Performance Analysis of Different Classifiers used in Detecting Benign and Malignant Cells of Breast Cancer

Taskin Noor Turna
Lecturer, ICE
Pabna University of Science and Technology

Mst. Alema Khatun
Lecturer (CSE), BBA
Dhaka Commerce College

ABSTRACT
Breast cancer is the most common disease now a days. To get an early detection the target is to find an efficient way to use scientific investigation, because early detection is the only way to remove cancer cell. To predict the accuracy of breast cancer detection, researchers have used different classification techniques. In this paper random forest, Support vector machine, XGBoost, Decision Tree, Naïve Bayes and AdaBoost have been used to analyze and compare the performance. A comparative study is done on these five classifiers using different accuracy measurements like performance, accuracy rate. This study shows that XGBoost gives the high performance among others.

Keywords
SVM, XGBoost, performance, classification, breast cancer

1. INTRODUCTION
In today's world, breast cancer is the most commonly occurred female cancer and increasing day by day. It develops in breast tissue when the bosom cells mutate and form a tumor. This tumor can be of two types, benign and malignant. Benign tumor is the initial stage and almost harmless. But the malignant stage is when the tumor spread successively in the other parts of breast through the lymph system. If the cancer could be detected at an early stage, it is totally curable. So the early detection of benign and malignant cells is very necessary to prevent the disease from being fatal.

In Bangladesh the rate of breast cancer occurrence is 22.5 per 100000 females aged between 15-44 years [1]. About 12764 women are detected with breast cancer every year in Bangladesh which is 8.5% of all cancer detection and 6844 of them die of [2]. In all over the world there were over 2 million new cases of breast cancer in 2018 [3].

In order to detect the benign and malignant cells of breast cancer at an early stage different machine learning techniques are used by the doctors and researchers. There are different types of algorithms already developed and improving day by day for detecting breast cancer such as machine learning, more specifically deep learning, Convolutional neural network based systems which use processing of image sources or data sets. In this paper, different classification techniques used in machine learning system for detecting breast cancer and their performance have been analyzed. The main objectives of this work is to study different types of breast cancer detecting system and analyze the performances of different types of classifiers in order to evaluate the accuracy and overall most convenient system that could be implemented.

This paper incudes related work in section 2. Section 3 includes the system model. Data preprocessing is explained in section 4. Section 5 includes the classification techniques. The result and discussion part is discussed in section 6 and section 7 includes the conclusion.

2. RELATED WORKS
Few research works [4, 10] are related to breast cancer detection and comparison using different machine learning algorithms and different classifiers.

In paper [4] a model is proposed to find out breast cancer mass detection by minimizing the overheads of manual analysis with the help of convolutional neural networks. The authors have evaluated the model to detect mass abnormality and then classified them into benign and malignant in mammogram images.The authors of paper [5] proposed a deep learning algorithm neural network for breast cancer diagnosis by using the Wisconsin Breast Cancer database. They have trained the model to find out the accuracy of deep learning algorithm comparison to other machine learning algorithms which is effective for human data analysis without any special medical knowledge. In Paper [6] breast cancer dataset is used to categorize threatening or nonthreatening cancer by comparing the accuracy measures of different classifiers like decision trees, Naïve Bayes, KNN, and SVM. Different accuracy measures like precision, recall and f1-score were used to find out the effectiveness of classifiers. After their comparative study of different classifiers they found that decision tree classifiers are best among all the classifiers to predict breast cancer. The objective of paper [7] is to summarize various review and technical articles on diagnosis and prognosis of breast cancer which gives an overview of the current research being carried out on various breast cancer datasets using the data mining techniques. The authors' analyzed data mining techniques with the help of the artificial neural network (ANN) and its accuracy is highly acceptable compare to other classification techniques.In paper [8] breast cancer recurrence by comparing the performance of three machine learning techniques is predicted, i.e., Decision Tree (C4.5), Support Vector Machine (SVM), and Artificial Neural Network(ANN). After their evaluation by using different types of parameters named sensitivity, specificity, and accuracy, they found that the accuracy of SVM is the highest among three machine learning techniques. In paper [9], a model is proposed using trained Artificial Neural Network (ANN) and Convolutional Neural Networks (CNN) to diagnose mass types of benign and malignant cells in mammograms. The authors estimated that the comparison between two methods lies in their segmentation technique. In ANN using a region growing algorithm is applied and as a result threshold is obtained. On the contrary in CNN genetic algorithm is implemented to find out the accurate features to diagnose apprehensive masses in mammograms.Paper [10] proposed a model which defines the performance of ANN to select best predicted parameters to find out the presence of breast cancer in thermography on the basis of mean temperature and standard deviation.

