Detection of Heart Disease Risk Utilizing Correlation Matrix, Random Forest and Permutation Feature Importance Approaches
Kalp Rahatsızlığı Riskinin Korelasyon Matrisi, Rastgele Ağaç ve Permütasyon Öznitelik Seçimi Yöntemleriyle Tespit Edilmesi

Sude Pehlivan[1], Yalcin Isler[2]
[1]Department of Biomedical Technologies, Izmir Katip Celebi University, Izmir, Turkey
sudepehlivan35@gmail.com
[2]Department of Biomedical Engineering, Izmir Katip Celebi University, Izmir, Turkey
islerya@yahoo.com

***Abstract*—Heart diseases are among the conditions that threaten the quality of human life. Researchers have been investigating the risk of having heart diseases using different metabolic measures, however, it is a difficult mission to forecast whether these measures related to heart diseases due to the deep and complicated relations of them with each other. In this study, it was aimed to illuminate the complex relationship of different metabolic measures utilizing different machine learning techniques with random forest, correlation matrix and permutation feature importance feature selection techniques. As a result of the adopted approaches, classifiers, and feature selection techniques that have been utilized in machine learning applications were found to be useful and it was considered that performance could be increased by expanding the dataset.**

***Keywords*—Heart disease; machine learning; metabolic measures.**

***Özetçe*—Kalp rahatsızlıkları birçok insanın yaşam kalitesini tehdit eden hastalıklar arasındadır. Araştırmacılar, farklı metabolik ölçümleri kullanarak insanların kalp rahatsızlığı bulundurma riskini araştırmaktadır ancak bu ölçümlerin kalp rahatsızlığı riskine katkıları olup olmadığını önceden tahmin etmek birbirleriyle olan derin ve karmaşık ilişkileri nedeniyle zorlu bir görevdir. Bu çalışmada, farklı metabolik ölçümlerin sahip olduğu karmaşık ilişkiler, rastgele orman, corelasyon matrisi ve permütasyon öznitelik seçimi yöntemleri kullanılarak değişik sınıflandırıcılar ile aydınlatılmaya çalışılmıştır. Uygulanan teknikler sonucunda makine öğrenmesi uygulamalarında kullanılan sınıflandırıcı ve öznitelik seçme teknikleri yararlı bulunmuş ve veri setinin genişletilmesiyle performansın artırılabileceği düşünülmüştür.**

***Anahtar Kelimeler*—Kalp rahatsızlığı; makine öğrenmesi; metabolik ölçümler.**

I. INTRODUCTION

Heart disease is a serious condition and several measures from the human body can be used for the detection of heart diseases. Metabolic Syndrome (MS) is a serious disease that includes different medical conditions at the same time. Diagnosis criteria for MS include waist circumference, triacylglycerol, and High-density Lipoprotein (HDL) cholesterol levels, Fasting Blood Sugar (FBS) level, and Blood Pressure (BP) [1]. It was pointed out that the risk of Cardiovascular Disease (CVD) increases with the appearance of MS [2]. Fox et al. reported that people with diabetes were at greater risk than those without diabetes in the case of CVD [3]. Chest pain is also an important indicator of cardiac diseases, however, it can be caused by either cardiac problems such as ischemic cardiac disease or non-cardiac problems such as esophageal disease or even panic disorder. The clinical name for ischemic heart disease caused chest pain is angina pectoris [4]. There are other types of chest pain such as exercise-induced angina and it was concluded that for patients suffer from coronary artery disease, myocardial stunning can occur posterior to exercise-induced angina [5]. Elamin et al. conducted a study with patients suffering from angina pectoris to analyze ST-segment depression related to Heart Rate (HR) increment while performing an exercise. It was reported that the maximum valued ST/HR slope could be used to detect myocardial ischemia which is a cardiac disease [6].

In this paper, different metabolic, electrocardiographic, and different types of measures such as age and gender were used as features to investigate the presence of CHD risk utilizing different machine learning classifiers and feature selection techniques.

