An Analytical Model for Prediction of Heart Disease using Machine Learning Classifiers

Diti Roy, Md. Ashiq Mahmood and Tamal Joyti Roy
Institute of Information and Communication Technology
Khulna University of Engineering & Technology (KUET)
Khulna-9203, Bangladesh
dity267@gmail.com,ashiqmahmoodbipu@gmail.com, tjroy13june@gmail.com

Abstract: *Heart Disease is the most dominating disease which is taking a large number of deaths every year. A report from WHO in 2016 portrayed that every year at least 17 million people die of heart disease. This number is gradually increasing day by day and WHO estimated that this death toll will reach the summit of 75 million by 2030. Despite having modern technology and health care system predicting heart disease is still beyond limitations. As the Machine Learning algorithm is a vital source predicting data from available data sets we have used a machine learning approach to predict heart disease. We have collected data from the UCI repository. In our study, we have used Random Forest, Zero R, Voted Perceptron, K star classifier. We have got the best result through the Random Forest classifier with an accuracy of 97.69.*

Keywords. Heart Disease, Prediction, Machine Learning, Kstar, Random Forest, Voted Perceptron, Zero R

I. INTRODUCTION

Cardiovascular disease which is termed heart disease is the number 1 cause of death in the whole world taking at least 17 million people's death every year [1]. Though at least three-quarters of death has occurred in the low and middle-income country rate of death is also alarming in developed countries. According to the Center for Disease Control and Prevention, at least 25 percent of death occurred due to heart disease in the USA. This situation is also depicted in another country of different ages, races, classes, etc. Though medical science has progressed tremendously all over the world but preventing different types of heart disease is yet to possible. In Bangladesh, from 1986 to 2006 death from heart disease increase at least 3527% whereas the death of dysentery and respiratory infection reduced by 79% and 86% [2]. The most alarming matter of heart disease is that most of the people are suffering from heart disease at most productive years of their life [3] stated that in India 50% percent of heart disease occurred before 50 years whereas at least 25 percent faces the same disease before 40 years. As the low-income countries are lacking basic health care facilities they don't get proper guidelines about heart disease which causes death, as well as huge costs in medication, lead every family towards poverty. Though heart disease is creating a catastrophic moment to both patient and health authority still vital challenge is to predict and detect its presence in the human body despite having different techniques [4]. That's why to decrease the death rate and protect every family from economic vulnerability prediction of heart disease is an important factor which will ultimately help policymakers to take appropriate step anent heart disease. Machine learning is a useful instrument to conclude a huge number of data regarding health, technology, business, etc. It can assist in increasing access and analysis of health care facilities in developing countries. The instrument of the decision tree, naïve bays, support vector machine can be used in predicting heart disease which will be more efficient than other techniques [5]. Therefore, in our study, we have used machine learning algorithms to predict heart disease.

II. RELATED WORKS

Heart disease is the most important issue and a common problem in the total world. Thousands of people died of heart disease every year. For this reason, many researchers are trying to predict this cardiovascular disease which is a critical challenge in the area of clinical data analysis. In this paper, Mohan et al [6] had proposed a unique method to predict heart disease by using machine learning techniques. This prediction model was done with different combinations of features which were known as classification techniques. They had used several classification methods like data pre-processing, feature selection and reduction, different classification modeling, decision trees, language model, support vector model, and random forest. Finally, with the hybrid random forest with a linear model (HRFLM) they had able to show the accuracy level of 88.7% through the prediction model for heart disease.Heart disease is one of the leading causes of death nowadays. As it is a complex task sometimes predicting the heart attack is more difficult for medical practitioners because of less knowledge and experience. Not only that sometimes the health sector hides some information that is needed for making decisions. Maestre et al [7] showed a model to predict heart disease. There are different data mining algorithms such as J48, Naïve Bayes, REPTREE, CART, and Bayes Net were used in this research for predicting heart attacks. Last of all the research result showed the prediction accuracy which was 99%. And this research also showed that data mining enabled the health sector to predict patterns in the dataset.From an investigation, Gavhane et al [8] constructed a model which could detect the symptoms which will be helpful to prevent heat stroke at an early age, and day by day its increasing rate had been developed. They proposed an application that would use to show the symptoms like age, sex, pulse rate, etc, and would able to predict heart disease. They had used machine learning algorithm neural networks to find the best accuracy of heart disease. Due to many reasons, heart disease is increasing rapidly. Although different health care centers and doctors collect data daily as they don't use machine learning and pattern machine techniques it is reducing their predictability. For this reason, Awan et al [9] showed a prediction model. In this paper, they had collected data and attributes for the UCI repository. By using this data, they had tried to predict heart disease. For this development, they had used several techniques in Artificial Neural Network (ANN). They had shown accuracy such as 94.7% for ANN but 97.7% accuracy rate for Principle Component Analysis (PCA).Yadav et al [10] collected the information for

