Different Approaches to Reducing Bias in Classification of Medical Data by Ensemble Learning Methods

Adem Doganer, Kahramanmaras Sutcu Imam University, Kahramanmaras, Turkey
https://orcid.org/0000-0002-0270-9350

ABSTRACT

In this study, different models were created to reduce bias by ensemble learning methods. Reducing the bias error will improve the classification performance. In order to increase the classification performance, the most appropriate ensemble learning method and ideal sample size were investigated. Bias values and learning performances of different ensemble learning methods were compared. AdaBoost ensemble learning method provided the lowest bias value with n: 250 sample size while Stacking ensemble learning method provided the lowest bias value with n: 500, n: 750, n: 1000, n: 2000, n: 4000, n: 6000, n: 8000, n: 10000, and n: 20000 sample sizes. When the learning performances were compared, AdaBoost ensemble learning method and RBF classifier achieved the best performance with n: 250 sample size (ACC = 0.956, AUC: 0.987). The AdaBoost ensemble learning method and REPTree classifier achieved the best performance with n: 20000 sample size (ACC = 0.990, AUC = 0.999). In conclusion, for reduction of bias, methods based on stacking displayed a higher performance compared to other methods.

KEYWORDS

Boosting, Deep Learning, Ensemble Learning, Machine Learning

INTRODUCTION

Machine learning methods have been widely used in the field of data mining in recent years. Machine learning algorithms based on the theoretical structure of statistics and computer science can provide high performance in data extraction, estimation and classification. With the development of technology in the field of health, there has been a rapid increase in data. Traditional statistical methods have been insufficient in terms of performance regarding data extraction and classification. Machine learning methods have been a powerful alternative to traditional methods because they both save time and provide high performance. Machine learning methods form the basis of many artificial intelligence applications. These methods are used in many medical fields such as diagnosis, early diagnosis and pattern recognition.

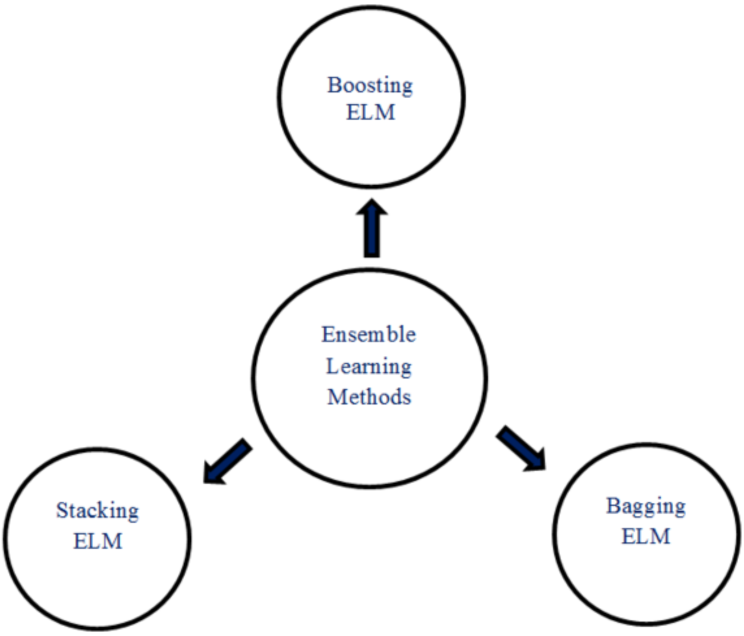
DOI: 10.4018/IJBDAH.20210701.oa2

Although machine learning methods are widely used and can be easily learned from data, there are cases where they cannot provide high classification performance in all conditions. In some cases, although there is a high performance learning from the training data set, a poor performance is given regarding test data sets. There are different reasons for this issue. Although the model provides high accuracy performance with the training data set, the main reason for the low accuracy performance with the test data set is the problem of overfitting. The problem of overfitting is an error caused by the model's memorizing the data instead of learning the pattern in the training data set. The model that memorizes the data during the training phase provides a high accuracy performance but gives a low accuracy performance with different data due to failure to learn the pattern in the testing phase. This error, which causes the problem of overfitting, is described as variance in machine learning. Variance is not the only error in machine learning. In the training phase, the model does not provide high accuracy performance with the training data set. A model that shows low accuracy performance during the training phase will also demonstrate low accuracy performance during the testing phase. The failure of the model to achieve the desired accuracy performance in the training data set is described as the problem of underfitting. The problem of underfitting occurs when the bias error is high. Bias is an error caused by the inability of the model to learn the pattern in the training data set. There are different reasons for the occurrence of bias. One of the reasons is that the correct model has not been selected for the training data set. Some models are not sufficient to classify the data set and learn the pattern. These models are weak classifiers. Therefore, strong classifiers are used to reduce bias. However, while strong classifiers provide high performance with training data sets, they might show a lower performance with test data sets. This issue induces the problem of overfitting. Another method to reduce bias is to increase model complexity. Increasing model complexity is important for reducing bias in the training data set. However, since increasing model complexity also increases variance, it results in a poor performance with the test data set.

Numerous studies have been conducted on the reduction of bias in the classification. Brain and Gwebb (1999) investigated the effect of sample size on bias and variance in classification. They stated that increasing sample size had no effect on bias. Brain and Gwebb (2002) have examined the suitability of algorithms used for the classification of small data sets for use for bigger data sets and their effects on bias and variance. Liu et al. (2016) proposed a different approach that includes selection in order to reduce bias in machine learning and classification. Cawley and Talbot (2010) conducted a performance assessment by examining the effects on bias and variance in case of excessive learning in model selection. In their study, Hainmueller and Hazzlet (2014) proposed the Kernel-based least squares method to reduce bias in machine learning methods. By using the Naive Bayes method in their study, Yang and Webb (2009) stated that classification error could be reduced by managing bias and variance. Suen et al. (2005) compared the performances of bias and variance reduction techniques by combining them in regression trees. Lee et al. (2010) produced results with lower bias by using weight tendency scores in machine learning methods in their study. In his study, Aminian (2005) aimed to reduce bias and variance by using the Jensen-Shannon divergence. Noh et al. (2014) stated in their study that by changing the distance metric, bias can be reduced in the nearest neighbor method. Varma and Simon (2006) conducted studies on the use of cross-validation in the prediction of bias in model selection. In their study, Perdue et al. (2018) aimed to reduce model bias in deep learning classifiers by utilizing reverse neural networks.

