A Comparative Study of Different Ensemble Learning Techniques using Wisconsin Breast Cancer dataset

Dr. Chandan Banerjee
Department of Information Technology
Netaji Subhash Engineering College
Kolkata, India
chandanbanerjee1@gmail.com

Sayak Paul
Department of Information Technology
Netaji Subhash Engineering College
Kolkata, India
spsayakpaul@gmail.com

Moinak Ghoshal
Department of Information Technology
Netaji Subhash Engineering College
Kolkata, India
moinakghoshal1@gmail.com

***Abstract*—The researches in the world Machine Learning and Artificial Intelligence are increasing as the modern day progresses. Finding manifold applications in wide range of fields the art of Machine Learning only promises to get better. Predictive models form the core of Machine Learning. Better the accuracy better the model is and so is the solution to a particular problem. Ensemble Learning algorithms are a set of algorithms which are used enhance the predictive accuracy of a predictive model. In this paper, a comparative study of different Ensemble Learning techniques has been presented using the Wisconsin Breast Cancer dataset. The primary objective is a classification task. This comparative study will help the researchers to find the most suitable Ensemble Learning technique for improving the results.**

Keywords— Ensemble Learning; Breast Cancer dataset; scikit-learn; Cross Validation; Predictive Models

I. INTRODUCTION (*HEADING 1*)

In the world of Statistics and Machine Learning, Ensemble Learning techniques attempt to make the performance of the predictive models better by improving their accuracy. Ensemble Learning is a process using which multiple models, such as classifiers are strategically constructed to solve a particular problem [1]. Ensemble Learning can not only improve the classification accuracy of a particular classifier but also it can be used to improve any function approximation, assign a confidence to the decision made by a model, select optimal (or approximately optimal) features. It also finds its use in Data Fusion, Incremental Learning also [2].

Classification belongs to the family of Supervised Learning problems. Ensemble Learning improves the performance of a particular classifier by forming multiple hypotheses for it which in turn produces a better hypothesis for the given problem. In other words, an Ensemble is a set of classifiers whose individual decisions are combined in some strategy (typically by weighted or unweighted voting) to make predictions. This is why Ensemble Learning models require more computation. Ensemble models can be trained and they can be used for making predictions which in turn makes them a part of Supervised Learning models. For an ensemble to function more efficiently than any of its individual classifiers there is one must followed condition and that is if the classifiers are accurate and diverse. Ensemble based systems can be incorporated in both of the two scenarios i.e. when data is of large volume and when data is too little.

An accurate classifier is one that has a lower error rate in making predictions. On the other hand, two classifiers are diverse the behaviour of their predictions changes on same population [3].

There exist several techniques for constructing Ensembles such as Bayesian Voting, Manipulation of Training Examples, Injection of randomness, Manipulation of Input Features, Manipulation of Output Targets etc. Ensembles are used in Unsupervised Learning problems also for example Consensus Clustering, Anomaly Detection etc [4].

This paper aims to present a comparative study of different existing Ensemble Learning techniques using the Wisconsin Breast Cancer dataset. For the experimental purpose, only classification based Ensemble Learning techniques has been shown and compared.

In Section II, related works have been discussed. Section III deals with the description of the dataset while Section IV presents the proposed work. In Section V, experimental results are shown and discussed. Conclusion and future work have been briefed in Section VI.

II. RELATED WORK

In 1990, Hansen and Salamon proposed the generalized performance of a neural network can be improved using an ensemble of similarly configured neural networks [5]. In 1990, Schapire proved that a strong classifier in probably approximately correct (PAC) sense can be generated by combining weak classifiers through boosting [6]. In 1991, Jacobs, Jordan, Nowlan, and Hinton, proposed an adaptive approach for mixing local experts for improving Neural Computation [7]. In 1992, Wolpert presented a Stack

Generalization approach for forming Ensembles [8]. Ho, Hull and Srihari showed how decisions of multiple classifiers can be combined for improving the performance of the overall system in 1994 [9]. In 1995, Cho and Kim combined several neural networks with a fuzzy based approach and indeed came up with a robust classification approach [10]. Breiman, presented a Bagging predictors based Ensemble approach in 1996 [11]. In 2000, Allwein, Schapire, and Singer formulated a unifying approach for margin classifiers which could be used to reduce multiclass classification problems to binary classification problems [12]. Giacinto and Roli proposed an approach to automatically design multiple classifier based framework in 2001 [13]. In 2001, Kuncheva and Whitaker came up with enumerative experiment where they took feature subsets for classifier combination [14]. In 2005, Fumera and Roli presented a theoretical analysis of linear combiners for multiple classifier systems [15]. In 2006, Polikar broadly presented Ensemble based systems for effective decision making [16].

