The Superiority of the Ensemble Classification Methods: A Comprehensive Review

Silas Nzuva (Corresponding author)
School of Computing and Information Technology
Jomo Kenyatta University of Agriculture and Technology
Nairobi, Kenya.
E-mail: nzuvah@gmail.com

Dr. Lawrence Nderu
School of Computing and Information Technology
Jomo Kenyatta University of Agriculture and Technology
Nairobi, Kenya.
E-mail: nderu@jkuat.ac.ke

Abstract

The modern technologies, which are characterized by cyber-physical systems and internet of things expose organizations to big data, which in turn can be processed to derive actionable knowledge. Machine learning techniques have vastly been employed in both supervised and unsupervised environments in an effort to develop systems that are capable of making feasible decisions in light of past data. In order to enhance the accuracy of supervised learning algorithms, various classification-based ensemble methods have been developed. Herein, we review the superiority exhibited by ensemble learning algorithms based on the past that has been carried out over the years. Moreover, we proceed to compare and discuss the common classification-based ensemble methods, with an emphasis on the boosting and bagging ensemble-learning models. We conclude by out setting the superiority of the ensemble learning models over individual base learners.

Keywords: Ensemble, supervised learning, Ensemble model, AdaBoost, Bagging, Randomization, Boosting, Strong learner, Weak learner, classifier fusion, classifier selection, Classifier combination.

DOI: 10.7176/JIEA/9-5-05
Publication date: August 31st 2019

1. Introduction

Essentially, the ensemble learning models entail the combination of predictions from a number of learners. As a result, they involve the aggregation of different classifier outputs in an effort to take advantage of their complementarity with the anticipation that the developed ensemble algorithm is far much accurate than the individual classifiers [23]. Nevertheless, it is important to note that the underlying assumption in the development of the ensemble models is that there is little or no correlation in the output of the individual models. The fact that the ensemble model is likely to give improved results as compared to the individual model is supported by the Bayesian framework, where the accuracy of the ensemble model can be determined through the calculation of the weighted average of all models; whereby in each model the posteriori probability and prediction weights are taken into consideration.

A wide array of classification algorithms have been developed over the years in search of a model that has high accuracy rates.Some of the ensemble-based algorithms include dynamic classifier selection, classifier fusion, and a combination of multiple classifiers, consensus aggregation, stacked generalization, a mixture of experts, composite classifier systems and bagging, among others. While some models may be suited for a specific context, others may not. Despite the advancement in the development of classification algorithms, the models developed are yet to achieve 100 percent accuracy rate; hence, the ensemble models come in handy as they compliment the individual models and boost the accuracy in the classification of different instances.

2. Developments in Classification Ensembles

Wittner and Denker came up with various strategies that can be applied when training the layered artificial neural network on how to classify new instances [32]. Boosting was later introduced by Schapire [1] and which is well discussed later in this paper. These studies and developments acted as a precursor to the research by Hansen and Salmon, which reviewed the applications and benefits of using ensemble learning model on an artificial neural network [37]. In the research towards a more powerful classification algorithm, the concept of stochastic discrimination was developed by Kleinberg, and which greatly aided in the separation of instances in a multidimensional space [15]. Stochastic discrimination perhaps contributed significantly to the improvement of the ensemble models, as the latter was primarily built on the concept of taking in poor solutions and generating good outcomes. Afterward, a random subspace learning technique was developed as a derivative of stochastic discrimination [2].

Stacked generalization was later on introduced by Wolpert; this technique was mainly aimed at improving the performance of the classifier through minimization of the generalization error rate for one or more of the learners [33]. Afterward, a research was done various methods that can be used in combining multiple classifiers in order to improve their accuracy. Particular, Xu et al. suggested the various methods that can be used to enhance handwriting recognition; specifically, the researchers recommended the combination of different algorithms in order to reduce the possible errors by a single classifier [34]. The researchers explained that in combining the different classifiers, three possible techniques could be used; Dempster-sharef formalism, Bayesian formalism, and voting. The findings by the researchers indicated that the use of the three techniques yielded higher accuracy than individual classifiers; nevertheless, the researchers commented that Dempster-Sharif formalism had a significantly higher accuracy with respect to handwriting recognition as compared to the other two [34].

The research by Xu et al. significantly contributed to research toward ensemble learning models. Ensemble model theoretical framework was later developed by Perrone and Coper, specifically in the construction of regression estimation models [24]. Consequently, a hierarchical mixture of experts model was developed by Jordan and Jacobs, to reduce the error rate in soft probabilistic splits experienced in tree-based models of classification and regression [13]. A ranking based multiple classifiers was afterward suggested by Ho et al. [11] and a similar model developed by Batiti and Cola [40]. After experimentation, the researchers argued that the combined performance resulting from the multiple classifiers was better compared to the best output that could be obtained from the individual classifiers at any given point.