3. SYSTEM MODEL

In this paper breast cancer Wisconsin dataset is used which is downloaded from UCI [11, 13]. Number of total instances in this dataset is 569 and attributes is 30. Each instance contains a specific class either malignant or benign which is 37% and 63% respectively. To find the final performance Fig. 1 work flow diagram has followed.

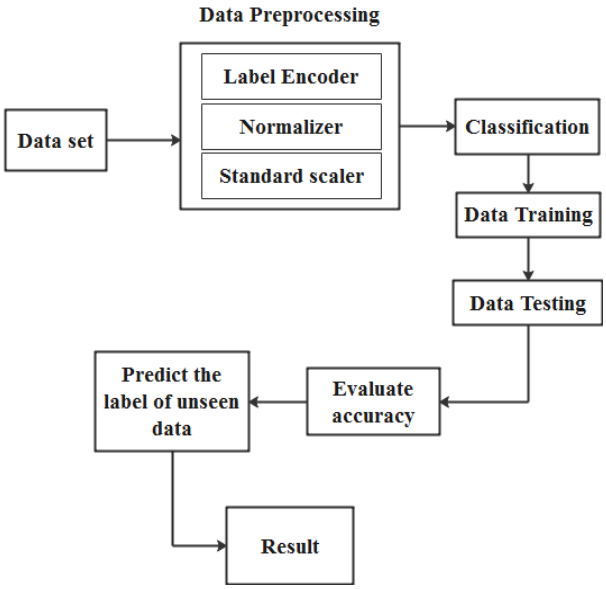

Fig 1: Work Flow Diagram

After getting the dataset is converted into csv file. Then data is preprocessed to prepare for the classification. Data preprocessing includes label encoding, normalizing and standard scaling. The next step is to follow the classification technique which includes training and testing the preprocessed data. In this work six classifiers, Random forest, SVM, XGBoost, Decision tree, Naïve Bayes and AdaBoost are used to compare the performance. After data training and testing, the system will evaluate the accuracy and predict the label of unseen data. Then finally it will give the result.

4. DATA PREPROCESSING

Data preprocessing is the process of transforming or encoding data into a machine understandable form. In this work, the collected data is preprocessed in three steps.

4.1 Label Encoding

To convert the labels of data into numeric form in order to decide a better of operating the labels by the machine learning algorithms is called label encoding. It converts the data in machine readable form by assigning unique number to each class of data [14].

In this work, benign and malignant cells are classified with 0 and 1. After this encoding, to achieve accuracy, neural network dataset is applied. But the accuracy is still not so good. Fig.2 shows the number of malignant and benign cells and the label encoding method.

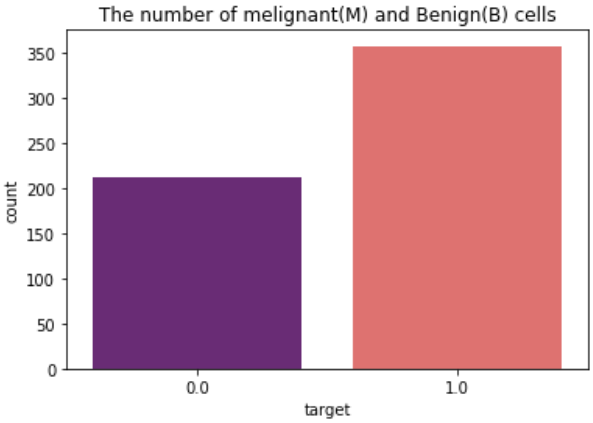

Fig 2: Label encoding

4.2 Normalizing

After label encoding, normalization is used to scale the data of an attribute so that it falls in a smaller range. When multiple attributes with values on different scales appear, this may lead poor performance. In this work, after label encoding data set is converted into numeric data set. Then it is normalized and neural network applied to achieve greater accuracy.

$$V' = \frac{V}{10^j} \text{ ---------- (1)}$$

Where j is the lowest integer.