III. DATASET DESCRIPTION

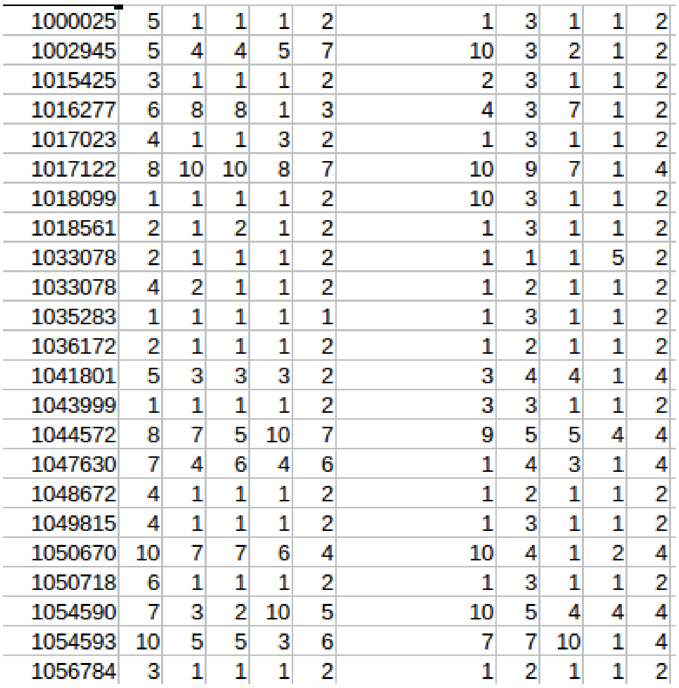
This dataset is popularly known as Wisconsin Breast Cancer dataset and is a standard dataset for applying Machine Learning algorithms and finds its use in Bioinformatics also. The dataset is chosen for the reason being the amount of data is less which makes it legit for constructing Ensemble Learning models.

The dataset contains a total number of 10 features labelled in either benign or malignant classes. The features have 699 instances out of which 16 feature values are missing. The dataset only contains numeric values. Table I presents the names of the features with their value ranges.

TABLE I: FEATURE NAMES AND THEIR VALUE RANGES

No.	Feature Name	Value Range
0	Sample code number	Id No.
1	Clump Thickness:	1 - 10
2	Uniformity of Cell Size	1 – 10
3	Uniformity of Cell Shape	1 – 10
4	Marginal Adhesion	1 – 10
5	Single Epithelial Cell Size	1 – 10
6	Bare Nuclei (Contains missing values)	1 – 10
7	Bland Chromatin	1 – 10
8	Normal Nucleoli	1 – 10
9	Mitoses	1 - 10
10	Class	2 for Benign and 4 for Malignant

In Figure 1, a snapshot of the dataset is shown [17].



1000025	5	1	1	1	2	1	3	1	1	2
1002945	5	4	4	5	7	10	3	2	1	2
1015425	3	1	1	1	2	2	3	1	1	2
1016277	6	8	8	1	3	4	3	7	1	2
1017023	4	1	1	3	2	1	3	1	1	2
1017122	8	10	10	8	7	10	9	7	1	4
1018099	1	1	1	1	2	10	3	1	1	2
1018561	2	1	2	1	2	1	3	1	1	2
1033078	2	1	1	1	2	1	1	1	5	2
1033078	4	2	1	1	2	1	2	1	1	2
1035283	1	1	1	1	1	1	3	1	1	2
1036172	2	1	1	1	2	1	2	1	1	2
1041801	5	3	3	3	2	3	4	4	1	4
1043999	1	1	1	1	2	3	3	1	1	2
1044572	8	7	5	10	7	9	5	5	4	4
1047630	7	4	6	4	6	1	4	3	1	4
1048672	4	1	1	1	2	1	2	1	1	2
1049815	4	1	1	1	2	1	3	1	1	2
1050670	10	7	7	6	4	10	4	1	2	4
1050718	6	1	1	1	2	1	3	1	1	2
1054590	7	3	2	10	5	10	5	4	4	4
1054593	10	5	5	3	6	7	7	10	1	4
1056784	3	1	1	1	2	1	2	1	1	2

Fig. 1 A snapshot of the above-mentioned dataset

IV. PROPOSED WORK

The main objective of this paper is to provide a comparative study of the most commonly used Ensemble Learning techniques to the researchers so that they can get an overview of the performance of different Ensemble Learning techniques to be discussed here and use them accordingly.