II. Materials and Method

A. Software Setup

To represent, analyze, and classify the data, Python programming language was used because of its open-source libraries and code readability. Integrated software of Python includes different Integrated Development Environments (IDEs) such as Spyder and applications such as web-based Jupyter Notebook. As programming platform, Jupyter Notebook was selected because of its advantage of executing one code line at a time and observing the output immediately which makes it useful to detect and solve errors sequentially. In addition, $markdown$ option allows user to write text and explain the purpose of commands Python packages as $Numpy$ to utilize numerical data handling, $Matplotlib$ to create 2-D graphs, $pandas$ to analyze and manipulate data, $seaborn$ to visualize data frame structures, $scikit-learn$ to implement machine learning algorithms and $eli5$ to implement permutation importance were used [7].

To utilize Python, along with its programming platforms and packages, Anaconda was used which is a free, open-source distribution and a package manager for Python. Anaconda was installed from its web page by selecting a Python version and operating system. It was important to add Anaconda into path variables and register it as the system for Python during installation. After the installation was completed, Anaconda Navigator was opened and $Not\ Installed$ category was selected to install relevant packages.

B. Data Pre-processing, Exploration and Representation

In this project, instead of the original dataset with 76 attributes, a different version of feature set with 14 attributes called Cleveland dataset was utilized. Age was given in years, gender was defined with sex (male:1, female:0), chest pain was defined with cp (Value 0:typical angina, Value 1:atypical angina, Value 2:non-anginal pain, Value 3:asymptomatic), resting blood pressure was defined with $trestbps$ (in mmHg), serum cholesterol was defined with $chol$ (in mg/dl), FBS was defined with fbs (>120 mg/dl:1, <120 mg/dl:0), resting ECG results was defined with $restecg$ (Value 0:normal, Value 1:having ST-T wave abnormality, Value 2:showing probable or definite left ventricular hypertrophy by Estes' criteria), maximum HR achieved was defined with $thalach$, exercise-induced angina was defined with $exang$ (yes:1, no:0), ST depression induced by exercise relative to rest was defined with $oldpeak$, the slope of peak exercise ST segment was defined with $slope$ (Value 0:upsloping, Value 1:flat Value 2:downsloping), number of major vessels colored by fluoroscopy was defined with ca (0-3), nuclear stress test result was defined with $thal$ (normal:0, fixed defect:1, reversible defect:2) and finally target was defined with $condition$ (disease:1, no disease:0) [8], [9]. Distributions of each attribute were plotted using several plotting functions as $countplot$, $catplot$, $jointplot$, and $boxplot$ to understand features illustrated in Figure 1 as an example. It was clear that some of the features were not related to each other or target. It was important to know which features were adding importance to the categorization because irrelevant data meant a decrease in the classification accuracy and loss of time. Visualizing the distribution of data was not enough for this purpose, thus feature selection was implemented.

As pre-processing steps, categorical features were converted into indicator variables with $pandas.get\_dummies()$ command, and data scaling was performed with $StandardScaler()$ from sklearn.preprocessing package.

C. Cross-validation

Cross-validation (CV) is a practical tool to evaluate the performance of different classifiers or a particular classifier with several parameters synchronously which can prevent overfitting. There are several CV techniques such as leave-one-out CV and K-fold CV. In the K-fold CV process, the dataset is divided into K portions and in every cycle, one of the portions is selected as the test set and other $K-1$ portions are selected as the training set. This cycle repeats until every portion is used as a test set. When the process is finished, the mean Accuracy (ACC) is given as the performance metric of the model [10]. In this paper, 5-fold CV was performed to evaluate the performance of classifiers and each classifier with different parameters.

D. Classification

In supervised machine learning algorithms, data must be divided into two subsets as $training\ set$ that model learns from and $test\ set$ to predict the class of provided samples based on the insight that it learned. Training and test sets include targets to inform the model about the class of the sample. Random division of the subsets is important to prevent the bias for the target of the data since data could be ordered considering the target class which can affect the performance of the model [11]. In this paper, $train\_test\_split()$ command was utilized to divide the dataset randomly. Following classifies were built, fitted and scored.

Support Vector Machine classifier maximizes the distance between classes by using separation boundaries called hyperplane and uses several kernels to classify non-linear data [12]. In this paper, $SVC()$ command with $sigmoid$ kernel was used since it provided the best performance results during the CV along with rbf kernel for only Permutation Importance (PI) case.

K-nearest Neighbors (KNN) classifier represents features with vectors and computes the distance between the sample vector given into the system and feature vectors. The sample is assigned to the class that K closest neighbors exist [13]. In this paper $KNeighborsClassifier()$ command with 6 neighbors was used since it provided the best performance results during CV.