AdaBoost was introduced by Freund and Schapire in 1995; this is a meta-algorithm that the researchers explained it had the potential of increasing the accuracy in classification of the categorical and binary dataset [7]. Developed from bosting, AdaBoost has undergone various development in its architecture over the years to enhance its classification efficiency. Further, the experimentation by the researchers using AdaBoost and bagging, led to the conclusion that boosting outperformed bagging specifically when combined with weak learners in categorical datasets [7]. The researchers proceeded to mention that when combined with c 4.5 algorithm, which is the advancement of ID3 decision-based algorithm, boosting had better prospects of accurate classification as compared to the bagging meta-algorithm[7].

Using fuzzy logic, Cho and Kim employed multiple neural networks in an effort to enhance the accuracy of the classification \cite{Cho}. In a study on the efficiency of the model combination, Lam and Suen categorically focused on weighted majority technique, Bayesian formulations, and the majority vote [18]. The study by the researchers led to the conclusion that the majority vote was the most efficient, reliable, and easy to use in datasets that lacked an explicit representative training set [18]. In their contribution towards ensemble models, Krogh and Vedelsby explained that when training ensemble models, the use of unlabeled data was very crucial and more advantageous as compared to the use of labeled dataset [16].

On a different point of view, research by Tumer and Ghosh indicated that the variance of the area around decision boundaries could be reduced through a linear combination of the neural network when trying to generate an optimum boundary [28]. The findings from this study goes hand in hand with the findings of woods et al., [31] who found improved accuracy in classification through the combination of several classifiers; this was after the researchers developed and tested a method that entailed combination of the classifiers that utilized local accuracy derived from individual classifier estimates with respect to surrounding feature space in unknown test dataset. The experiment by the researchers revealed that leveraging on the individual classifiers local accuracy produced more effective and reliable results as compared to results from behavior knowledge space algorithm, modified classifier rank algorithm and classifier rank algorithm [31]. Of importance, the researchers commented that a combination of various classifiers was much suited to complex and large datasets, which rendered individual classifiers less useful [31]. This finding is in line with the conclusion by Leakey and Croft, who found improved performance in text categorization through the combination of different classifiers [19]. In yet another experiment regarding the combination of different classifiers to exploit their individual strengths, Ho et al. developed and tested the random space method, which the researcher mentioned that is suitable for the construction of decision forests [11]. After several tests, the researcher found the method to be best suited for large and complex datasets, which are characterized by a high of features [11].

In the contribution towards the development of ensembles, Kittler et al. successfully developed a theoretical framework that made it possible to combine classifiers, specifically in distinct representation fusion scenario and identical resonations scenario [14]. For the identical fusion scenario, the aim of the combination was to develop better probabilities of the posteriori class. The findings from the examination of the classifier combination in the identical fusion scenario and the distinct representation fusion scenario using a common theoretical model established that the sum rule combination scheme outperformed the majority voting, median rule, max rule, min rule, and sum rule. However, the fact that the sum rule classifier combination method seemed to perform better based on the evaluation framework developed by the researchers did not imply that it is the best in all scenarios. This is supported by a study by Krawczyk et al., who found the majority voting technique to be highly applicable in pattern recognition [41].

Though different ensemble techniques and associated algorithms have been developed, the accuracy of these models is yet to be determined in various contexts, as one ensemble may be more suited in a specific context than another. In line with this, Opittz and Maclin did a comparison of Arcing and the AdaBoost ensemble methods. The study by the researchers led to the conclusion that AdaBoost was more suited in datasets that had low noise level, and outperformed arching in such conditions \cite{Opittz}. However, Arcing was found to be most applicable in different contexts and specifically outperformed the AdaBoost Meta algorithm in datasets that had a high noise level. This new development perhaps triggered the research by Miller and Yan, who came up with a critic driven framework for the evaluation of ensembles; the data noise factor was also mentioned to be a critical point to consider in the selection of the ensemble method to use [22].

A different study by Jain et al. did outset the various reasons for the use of combined classifiers [12]. According to the researchers, individual classifiers may have different classification sessions, different classification methods, different training sets, and different feature selection aspects; all these aspects when combined in an algorithm make the classifier strong enough and greatly enhance accuracy [12]. A critical conclusion drawn by the researcher from the study is that the combination of different classifiers on the same feature set yields no additional benefit, while the combination of different classifiers over different features yields better results. Overall, best results can, therefore, be obtained through by applying the nearest mean method to the results obtained form combined classifiers for every feature [12].

On a different perspective, algometric implementation of stochastic discrimination by Kleinberhg indicated a higher performance of randomization as compared to bagging and boosting [39]. Similar research by Kuncheva concerning the implementation of ensembles found an ambivalent relationship between the combined algorithm performance and the individual classifier independence [17]. The findings by the researcher showed that the combination of classifiers that had negative independence relationship performed better than those with positive independence [17]. These findings are as well supported by the evolutionary ensemble learning models that had a negative correlation with respect to the individual algorithms. Dietterich comparison of the bagging, boosting and randomization ensemble techniques in improving the accuracy of decision tree established that randomization was more efficient and quite superior than bagging specifically in datasets that had no or little noise; however, boosting outperformed both randomization and bagging [5].