In Figure 2, a flowchart has been shown to describe the initial phase of the experimental work.

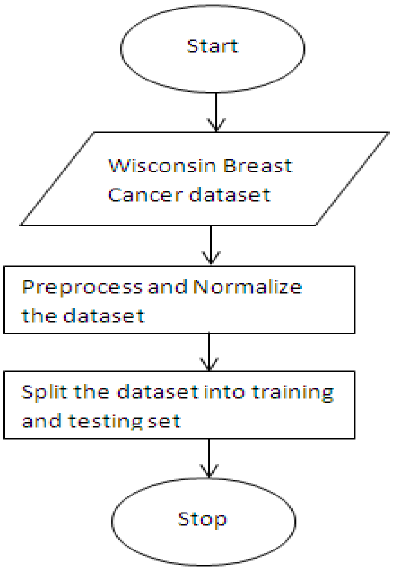

Start
Wisconsin Breast Cancer dataset
Preprocess and Normalize the dataset
Split the dataset into training and testing set
Stop

Fig. 2 Flowchart showing the initial phases of the experimental work

Figure 3 presents the core of the experimental work.

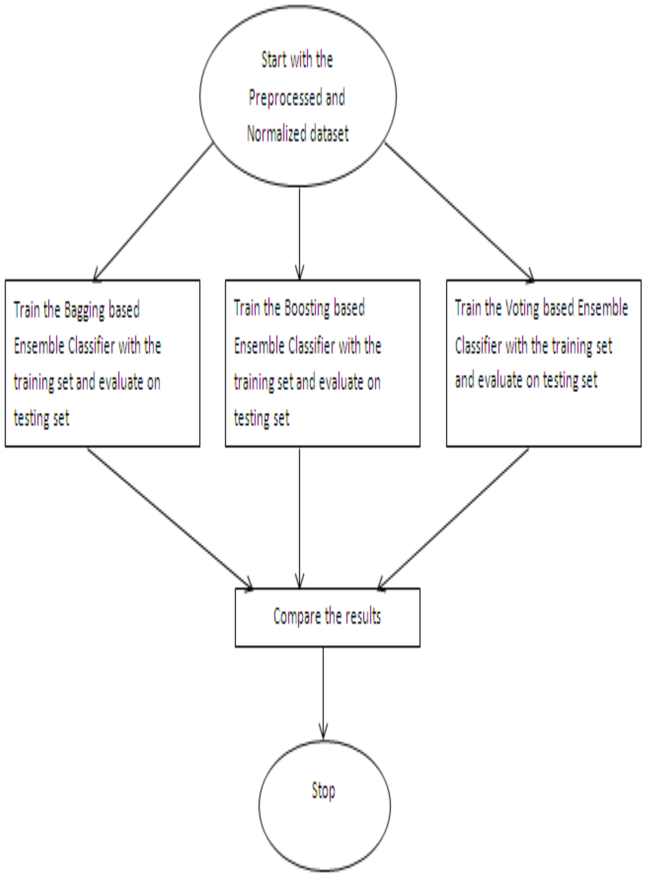

Start with the Preprocessed and Normalized dataset
Train the Bagging based Ensemble Classifier with the training set and evaluate on testing set
Train the Boosting based Ensemble Classifier with the training set and evaluate on testing set
Train the Voting based Ensemble Classifier with the training set and evaluate on testing set
Compare the results
Stop

Fig. 3 Flowchart showing the core activities of the experimental work

A. Preprocessing the dataset

Machine Learning models often produce unexpected results if they are fed with data that has missing values. The dataset that has been used contains missing values. So, to address this issue the missing values in the dataset has been imputed using Mean Imputation technique which is a very popular statistical technique [18]. The Sample Code No. feature has no influence on the class variable. So, to reduce the dimensionality of the dataset and to avoid incorporating insignificant features, Same Code No. column has been discarded which concludes Data Preprocessing step.

B. Normalizing the dataset

After the dataset is preprocessed, the dataset is normalized to fit our calculations. The Normalization is done using MinMax technique [19].