prediction from the UCI repository. 1025 Instances with14 attributes dataset were used for this prediction model. After accomplishing this research, they had proposed a model and analyzed classification accuracy, precision, and sensitivity by four tree-based classification algorithms like M5P, random Tree, and Reduced Error Pruning with the Random forest ensemble method. After the feature selection of the heart patient's dataset, all the prediction algorithms were used. They had used three features-based algorithms like Pearson Correlation, Recursive Features Elimination, and Lasso Regularization. Three experimental setups were used to finish this analysis. Pearson Correlation on M5P, random Tree, Reduced Error Pruning, and Random forest ensemble method was applied for the first experiment. In the second experiment, Recursive Features Elimination and application on the above four tree-based algorithms were used. And for the third experiment Lasso Regularization and applied on as above tree-based algorithms were used. After completing this experiment, they had analyzed and calculated classification accuracy, precision, and sensitivity. Finally, they were capable to show the best accuracy that was 99% and it is conducted by feature selection methods Pearson correlation and Lasso Regularization with random forest ensemble method. Heart disease is an important issue in the whole world. In this paper, Sowmiya et al [11] proposed a model with novel feature selection and classification techniques to predict impermanence in overflowing heart failure patients with a view to decreasing the death rate due to heart disease. For selecting the best feature for hybrid K-nearest neighbor (KNN) classifier the ant colony optimization (ACO) algorithm was utilized. Their suggested approach was contrasted with the prior classification techniques such as the Support vector machine, Naïve Bayes, KNN, C4.5, and decision tree. For implementation UCI Cleveland dataset was utilized. Finally, they had found the best result with accuracy 99.2% by Using the Netbeans IDE. Hasan et al [12] showed a model by collecting information about heart disease and used different techniques like feature selection technique and removing unnecessary features, different classification techniques such that KNN, Decision Tree (ID3), Gaussian Naïve Bayes, Logistic Regression and Random for better prediction on heart disease. Several performance measurement elements like accuracy, ROC curve, precision, recall, sensitivity, specificity, and F1-score were regarded to find out the performance of the classification techniques. Among them, they had found the best result by using Logistic and the classification accuracy was 92.76%.Heart disease is the most common problem in the world and thousands of people are suffering from this problem. Death rates are increasing day by day because of heart disease. Many researchers are trying to predict this disease by using different techniques. A research was made by Singhal et al [13] to design a method with Convolutional Neural Networks (CNNs) to predict this disease. They had used 13 clinical features as input to CNN. They had used modified back propagation training method to train the CNN. During testing, they had found that accuracy was 95% by using CNN for predicting absence and presence of heart disease.Machine learning (ML) is very efficient in making decisions and prediction. Now-a –days, many researchers use machine learning for prediction purpose and others. In this paper, Vindhya et al [14] also proposed a model with ML. They had tried to find out best result by using Machine Learning techniques. They also used various combinations of classification and feature techniques which finally showed the higher accuracy of 88.7% using hybrid random forest with linear model (HRFLM).Data mining plays an important role in various sectors and it plays an important role for prediction purpose. For detecting a disease many tests are needed and sometimes it is very risky. But using data mining technique some number of tests can be reduced. Using data mining approach Chatterjee et al [15] proposed a model by using patient's dataset and then applied different algorithms like IBK, Classification Tree, Naïve Bayes and KNN to predict the heart disease and found the different accuracy. They had showed the accuracy that was 99.19%, 93%, 97.13%, and 99% using IBK, Classification Tree, Naïve Bayes and KNN. Among this algorithm KNN showed the highest accuracy and it was 99%.