The work of Skrurichina perhaps shed light on the weak classifier stabilization in addition to the comparison of the different ensemble models, with the main focus being the randomization, boosting, and bagging [27]. The researcher concluded that bagging is best suited for unstable and weak classifiers that exhibit training sample size that is critical and has a non-decreasing learning curve. On the other hand, the random subspace technique was found to be useful for unstable and weak classifiers that exhibit critical training sample that is small and has a decreasing learning curve [27]. Boosting was found to be best suited for weak classifiers that are not only weak but also exhibit a non-decreasing curve [27]. Further, large datasets for training was found to be a necessity as well. The findings by Skurichina corroborate those of the Skurichina and Robert With respect to the application of ensembles on linear discriminant analysis. The researchers established that for training sample sizes that were critical, the random subspace method and bagging were more useful while boosting was found to be quite useful in large datasets training [26].

It is literally impossible to talk about able learning without touching on classifier fusion. In line with this, various fusion approaches have been developed over the decades, with the common ones being weighted average and a simple average of the outputs of the individual classifiers. In a study to examine classifier fusion, Fumera and Roli studied the two techniques and concluded that for imbalanced classifiers, the weighted average was more appropriate [9]. The majority vote vs. sum remains a subject of contention to date; various studies have been conducted on the same. The study by Kittler and Alkoot unveiled that the sum technique performed better than majority voting regarding Gaussian estimation error distribution [38]. However, the researchers also established a better performance of majority voting over summation with respect to the heavy tail distribution. Further, with respect to the majority vote, the researchers developed lower and upper limits for the classifiers and concluded that the same negative pairwise dependence is quite beneficial [38]. The research, therefore, underpins the importance of diversity in the development of ensemble methods for handling modern-day problems relative to [pattern recognition. In 2005 Furemera and Roli studied linear classifier combination with respect to simple average rule and weighted average rule [8]. The findings by the researchers pointed out that the performance of the resultant combined classifier dependent on the correlation between the outputs of the individual classifiers as well as the performance of the individual classifiers.[8] This implied if the overall performance of the individual classifiers was low; then the resulting combined classifier was also poor.

On a different note, the work of Džeroski and Zenko explain the essence of the stacking technique in developing ensemble models and explain that the developed models often work better than the best individual classifier via cross-validation [36]. The researchers applied multi-response linear regression and probability distribution while exploring stacking ensemble approach. Based on the researchers' findings, stacking enhanced the accuracy of the output compared to the individual outputs [36]. By experimenting on distributed environment data, Chawla et al. developed a framework that could be used to develop a number of classifiers on small data subsets [3].

Errors of ensemble models have as well received a fair share of studies from different researchers. In line with this, Evgeniou et al. investigated the generalization and leave-one-out errors concerning kernel machine, and specifically the support vector machines [6]. The researchers concluded that a combination of the support vector machines did not result in improved performance as compared to the individual classifiers; nevertheless, the researchers also noted that bagging ensemble technique increased the learning stability and performance of the weak learners. In the same realm of support vector machine-based ensembles, Valentini and Dietterich studied the possible bias-variance errors and argued that to solve the issues, it was critical to utilize low-bias support vector machines when using the bagging ensemble technique, as well as the use of a diverse set of heterogeneous low-bias classifiers [29].

In the urge to enhance the accuracy of the ensembles, Diverse Ensemble Creation by Oppositional Relabeling of Artificial Training Examples (DECORATE) was developed by Melville and Mooney [21]. The technique uses additional training examples that are artificially trained to construct diverse hypotheses. Their technique had comparable performance to AdaBoost on large training datasets but outperformed AdaBoost, bagging and random forests in training datasets that were small.

Reyzin and Schapire study led to the conclusion that the boosting ensemble technique bears the ability to increase the classifiers' complexity [25]. The researchers then concluded that while boosting is a critical ensemble that can aid in the maximization of the margins; it is not best suited in scenarios that are characterized by an increase in the complexity of the base classifier. This conclusion is supported by the findings by Hadjitodorov et al., whose study established that best cluster ensembles were found in ensembles that exhibited a limited level of base classifier complexity [10]. On the issue of random initialization with respect to k-means based clustering ensemble, the researchers established that the dataset was the main factor that affected the accuracy and stability of the clusters.

On a different point of view, various versions of the ensemble models have been developed to obtain ensembles that are more accurate. To begin with, Zhang and Zhang used Adobost's used boosting by resampling version to develop a local boosting algorithm, which they experimented on and found to be highly reliable and accurate compared to AdaBoost [35]. On the same note, Mease and Wyner study found boosting to be more preferable, and categorically mentioned the need to try larger trees in scenarios where the stumps cause overfitting [20]. In explaining this, the researchers mention the need to try higher complexity base learner, whenever the low complexity base learner is found to cause overfitting.

A study by Wang et al. revealed that ensemble learning methods could be used in sentiment classification specifically on content generated by the users online. On the same, the researchers argue that Random subspace bears more reliable results as compared to boosting and bagging [30]. This conclusion was drawn from the 1200 comparative experiments that were performed with respect to sentiment classification by the three ensembles [30].