4.3 Standard Scaling

Standard scaling method includes the process of rescaling attributes so that they have a mean value of 0 and standard deviation of 1. After scaling data in standard scaling method, it is seen that the accuracy is greater.

$$\frac{x_i - mean\ (x)}{stdev\ (x)} \text{ ---------- (2)}$$

5. CLASSIFICATION TECHNIQUES

5.1 Random Forest Classifier

The random forest classifier is an ensemble tree based learning algorithm. It creates a randomly selected subset of training set and form a set of decision trees. To decide the final class of the test object it amalgamates the choices from different decision trees. The ensemble algorithm combines more than one algorithms of same or different kind for classifying objects. It is a highly accurate classifier and runs efficiently on large databases [15].

5.2 Support Vector Machine Classifier

SVM can work on both linear and nonlinear types of data by the conversion of the data fixed for training into a data of higher dimension dataset. It splits the data by a hyperplane or decision boundary to identify different data classes [16]. It gives a better accuracy with less computational power consumption. Following figure shows the data separation process by hyperplane.

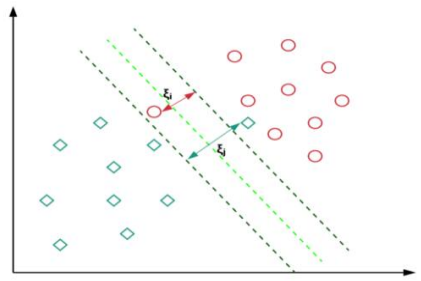

Fig 3: Data separated by hyperplane [17].

5.3 XGboost Classifier

XGboost provides a high performance parallel tree boosting to classify. A single model of dataset is simply trained and then boosted. It takes an iterative approach. Rather training all models, it uses single model succession with new model being trained to correct previous models error [18].

5.4 Decission Tree Classifier

Decision Tree is a supervised learning technique that can be used for both classification and Regression problems, but mostly it is preferred for solving Classification problems. It is a tree-structured classifier, where internal nodes represent the features of a dataset, branches represent the decision rules and each leaf node represents the outcome. In a Decision tree, there are two nodes, which are the Decision Node and Leaf Node. Decision nodes are used to make any decision and have multiple branches, whereas Leaf nodes are the output of those decisions and do not contain any further branches [19].

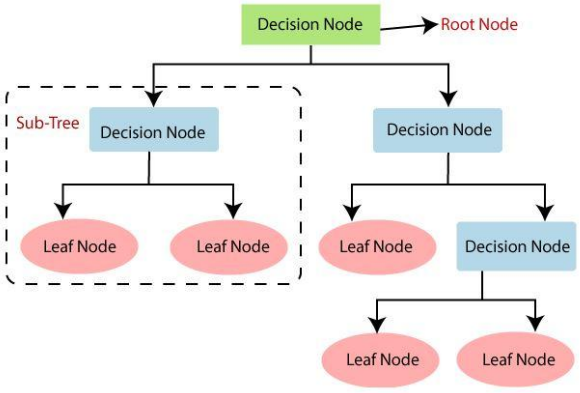

Fig 4: Decision Tree [19].

5.5 Naïve Bayes Classifier

Naive Bayes classifiers are a collection of classification algorithms based on Bayes' Theorem. It is not a single algorithm but a family of algorithms where all of them share a common principle, i.e. every pair of features being classified is independent of each other [20].

5.6 AdaBoost Classifier

An AdaBoost classifier is a meta-estimator that begins by fitting a classifier on the original dataset and then fits additional copies of the classifier on the same dataset but where the weights of incorrectly classified instances are adjusted such that subsequent classifiers focus more on difficult cases. [21]

6. ANALYSIS AND RESULT

The csv file of wiscosin data set is used to process, train and test. After converting it to data frame the some features like mean smoothness, mean area, mean perimeter, mean texture, and mean radius extracted from the frame is pictorially represented in the following figure 5.