III. PROPOSED METHODOLOGY

. Our proposed methodology showing in Fig. 1 the flow chart. The steps by steps approach are discussed. We use the UCI data repository for our machine learning approach. Data set managing, collecting the data set features, pre-process the data set, choosing the feature, classify the instances, measure the performance of the classifiers, compare the accuracy and last the result is acquired. Four machine learning techniques are applied to examine the accuracy rate for our heart data set. Evaluating the performance we tuned for improve the accuracy rate. Later in that, the confusion matrix for each machine learning technique has been visualized for the validity of the experimental model. After preprocessing the data set and cleaning the data we use nine attributes that make sense for our experiment. There were fourteen attributes but did not take all of those to grant as some of the attributes did not make any sense at all. Table I and II showing the attributes and data set we have used for our experiment. Table I showing the basic attributes with the data set , There are all

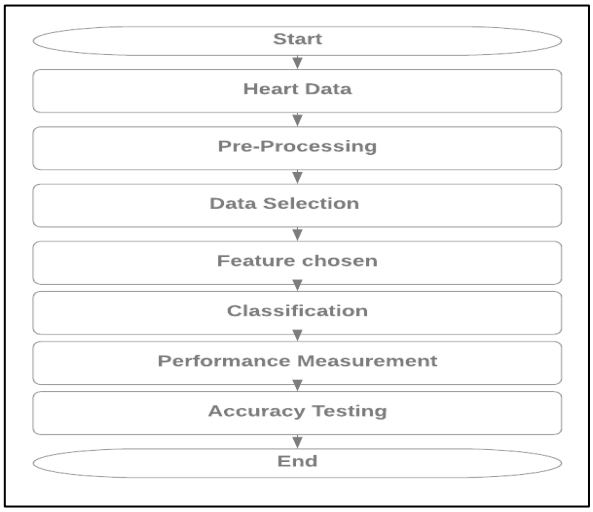

Fig. 1. Overall Flow Chart

together three hundred three data with which we did our experimental analysis. Fig. 3 showing how our machine learning experimented. The test and training data set worked independently. For avoiding data misbalancing proper rending is done to our data set. For each execution the parameter of float value is avoided so there were no decimal values in our data. We divided our data set into two sub subsets is training and another is tested. The ratio between

training and test data set was 70:30. We used four machine learning classifier algorithms for the application of the experiment. All four of the used in training and test data set and predicted the result and accuracy. Our machine learning classifiers were Random Forest, ZeroR, Voted Perceptron, and Kstar. A detailed explanation of the four algorithms also showed in this paper. We used different machine learning techniques to evaluate the right accuracy and also we tested the training sets several times. To training a machine is the toughest part of the experiment, simple mistakes would shatter the experiment.

TABLE I. Data sets description

Age	Sex	Chest Pain Type	Resting Blood Pressure	Serum Cholesterol mg/dl	Fasting Blood Sugar	Resting Electrocardiographic	Maximum Heart Rate	Class
71	0	2	110	265	1	0	130	1
54	1	1	108	309	0	1	156	1
52	1	3	118	186	0	0	190	0
41	1	1	135	203	0	1	132	0
58	1	2	140	211	1	0	165	1
35	0	0	138	183	0	1	182	0
51	1	2	100	222	0	1	143	0

TABLE II. Data Set Description

Attributes	Description
Age	This means the age of a particular person
Sex	If the value of Sex=1 the person is female and Sex=0 is male
Chest Pain Type	How much the chest pain is.0 is no pain to 3 is the highest pain
Resting Blood Pressure	Blood pressure when the person is in resting
Serum Cholesterol mg/dl	It is the bad cholesterol that indirectly causes a heart attack.
Fasting Blood Sugar	If the attribute =1 then the person is in fasting, 0 means negative
Resting Electrocardiographic	The resting ECG result
Maximum Heart Rate	The maximum heart rate value
Class	If the class=1 that means the person is at risk for heart disease or had an attack or likely to suffer, 0 means no risk, completely healthy person.

Table II showing the descriptiveness of our attributes. For the experiment, the data sets in Table I preprocessed for getting the desired accuracy several times. The interdependent variables are used for the prediction of the accuracy of our dataset. Column in Table I "class" divided the risk factor with variable "0" and "1". 0 is healthy and no risk factor of heat-related illness and 1 is more likely to suffer in heat-related illness. The main data set was in CSV format. We convert it to arff and performed the unsupervised machine learning filter "Numeric to Nominal", that we visualize our data sets with the fullest sense of analytical procedure. So the data set we deal with was nominal for the experimental purposes. We got rid of the NAN [16] values as it would hamper the experimental results. Fig. 2 showing the male and female percentages in our data set.