C. Splitting the dataset

The dataset is not available in the form of separate training and testing sets. So, to facilitate efficient evaluation of the models the dataset is split into separate training and testing sets [20].

D. Constructing three different ensemble models

1) *Bagging based Ensemble Classifier*: Bagging is one of the Ensemble Construction techniques which is also known as Bootstrap Aggregation. Bootstrap establishes the foundation of Bagging technique. Bootstrap is a sampling technique in which we select “n” observations out of a population of “n” observations. But the selection is completely random i.e. each observation can be selected from the original population so that each observation is equally likely to selected in each iteration of the Bootstrapping process. After the Bootstrapped samples are formed, separate models are trained with the Bootstrapped samples. In the experiment the Bootstrapped samples are drawn from the training set and the sub-models are tested using the testing set. The final output prediction is combined across the predictions of all of the sub-models. For the experimental purpose, a Decision Tree based classifier model is chosen.

2) *Boosting based Ensemble Classifier:* Boosting is a form of sequential learning technique. The algorithm works by training a model with the entire training set and subsequent models are constructed by fitting the residual error values of the initial model. In this way, Boosting attempts to give higher weight to those observations that were poorly estimated by the previous model. Once the sequence of the models is created the predictions made by models are weighted by their accuracy scores and the results are combined to create a final estimation. Models that are typically used in Boosting technique are XGBoost, GBM, ADABoost etc. ADABoost is used for the experimental purpose.

3) Voting based Ensemble Classifier: Voting is one of the simplest Ensemble Learning techniques in which predictions from multiple models are combined. The method starts with creating two or more separate models with same dataset. Then a Voting based Ensemble Model can be used to wrap the previous models and aggregate the predictions of those models. After the Voting based Ensemble Model is constructed, it can be used to make prediction on new data.

The predictions made by the sub-models can be assigned weights. Stacked Aggregation is a technique which can be used to learn how to weigh these predictions in the best possible way. For the experiment, Logistic Regression, Decision Tree and linear Support Vector Machine have been used as the separate baseline models [16].

E. Comparing the results generated from each technique

In this step, all the results yielded by the above mentioned techniques are compared with respect to their accuracy scores. In this context, accuracy is the measurement of how well the prediction of a model is on the new data. The time needed to execute the techniques has also been compared.

V. EXPERIMENTAL RESULTS

All the techniques discussed above is experimented using the “scikit-learn” Python library. In Figure 4, a graph is shown to compare the execution time of the different Ensemble techniques [21].

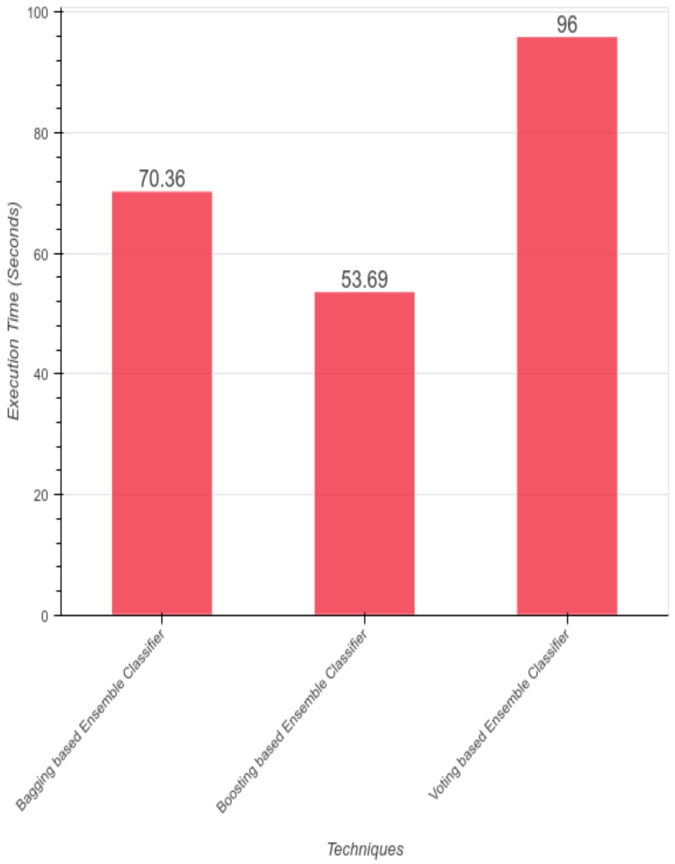


Fig. 4 Graph showing a comparison of the execution time of the above mentioned Ensemble techniques

Figure 5 presents the comparison between the accuracy scores of the techniques.

