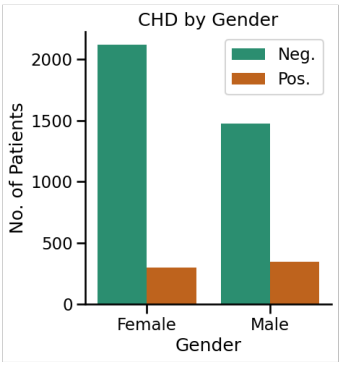

Figure 1: Count Outcome Heart Disease with respect to Gender - The target data which is a 10-year risk of coronary heart disease CHD has 3645 samples of class 0 and 617 samples of class 1. Therefore, this makes the dataset highly imbalance.

The class wise distribution of class 0 (Negative) and 1 (Positive) is shown in Figure 2.

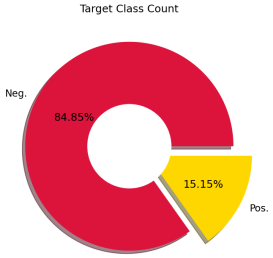

Figure 2: Class Wise Percentage - This pie chart depicts that the dataset contains almost 85% of negative samples and 15% of positive samples.

3.2 Proposed Architecture

The repository is accessed first, and the data is then analyzed. We search the data for each characteristic's range and any probable examples of missing values. Outliers are removed and missing values are located and replaced with that column's median after the categorical characteristics, which had previously been coded, are detected. Re sampling approach is used to balance the data before further feature engineering. 5 different varieties of machine learning algorithms are used. The dataset was used to train algorithms utilizing 3 Fold Cross Fold Validation, RandomizedSearchCV, and GridSearchCV techniques. The output is determined by the top-performing algorithm, and LIME is then utilized to determine the most crucial feature. Finally, web crawling is used to get data from the internet by most important feature found out in the earlier step.

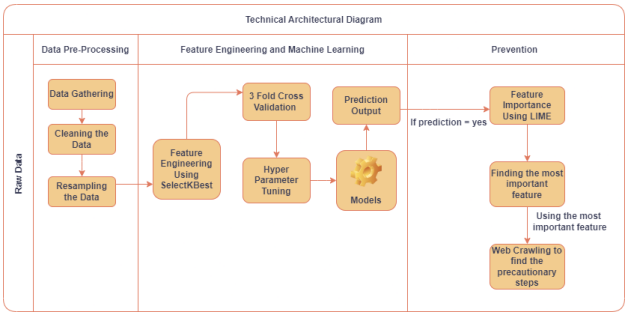

Figure 3: Technical Architecture Describing End to End Process

3.3 Data Pre-Processing

The first step in the pre-processing of raw data was data visualization using descriptive statistics tables, skewness, and other descriptions like min, max, percentile values, and mean. It also entails turning categorical data (such as the values in the sex, Current Smoker, and Diabetes columns) into numbers and identifying and removing any missing values. The missing values for cigsPerDay, totalChol, BMI, glucose, and heart rate were filled in using the medians of each column.

Further, outliers are removed from the data from 2 different columns "Tot Chol" and "Sys BP" by finding the maximum and minimum from the data columns. The dataset is further segregated based on the age group. The age group under the value 40 is regarded as Adults; the age group between 40 and 55 is regarded as Middle Aged; and, finally, the age group over 55 is regarded as Senior.

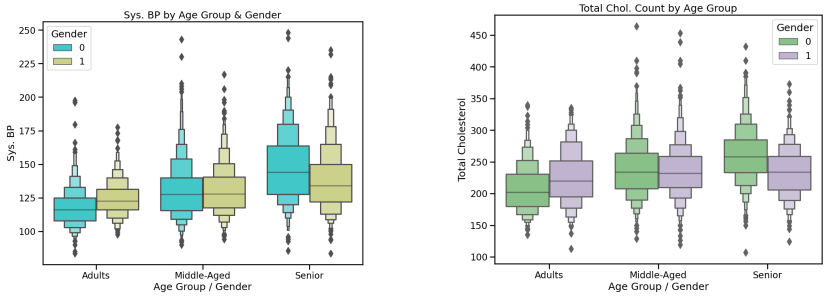

Figure 4: Plot showing outliers - This plot depicts that there are a few outliers in Sys BP and Total Chol columns of the dataset.

The correlation matrix illustrates how the features are related to one another or to the target variable. Further analysis of the data also shows that the column education is negatively correlated with respect to the output variable and it does not provide any significant value to the model, hence it has been removed from the column.