Logistic Regression (LR) returns a binary output by using a logistic function called sigmoid and the output of the function for any input is always between 0 and 1. After implementing thresholding for the logistic function, the model maps given inputs to a dichotomous target [14]. $LogisticRegression()$ command was used with default parameters.

RF classifies the data by dividing it into branches and groups them based on a set of decision rules and in this paper $RandomForestClassifier()$ command was used [15].

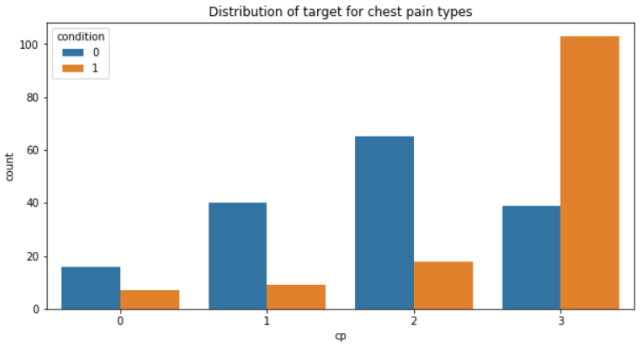


Figure 1: Distribution of Target Class among Chest Pain Types: For typical angina (cp value0), atypical angina (cp value1), and non-anginal pain (cp value2) absense of heart disease (condition value0) is more frequent whereas for asymptomatic (cp value3) the precense of heart disease (condition value1) is more frequent

E. Feature Selection

In the real world, several features are provided in the classification process. Considering the fact that they are not equally important, processing unnecessary features causes loss of time and affects the model complexity. To simplify the model, a feature selection process is performed [16]. There are several approaches for feature selection and in this paper, Correlation Matrix (CM) feature importance, RF feature importance, and PI techniques were utilized.

A CM is the representation of feature redundancy which is the relation of each feature with the other features and feature relevancy which is the relation of the target with other features [17]. The correlation matrix of this dataset is given in Figure 2. As a result of setting threshold into 0.41, features that were highly correlated with target identified as $exang, thalach, oldpeak, ca, thal$. The classification was re-performed with only CM based selected features.

RF also used for feature selection purposes with different measures and one of them is the Mean Decrease Impurity (MDI) which is sometimes called Gini Importance (GI). Gini impurity which is a metric that measures the frequency of incorrect labeling of a sample by choosing a label in the group that sample is belonged to, should be explained to understand GI [15]. The mean decrease in Gini impurity is the measure of feature importance for RF. In RF feature importance case $ca, thal, cp, trestbps, thalach, oldpeak$ were selected as most important features as shown in Figure 3.

In this paper, $PermutationImportance()$ from eli5 library was utilized as feature selection method. This command calculates the Mean Decrease Accuracy (MDA) which means the importance of the feature is assigned by measuring the decrement of the accuracy in the absence of selected features. It was introduced for RF by Breiman, however, it can also be used with different classifiers [18], [19]. In this case, RF was utilized to define most relevant features with PI which were $ca, cp, thal, exang$ and weights of the features are given in Table I.

Weight	Feature
0.0422 ± 0.0475	thal
0.0378 ± 0.0412	ca
0.0267 ± 0.0412	cp
0.0156 ± 0.0178	exang
0.0089 ± 0.0327	oldpeak
0.0067 ± 0.0412	thalach
0.0022 ± 0.0166	sex
-0.0067 ± 0.0178	fbs
-0.0067 ± 0.0109	restecg
-0.0089 ± 0.0295	trestbps
-0.0111 ± 0.0199	age
-0.0222 ± 0.0243	slope
-0.0244 ± 0.0327	chol

Table I: Permutation Importance weights of Features: First column describes the decrease in the performance when the corresponding feature is extracted

III. Results and Discussion

In this project, RF, LR, KNN and SVM classifiers with CM feature importance, RF feature importance, and permutation importance were utilized using metabolic, ECG related and categorical features to detect heart disease risk.