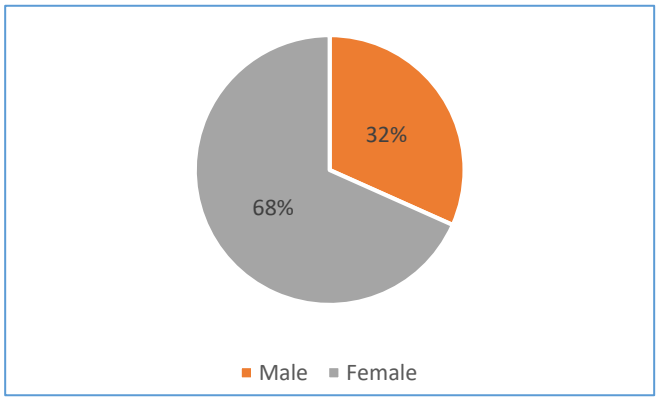

Fig. 2. Male and Female percentage

We performed a correlation test among the attributes. The result we got is showing in Table III.

TABLE III. Correlation among attributes

age	Correlation Coefficient	1.000	-.087
	Significant level (2-tailed)	.	.129
	N	303	303
cheastpaintype	Correlation Coefficient	-.087	1.000
	Significant level (2-tailed)	.129	.

	N	303	303
maximum_ heart_rate_ achieved	Correlation Coefficient	-.398	.324
	Significant level (2-tailed)	.000	.000
	N	303	303

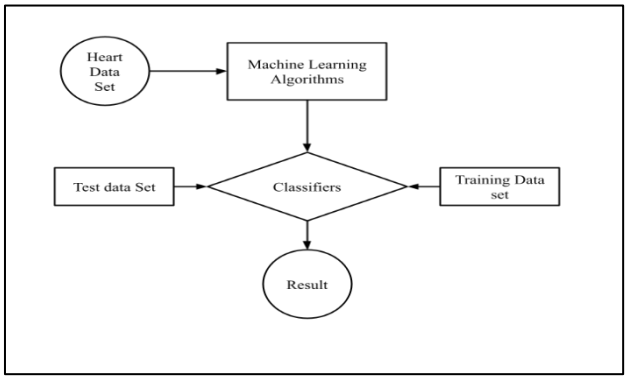

Fig. 3. Machine learning flow chart

Random Forest: It is a supervised machine learning algorithm, it makes decision trees, it highly depends on the trained data set. It is a learning model for getting the results. Plotting trees randomly for getting the highest and stable predictions.

Algorithm (RandomForest):
Training Set :D=(a1,b1)….(an,bn)
Feature is :F
Number of trees:N
Function<-RANDOM(D,F)
H<-0
for i from D(ai,bi)….to N do
D(i)<-Sample from D
Hk<-RANDOMLEARN(D(i),F)
end

ZeroR: It is the least difficult order strategy which depends on the objective and overlooks all indicators. ZeroR classifier predicts the larger part classification (class). Even though there is no consistency power in ZeroR, it helps decide a pattern execution as a benchmark for other order strategies.

Voted Perceptron: The voted perceptron method is based on the perceptron algorithm of Rosenblatt and Frank [17]. The calculation exploits information that is directly distinct with huge edges. This strategy is less complex to execute, and considerably more proficient as far as calculation time when contrasted with Vapnik's SVM. The calculation can likewise be utilized in high dimensional spaces utilizing piece capacities.

K-Star: It is used for finding the depth in the field or this case the depth in the accuracy. Model-based upon two training sets, predicts the values in the nearest first looking mode. It works instantly, super-fast classification for the training sets.

IV. EXPERIMENTAL ANALYSIS

The experimental results were obtained through different analyses. Thee were test data set which was roughly 30% and the training consisted with 70%, a total of 100% data is converged with the experiment. Table IV showing the different accuracy from the test data set.

TABLE IV. Test dataset results

Classifier name	Accuracy(%)
Random Forest	96.703
ZeroR	90.10
Voted Perceptron	94.20
Kstar	97.80

Table V showing the training result of our data set.