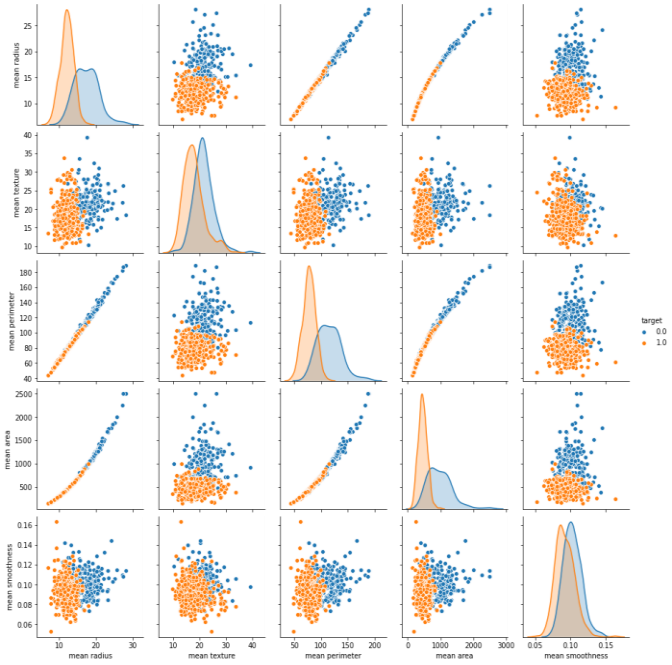

Fig 5:Sample features

After data preprocessing the preprocessed data has been trained and tested. Finally the accuracy rate and the confusion matrix are resulted as the output. The comparison after normalization and standard scaling for different classifiers is stated in the following Table.1.

Table 1. Comparison after normalization and standard scaling

Classifier name	After normalization	After standard scaling
Random forest	0.97	0.75
SVM	0.93	0.96
XGBoost	0.98	0.98
Decision Tree	0.94	0.75
Naïve Bayes	0.94	0.93
AdaBoost	0.94	0.94

Fig. 6, Fig. 7, Fig. 8, Fig. 9, Fig. 10 and Fig.11 represents the performance as the confusion matrix for random forest, SVM, XGBoost, Decision Tree, Naïve Bayes and AdaBoost respectively.

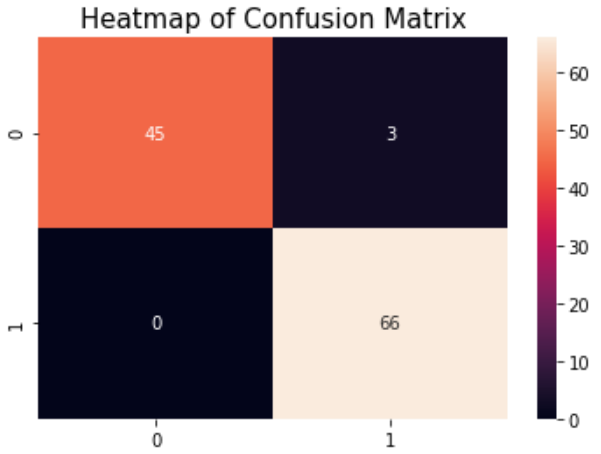

Fig 6: Confusion matrix for random forest

The confusion matrix is a table that shows the performance of a classification model. The upper left and upper right portion of the matrix indicates the true positive and false positive results respectively. The lower left and lower right portion state false negative and true negative results respectively.