3. Bagging vs. Boosting

3.1 Bootstrap Aggregation (Bagging)

Commonly known as bagging, this is an ensemble-learning model that is fundamentally designed to foster accuracy and stability of the supervised regression and classification based algorithms [42]. A study by Kong and Dietterich indicates that the Meta algorithm also aids in the reduction of overfitting and variance, hence contributing to a positive classification of the new instances [43]. Developed by Beriman in 1996, this Meta algorithm performs random variable selection and fitting in a given linear model. As such, the fundamental logic that this model operates on is quite simple, as it entails the construction of numerous instances of black-box estimator based on random subsets of data with replacement sampled from the training set using bootstrap sampling; the prediction performance of the numerous instances is then aggregated to determine how a new instance will be classified [44].

The bootstrap sampling and aggregation aids in the reduction of the variance associated with treating the training set as a single sample. Owing to the simplicity associated with this model, it works almost best with all models, irrespective of whether they are strong or weak.

3.2 How Bagging Works

For every iteration, T1, T2, T3….TN

DO:

- Sample out N samples randomly with replacement
- Train the k estimators of the supervised learning model using the samples
- Average the resulting predictions in case of regression
- Perform majority voting in case of classification

TEST

- Predict how to classify a new instance, based on the aggregated results

As a result, by simply using numerous copies of one model, the overall accuracy can be improved. The averaging of the data split classification by the different instance of black-box estimator warrants improved predictive performance [44],[45]. The prediction in the regression is achieved through averaging while in classification; it is attained through majority vote [42],[44]. The bagging meta-algorithm is however best suited in base models that are characterized by underfitting and overfitting due to its ability to reduce the base model variance through averaging or voting without doing any significant change on the bias and high

Experimentation by various researchers has established that bagging ensemble to be quite superior to their associated base models [47],[48],[49]. The researchers continue to assert that the ability of bagging to minimize prediction errors as well as optimize prediction accuracy can be utilized in High-Performance Concrete (HPC) slump flow modeling to attain high accuracy [47]. As identified by the researchers, HPC remains a complex field that can potentially benefit from bagging ensembles, as the latter is often associated with significant noise in the dataset [47]. In the banking industry, Erdal and Karahanoğlu experimented the variants of bagging and specifically Reduced Error Pruning Tree (REPTree), Random Tree (RTree) and Decision Stump (DStump) on financial data and concluded that bagging could be potentially used for prediction of the Turkish Development and Investment Banks profitability determinants [49]. Ekinci and Erdal study found bagging to be capable of optimizing the forecasting crude oil process on a monthly bases, as the latter had higher accuracy compared to the base learner models [48]. Weat et al. study that entails the analysis of the credit scoring through the use of ensemble learning models constructed using k‐nearest neighbors, multilayer perceptron and decision tree unveiled that the bagging enables were more accurate compared to the results of the individual base models [56]. Additionally, the researchers also note that k‐nearest neighbor based bagging algorithm gave the best results and hence commented that in cases involving big unbalanced credit coring datasets [56]. Other applications of bagging include network intrusion detection to enhance true positives and reduce false negatives [50], credit card fraud detection [51], medical diagnosis of arrhythmia beats [52], Urban traffic flow forecasting [53], forecasting of wind and solar power [54] and imbalanced data classification using Evolutionary under-sampling bootstrap aggregation models [55].

3.3 Boosting

Boosting ensemble is another common metal algorithm. The algorithm was developed from the Probably Approximately Correct (PAC) framework coined by Kearns and Valiant in an effort to enhance the accuracy of classifications [59]. Essentially, boosting was developed in an effort to convert the weak learners into strong learners to bolster their overall classification accuracy [44]. This algorithm works by reducing classification variance and bias. Essentially, though weak learners perform better than algorithms that rely on random guessing, they still have low accuracy rates. Hence, boosting allows for the enhancement of the classifier correlation with true classification. While there exist variants of the booting algorithm, they all often work through iterative learning of the weak classifiers to form strong classifiers. Boosting capitalizes in the ability to re-classify the previously misclassified instances through the readjustments of the data weights whenever a new weak learner is added [58].

3.4 AdaBoost

The AdaBoost algorithm is a meta-algorithm that is built on the boosting model and which has widely been used in instances that require strengthening of each classifier [60]. The algorithm was developed by Freund and Schapire in 1995 and has greatly aided in solving a wide array of classification problems [61],[62]. Its adaptive stature hails from the fact that the classifiers in execution are adjusted based on the output of the previous classifiers, specifically concerning the wrongly classified instances. Persistent iterations when reclassifying the wrongly classified instances results in increased accuracy of the classifier for the test instances. Meta-algorithm is, therefore, sensitive to noise in the data and incorrect attributes for the instances [63]. In every iteration by the algorithm, the weight of the correctly classified training instances reduces while the weight of the incorrectly classified instances rises [63]. As the iteration continues, the classifiers are required to focus on reducing the high weights recorded in the wrongly classified cases [61],[62],[63]. To improve performance, the meta-algorithm can be used in combination with other classification algorithms to bolster its performance. While there are variations of the Boosting algorithm, Schapire recommends the use of the AdaBoost, an award-winning adaptive boosting algorithm that greatly enhances accurate classification in addition to its history of being able to adapt to weak learners [62]. Schapire explains that the adaptive nature of the AdaBoost algorithm, as well as its sequential approach to learning, make it is well suited for large datasets [62].