TABLE V. Training data set results

Classifier name	Accuracy(%)
Random Forest	97.69
ZeroR	85.14
Voted Perceptron	94.39
Kstar	94.05

The result from test data and training data varies. Fig. 5 clearly showing that there is a relation between age and maximum heart rate. Heart rate increase with age. Table VI shows the confusion matrix of each classifier. The confusion matrix results were acquired from the training data set as the test data set worked with only 30% of the total data set.
After the confusion matrix result, we can surely say that the result of Random Forest is higher than the rest of the four algorithms. The true positive rate of Random Forest is 0.967 and the false positive is 0.300. With having the roc and PRC area in both cases is 1.

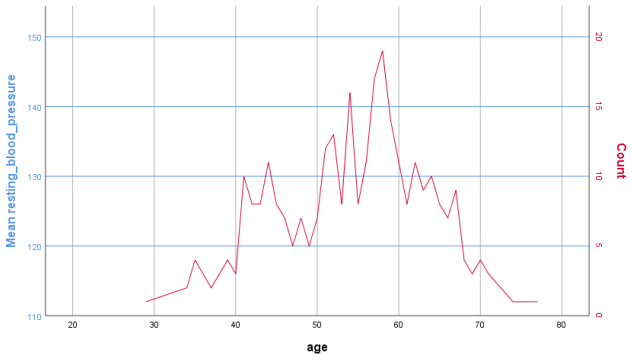

Fig. 4. Age-High Blood Pressure Relation

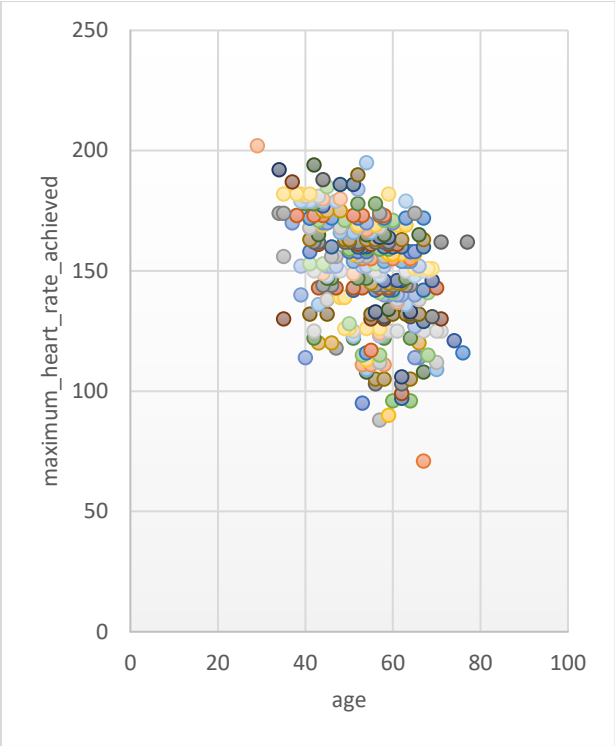

Fig. 5. Scatter point among age and maximum heart rate

Fig.4 showing a relation between age and high blood pressure. Our finding showed there's a relation between age and high blood pressure. People are likely to develop high pressure in their 40s or 50s. So better to follow up with a personal physician after the late 30s.

TABLE VI. Confusion matrix of the classifiers

Random Forest		**ZeroR**	
258	0	258	0
7	38	45	0
Kstar		**VotedPerceptron**	
258	0	256	2
18	27	1	44

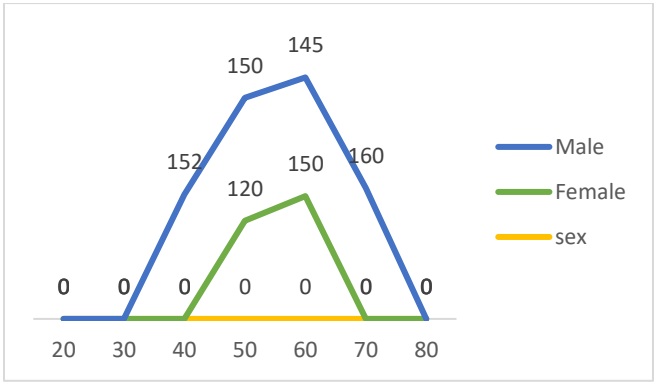

Fig. 6. Resting blood sugar and the relation with age and sex

Fig.6 showing a visualization between sex, age, and resting blood sugar. The resting blood sugar is on the higher side if the age increase and vice versa.