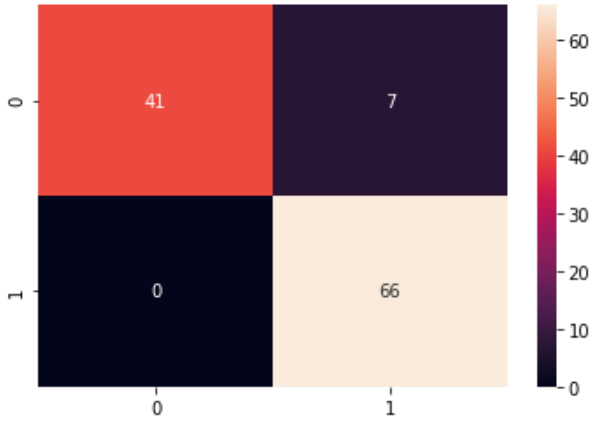

Fig 7: Confusion matrix for SVM

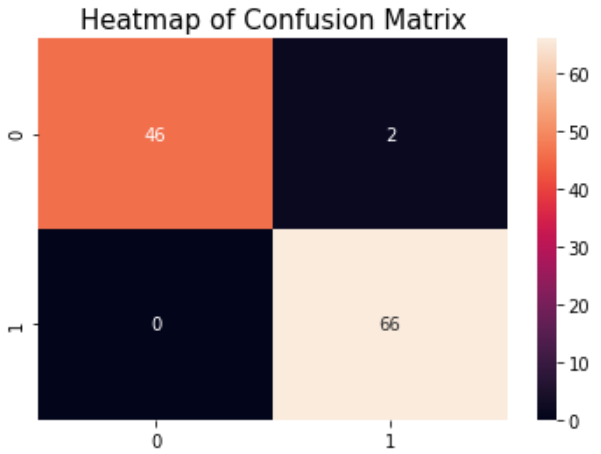

Fig 8: Confusion matrix for XGboost

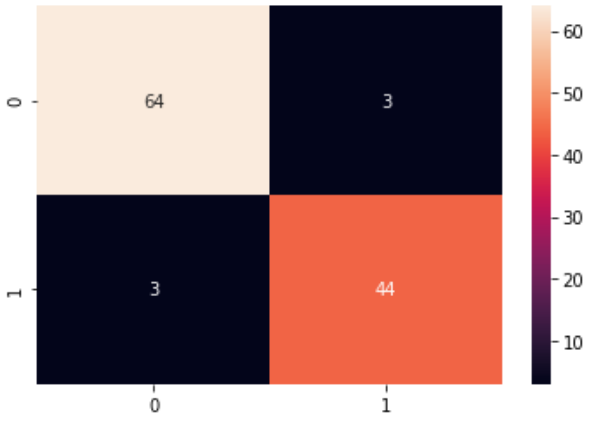

Fig 9: Confusion matrix for Decision Tree

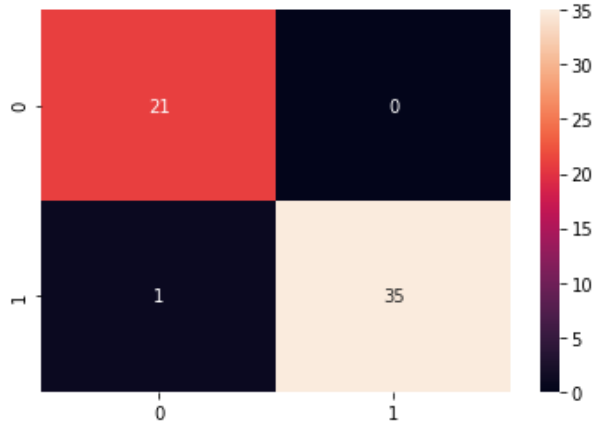

Fig 10: Confusion matrix for Naïve Bayes

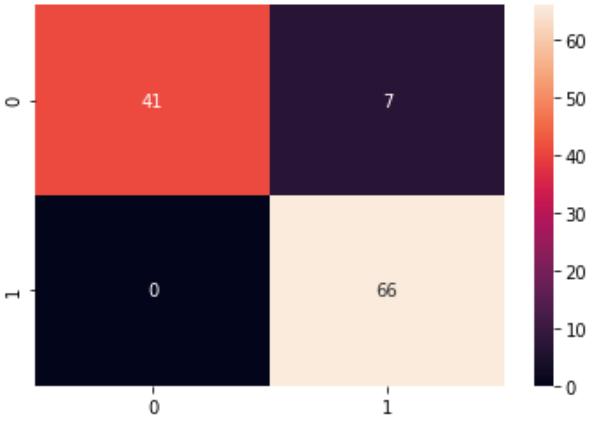

Fig 11: Confusion Matrix for AdaBoost

The Accuracy measurement formula is

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad \ldots\ldots\ldots (3)$$

Where, TP= Number of True positive
TN= Number of True negative
FP= Number of False positive
FN= Number of False negative

Table 2 states the accuracy comparison for random forest, SVM, XGBoost, Decision Tree, Naïve Bayes and AdaBoost.

Table 2. Accuracy Comparison

Classifier	Accuracy
Random Forest	97%
SVM	93%
XGBoost	98%
Decision Tree	95%
Naïve Bayes	97%
AdaBoost	93%

The result shows that the accuracy rate of random forest, SVM, XGBoost, Decision Tree, Naïve Bayes and AdaBoost are 97%, 93%, 98%, 95%, 97% and 93% respectively. XGBoost gives the best accuracy rate.

