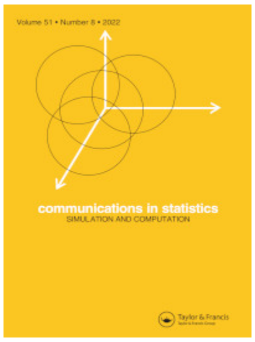

Taylor & Francis
Taylor & Francis Group

Communications in Statistics - Simulation and Computation

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/lssp20

Effect of data preprocessing on ensemble learning for classification in disease diagnosis

Yüksel Özkan, Mert Demirarslan & Aslı Suner

To cite this article: Yüksel Özkan, Mert Demirarslan & Aslı Suner (2022): Effect of data preprocessing on ensemble learning for classification in disease diagnosis, Communications in Statistics - Simulation and Computation, DOI: 10.1080/03610918.2022.2053717

To link to this article: https://doi.org/10.1080/03610918.2022.2053717

Published online: 24 Mar 2022.

COMMUNICATIONS IN STATISTICS - SIMULATION AND COMPUTATION®
https://doi.org/10.1080/03610918.2022.2053717

Effect of data preprocessing on ensemble learning for classification in disease diagnosis

Yüksel Özkan, Mert Demirarslan, and Aslı Suner

Department of Biostatistics and Medical Informatics, Faculty of Medicine, Ege University, Izmir, Turkey

ABSTRACT

In recent years, supervised machine learning methods have increased attention to extracting clinically relevant information from complex health data. Ensemble learning methods enable the establishment of more successful models by training multiple learners jointly to solve the same problem. Herein, we aimed to compare the performance of classification algorithms after data preprocessing to problems such as missing data, class noise, and class imbalance that may be encountered in the datasets used to make an accurate disease diagnosis. To this end, we used random forest and weighted subspace random forest as bagging algorithms while additive logistic regression and gradient boosted machines algorithms were used as boosting algorithms. The performance and running time of the algorithms were also calculated. Our findings indicated that the performance of algorithms increased after data preprocessing and the performance of boosting algorithms yielded higher results than the bagging algorithms. We also observed that the boosting algorithms were the longest-running ones. In conclusion, complementing existing studies, our work highlights the importance and effect of using multiple data preprocessing methods together.

ARTICLE HISTORY
Received 4 February 2021
Accepted 9 March 2022

KEYWORDS
Diagnosis; Data preprocessing; Missing data; Class noise; Class imbalance; Ensemble learning

MATHEMATICAL SUBJECT CLASSIFICATION
68Q25; 62-08

1. Introduction

Thanks to artificial intelligence, which paves the way for technological developments that could not be possible until today, the analysis of several complex data in different fields have been efficiently performed. In particular, the industries interested in healthcare work continue to overcome problems specific to this field through machine learning methods (Salcedo-Bernal, Villamil-Giraldo, and Moreno-Barbosa 2016). The number of studies on the application of learning algorithms in the field of health has increased since the existence of very large medical datasets (Thottakkara et al. 2016). Few studies have contributed significantly to clinical care, although there are thousands of studies applying machine learning algorithms to medical data (Kononenko 2001). Classification tasks in the medical domain are part of the health studies that provide the basis for disease recognition and investigation. In this way, it helps physicians decide how to divide complex data into meaningful pieces (Jutel 2011).

In parallel with the rapid increase in information produced in the medical domain, there is also an increase in the number of decision support applications. In this context, the applications of machine learning in medical diagnosis point to the emergence of algorithms, systems, and methodologies that enable advanced and complex data analysis (Kononenko 2001). Soon, it is widely foreseen that machine learning will be used to process large amounts of information

CONTACT Aslı SUNER asli.suner@ege.edu.tr Department of Biostatistics and Medical Informatics, Faculty of Medicine, Ege University, Izmir 35100, Turkey.

produced and stored by modern technology (Chen and Asch 2017). Currently, existing machine learning methods have considerable potential in uncovering undefined relationships and patterns of data in the biomedical field. Besides, machine learning can help healthcare professionals automatically make meaningful and interpretable large datasets that enable them to switch to a personalized care service known as precision medicine (Azencott 2018).