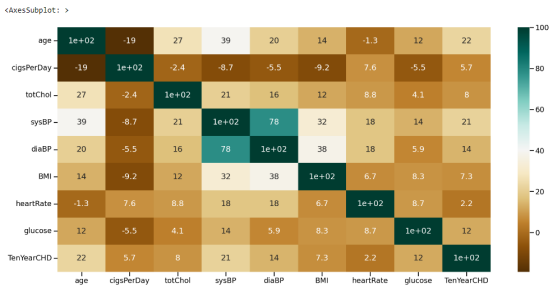

Figure 5: Correlation Matrix of All Numerical Features.
The correlation heat map shows that Sys BP and dia BP are highly correlated

The original dataset was imbalanced in nature and hence the re sampling technique is used, which increased the number of cases in the dataset in a balanced way. That is, it made new instances from minority instances that already existed. The overall findings for the output of the data re sampling is shown in Fig 6 below.

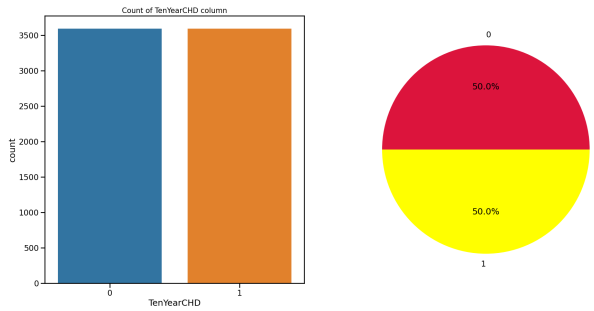

Figure 6: Count Outcome of Heart Disease After Re sampling. This can be seen as an improvement with previous data sampling as shown in figure 2

3.4 Feature Engineering

A feature selection method based on filters was applied to the dataset. It uses statistical measurements to score the correlation or dependence between the input variables, which can then be filtered to select the most important features. In particular, the SelectKBest feature selection approach was taken into account when using the Mutual Information Feature Selection feature extraction or selection technique. When the value of the second variable is known, it determines the relationship between the first and second variables and quantifies the uncertainty reduction for the first variable (16). Figure 7 show the pertinent elements' scores and plots, respectively.

	Specs	Score
1	age	8570.915050
3	cigsPerDay	3881.387188
9	sysBP	3645.003649
8	totChol	2349.933813
13	glucose	901.802363
10	diaBP	387.371066
6	prevalentHyp	358.977899
2	currentSmoker	154.764025
4	BPMeds	112.701283
7	diabetes	45.003967
5	prevalentStroke	43.559786
11	BMI	34.701806
12	heartRate	28.705739
0	male	2.250710

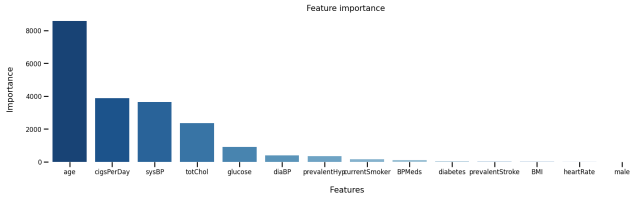

Figure 7: SelectKBest Plots. Out all the columns of the data set, according to the figure, top 12 columns are selected as features for further processing.

4 Model Development

The creation of a model using various machine learning techniques is the last step in the prediction section. k-nearest neighbor (KNN), support vector machines (SVM), decision tree (CART), logistic regression (LR), Gradient Boosting (GBC), and random forest (RF) are some of the methods used in this research. Following feature selection, the dataset was split into training and test sets (75:25), and the a fore mentioned algorithms were used to train the training sets. To improve training and testing hyper parameter tuning has been performed using RandomisedSearchCV and GridSearchCV techniques. The distribution of data samples in the raw data was uneven, as was previously indicated in the data analysis. This may affect the accuracy and significance of a feature in ML models. In order to enhance this distribution and increase the number of records, the original dataset was resampled.