Different methods have been developed to improve classification performance and reduce errors in machine learning. One of those methods is ensemble learning. The ensemble learning method is based on the assumption that joint estimation of estimates obtained by multiple classifiers may have a higher accuracy than the estimation of a single classifier (Zhang and Ma, 2012). In this sense, ensemble learning methods have proven to be popular and powerful method among machine learning methods. Ensemble learning methods are showed Figure 1. Wang et al. (2014) used ensemble learning methods in their study for the classification of emotions. Shi et al. (2011) used ensemble learning methods in their study for text classification. Han and Liu (2011) made use of different ensemble learning

Figure 1. Ensemble Learning Methods



methods for remote image classification. Hsieh et al. (2012) utilized ensemble learning methods in their study for early detection of breast cancer. Sun (2007) utilized ensemble learning methods in his study for the classification of EEG signals.

Ensemble learning methods are widely used for early diagnosis of diseases. Ensemble learning methods provide high classification performance. Kazemi and Mirrashondel (2018) used ensemble learning methods to predict kidney stones. Tadepalli and Lakshmi (2019) proposed a machine learning-based model for IVF prediction. Farahani et al. (2018) proposed a ensemble learning-based hybrid model to detect lung nodules from CT images. Fitriyani et al. (2019) proposed an ensemble learning-based prediction model for predicting diabetes and hypertension diseases. Wang et al. (2019) proposed a model based on stacking ensemble learning method for detection of prostate cancer.

In this study, different approaches were taken into consideration to reduce the bias observed in the machine learning methods during the training phase. There are many studies on overfitting and variance errors in the literature. However, there are very few studies aimed at reducing bias during the training phase. In our study, different methods that provide the highest performance for bias reduction were compared. In order to examine the effect of sample size on bias, data sets with different sample sizes were studied. In this study, the aim was to develop the most appropriate model in order to reduce bias and minimize the errors in the training phase. It is aimed to minimize the bias error and increase the classification performance. In order to increase the classification performance, the most appropriate ensemble learning method and ideal sample size were investigated. In this study, it is aimed to determine the most successful ensemble learning method in order to reduce the bias error. The effects of sample size on classification performance were investigated. In the model, the performances of boosting, voting stacking, bagging methods and SVM, SGD, NB, RBF, REPTREE classifiers were compared.

MATERIAL AND METHODS

Data Set

The data set of this study was produced hypothetically with simulation by considering the statistical parameters in the studies conducted on hemogram values of patients with coagulase infection and healthy individuals. Data production with simulation was carried out in accordance with the normal distribution, taking into account the mean and standard deviation parameters of the variables. Compliance of simulated data to normal distribution was checked with Q-Q graphs. 10 data sets were created in different sample sizes (n: 250, n: 500, n: 750, n: 1000, n: 2000, n: 4000, n: 6000 n: 8000, n: 10000 and n: 20000). The study included 1 target (Group) and 14 predictors (Gender, Age, Eosinophil, Monocyte, Hemoglobin, Hematocrit, Platelet, Neutrophil, Lymphocyte, RBC, WBC, CRP, MCV and IG). The definition of variables is summarized in Table 1. WEKA (Waikato Environment for Knowledge Analysis) version 3.9 was used for the evaluation of the data.

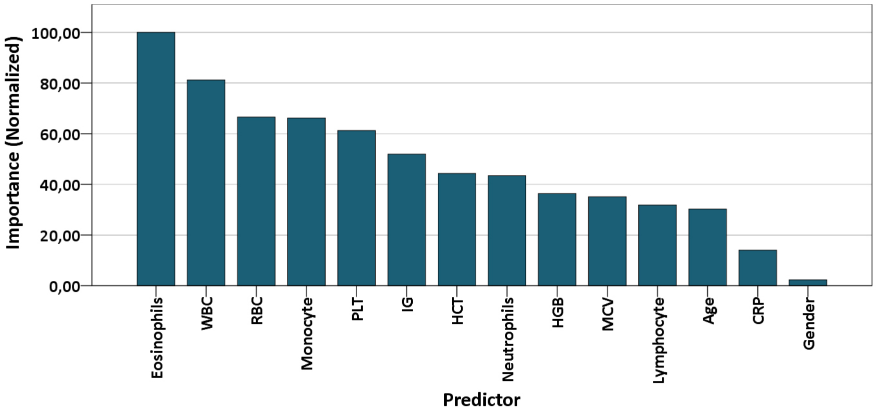
Pre-Analysis Dataset and Feature Selection

Outlier and extreme values in the data set were examined by Local Outlier Factor (LOF) algorithm. The LOF Algorithm scales the distance of objects close to it for each data depending on the local densities of the data. It is a powerful method used to detect local outlier, outlier and outlier observations (Breunig et al.2000; Lee et al. 2011). Outlier and extreme values were excluded from the study. Standardization was applied to quantitative data. In order to provide the best classification performance and to reduce the noisy data by identifying the variables that are not related to the model, feature selection was applied to the predictor variables. As a result of feature selection, CRP and gender variable, which had the least contribution in explaining the target variable, were excluded from the model. The feature selection process is shown in Figure 2.

Table 1. The definition of the Variables in the Current Study

Variables	Definition	Role
Group	Patient/Control	Target
Gender	Male/Female	Predictor
Eosinophil	Integer	Predictor
Monocytes	Integer	Predictor
Hemoglobin	Integer	Predictor
Hematocrit	Integer	Predictor
Platelets	Integer	Predictor
Neutrophil	Integer	Predictor
Lymphocytes	Integer	Predictor
RBC	Integer	Predictor
WBC	Integer	Predictor
Age	Integer	Predictor
CRP	Integer	Predictor
MCV	Integer	Predictor
IG	Integer	Predictor

Figure 2. Importance Values of Predictor Variables



Machine Learning Experimentation Methods

In this study, 10 data sets consisting of different sample sizes were trained to evaluate the classification performances with different classifiers and ensemble methods. Data sets were trained with Support vector Machine (SVM), Stochastic Gradient Descent (SGD), Radial Basis Function Classifier (RBF), Naive Bayes, REPTree, Random Forest and K-NN classifiers. Random Forest and K-NN classifiers were excluded from the model because they had overfitting problems. The performances and bias rates of Bagging, Boosting, Voting and Stacking ensemble methods were evaluated in addition to the performances of each basic classifier in the model. AdaBoost method was applied for Boosting ensemble method. Random Forest classifier was utilized as meta classifier in the Stacking method. In the comparison of the bias rates, the mean of the bias rates of the classifiers in the non-ensemble method and the bias rates of the Boosting, Bagging ensemble methods were compared. Data sets were trained with K-10 fold cross-validation and by division into 70% training and 30% testing. Performance comparisons were performed with Accuracy, Precision, Recall, AUC and Kappa metrics. The bias values of the ensemble learning methods during training were also compared. The confusion matrix for metrics are given Table 2. Main model of study and process are shown in Figure 3.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