3.5 How AdaBoost Works

Pseudocode of AdaBoost:

```
Set uniform example weights.
FOR Each base-learner do:
•       Train base-learner with a weighted sample.
•       Test base-learner on all data
•       Set learner weight with a weighted error.
•       Set example weights based on ensemble predictions.
END FOR
```

With respect to applications, Kudo explains that boosting meta-algorithms can be used in semi-structured texts classification, which in turn can foster the organization of the Emails, internet news, and World Wide Web [65]. Experimentation various researchers shows that the AdaBoost eta algorithm and the support vector machines can aid in achieving real-time facial detection and expression recognition [66]. This can be used to improve human-computer interaction in designing user interfaces. We unveiled that the Adaboost Meta algorithm can be sued to foster the accuracy of multimedia classification through the use of an exponential loss function and reduction of multiclass classification into a two-class problem [67]. Further, AdaBoost meta-algorithm can as well be used for the reduction of financial distress [55], through foresting of future economic trends, accurate prediction of future financial distress for business organizations [69], and increased accuracy in bankruptcy forecasting [70],[71].

4. Ensemble Algorithm Selection

The different ensemble-based algorithms are fundamentally different from each other on a number of factors including how the rules are combined to get the ensemble decision, the procedure that is used to generate the ensemble members, and how the training set for the individual classifier is selected. According to Mangai, Samanta, Das, and Chowdhury, two critical settings must be well defined and understood in the selection of ensemble-based system; these in include the classifier selection and classifier fusion [56].

With respect to classifier selection, the feature space is divided into local neighborhoods, and the classifiers trained as local experts in respective neighborhoods in the learning phase [56]. In the test phase, whenever a new instance occurs, each of the classifiers check the new instance against the learning data in the local vicinity, and the classifier that was trained with data similar to or close to the distance as compared to the other classifiers is given the first priority or weight in contribution towards the final decision, or it can as well make the final decision [72]. In classifier fusion, the training of the classifiers is done in the entire feature space then put together to get a combined classifier that has little variance, and ultimately, lower rate.

Classifier fusion can be based on various reasons, including but not limited to continuous class-specific output or labels only [73]. However, for the class-specific, normalization of the classifier output has to be done to [1,0] interval, which then can be interpreted as the classifier support to various classes. This process touches on various aspects such as decision templates [74], the Dempster–Shafer framework [75] and rules concerning algebraic combination, including but not limited to product, sum, minimum, maximum, weighted or simple majority voting among other class-specific output algebraic combinations [56].

5. Conclusion

We conclude that the ensemble learning methods can be grouped into two; those that focus on establishing ways to integrate the single classification models to obtain a hybrid model that is highly accurate, and those that are oriented towards developing new training datasets from the original training dataset for training the different single models. A pertinent question when choosing an ensemble is how the individual classifiers can be blended to enhance their accuracy to improve the overall performance of the learning model, more so in multiple classification problems.

The use of ensemble models in supervised learning furthers the ideology that the classification is done in light of the knowledge gained using several models. This implies that the output obtained from ensemble models is far much accurate and reliable as compared to the output generated by the individual classifiers.Different classification models are more appropriate in different contexts; as such, there is no single model that performs exceptionally well than others in all contexts.

Ensemble methods have been argued to provide optimal solutions in machine learning. “A composite classifier system design” by Dasarathy and Sheela on the use of multiple classifiers in feature space partitioning is one of the pioneering examples of ensemble model development. An improvement in the classification performance could be made by applying a similar ensemble configuration on different classification-based algorithms. Based on the findings of this review, it possible to construct an arbitrary low-error classifier in a binary classification problem using ensemble models, which in turn can be used in solving regression and multiple class problems.

The ensemble based algorithms have gained popularity in both research and application, due to their ability to enhance the accuracy of the classifiers. The AdaBoost, which is one of the common ensemble models, builds iterative models through variance of the case weights by down-weighting of the cases correctly estimated and up-weighting of the cases that have significantly large errors, and the use of the weighted sum of the models’ sequence. The Random forest, which is a variation of the bragging algorithm, focusses more on enhancing diversity to increase the accuracy of the output by adding a stochastic element on the decision trees being combined. Bootstrap aggregating utilizes the average of the estimates or the majority vote after bootstrapping the training dataset.

6. References

[1] Schapire, Robert E. "The strength of weak learnability." Machine learning 5, no. 2 (1990): 197-227.

[2] Barandiaran, Iñigo. "The random subspace method for constructing decision forests." IEEE transactions on pattern analysis and machine intelligence 20, no. 8 (1998).