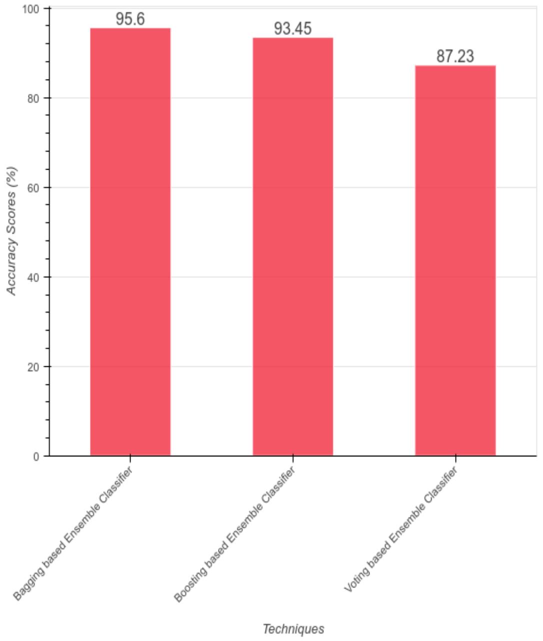


Fig. 5 Graph showing the accuracy scores of the Ensemble techniques

All the sub-models used in the experiments are trained with default parameters as set by the “scikit-learn” library. In Figure 6, the graph shows the accuracy scores of two models, one is trained with Ensemble Learning (Bagging based) and other one is trained without Ensemble Learning (only a simple Decision Tree based model) using the same Wisconsin Breast Cancer dataset.

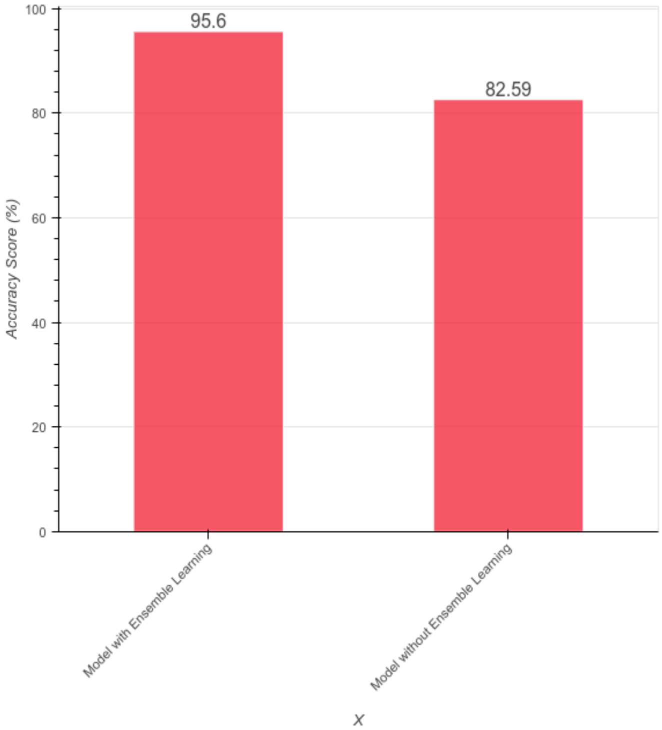


Fig. 6 Graph showing the accuracy score of two models. Among them, one is trained using Ensemble Learning and the other one is trained without Ensemble Learning

Figure 7 presents the execution time of the same two models mentioned in Figure 6.

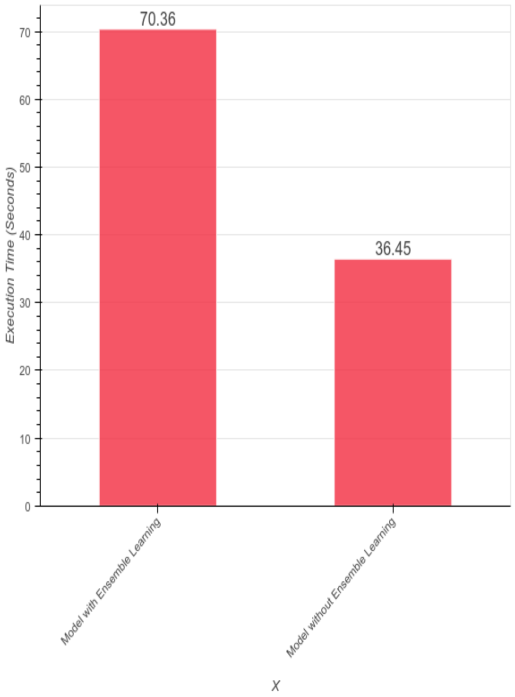


Fig. 7 Bar graph showing the execution time of the two models mentioned in Figure 6