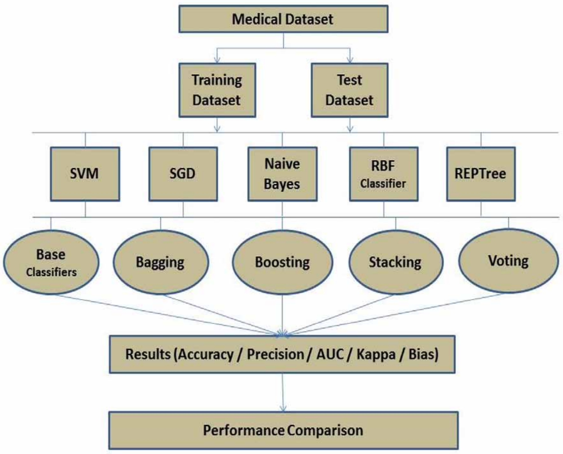
$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

Table 2. Confusion Matrix

		Actual Condition	
		Positive	Negative
Test Result	Positive	True Positive (TP)	False Positive (FP)
	Negative	False Negative (FN)	True Negative (TN)

Figure 3. Classifiers, Ensemble Methods and Main Model of Study



Bagging

Bagging (Bootstrap Aggregation), one of the first ensemble learning methods, is a simple and effective method. Developed by Breiman (Breiman, 1996) this method is based on the training of different subsets of a data set resampled by Bootstrap method in parallel with separate classifiers and testing in the same data set. The ensemble decision is determined by applying the majority voting method to the estimates obtained from each classifier. (Zhang and Ma, 2012).

Boosting

Boosting is a powerful ensemble method based on weighting that offers a more precise and powerful classification chance by focusing on iterative steps and errors in the previous training at each step (Schapire, 1990). Although it is structurally similar to the Bagging method, in Bagging method, the majority decision is determined by the majority vote for the estimates produced in separate classifiers at the same time. In the Boosting method, the processes proceed in an iterative way and not simultaneously. In each iterative process, a strong classifier is obtained by Boosting method in order to prevent the same errors from occurring by considering the errors in the previous estimation. (Zhou, 2012). One of the most powerful Boosting algorithms is the AdaBoost algorithm. AdaBoost algorithm minimizes the exponential loss function (Freund and Schapire, 1996).

Voting

Voting, one of the widely used ensemble learning methods, is based on combining estimates obtained from different classifiers. In the Voting process, several different classifiers train the same data set in parallel. The Voting process is applied to all estimates obtained by each classifier and the estimate of the majority is accepted as the estimate of the ensemble (Hansen and Salamon, 1990; Zhang and Ma, 2012).

Stacking

The Stacking ensemble learning method accepts estimates that different classifiers obtain from the training data set as input for a meta classifier. These input data, consisting of estimates, are re-trained in the meta classifier to obtain ensemble estimation (Wolpert, 1992; Zhou, 2012).

Classifiers

Support Vector Machine

It is a powerful classifier that is widely used in classification processes. In the linear plane, two hyperlines are formed as a boundary to separate the data belonging to the two groups. These lines aim to keep the distance between the lines at the highest level in an optimal way to separate the data of the two groups. It is a supervised machine learning algorithm that produces successful results for linear and nonlinear classification problems. Hyperparameter optimization is performed to ensure the best classification values. Support vector machines can be implemented with different core functions (Vapnik, 2013; Cortes and Vapnik, 1995).

Stochastic Gradient Descent

Stochastic gradient descent applies a classification process that supports loss functions and penalties in linear models. Its successful performance with high data sizes makes this classifier popular. While a normal gradient descent model considers all values in the data set and iteratively works on the entire data set when changing weight, in the stochastic gradient descent model a single point is taken into account. This provides rapidity compared to other methods (Bottou, 2010; Çürük et al. 2018).

Naive Bayes

Naive Bayes is a supervised classifier based on probabilistic calculations grounded on Bayes theorem. It can produce successful results in big data. The properties in the data set are independent of each other. The Naive Bayes classifier estimates which class the samples in the data set are included in (Rish, 2001). Naive Bayes classifier is very successful in the classification of categorical data such as text mining (Serrano-Guerrero et al. 2015).

Radial Basis Function Classifier

The radial basis function classifier is based on radial-based function neural networks and provides the least squares error in the optimization process using the BFGS method. In this method that performs supervised learning, normalization is applied to all features [0,1] (Frank, 2014).

Reduced Error Pruning Tree (REPTree)

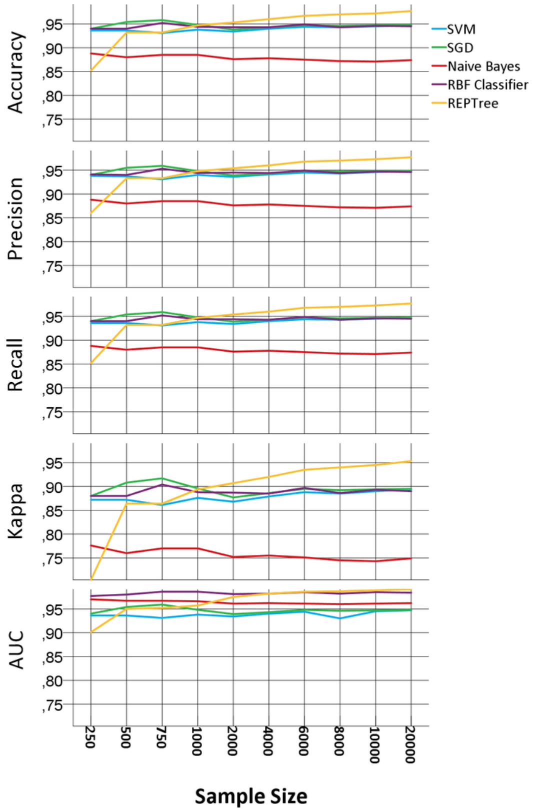
The Reduced Error Pruning Tree (REPTree) classifier determines the best tree among the many decision trees it produces with different iterations. The average squares error generated by the tree is used as a measure for pruning with the aim of estimation. Knowledge acquisition and variance are used to classify and form a decision tree. One of the most important aspects of this classifier is that it is a fast classifier. (Srinivasan and Mekala, 2014; Kalmegh, 2015).

Results

In the first stage, the performances of 10 data sets with different sample sizes were evaluated in base classifiers without using any ensemble learning methods. Accuracy, Precision, Recall, AUC and Kappa metrics were compared. The findings of the comparison are given in Figure 4. According to the findings in Figure 4, SGD classifier provided the highest performance with n: 250 sample size in accuracy metric. As for the n: 20000 sample size, the REPTree classifier provided the highest performance. In all metrics, the REPTree algorithm showed a performance increase parallel to the increase in sample size. The performances of the classifiers according to the sample sizes are shown in Figure 4.