4.1 Comparison between different models implemented

During model training, our main objective was to select a model based on its accuracy and its diagnostic ability which is determined by ROC Curves. The accuracy of the first model we tried i.e., logistic regression was good for the raw data but had a very back ROC curve as compared to that of resampled data with less accuracy , which determines that the model has a high diagnostic ability after re sampling of the data. Similarly in the SVM classifier, the accuracy of the data after re sampling decreased. We found that this was due to the uneven distribution of the data, the model can tilt more toward the majority side after looking at the confusion matrix.
The tree-based Random Forest Classifier was the next model we examined. We anticipated performing three fitting folds for each of the 100 candidates in the randomized search CV, for a total of 300 fitting folds. In this case, re sampling greatly improved accuracy with a higher diagnostic score. Next, we attempted to train using K-Nearest Neighbor KNN, one of the most simple machine learning algorithms. However, this method had the issue of taking a long time and requiring a lot of memory during training. Hence this algorithm was not used in the further processing. The Gradient Boosting Classifier, which updates the classifiers while utilizing weighted minimization and weighted inputs in addition to the Ada Boosting method, was then used. Here, the accuracy using the resampled data significantly outperformed all the models which were implemented and it also had the maximum ROC score. We also tested decision tree learning because we found that tree-based models had the highest accuracy. Although the accuracy was higher, we discovered that the Gradient Boosting Classifier performed the best in terms of accuracy as well as ROC scores.

4.2 Results

Before Resampling the Data

Logistic Regression	Random Forest	K-Nearest Neighbor	Gradient Boosting Classifier	Support Vector Classifier	Decision Tree
85%	84%	84%	76%	85%	83%

Table 1: Results table (Before Re sampling)

After Resampling the Data

Logistic Regression	Random Forest	K-Nearest Neighbor	Gradient Boosting Classifier	Support Vector Classifier	Decision Tree
67%	96%	91%	98%	71%	91%

Table 2: Results table (After Resampling)

The use of ROC Receiver Operating Characteristics analysis as a tool for evaluating the performance of classification models in machine learning has been increasing in the last decade. The below figure depicts the ROC Curves for all the 5 different machine learning algorithms implemented on the processed dataset and it depicts that Gradient Boosting Classifier outperformed the other models with overall accuracy of roughly 98%.

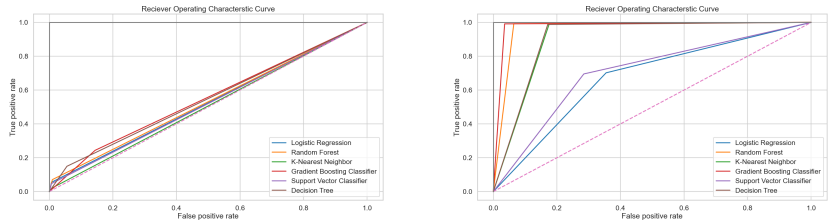

Figure 8: ROC Curves

5 System - Oriented Interpretation of Machine Learning Models

5.1 SHAP

The SHAP framework finds the class of additive feature importance methods, which also includes the six earlier methods, and demonstrates that this class contains a single solution that complies to the desired qualities.(17) The central notion of Shapley value based explanations of machine learning models is to distribute credit for a model's output among its input features using fair allocation findings from cooperative game theory.

Figure 9: SHAP using one instance

The feature values are compared to the base values in the model. Here, the anticipated value is 0.05, and the base value is 0.1521. In contrast to the reasons why the prediction was reduced in blue, the features presented in pink are the features that increase the prediction, and the bar shows the impact.

Figure 10: Testing 200 Instances using SHAP

5.2 LIME

LIME determines the assumptions behind the interpretive representation of these sample points, then creates a linear weight by minimizing complexity and loss. method of spatial definition. Each prediction is interpreted by LIME by reading the local translation model. The idea underlying LIME is that an interpretive manifestation of the initial input is a sampling of occurrences from nearby and distant regions. After that, LIME finds the assumptions behind the interpretive representation of these sample points and creates a linear weight by cutting complexity and loss. The samples' weight is determined by how far they are from their starting point. As points are extended, their weight decreases. The description serves as a forecast for local occurrences because it is trustworthy locally.
Lime Explanations:

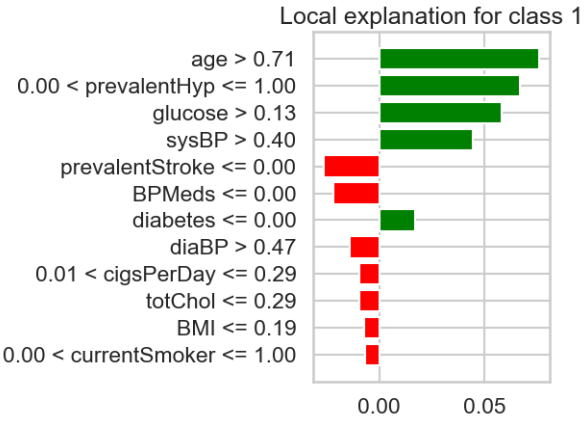

Figure 11: LIME using Gradient Boosting Classifier

We have used LIME in the Gradient Boosting Algorithm. The algorithm predicts that the instance belongs to Class 1 (Positive) with most important features contributing to it being PrevalentHyp, age, sysBP,glucose, in the descending order.