Medical diagnosis is a critical decision-making process for healthcare professionals. Classification algorithms are used to make the decisions of physicians more effective in making medical diagnosis (Gammerman 2010). However, several problems may cause diagnostic errors such as cognitive errors resulting from insufficient information, incorrect data collection and/or verification, missing data problem, unbalanced class problem, unrelated variable, response variable, and problems encountered in the definition of explanatory variables (Kübler, Liu, and Sayyed 2017). The performance of classification algorithms are affected by the aforementioned issues and it is necessary to develop new algorithms and tools are essential to handle such problems. Additionally, presenting these algorithm and tools with the simplicity to medical experts is key (Obermeyer and Emanuel 2016). Thus, researchers in diverse fields such as patient safety, quality improvement, decision making, and problem-solving can safely use these supportive algorithms and tools to make medical diagnosis in reliable manner (Shah et al. 2019).

Herein, we aimed to apply machine learning classification algorithms and compare the performances and running time of the algorithms after data preprocessing were performed on the medical datasets.

2. Method

During the application phase of the study, a computer with Intel (R) Core (TM) i5-6200U CPU @ 2.3 GHz 2.40 GHz processor feature, 8.00 GB of installed memory (RAM), and 64-bit Operating System, and x64-based processor system were used. Also, all analyzes in the study were performed in RStudio 1.2.1335 – Windows 7+ (64-bit) program. In some cases, the terms “preprocessing” and “data cleansing or data cleaning” are used interchangeably. Data cleansing is also known as the process of detecting and correcting missing or inaccurate values (e.g., handling missing data or filtering unwanted outliers or removing irrelevant variables) from a given dataset. In our work, we used the term “preprocessing” to term to refer “data cleansing.” First, the data preprocessing phase was completed for possible problems in the original downloaded data, then the processed data were recorded, and ensemble learning methods were employed again. Finally, the most successful model was decided by comparing the performance of both original and processed data and calculating the running time of algorithms. Because the accuracy values of the algorithms before and after data preprocessing were normally distributed, the paired samples t-test was used to compare the accuracy values. A general flow chart of the study was given in Figure 1.

2.1. Datasets used in the study

Cleveland, Heart, Hepatitis, Lymphography, Mammography, Newthyroid, Pima, and Thyroid datasets were obtained from a database designed with the General Public License version 3 (GPLv3), an open-source Java software tool called Evolutionary Learning Information Extraction (KEEL). We particularly selected aforementioned datasets to make directly comparable our findings in an unbiased manner to the results of the previous studies (Verma and Hassan 2011; Fernández et al. 2013; Kumar, Kongara, and Ramachandra 2013; Alanis-Tamez, Villuendas-Rey, and Yáñez-Márquez 2017) which used the same datasets and performed preprocessing before validation step.