In addition to the performances of base classifiers, their performances with ensemble methods were also evaluated in the study. accuracy performances of classifiers in different sample sizes with the Boosting (AdaBoost), Bagging ensemble methods and without any ensemble method were compared. The highest performance in terms of accuracy metric was observed in the AdaBoost (Boosting)

Figure 4. Performances of Base Classifiers at Different Sample Sizes



ensemble method. AdaBoost method significantly improved performance for Naive Bayes, RBF and REPTree classifiers. The ensemble methods did not provide a significant increase for SVM and SGD classifiers. The highest performance was observed in the REPTree classifier in the AdaBoost ensemble learning method. Additionally, increasing sample size in REPTree classifier contributed to increase in performance. The accuracy performances of ensemble methods and classifiers at different sample sizes are shown in Figure 5. In terms of Precision, Recall and Kappa metrics, the performance results of the methods and classifiers are similar to the accuracy metric results. Again, the best performance in these metrics was achieved with the AdaBoost ensemble learning method. The performance of methods and classifiers in terms of Precision, Recall and Kappa metrics are shown in Figure 6, Figure 7 and Figure 8 respectively. In terms of AUC metrics, the AdaBoost ensemble learning method was the best performing method for all classifiers. The highest performance was achieved with n: 20000 sample size with the REPTree classifier and the AdaBoost ensemble method. The performances of the methods and classifiers in terms of AUC metric are given in Figure 9.

In this study, classification performances of ensemble learning methods at n: 250 and n: 20000 sample sizes were evaluated. The highest classification performance in the n: 250 sample size was demonstrated by the RBF classifier with AdaBoost ensemble learning method. The highest classification performance in the n: 20000 sample size was demonstrated by the REPTree classifier with AdaBoost ensemble learning method. The classification performances of the methods are shown in Table 3.

The main findings of the study are the bias values that emerged in the training data set of ensemble learning methods. At this stage, the bias values the ensemble learning methods revealed