7. CONCLUSION

In this paper, the accuracy measurement of different classification technique like random forest, SVM, XGBoost, Decision tree, Naïve Bayes and AdaBoost are compared which are used to classify and detect breast cancer. This research is conducted on Wisconsin data set. The result shows that XGBoost classifier is the best among all that can be used to predict breast cancer for benign and malignant cells. In future more classifiers can be used for finding the best accuracy rate and also new classifiers can be proposed that can give a better result.

8. REFERENCES

[1] Breast Cancer Definition: https://www.longdom.org/ (accessed 16.06.21).

[2] Breast Cancer Statistics: https://www.thefinancialexpress.com.bd/health/breast-cancer-takes-6844-lives-in-bangladesh-every-year-1570707616/ (accessed 17.06.21).

[3] Breast Cancer Statistics: https://www.wcrf.org/dietandcancer/cancer-trends/breast-cancer-statistics/ (accessed 18.06.21).

[4] S.H. Nrea, Y.G. Gezahegn, A.S. Boltena, G. Hagos, Breast cancer detection using convolutional neural networks, In: Workshop paper at AI4AH, ICLR (2020).

[5] N. Khuriwal, N. Mishra, Breast Cancer Diagnosis Using Deep Learning Algorithm. In: International Conference on Advances in Computing, Communication Control and Networking (2018) 98-103.

[6] Maheshwar, G. Kumar, Breast Cancer Detection Using Decision Tree, Naïve Bayes, KNN and SVM Classifiers: A Comparative Study. In: 2nd International Conference on Smart Systems and Inventive Technology, 683-686 (2019).

[7] Gupta, S., Kumar, D., Sharma, A.: Data Mining Classification Techniques Applied For Breast Cancer Diagnosis And Prognosis. Indian Journal of Computer Science and Engineering 2(2), (2011) 188-195.

[8] L.G. Ahmad, A.T. Eshlaghy, A. Poorebrahimi, M. Ebrahimi, A.R. Razavi, Using Three Machine Learning Techniques for Predicting Breast Cancer Recurrence. Journal of Health & Medical Informatics 4(2), (2013).

[9] R. Rouhi, M. Jafari, S. Kasaei, P. Keshavarzian, Benign and Malignant Breast Tumors Classification Based On Region Growing and CNN Segmentation. Expert Systems with Application xxx (2014) xxx-xxx, 1-13.

[10] J. Koay, C. Herry, M. Frize, Analysis of Breast Thermography with an Artificial Neural Network. Proceedings of the 26th Annual International Conference of the IEEE EMBS San Francisco, CA, USA, (2004) 1159-1162.

[11] W.N. Street, W.H. Wolberg, O.L. Mangasarian, Nuclear feature extraction for breast tumor diagnosis. In: International Symposium on Electronic Imaging: Science and Technology, San Jose, CA, vol. 1905, (1993) 861-870.

[12] O.L. Mangasarian, W.N. Street, W.H. Wolberg, Breast cancer diagnosis and prognosis via linear programming. Operations Research 43(4), (1995) 570-577.

[13] W.H. Wolberg, W.N. Street, O.L. Mangasarian, Machine learning techniques to diagnose breast cancer from fine-needle aspirates. Cancer Letters 77 (1994).

[14] Label encoding:www.geeksforgeeks.org/ (accessed 24.06.21).

[15] Random forest classification: https://towardsdatascience.com/random-forest-classification-and-its-implementation-d5d840dbead0/ (accessed 25.06.21).

[16] Support vector machine classifier: https://www.javatpoint.com/machine-learning-support-vector-machine-algorithm?fbclid=IwAR3iZFe32lumxzUcnwTJZU_xXErtLI2Nd_p6JC9y5otqJcLk_k-1iIxR_-0/(accessed 26.06.21).

[17] Support Vector Machines Soft Margin Formulation and Kernel Trick: https://towardsdatascience.com/support-vector-machines-soft-margin-formulation-and-kernel-trick-4c9729dc8efe/ (accessed 27.06.21).

[18] XGBoost Classifier: https://towardsdatascience.com/a-beginners-guide-to-xgboost-87f5d4c30ed7/ (accessed 28.06.21).

[19] Decision tree classification algorithm: http://www.javapoint.com (accessed 30.06.21).

[20] Naïve Bayes Classifier: www.GeeksforGeeks.org (accessed 09.07. 21).

[21] AdaBoost Classifier: www.Scikit-learn.org (accessed 10.7.21)