V. CONCLUSION

Heart disease is a crucial health problem all over the world. A proper and scientific prediction approach can mitigate the loss of heart disease. we have constructed a technique to predict heart disease by using machine learning algorithms. In our study, we have used Random Forest, Kstar, Zeror, Voted Perceptron classifier. The accuracy we have got from our study that Random Forest classifier is 97.69%, ZeroR is 85.14%, Voted Perceptron is 94.39%, Kstar is 94.05%. Among the classifier we have used in our study, we have found that the Random Forest classifier has produced the best result with an accuracy of 97.69 percent. In our future work, we will use a more accurate data set to get better results with more technological and scientific knowledge in another branch of medical science.

VI. REFERENCES

[1] World Health Organization/health-topics/cardiovascular-diseases
[2] Ahsan Karar, Z., Alam, N., & Kim Streatfield, P. (2009). Epidemiological transition in rural Bangladesh, 1986–2006. Global health action, 2(1), 1904.
[3] Mittal, R. A. (2017). Increasing heart attacks in young Indians'. The Times of India.
[4] M. A. Jabbar, P. Chandra, and B. L. Deekshatulu, "Prediction of risk score for heart disease using associative classification and hybrid feature subset selection," Int. Conf. Intell. Syst. Des. Appl. ISDA, pp. 628–634, 2012.
[5] 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.
[6] Mohan, S., Thirumalai, C., & Srivastava, G. (2019). Effective heart disease prediction using hybrid machine learning techniques. IEEE Access, 7, 81542-81554.
[7] Masethe, H. D., & Masethe, M. A. (2014, October). Prediction of heart disease using classification algorithms. In Proceedings of the world Congress on Engineering and computer Science (Vol. 2, pp. 22-24).
[8] Gavhane, A., Kokkula, G., Pandya, I., & Devadkar, K. (2018, March). Prediction of heart disease using machine learning. In 2018 Second International Conference on Electronics, Communication and Aerospace Technology (ICECA) (pp. 1275-1278). IEEE.

[9] Awan, S. M., Riaz, M. U., & Khan, A. G. (2018). Prediction of heart disease using artificial neural network. VFAST Transactions on Software Engineering, 6(1), 51-61.
[10] Yadav, D. C., & Pal, S. A. U. R. A. B. H. (2020). Prediction of heart disease using feature selection and random forest ensemble method. International Journal of Pharmaceutical Research, 12(4).
[11] Sowmiya, C., & Sumitra, P. (2020). A hybrid approach for mortality prediction for heart patients using ACO-HKNN. Journal of Ambient Intelligence and Humanized Computing, 1-8.
[12] Hasan, S. M. M., Mamun, M. A., Uddin, M. P., & Hossain, M. A. (2018, February). Comparative analysis of classification approaches for heart disease prediction. In 2018 International Conference on Computer, Communication, Chemical, Material and Electronic Engineering (IC4ME2) (pp. 1-4). IEEE.
[13] Singhal, S., Kumar, H., & Passricha, V. (2018). Prediction of heart disease using CNN. Am. Int. J. Res. Sci. Technol. Eng. Math, 23(1), 257-261.
[14] Vindhya, L., Beliray, P. A., Sravani, C. R., & Divya, D. R. (2020). Prediction of Heart Disease Using Machine Learning Techniques. International Journal of Research in Engineering, Science and Management, 3(8), 325-326.
[15] Chatterjee, S., Jaggi, Y., & Sowmiya, B. (2019, February). Survey on Prediction of Heart Disease Using Data Mining. In 2019 International Conference on Intelligent Sustainable Systems (ICISS) (pp. 341-344). IEEE.
[16] "Wikipedia, the free encyclopedia," [Online]. Available: https://en.wikipedia.org/wiki/NaN#:~:text=In%20computing%2C%20NaN%20(%2Fn,especially%20in%20floating%2Dpoint%20arithmetic.&text=Quiet%20NaNs%20are%20used%20to,from%20invalid%20operations%20or%20values.. [Accessed 13 May 2021].
[17] "curtis," 29 October 2011. [Online]. Available: http://curtis.ml.cmu.edu/w/courses/index.php/Voted_Perceptron. [Accessed 13 May 2021].