Figure 5. Accuracy Performances of Ensemble Methods

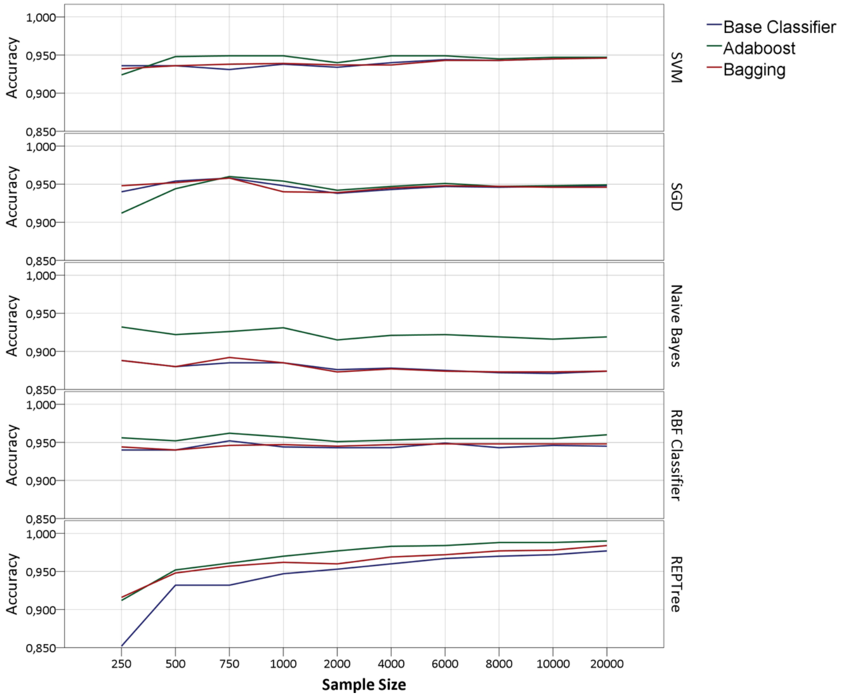


Figure 6. Precision Performances of Ensemble Methods

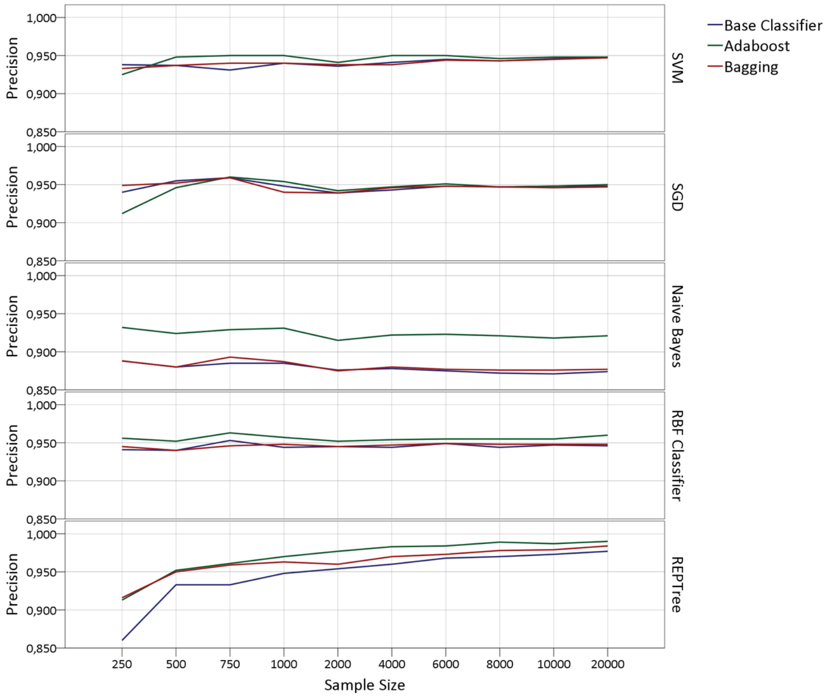


Figure 7. Recall Performances of Ensemble Methods

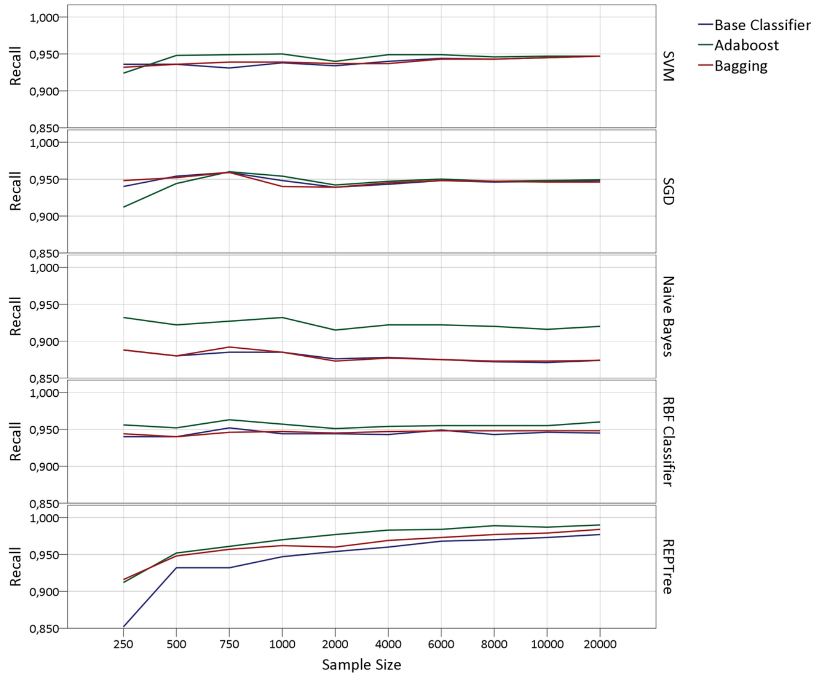


Figure 8. Kappa Performances of Ensemble Methods

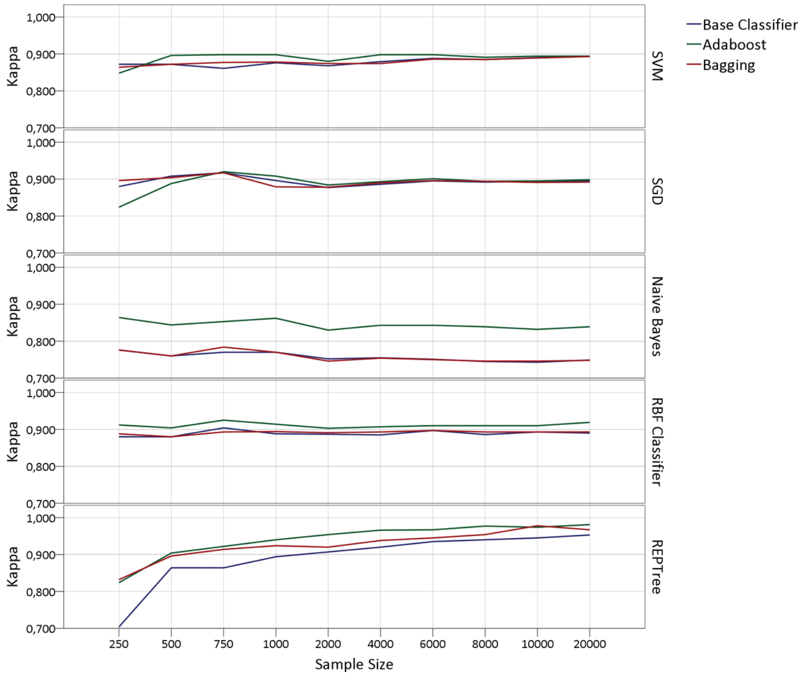


Figure 9. AUC Performances of Ensemble Methods

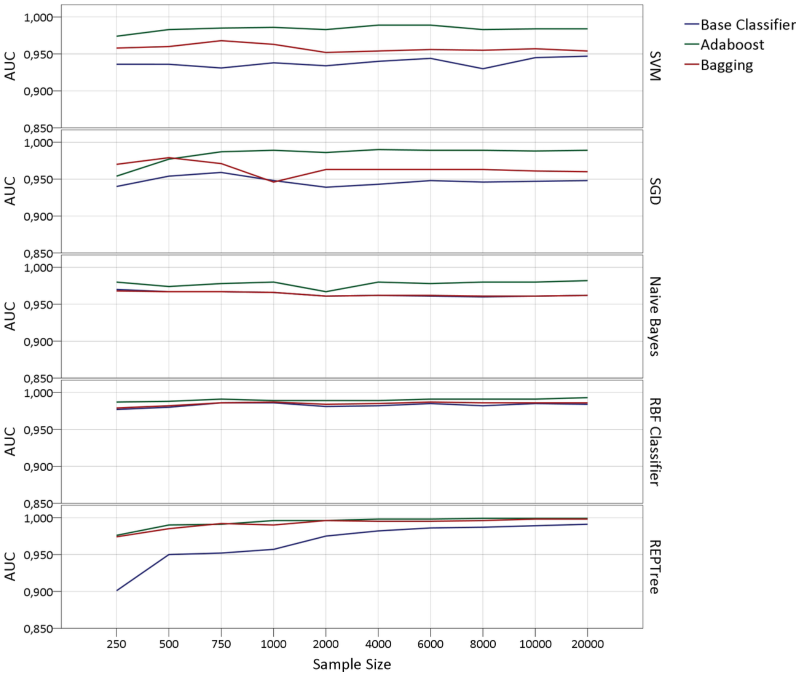


Table 3. All metrics performances of ensemble methods at n:250 and n:20000 sample sizes

		n: 250					n:20000				
		Acc.	**Prec.**	**Recall**	**AUC**	**Kappa**	**Acc.**	**Prec.**	**Recall**	**AUC**	**Kappa**
Base Classifier	**SVM**	0,936	0,938	0,936	0,936	0,872	0,947	0,948	0,947	0,947	0,894
	SGD	0,940	0,940	0,940	0,940	0,880	0,948	0,948	0,948	0,948	0,895
	Naive Bayes	0,888	0,888	0,888	0,970	0,776	0,874	0,874	0,874	0,962	0,749
	RBF Classifier	0,940	0,941	0,940	0,977	0,880	0,945	0,946	0,945	0,984	0,890
	REPTree	0,852	0,860	0,852	0,901	0,704	0,977	0,977	0,977	0,991	0,953
Adaboost	**SVM**	0,924	0,925	0,924	0,974	0,848	0,947	0,948	0,947	0,984	0,894
	SGD	0,912	0,912	0,912	0,954	0,824	0,949	0,950	0,949	0,989	0,898
	Naive Bayes	0,932	0,932	0,932	0,980	0,864	0,919	0,921	0,920	0,982	0,839
	RBF Classifier	**0,956**	**0,956**	**0,956**	**0,987**	**0,912**	0,960	0,960	0,960	0,993	0,919
	REPTree	0,912	0,913	0,912	0,976	0,824	**0,990**	**0,990**	**0,990**	**0,999**	**0,981**
Bagging	**SVM**	0,932	0,933	0,932	0,958	0,864	0,946	0,947	0,947	0,954	0,893
	SGD	0,948	0,949	0,948	0,970	0,896	0,946	0,947	0,946	0,960	0,892
	Naive Bayes	0,888	0,888	0,888	0,968	0,776	0,874	0,877	0,874	0,962	0,748
	RBF Classifier	0,944	0,945	0,944	0,979	0,888	0,948	0,948	0,948	0,986	0,893
	REPTree	0,916	0,916	0,916	0,974	0,832	0,984	0,984	0,984	0,998	0,967
Stacking		0,944	0,944	0,944	0,975	0,888	0,978	0,978	0,978	0,994	0,955
Voting		0,944	0,945	0,945	0,978	0,888	0,955	0,955	0,955	0,992	0,909