[3] Chawla, Nitesh V., Lawrence O. Hall, Kevin W. Bowyer, and W. Philip Kegelmeyer. "Learning ensembles from bites: A scalable and accurate approach." Journal of Machine Learning Research 5, no. Apr (2004): 421-451.

[4] Cho, Sung-Bae, and Jin H. Kim. "Multiple network fusion using fuzzy logic." IEEE Transactions on Neural Networks 6, no. 2 (1995): 497-501.

[5] Dietterich, Thomas G. "An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting, and randomization." Machine learning 40, no. 2 (2000): 139-157.

[6] Evgeniou, Theodoros, Massimiliano Pontil, and André Elisseeff. "Leave one out error, stability, and generalization of voting combinations of classifiers." Machine learning 55, no. 1 (2004): 71-97.

[7] Freund, Yoav, and Robert E. Schapire. "Experiments with a new boosting algorithm." In Icml, vol. 96, pp. 148-156. 1996.

[8] Fumera, Giorgio, and Fabio Roli. "A theoretical and experimental analysis of linear combiners for multiple classifier systems." IEEE Transactions on Pattern Analysis and Machine Intelligence 27, no. 6 (2005): 942-956.

[9] Fumera, Giorgio, and Fabio Roli. "Performance analysis and comparison of linear combiners for classifier fusion." In Joint IAPR International Workshops on Statistical Techniques in Pattern Recognition (SPR) and Structural and Syntactic Pattern Recognition (SSPR), pp. 424-432. Springer, Berlin, Heidelberg, 2002.

[10] Hadjitodorov, Stefan T., Ludmila I. Kuncheva, and Ludmila P. Todorova. "Moderate diversity for better cluster ensembles." Information Fusion 7, no. 3 (2006): 264-275.

[11] Ho, Tin Kam, Jonathan J. Hull, and Sargur N. Srihari. "Decision combination in multiple classifier systems." IEEE transactions on pattern analysis and machine intelligence 16, no. 1 (1994): 66-75.

[12] Jain, Anil K., Robert PW Duin, and Jianchang Mao. "Statistical pattern recognition: A review." IEEE Transactions on pattern analysis and machine intelligence 22, no. 1 (2000): 4-37.

[13] Jordan, Michael I., and Robert A. Jacobs. "Hierarchical mixtures of experts and the EM algorithm." Neural computation 6, no. 2 (1994): 181-214.

[14] Kittler, Josef, Mohamad Hatef, Robert PW Duin, and Jiri Matas. "On combining classifiers." IEEE transactions on pattern analysis and machine intelligence 20, no. 3 (1998): 226-239.

[15] Kleinberg, E. M. (1990). Stochastic discrimination. Annals of Mathematics and Artificial intelligence, 1(1), 207-239.

[16] Krogh, Anders, and Jesper Vedelsby. "Neural network ensembles, cross validation, and active learning." In Advances in neural information processing systems, pp. 231-238. 1995.

[17] Kuncheva, Ludmila I. "A theoretical study on six classifier fusion strategies." IEEE Transactions on Pattern Analysis and Machine Intelligence 2 (2002): 281-286.

[18] Lam, Louisa, and Ching Y. Suen. "Optimal combinations of pattern classifiers." Pattern Recognition Letters 16, no. 9 (1995): 945-954.

[19] Larkey, Leah S., and W. Bruce Croft. "Combining classifiers in text categorization." In Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval, pp. 289-297. ACM, 1996.

[20] Mease, David, and Abraham Wyner. "Evidence contrary to the statistical view of boosting." Journal of Machine Learning Research 9, no. Feb (2008): 131-156.

[21] Melville, Prem, and Raymond J. Mooney. "Creating diversity in ensembles using artificial data." Information Fusion 6, no. 1 (2005): 99-111.

[22] Miller, David J., and Lian Yan. "Critic-driven ensemble classification." IEEE Transactions on Signal Processing 47, no. 10 (1999): 2833-2844.

[23] Opitz, David, and Richard Maclin. "Popular ensemble methods: An empirical study." Journal of artificial intelligence research 11 (1999): 169-198.

[24] Perrone, Michael P., and Leon N. Cooper. "When networks disagree: Ensemble methods for hybrid neural networks." In How We Learn; How We Remember: Toward an Understanding of Brain and Neural Systems: Selected Papers of Leon N Cooper, pp. 342-358. 1995.

[25] Reyzin, Lev, and Robert E. Schapire. "How boosting the margin can also boost classifier complexity." In Proceedings of the 23rd international conference on Machine learning, pp. 753-760. ACM, 2006.

[26] Skurichina, Marina, and Robert PW Duin. "Bagging, boosting and the random subspace method for linear classifiers." Pattern Analysis and Applications 5, no. 2 (2002): 121-135.

[27] Skurichina, Marina. "Stabilizing weak classifiers: Regularization and combining techniques in

discriminant analysis." PhD diss., TU Delft, Delft University of Technology, 2001.

[28] Tumer, Kagan, and Joydeep Ghosh. "Analysis of decision boundaries in linearly combined neural classifiers." Pattern Recognition 29, no. 2 (1996): 341-348.