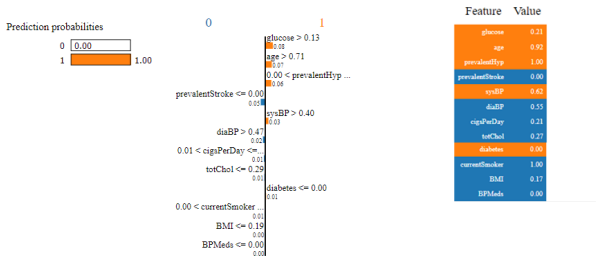

Figure 12: Lime Using Gradient Boosting Classifier

In the above figure 12 depicts that there is 100% probability that the instance belongs to Class 1 (Positive) data sample with PrevalentHyp i.e., Hypertension being the most important contributing feature.

6 Predicting Remedies for the predicted Cause

Once we identified the features that had the greatest impact on the prediction, we can infer that these are the factors that contribute to the person's poor heart health. The next step was to identify the appropriate treatments that can be utilized to lower the risk of a heart attack.
We used https://my.clevelandclinic.org/health/diseases/ to find treatment recommendations for heart attack-related conditions. Once indicated, these cures can also be confirmed by the patient's doctors. The following websites in the below figure 13 were mapped as being utilized for web crawling for comparable causes:

```
disease_mapping = {'Age' : 'https://my.clevelandclinic.org/health/diseases/16891-heart-disease-adult-congenital-heart-disease' ,
                   'BP Medication' : 'https://my.clevelandclinic.org/health/diseases/4314-hypertension-high-blood-pressure' ,
                   'Hypertension' : 'https://my.clevelandclinic.org/health/diseases/4314-hypertension-high-blood-pressure' ,
                   'Diabetes' : 'https://my.clevelandclinic.org/health/diseases/9812-diabetes-and-stroke' ,
                   'Cholesterol' : 'https://my.clevelandclinic.org/health/diseases/21656-hyperlipidemia' ,
                   'Systolic Blood Pressure' : 'https://my.clevelandclinic.org/health/diseases/4314-high-blood-pressure' ,
                   'Diastolic Blood Pressure' : 'https://my.clevelandclinic.org/health/diseases/4314-high-blood-pressure' ,
                    'Stroke' : 'https://my.clevelandclinic.org/health/diseases/5601-stroke' ,
                   'Glucose' : 'https://my.clevelandclinic.org/health/diseases/9815-hyperglycemia-high-blood-sugar'}
```

Figure 13: Websites used for data retrieval

The relevant solutions are then suggested using the crawled content. Due to a paucity of data studies, we predict that the proper precision of heart disease prevention cannot be determined. However, it might be covered in upcoming research in the same field.

7 Conclusion

The present study developed multiple classification models for the prediction of coronary heart disease using different machine learning algorithms. A re sampling technique is introduced to transform the initial data by balancing the classes. 3 Fold Cross Validation technique is used to train the data along with 2 performance metrics i.e., accuracy and ROC Curves.
The results indicated that for balanced data Gradient Boosting Classifier Algorithm outperforms all the other models with accuracy being roughly around 98%. On the other hand logistic regression performs the worst with accuracy being around 67%. Hyper parameter tuning for all the models was done to improve the training and testing of the model. Furthermore, to find out the most important feature contributing to the test instance two different interpretative machine learning techniques are used i.e., Local Interpretable Model-agnostic Explanations (LIME) and SHapley Additive exPlanations (SHAP) on Gradient Boosting Classifier.We chose LIME as our main interpretable model because the dataset comprised fewer samples and low-dimensional data with low computational power. In conclusion, LIME is used to identify the most crucial component, which is then transmitted to the web crawler so that it can discover treatments for that specific issue on the web.