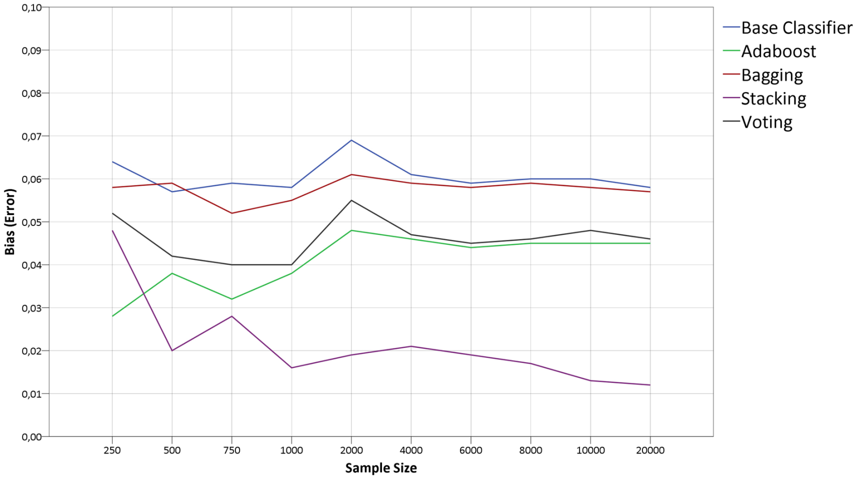
in the training data set were compared. In the comparison, the mean values of bias obtained from single classifiers and classifiers in the AdaBoost and bagging methods were calculated. According to the comparison results, the lowest bias (bias error) value was obtained by AdaBoost method with n: 250 sample size. On the other hand, in all other sample sizes (n: 500, n: 750, n: 1000, n: 2000, n: 4000, n: 6000, n: 8000, n: 10000, n: 20000), the lowest bias value was provided by the Stacking ensemble learning method. It was observed that bias value decreased with increasing sample size in the Stacking ensemble learning method. The performances of the ensemble methods with their bias values are given in Figure 10.

DISCUSSION

The development of technology has enabled significant achievements to be achieved in human health. Technology has made people's lives easier. Diseases could be early diagnosed (So et al. 2017; Parisi et al. 2018). Access to health information has become easier (Beam and Kohane, 2018). New treatment methods have been developed (Scheeder et al 2018). Telemedicine services and home nursing services have reached even more widespread use (Ramkumar et al. 2018). Important improvements have been made in elderly care and elderly health (Özsungur, 2019a; Özsungur, 2019b). The accuracy performance of medical imaging reports and laboratory results has increased (Suzuki, 2017). Artificial intelligence and machine learning methods form the basis of technological developments in the health system.

Behind the artificial intelligence technologies that have been rapidly in recent years are powerful machine learning methods. The high performance of classifiers plays an important role in the popularity of machine learning methods. The first condition for the high success of classification in machine learning methods depends on minimizing the bias in the training data sets of classification algorithms and the model learning the pattern correctly. Different methods have been developed in order to have high classification performance in machine learning methods. Some of those methods are ensemble learning methods. Ensemble learning methods are based on the principle that the performance of multiple classifiers will be better than the performance of a single classifier. In their study, Wang et al. (2011) compared the performances of the credit scores of different countries by classifying them with basic classifiers, Bagging, Boosting and Stacking ensemble methods. Accuracy, type 1 and type 2 error metrics were evaluated in the comparison. They achieved the best performance with Bagging

Figure 10. Bias Performances of Ensemble Methods



and Stacking ensemble methods and Decision Tree classifier. They stated that the Bagging method showed a better performance than the Boosting method. In our study, the best performances were provided by RBF classifier in the Boosting ensemble method and REPTree classifier in large sample size. Tan and Gilbert (2003) compared the performance of C4.5 basic classifier, Bagging and Boosting ensemble learning methods for cancer classification in their study. In the study, it was observed that Bagging and Boosting ensemble methods outperformed the basic classifiers. In our study, ensemble methods provided higher accuracy performance as well. Sun (2002) used ensemble learning methods to predict sound accents in a study. The ensemble learning methods were observed to provide better results than basic learners. In their study, Kaur and Kaur (2014) compared the performances of different classifiers in ensemble learning methods. As a result of the comparison, they observed that ensemble methods increased the performance compared to the basic classifiers. Das and Sengur (2010) compared the performances of ensemble learning methods in the diagnosis of heart disorders in their study. They stated that AdaBoost ensemble method provides higher accuracy performance than basic classifier, Bagging and random Subspace ensemble methods. The AdaBoost method provided the highest accuracy according to the findings of our study as well. Our study demonstrated the effect of ensemble learning methods on bias in training data sets. In their study, Liu and Cocea (2017) made a recommendation based on granular calculation to reduce bias in ensemble methods. Lu et al. (2017) used the ensemble learning methods in their study to reduce the bias in the new computer-enhanced diagnosis systems used for the detection of breast cancer.

CONCLUSION

In our study, the ensemble learning methods were observed to reduce bias in the training data set. Stacking ensemble learning method has been identified as the most successful method in reducing bias. On the other hand, increasing sample size in Voting, Bagging and Boosting ensemble methods does not contribute to decreasing bias. It has been observed that increasing sample size in the Stacking ensemble method contributes to reducing bias.

Ensemble learning methods make a significant contribution to reducing bias in accuracy performance and training data sets when used with appropriate data and appropriate classifiers. The use of the Stacking ensemble learning method contributes significantly to the reduction of bias. Not each classifier or ensemble method provides an increase in performance despite increasing sample size. The AdaBoost (boosting) ensemble learning method provides high performance with RBF classifier in small sample size data set and REPTree classifier in large sample size data set in accuracy performance.

In future studies, it is planned to create hybrid models based on the stacking ensemble learning method. Hybrid models based on ensemble learning model can be applied for early diagnosis of diseases. Ensemble learning-based hybrid models are predicted to be classified with high performance.