[29] Valentini, Giorgio, and Thomas G. Dietterich. "Bias-variance analysis of support vector machines for the development of SVM-based ensemble methods." Journal of Machine Learning Research 5, no. Jul (2004): 725-775.

[30] Wang, Gang, Jianshan Sun, Jian Ma, Kaiquan Xu, and Jibao Gu. "Sentiment classification: The contribution of ensemble learning." Decision support systems 57 (2014): 77-93.

[31] Woods, Kevin, W. Philip Kegelmeyer, and Kevin Bowyer. "Combination of multiple classifiers using local accuracy estimates." IEEE transactions on pattern analysis and machine intelligence 19, no. 4 (1997): 405-410.

[32] Wittner, B. S., and Denker, J. S. (1988). Strategies for teaching layered networks classification tasks. In Neural information processing systems (pp. 850-859).

[33] Wolpert, David H. "Stacked generalization." Neural networks5, no. 2 (1992): 241-259.

[34] Xu, Lei, Adam Krzyzak, and Ching Y. Suen. "Methods of combining multiple classifiers and their applications to handwriting recognition." IEEE transactions on systems, man, and cybernetics 22, no. 3 (1992): 418-435.

[35] Zhang, Chun-Xia, and Jiang-She Zhang. "A local boosting algorithm for solving classification problems." Computational Statistics and Data Analysis 52, no. 4 (2008): 1928-1941.

[36] Džeroski, S., and Ženko, B. (2004). Is combining classifiers with stacking better than selecting the best one?. Machine learning, 54(3), 255-273.

[37] Hansen, L. K., and Salamon, P. (1990). Neural network ensembles. IEEE transactions on pattern analysis and machine intelligence, 12(10), 993-1001.

[38] Kittler, Josef, and Fuad M. Alkoot. "Sum versus vote fusion in multiple classifier systems." IEEE Transactions on Pattern Analysis and Machine Intelligence 1 (2003): 110-115.

[39] Kleinberg, Eugene M. "On the algorithmic implementation of stochastic discrimination." IEEE Transactions on Pattern Analysis and Machine Intelligence 5 (2000): 473-490.

[40] Battiti, Roberto, and Anna Maria Colla. "Democracy in neural nets: Voting schemes for classification." Neural Networks 7, no. 4 (1994): 691-707.

[41] Krawczyk, Bartosz, Leandro L. Minku, João Gama, Jerzy Stefanowski, and Michał Woźniak. "Ensemble learning for data stream analysis: A survey." Information Fusion 37 (2017): 132-156.

[42] Hothorn, Torsten, and Berthold Lausen. "Double-bagging: Combining classifiers by bootstrap aggregation." Pattern Recognition 36, no. 6 (2003): 1303-1309.

[43] Kong, Eun Bae, and Thomas G. Dietterich. "Error-correcting output coding corrects bias and variance." In Machine Learning Proceedings 1995, pp. 313-321. 1995.

[44] Bühlmann, Peter, and Bin Yu. "Analyzing bagging." The Annals of Statistics 30, no. 4 (2002): 927-961.

[45] Ganjisaffar, Yasser, Rich Caruana, and Cristina Videira Lopes. "Bagging gradient-boosted trees for high precision, low variance ranking models." In Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval, pp. 85-94. ACM, 2011.

[46] Buja, Andreas, and Werner Stuetzle. "Observations on bagging." Statistica Sinica (2006): 323-351.

[47] Aydogmus, Hacer Yumurtacı, H. İ. Erdal, Onur Karakurt, Ersin Namli, Yusuf S. Turkan, and Hamit Erdal. "A comparative assessment of bagging ensemble models for modeling concrete slump flow." Computers and Concrete 16, no. 5 (2015): 741-757.

[48] Ekinci, Aykut, and Hamit Erdal. "Optimizing the monthly crude oil price forecasting accuracy via bagging ensemble models." Journal of Economics and International Finance 7, no. 5 (2015): 127-136.

[49] Erdal, Hamit, and İlhami Karahanoğlu. "Bagging ensemble models for bank profitability: An emprical research on Turkish development and investment banks." Applied Soft Computing49 (2016): 861-867.

[50] Gaikwad, D. P., and Ravindra C. Thool. "Intrusion detection system using bagging ensemble method of machine learning." In Computing Communication Control and Automation (ICCUBEA), 2015 International

Conference on, pp. 291-295. IEEE, 2015.

[51] Zareapoor, Masoumeh, and Pourya Shamsolmoali. "Application of credit card fraud detection: Based on bagging ensemble classifier." Procedia Computer Science 48, no. 2015 (2015): 679-685.

[52] Mert, Ahmet, Niyazi Kılıç, and Aydın Akan. "Evaluation of bagging ensemble method with time-domain feature extraction for diagnosing of arrhythmia beats." Neural Computing and Applications 24, no. 2 (2014): 317-326.

[53] Moretti, Fabio, Stefano Pizzuti, Stefano Panzieri, and Mauro Annunziato. "Urban traffic flow forecasting through statistical and neural network bagging ensemble hybrid modeling." Neurocomputing 167 (2015): 3-7.