So, from Figure 4, 5, 6 and 7 it can be said that the rightful Ensemble should be chosen considering few aspects with respect to the given task. If a marginal accuracy score with lesser execution time is needed then a simple model can do the work. But if the accuracy is the utmost concern then Bagging based Ensemble model should be chosen as it yields the highest accuracy score and its execution time is also favourable with respect to the other models.

VI. CONCLUSION AND FUTURE WORK

In this paper, we presented a comparative study of different Ensemble Learning techniques. The techniques were applied only on the Classification based models for the purpose. All the experiments were done using the Wisconsin Breast Cancer dataset which is a standard dataset for Machine Learning research. This comparative study will help researchers to find the apt Ensemble model for their purposes which will further help in the improvement of the results. The one limitation of the paper is we took the default settings of the Classification models. In future, we wish to eliminate this limitation by fine tuning the base models to achieve even better results.

REFERENCES

[1] D. Opitz, R. Maclin, “Popular Ensemble Methods: An Empirical Study”, Journal Artificial Intelligence Research, vol. 11, pp. 169 – 198, 1999.

[2] S. Sheet, A. Ghosh and S. B. Mandal, “Selection of genes mediating human leukemia, using an Artificial Neural Network approach”, Third International Conference on Advances in Electrical, Electronics, Information, Communication and Bio-Informatics (AEEICB), Chennai, India, pp. 210-214, 2017.

[3] .M. Bishop, "Pattern Recognition and Machine Learning (Information Science and Statistics)", Springer Book, New York, NY, USA, pp. 130-131, 2007.

[4] L. K. Hansen and P. Salamon, “Neural Network Ensembles”, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 12, pp. 993-1001, 1990.

[5] R. E. Schapire, “The Strength of Weak Learnability”, Machine Learning, vol. 5, pp. 197-227, 1990.

[6] R.A. Jacobs, M. Jordan, S.J. Nowlan, and G.E. Hinton, “Adaptively mixtures of local experts”, Neural Computation, vol. 3, pp. 79-87, 1991.

[7] D. H. Wolpert, “Original Contribution: Stacked generalization”, Neural Networks, vol. 5, pp. 241-259, 1992.

[8] T. K. Ho, J. J. Hull, and S. N. Srihari, "Decision combination in multiple classifier systems," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 16, pp. 66-75, 1994.

[9] S. B. Cho and J. H. Kim, "Combining multiple neural networks by fuzzy integral for robust classification," IEEE Transactions on Systems, Man and Cybernetics, vol. 25, pp. 380-384, 1995.

[10] L. Breiman, "Bagging predictors," Machine Learning, vol. 24, pp. 123-140, 1996.

[11] Han J., Kamber M. and Pei J., Data Mining: Concepts and Techniques (3rd Ed.), Morgan Kaufmann Publishers Inc. San Francisco, pp. 259 - 260, 2011.

[12] G. Giacinto and F. Roli, "Approach to the automatic design of multiple classifier systems," Pattern Recognition Letters, vol. 22, pp. 25-33, 2001.

[13] L. I. Kuncheva, "Switching between selection and fusion in combining classifiers: An experiment," IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics, vol. 32, pp. 146-156, 2002.

[14] G. Fumera and F. Roli, "A Theoretical and Experimental Analysis of Linear Combiners for Multiple Classifier Systems," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 27, pp. 942-956, 2005.

[15] R. Polikar, “Ensemble based systems in decision making,” IEEE Circuits and Systems Magazine, vol. 6, pp. 21-45, 2006.

[16] Wisconsin Breast Cancer dataset available on: https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Original%29

[17] Mean Imputation technique described in scikit-learn official documentation available on: http://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.Imputer.html

[18] MinMax technique described in scikit-learn official documentation available on: http://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.MinMaxScaler.html

[19] Train test split official scikit-learn documentation available on: http://scikit-

learn.org/stable/modules/generated/sklearn.model_selection.train_test_split.html

[20] Scikit-learn official documentation available on: https://scikit-learn.org