ACKNOWLEDGMENT

This study was presented as an oral Presentation in 11 International Statistics Congress (4-8 October 2019). Data on humans and animals were not used in this study. The data were hypothetically generated by simulation. Therefore, ethics committee approval was not required. No funding for this study.

REFERENCES

Aminian, A. (2005). Active learning for reducing bias and variance of a classifier using Jensen-Shannon divergence. In *Fourth International Conference on Machine Learning and Applications (ICMLA'05)* (pp. 6-pp). IEEE. doi:10.1109/ICMLA.2005.7

Beam, A. L., & Kohane, I. S. (2018). Big data and machine learning in health care. *Journal of the American Medical Association*, *319*(13), 1317–1318. doi:10.1001/jama.2017.18391 PMID:29532063

Bottou, L. (2010). Large-scale machine learning with stochastic gradient descent. *Proceedings of COMPSTAT*, *2010*, 177–186. doi:10.1007/978-3-7908-2604-3_16

Brain, D., & Webb, G. (1999). On the effect of data set size on bias and variance in classification learning. In *Proceedings of the Fourth Australian Knowledge Acquisition Workshop*, *University of New South Wales* (pp. 117-128). Academic Press.

Brain, D., & Webb, G. I. (2002, August). The need for low bias algorithms in classification learning from large data sets. In *European Conference on Principles of Data Mining and Knowledge Discovery* (pp. 62-73). Springer. doi:10.1007/3-540-45681-3_6

Breiman, L. (1996). Bagging predictors. *Machine Learning*, *24*(2), 123–140. doi:10.1007/BF00058655

Breunig, M. M., Kriegel, H. P., Ng, R. T., & Sander, J. (2000, May). LOF: identifying density-based local outliers. In *Proceedings of the 2000 ACM SIGMOD international conference on Management of data* (pp. 93-104). doi:10.1145/342009.335388

Cawley, G. C., & Talbot, N. L. (2010). On over-fitting in model selection and subsequent selection bias in performance evaluation. *Journal of Machine Learning Research*, *11*(Jul), 2079–2107.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, *20*(3), 273–297. doi:10.1007/BF00994018

Çürük, E., Acı, Ç., & Eşsiz, E. S. (2018, September). Performance Analysis of Artificial Neural Network Based Classfiers for Cyberbulling Detection. In *2018 3rd International Conference on Computer Science and Engineering (UBMK)* (pp. 1-5). IEEE. doi:10.1109/UBMK.2018.8566566

Das, R., & Sengur, A. (2010). Evaluation of ensemble methods for diagnosing of valvular heart disease. *Expert Systems with Applications*, *37*(7), 5110–5115. doi:10.1016/j.eswa.2009.12.085

Farahani, F. V., Ahmadi, A., & Zarandi, M. H. F. (2018). Hybrid intelligent approach for diagnosis of the lung nodule from CT images using spatial kernelized fuzzy c-means and ensemble learning. *Mathematics and Computers in Simulation*, *149*, 48–68. doi:10.1016/j.matcom.2018.02.001

Fitriyani, N. L., Syafrudin, M., Alfian, G., & Rhee, J. (2019). Development of disease prediction model based on ensemble learning approach for diabetes and hypertension. *IEEE Access: Practical Innovations, Open Solutions*, *7*, 144777–144789. doi:10.1109/ACCESS.2019.2945129

Frank, E. (2014). *Fully supervised training of Gaussian radial basis function networks in WEKA* (Computer Science Working Papers, 04/2014). Hamilton, New Zealand: Department of Computer Science, The University of Waikato.

Freund, Y., & Schapire, R. E. (1996, July). Experiments with a new boosting algorithm. In ICML (Vol. 96, pp. 148-156). Academic Press.

Hainmueller, J., & Hazlett, C. (2014). Kernel regularized least squares: Reducing misspecification bias with a flexible and interpretable machine learning approach. *Political Analysis*, *22*(2), 143–168. doi:10.1093/pan/mpt019

Han, M., & Liu, B. (2015). Ensemble of extreme learning machine for remote sensing image classification. *Neurocomputing*, *149*, 65–70. doi:10.1016/j.neucom.2013.09.070

Hansen, L. K., & Salamon, P. (1990). Neural network ensembles. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *12*(10), 993–1001.

Hsieh, S. L., Hsieh, S. H., Cheng, P. H., Chen, C. H., Hsu, K. P., Lee, I. S., Wang, Z., & Lai, F. (2012). Design ensemble machine learning model for breast cancer diagnosis. *Journal of Medical Systems*, *36*(5), 2841–2847. doi:10.1007/s10916-011-9762-6 PMID:21811801

Kalmegh, S. (2015). Analysis of weka data mining algorithm reptree, simple cart and randomtree for classification of indian news. *International Journal of Innovative Science. Engineering & Technology*, *2*(2), 438–446.

Kaur, A., & Kaur, K. (2014, September). Performance analysis of ensemble learning for predicting defects in open source software. In 2014 international conference on advances in computing, communications and informatics (ICACCI) (pp. 219-225). IEEE. doi:10.1109/ICACCI.2014.6968438

Kazemi, Y., & Mirroshandel, S. A. (2018). A novel method for predicting kidney stone type using ensemble learning. *Artificial Intelligence in Medicine*, *84*, 117–126. doi:10.1016/j.artmed.2017.12.001 PMID:29241659

Lee, B. K., Lessler, J., & Stuart, E. A. (2010). Improving propensity score weighting using machine learning. *Statistics in Medicine*, *29*(3), 337–346. doi:10.1002/sim.3782 PMID:19960510

Lee, J., Kang, B., & Kang, S. H. (2011). Integrating independent component analysis and local outlier factor for plant-wide process monitoring. *Journal of Process Control*, *21*(7), 1011–1021. doi:10.1016/j.jprocont.2011.06.004

Liu, H., & Cocea, M. (2017). Granular computing-based approach for classification towards reduction of bias in ensemble learning. *Granular Computing*, *2*(3), 131–139.

Liu, H., Gegov, A., & Cocea, M. (2016). Nature and biology inspired approach of classification towards reduction of bias in machine learning. In *2016 International Conference on Machine Learning and Cybernetics (ICMLC)* (Vol. 2, pp. 588-593). IEEE. doi:10.1109/ICMLC.2016.7872953

Lu, W., Li, Z., & Chu, J. (2017). A novel computer-aided diagnosis system for breast MRI based on feature selection and ensemble learning. *Computers in Biology and Medicine*, *83*, 157–165. doi:10.1016/j.compbiomed.2017.03.002 PMID:28282591

Noh, Y. K., Sugiyama, M., Liu, S., Plessis, M. C., Park, F. C., & Lee, D. D. (2014, April). Bias reduction and metric learning for nearest-neighbor estimation of Kullback-Leibler divergence. In Artificial Intelligence and Statistics (pp. 669-677). Academic Press.