[54] Ren, Ye, P. N. Suganthan, and N. Srikanth. "Ensemble methods for wind and solar power forecasting—A state-of-the-art review." Renewable and Sustainable Energy Reviews 50 (2015): 82-91.

[55] Sun, Bo, Haiyan Chen, Jiandong Wang, and Hua Xie. "Evolutionary under-sampling based bagging ensemble method for imbalanced data classification." Frontiers of Computer Science 12, no. 2 (2018): 331-350.

[56] West, David, Scott Dellana, and Jingxia Qian. "Neural network ensemble strategies for financial decision applications." Computers and operations research 32, no. 10 (2005): 2543-2559.

[57] Bühlmann, Peter, and Torsten Hothorn. "Twin boosting: improved feature selection and prediction." Statistics and Computing 20, no. 2 (2010): 119-138..

[58] Bühlmann, Peter, and Bin Yu. "Boosting." Wiley Interdisciplinary Reviews: Computational Statistics 2, no. 1 (2010): 69-74.

[59] Martinez, Waldyn, and J. Brian Gray. "Noise peeling methods to improve boosting algorithms." Computational Statistics and Data Analysis 93 (2016): 483-497.

[60] Tanha, Jafar, Maarten van Someren, and Hamideh Afsarmanesh. "An adaboost algorithm for multiclass semi-supervised learning." In Data Mining (ICDM), 2012 IEEE 12th International Conference on, pp. 1116-1121. IEEE, 2012.

[61] Mathanker, S. K., P. R. Weckler, T. J. Bowser, N. Wang, and N. O. Maness. "AdaBoost classifiers for pecan defect classification." computers and electronics in Agriculture 77, no. 1 (2011): 60-68.

[62] Barrow, Devon K., and Sven F. Crone. "A comparison of AdaBoost algorithms for time series forecast combination." International Journal of Forecasting 32, no. 4 (2016): 1103-1119.

[63] Sen, Sanjay Kumar, and Sujatha Dash. "Meta learning algorithms for credit card fraud detection." International Journal of Engineering Research and Development 6, no. 6 (2013): 16-20.

[64] Iwakura, Tomoya. "A Boosting-based Algorithm for Classification of Semi-Structured Text using the Frequency of Substructures." In Proceedings of the International Conference Recent Advances in Natural Language Processing RANLP 2013, pp. 319-326. 2013.

[65] Kudo, Taku, and Yuji Matsumoto. "A boosting algorithm for classification of semi-structured text." In Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing. 2004.

[66] Ghimire, Deepak, and Joonwhoan Lee. "Geometric feature-based facial expression recognition in image sequences using multi-class adaboost and support vector machines." Sensors13, no. 6 (2013): 7714-7734.

[67] Hao, Wei, and Jiebo Luo. "Generalized multiclass adaboost and its applications to multimedia classification." In Computer Vision and Pattern Recognition Workshop, 2006. CVPRW'06. Conference on, pp. 113-113. IEEE, 2006.

[68] Sun, Jie, Ming-Yue Jia, and Hui Li. "AdaBoost ensemble for financial distress prediction: An empirical comparison with data from Chinese listed companies." Expert Systems with Applications 38, no. 8 (2011): 9305-9312.

[69] Kim, Soo Y., and Arun Upneja. "Predicting restaurant financial distress using decision tree and AdaBoosted decision tree models." Economic Modelling 36 (2014): 354-362.

[70] Alfaro, Esteban, Noelia García, Matías Gámez, and David Elizondo. "Bankruptcy forecasting: An empirical comparison of AdaBoost and neural networks." Decision Support Systems45, no. 1 (2008): 110-122.

[71] Vieira, Armando S., João Duarte, Bernardete Ribeiro, and João C. Neves. "Accurate prediction of financial distress of companies with machine learning algorithms." In International Conference on Adaptive and

Natural Computing Algorithms, pp. 569-576. Springer, Berlin, Heidelberg, 2009.

[72] Woźniak, Michał, Manuel Graña, and Emilio Corchado. "A survey of multiple classifier systems as hybrid systems." Information Fusion 16 (2014): 3-17.

[73] García-Pedrajas, Nicolás, and Domingo Ortiz-Boyer. "An empirical study of binary classifier fusion methods for multiclass classification." Information Fusion 12, no. 2 (2011): 111-130.

[74] Walter, Steffen, Stefan Scherer, Martin Schels, Michael Glodek, David Hrabal, Miriam Schmidt, Ronald Böck, Kerstin Limbrecht, Harald C. Traue, and Friedhelm Schwenker. "Multimodal emotion classification in naturalistic user behavior." In International Conference on Human-Computer Interaction, pp. 603-611. Springer, Berlin, Heidelberg, 2011.

[75] Denoeux, Thierry, Nicole El Zoghby, Véronique Cherfaoui, and Antoine Jouglet. "Optimal object association in the Dempster–Shafer framework." IEEE transactions on cybernetics 44, no. 12 (2014): 2521-2531.