As shown in Figure 4 (a) , LR, KNN, and SVM classifiers had 85.56% accuracy when the entire feature set with 13 attributes were utilized. In the case of accuracy, it seemed that using all feature set was the most successful approach with the mean value of 85.00% was using entire feature set followed by permutation importance with 84.45%, RF feature importance with 82.78% and finally CM feature importance with 79.45%. However, the amount of the decrease of the accuracy was negligible for the aim of reducing the complexity of the model and preventing the loss of time since entire feature set was utilizing 13 attributes and CM feature importance was using 5, RF feature importance was using 6 and permutation importance was using only 4 features. On the other hand, accuracy was not sufficient to evaluate the performance, thus sensitivity (SN) and specificity (SP) were calculated and given in Figure 4 (b) and Figure 4 (c). In the case of sensitivity, the highest values were provided with 89.74% for RF classifier utilizing the entire feature set and RF feature importance, 87.18% for LR and SVM utilizing CM feature importance, and permutation importance. The most successful approach was using the entire feature set with the mean value of 87.32% followed by permutation importance with 84.62%, RF feature importance with 83.98%, and CM feature importance with 81.41%. As the last performance metric, highest values of specificity were calculated as 84.31% for KNN and SVM using the entire feature set, 90.20% for KNN using CM feature importance, 86.27% for SVM using RF feature importance, and 90.20% for KNN using permutation importance. The best technique was chosen as permutation importance with 84.31% mean value and using the entire feature set was close with 83.33% mean value of utilizing the entire feature set. The most inadequate approach was CM feature importance with 77.94% mean of specificity. Considering the obtained results, KNN and SVM were the most successful classifiers and permutation importance was the most appropriate feature

Figure 2: Correlation Matrix of the Dataset

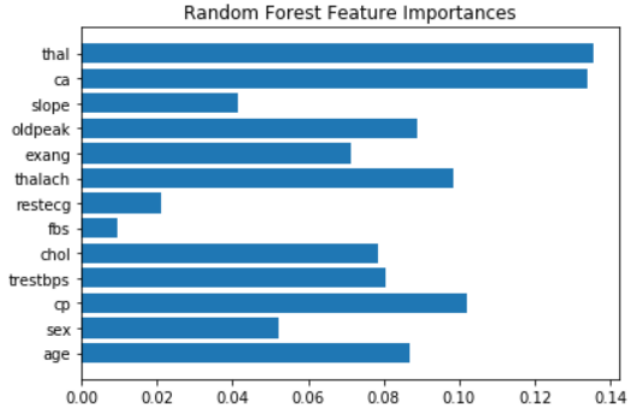


Figure 3: Random Forest Feature Importance Coefficients: The RF feature importance selected attributes were thal as the most important feature followed by ca, cp, thalach, oldpeak and trestbps

selection technique both with high- performance results and utilizing only 4 features.

In a paper published by Chen et al., a tool called Heart Disease Prediction System (HDPS) was designed in order to assist professionals by using a different version of the dataset utilized in this paper. Artificial Neural Network (ANN) was designed for this purpose and 80% accuracy along with 85% sensitivity and 70% specificity were achieved [20].

In another study, Patel et al. utilized the same 14 attributes to predict heart disease with Naive Bayes, Decision Tree, and Clustering data mining methods. The genetic search was adopted as feature selection technique and 6 attributes as $trestbps, oldpeak, cp, ca, exang, thalach$ was selected to conduct classification with Weka tool. Decision Tree provided 99.2% accuracy, Naive Bayes provided 96.5% accuracy, and Clustering provided 88.3% accuracy. The only difference between the selected features of Patel and this paper with RF feature importance was that they utilized $exang$ instead of $thal$ [21].

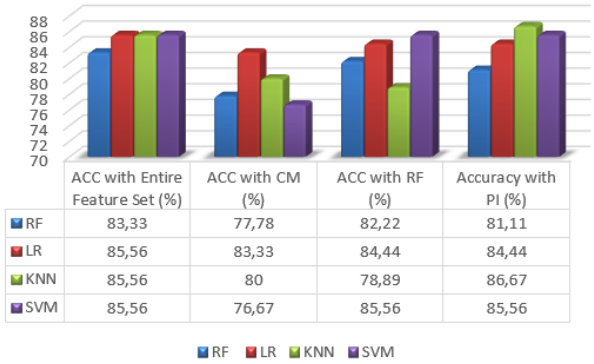
Usha Rani conducted a study with ANN using the Cleveland database with target class having multiple instances instead of binary condition as $normal$: 0, $firststroke$: 1, $secondstroke$: 2 and $endoflife$: 3. While single layer with 100 training samples provided 76% efficiency, 350 training samples provided 90.6% efficiency. In the case of multiple layers, 100 training samples resulted in 82%, and 350 training samples resulted in 94% efficiencies. Regardless of the complexity of an ANN, multilayer ANN provided valid classification results [22].