Özsungur, F. (2019a). Gerontechnological factors affecting successful aging of elderly. *The Aging Male*, 1–13. doi:10.1080/13685538.2018.1539963 PMID:30741066

Özsungur, F. (2019b). A Research on the Effects of Successful Aging on the Acceptance and Use of Technology of the Elderly. *Assistive Technology*, 1–14. doi:10.1080/10400435.2019.1691085 PMID:31710261

Parisi, L., RaviChandran, N., & Manaog, M. L. (2018). Feature-driven machine learning to improve early diagnosis of Parkinson's disease. *Expert Systems with Applications*, *110*, 182–190.

Perdue, G. N., Ghosh, A., Wospakrik, M., Akbar, F., Andrade, D. A., Ascencio, M., & Cai, T. et al. (2018). Reducing model bias in a deep learning classifier using domain adversarial neural networks in the MINERvA experiment. *Journal of Instrumentation: An IOP and SISSA Journal*, *13*(11), P11020. doi:10.1088/1748-0221/13/11/P11020

Ramkumar, P. N., Haeberle, H. S., Navarro, S. M., Sultan, A. A., Mont, M. A., Ricchetti, E. T., Schickendantz, M. S., & Iannotti, J. P. (2018). Mobile technology and telemedicine for shoulder range of motion: Validation of a motion-based machine-learning software development kit. *Journal of Shoulder and Elbow Surgery*, *27*(7), 1198–1204. doi:10.1016/j.jse.2018.01.013 PMID:29525490

Rish, I. (2001, August). An empirical study of the naive Bayes classifier. In IJCAI 2001 workshop on empirical methods in artificial intelligence (Vol. 3, No. 22, pp. 41-46). Academic Press.

Schapire, R. E. (1990). The strength of weak learnability. *Machine Learning*, *5*(2), 197–227. doi:10.1007/BF00116037

Scheeder, C., Heigwer, F., & Boutros, M. (2018). Machine learning and image-based profiling in drug discovery. *Current Opinion in Systems Biology*, *10*, 43–52. doi:10.1016/j.coisb.2018.05.004 PMID:30159406

Serrano-Guerrero, J., Olivas, J. A., Romero, F. P., & Herrera-Viedma, E. (2015). Sentiment analysis: A review and comparative analysis of web services. *Information Sciences*, *311*, 18–38. doi:10.1016/j.ins.2015.03.040

Shi, L., Ma, X., Xi, L., Duan, Q., & Zhao, J. (2011). Rough set and ensemble learning based semi-supervised algorithm for text classification. *Expert Systems with Applications*, *38*(5), 6300–6306. doi:10.1016/j.eswa.2010.11.069

So, A., Hooshyar, D., Park, K. W., & Lim, H. S. (2017). Early diagnosis of dementia from clinical data by machine learning techniques. *Applied Sciences (Basel, Switzerland)*, *7*(7), 651. doi:10.3390/app7070651

Srinivasan, D. B., & Mekala, P. (2014). Mining social networking data for classification using reptree. *International Journal of Advance Research in Computer Science and Management Studies*, *2*(10).

Suen, Y. L., Melville, P., & Mooney, R. J. (2005). Combining bias and variance reduction techniques for regression trees. In *European Conference on Machine Learning* (pp. 741-749). Springer. doi:10.1007/11564096_76

Sun, S. (2007, May). Ensemble learning methods for classifying EEG signals. In *International Workshop on Multiple Classifier Systems* (pp. 113-120). Springer. doi:10.1007/978-3-540-72523-7_12

Sun, X. (2002). Pitch accent prediction using ensemble machine learning. In *Seventh international conference on spoken language processing*. Academic Press.

Suzuki, K. (2017). Overview of deep learning in medical imaging. *Radiological Physics and Technology*, *10*(3), 257–273. doi:10.1007/s12194-017-0406-5 PMID:28689314

Tadepalli, S. K., & Lakshmi, P. V. (2019). Application of Machine Learning and Artificial Intelligence Techniques for IVF Analysis and Prediction. *International Journal of Big Data and Analytics in Healthcare*, *4*(2), 21–33. doi:10.4018/IJBDAH.2019070102

Tan, A. C., & Gilbert, D. (2003). *Ensemble machine learning on gene expression data for cancer classification*. Academic Press.

Vapnik, V. (2013). *The nature of statistical learning theory*. Springer science & business media.

Varma, S., & Simon, R. (2006). Bias in error estimation when using cross-validation for model selection. *BMC Bioinformatics*, *7*(1), 91. doi:10.1186/1471-2105-7-91 PMID:16504092

Wang, G., Hao, J., Ma, J., & Jiang, H. (2011). A comparative assessment of ensemble learning for credit scoring. *Expert Systems with Applications*, *38*(1), 223–230. doi:10.1016/j.eswa.2010.06.048

Wang, G., Sun, J., Ma, J., Xu, K., & Gu, J. (2014). Sentiment classification: The contribution of ensemble learning. *Decision Support Systems*, *57*, 77–93. doi:10.1016/j.dss.2013.08.002

Wang, Y., Wang, D., Geng, N., Wang, Y., Yin, Y., & Jin, Y. (2019). Stacking-based ensemble learning of decision trees for interpretable prostate cancer detection. *Applied Soft Computing*, *77*, 188–204. doi:10.1016/j.asoc.2019.01.015

Wolpert, D. H. (1992). Stacked generalization. *Neural Networks*, *5*(2), 241–259. doi:10.1016/S0893-6080(05)80023-1 PMID:18276425

Yang, Y., & Webb, G. I. (2009). Discretization for naive-Bayes learning: Managing discretization bias and variance. *Machine Learning*, *74*(1), 39–74. doi:10.1007/s10994-008-5083-5

Zhang, C., & Ma, Y. (Eds.). (2012). *Ensemble machine learning: methods and applications*. Springer Science & Business Media. doi:10.1007/978-1-4419-9326-7

Zhou, Z. H. (2012). *Ensemble methods: foundations and algorithms*. Chapman and Hall/CRC. doi:10.1201/b12207

Adem Doganer has a Bachelor's degree in Statistics, Master' s degree in Statistical Information Systems and PhD on Statistical Information Systems and Modeling. He has a second PhD on Biostatistics and Medical Informatics. Dr Adem Doganer is a faculty member in Faculty of Medicine at Kahramanmaras Sutcu Imam University. He has journal articles about statistics, artificial intelligence and data mining. He is statistics editor of KSU Medical Journal.