8 Future Scope

The population of Framingham Village served as the basis for the dataset utilized in this study. In the future, we can investigate whether same training models can be applied directly to other data sets, assuming that the causes of heart disease are constant across all geographical contexts. This study used 3 Fold Cross Validation for training the models whereas one can also make use of Stratified K Fold Cross Validation Techniques along with Synthetic Minority Oversampling Technique (SMOTE) for re sampling imbalanced data. The Permutation Importance ELI5 algorithm can be used in research in the future in addition to LIME and SHAP, and its accuracy and time complexity can be evaluated. Additionally, if a dataset for preventing heart disease is easily accessible, one can utilize various Natural Language Processing Algorithms to successfully recommend therapies. Furthermore, knowledge graphs can be used extensively to represent most appropriate remedies in a pictorial model.

References

[1] "Cardiovascular diseases (cvds)," World Health Organization. [Online]. Available: https://www.who.int/news-room/fact sheets/detail/cardiovascular-diseases-(cvds).[Accessed: 24-Nov2021].).

[2] Sharma, H., Rizvi, M. A. (2017). Prediction of heart disease using machine learning algorithms: A survey. International Journal on Recent and Innovation Trends in Computing and Communication, 5(8), 99-104.

[3] Alaa, A. M., Bolton, T., Di Angelantonio, E., Rudd, J. H., van derSchaar, M. (2019). Cardiovascular disease risk prediction using automated machine learning: a prospective study of 423,604 UK Biobank participants. PloS one, 14(5), e0213653.

[4] Dubey, A. K., Choudhary, K., Sharma, R. (2021). Predicting Heart Disease Based on Influential Features with Machine Learning. INTELLIGENT AUTOMATION AND SOFT COMPUTING, 30(3), 929-943.

[5] Saw, M., Saxena, T., Kaithwas, S., Yadav, R., Lal, N. (2020, January). Estimation of Prediction for Getting Heart Disease Using Logistic Regression Model of Machine Learning. In 2020 International Conference on Computer Communication and Informatics (ICCCI) (pp. 1-6). IEEE.

[6] Rahman, A., Tabassum, A. (2020). Model to assess the factors of 10-year future risk of coronary heart disease among people of Framingham, Massachusetts. International Journal of Public Health, 9(3), 259-266.

[7] Kwakye, K., Seong, Y., Yi, S. (2020, August). An Android-based mobile paratransit application for vulnerable road users. In Proceedings of the 24th Symposium on International Database Engineering Applications (pp. 1-5).

[8] Aboah, A. (2021). A vision-based system for traffic anomaly detection using deep learning and decision trees. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 4207-4212).

[9] Shoman, M., Aboah, A., Adu-Gyamfi, Y. (2020). Deep Learning Framework for Predicting Bus Delays on Multiple Routes Using Heterogenous Datasets. Journal of Big Data Analytics in Transportation, 2(3), 275-290.

[10] Aboah, A., Adu-Gyamfi, Y. (2020). Smartphone-Based Pavement Roughness Estimation Using Deep Learning with Entity Embedding. Advances in Data Science and Adaptive Analysis, 12(03n04), 2050007.

[11] Aboah, A., Boeding, M., Adu-Gyamfi, Y. (2021). Mobile Sensing for Multipurpose Applications in Transportation. arXiv preprint arXiv:2106.10733.

[12] Aboah, A., Arthur, E. (2021). Comparative Analysis of Machine Learning Models for Predicting Travel Time. arXiv preprint arXiv:2111.08226.

[13] Dadzie, E., Kwakye, K. (2021). Developing a Machine Learning Algorithm-Based Classification Models for the Detection of High- Energy Gamma Particles. arXiv preprint arXiv:2111.09496.

[14] Palaniappan, S., Awang, R. (2008, March). Intelligent heart disease prediction system using data mining techniques. In 2008 IEEE/ACS international conference on computer systems and applications (pp. 108-115). IEEE.

[15] Mahmoud, W. A., Aborizka, M., Amer, F. A. E. (2021). Heart Disease Prediction Using Machine Learning and Data Mining Techniques: Application of Framingham Dataset. Turkish Journal of Computer and Mathematics Education (TURCOMAT), 12(14), 4864-4870.

[16] Brownlee, J. (2019, October 15). Information Gain and Mutual Information for Machine Learning. Machine Learning Mastery. https://machinelearningmastery.com/information-gain-and-mutualinformation/.

[17] Scott M. Lundberg, Su-In Lee, A Unified Approach to Interpreting Model Predictions, Paul G Allen School of Computer Science, University of Washington.