Sharma et al. used the Cleveland database to utilize Decision Tree, Multivariate Adaptive Regression Splines (MARS), RF, and Tree-model from Genetic Algortihm (TMGA) data mining techniques. Decision Tree with 93.24% accuracy was the best approach while MARS provided 91.04%, RF provided 89.95%, and TMGA 88.85% accuracy [23].

These papers suggested that the choice of data mining technique and the number of training samples were extremly important to improve the accuracy of the classifiers. The disadvantage of this paper was utilizing limited number of classification and feature selection techniques. However, altough feature selection techniques decreased the performance metrics in general, it was still a convinient approach to adopt considering the obtained results. It should be noted that there are two possibilities as an explanation for obtaining 85.56% accuracy as the highest value while utilizing the entire feature set. One of them is that the processed dataset was not reflecting the true nature of people at the risk of having a heart disease as the original one or by acknowleding the fact that other papers obtained up to 94% accuracy, second possibility is that selection of data mining and feature selection techniques along with classifier parameters was creating a significant amount of difference.

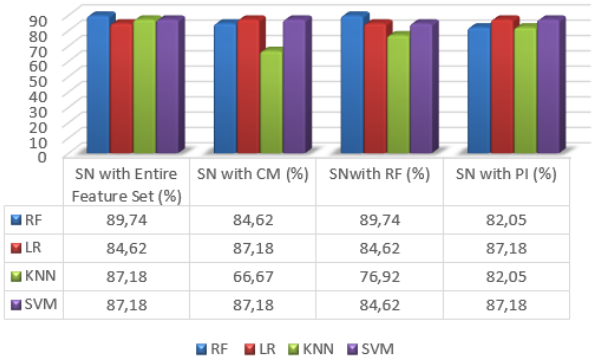
IV. Conclusion

CM, RF, and permutation feature importance were utilized as feature selection techniques with RF, LR, KNN and SVM classifiers on a dataset that contains several metabolic measures to identify the risk of heart disease. It was observed that in



	ACC with Entire Feature Set (%)	ACC with CM (%)	ACC with RF (%)	Accuracy with PI (%)
RF	83,33	77,78	82,22	81,11
LR	85,56	83,33	84,44	84,44
KNN	85,56	80	78,89	86,67
SVM	85,56	76,67	85,56	85,56

(a) Accuracy values



	SN with Entire Feature Set (%)	SN with CM (%)	SNwith RF (%)	SN with PI (%)
RF	89,74	84,62	89,74	82,05
LR	84,62	87,18	84,62	87,18
KNN	87,18	66,67	76,92	82,05
SVM	87,18	87,18	84,62	87,18

(b) Sensitivity values

	SP with Entire Feature Set (%)	SP with CM (%)	SP with RF (%)	SP with PI (%)
RF	78,43	72,55	76,47	80,39
LR	86,27	80,39	84,31	82,35
KNN	84,31	90,2	80,39	90,2
SVM	84,31	68,63	86,27	84,31

(c) Specificity values

Figure 4: Classifier performances of accuracy, sensitivity, and specificity values

some cases, while the accuracy dropped, the specificity or sensitivity of some classifiers increased. This inference suggested that the perception of choosing a model and feature selection technique was not a straightforward path, on the contrary, the most important metric among accuracy, specificity, and sensitivity should be decided as specific to the case.

In conclusion, disregarding the fact that the dataset required generalization in order to represent the target population sufficiently, RF, LR, KNN, and SVM were applicable approachs to adopt in order to detect heart disease risk since the results presented the effectiveness of machine learning classifiers. Besides, results showed that the model complexity could be reduced by narrowing feature set down such that the amount of the decrease in the number of features was more significant than the decrease in the accuracy which made feature selection techniques acceptable, especially permutation importance. The need of generalization of the classification and feature selection techniques is a serious gap that should be closed since heart disease is a critical condition. It is clear that further investigation and extension of the dataset should be considered in order to increase the efficiency and performance of both classifiers and feature selection techniques.

REFERENCES

[1] Després, J. P., & Lemieux, I. (2006). Abdominal obesity and metabolic syndrome. Nature, 444(7121), 881-887.

[2] Galassi, A., Reynolds, K., & He, J. (2006). Metabolic syndrome and risk of cardiovascular disease: a meta-analysis. The American journal of medicine, 119(10), 812-819.

[3] Fox, C. S., Coady, S., Sorlie, P. D., Levy, D., Meigs, J. B., D'Agostino, R. B., ... & Savage, P. J. (2004). Trends in cardiovascular complications of diabetes. Jama, 292(20), 2495-2499.

[4] Lenfant, C. (2010). Chest pain of cardiac and noncardiac origin. Metabolism, 59, S41-S46.

[5] Ambrosio, G., Betocchi, S., Pace, L., Losi, M. A., Perrone-Filardi, P., Soricelli, A., ... & Weiss, J. L. (1996). Prolonged impairment of regional contractile function after resolution of exercise-induced angina: evidence of myocardial stunning in patients with coronary artery disease. Circulation, 94(10), 2455-2464.

[6] Elamin, M. S., Mary, D. A. S. G., Smith, D. R., & Linden, R. J. (1980). Prediction of severity of coronary artery disease using slope of submaximal ST segment/heart rate relationship. Cardiovascular research, 14(12), 681-691.

[7] Johansson, R. (2018). Numerical Python: Scientific Computing and Data Science Applications with Numpy, SciPy and Matplotlib. Apress.

[8] (n.d.). Retrieved April 5, 2020, from http://archive.ics.uci.edu/ml/datasets/Heart Disease

[9] Cherngs. (2020, March 29). Heart Disease Cleveland UCI. Retrieved April 15, 2020, from https://www.kaggle.com/cherngs/heart-disease-cleveland-uci

[10] Wong, T. T. (2015). Performance evaluation of classification algorithms by k-fold and leave-one-out cross validation. Pattern Recognition, 48(9), 2839-2846.

[11] Garreta, R., & Moncecchi, G. (2013). Learning scikit-learn: machine learning in python. Packt Publishing Ltd.

[12] Cortes, C., & Vapnik, V. (1995). Support-vector networks. Machine learning, 20(3), 273-297.

[13] Saini, I., Singh, D., & Khosla, A. (2013). QRS detection using K-Nearest Neighbor algorithm (KNN) and evaluation on standard ECG databases. Journal of advanced research, 4(4), 331-344.

[14] Saini, I., Singh, D., & Khosla, A. (2013). QRS detection using K-Nearest Neighbor algorithm (KNN) and evaluation on standard ECG databases. Journal of advanced research, 4(4), 331-344.

[15] Qi, Y. (2012). Random forest for bioinformatics. In Ensemble machine learning (pp. 307-323). Springer, Boston, MA.

[16] Dash, M., & Liu, H. (1997). Feature selection for classification. Intelligent data analysis, 1(3), 131-156.

[17] Roobaert, D., Karakoulas, G., & Chawla, N. V. (2006). Information gain, correlation and support vector machines. In Feature extraction (pp. 463-470). Springer, Berlin, Heidelberg.

[18] Gómez-Ramírez, J., Ávila-Villanueva, M., & Fernández-Blázquez, M. Á. (2019). Selecting the most important self-assessed features for predicting conversion to Mild Cognitive Impairment with Random Forest and Permutation-based methods. bioRxiv, 785519.

[19] Breiman, L. (2001). Random forests. Machine learning, 45(1), 5-32.

[20] Chen, A. H., Huang, S. Y., Hong, P. S., Cheng, C. H., & Lin, E. J. (2011, September). HDPS: Heart disease prediction system. In 2011 Computing in Cardiology (pp. 557-560). IEEE.

[21] Patel, S. B., Yadav, P. K., & Shukla, D. P. (2013). Predict the diagnosis of heart disease patients using classification mining techniques. IOSR Journal of Agriculture and Veterinary Science (IOSR-JAVS), 4(2), 61-64.

[22] Rani, K. U. (2011). Analysis of heart diseases dataset using neural network approach. arXiv preprint arXiv:1110.2626.

[23] Dataset, C. (2017). Prediction of Heart Disease Using Cleveland Dataset: A Machine Learning Approach.