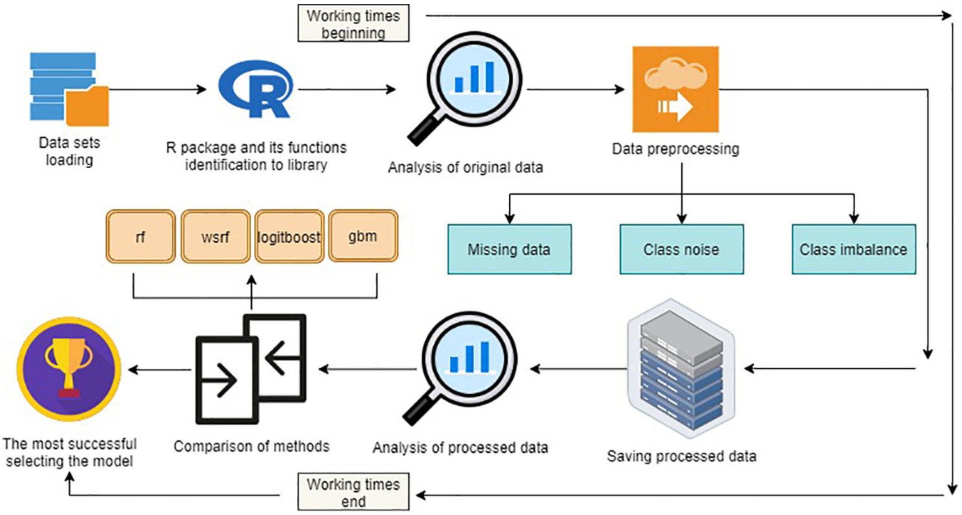

Figure 1. Flow chart showing the procedures for this study.

2.2. Problems encountered in the classification of the disease diagnosis

Classification algorithms are used to make physicians' decisions more effective in making a medical diagnosis. However, the performance of algorithms is affected and misdiagnosed because of the problems such as missing data, noisy data, and class imbalance/unbalance. For this reason, data should be preprocessed first. Some frequently occurring problems are described in detail below and solution suggestions for these problems are presented.

2.2.1. Problem 1: Missing data

Even in a well-designed and controlled study in health research, unknown or missing data may appear. Since an observation in the variable does not contain any value, the cells for that observation remain empty. In this case, assuming there is no missing information in all variables, problems occur when the analysis is made with standard statistical methods. Due to the missing data problem, some variables have relatively few observations, while sample width can decrease at the same rate. In this case, confidence intervals may be less reliable, statistical power lower and parameter estimates biased. How to apply different methods proposed to identify the type of missing data and cope with this problem has been subject to discussion in the literature (Little and Rubin 1987; Dong and Peng 2013).

It is stated that the presence of missing data can reduce statistical power (Kang 2013). Besides, it was emphasized that population parameters can be estimated as biased and the representative power of the sample may decrease. Missing observation takes place in two cases: not being answered at the unit level or not at the item level. For example, in a health study, if a participant refused to participate in a study or was not present at the time of the study, there was no response at the unit level since no information was received from a participant. In the other case, missing data occur due to the participants' inability to obtain information on some variables or features desired to be measured with a variable, such as not being answered at item level (Little and Rubin 1987). In the health researches, determining which solution suggestions to be presented and imputation methods to be used according to the type of missing observation to be encountered depends on missing data mechanisms. The mechanisms in the analysis of missing data were formulated in Rubin's (1987) theory (Little and Rubin 1987). Missing observation indicators are processed as random variables, assigned to a distribution. Missing data types are analyzed in three classes based on the relationship between missing and observed values. The three mechanisms are summarized below:

1. *Missing at random-MAR*: The case where the probability of the missing observation depends on the subject of research is called missing at random (MAR). Deficiency in the MAR mechanism is related to other observed variables. In addition, this mechanism is defined as whether it is negligible or not. There are methods to deal with negligible missing data (Little and Rubin 1991). When missing data mechanism is defined as MAR, expectation-maximization, regression imputation, and multiple imputation methods are appropriate imputation methods used in the solution of missing data problem (Sinharay, Stern, and Russell 2001).
2. *Missing completely at random-MCAR*: If the presence of missing observation arises from the variable itself rather than other variable values, this condition is defined as the missing completely at random (MCAR) mechanism (Little and Rubin 1987). To provide the assumption of randomization of missing data, missing variable values should not be related to the values of any variable. When the MCAR assumption is provided, it is provided to use full observations with listwise deletion or pairwise deletion (Acock 2005). The simple use of these methods and the short calculation times are among the reasons for choosing methods. In case the missing data mechanism does not have an MCAR feature, biased solutions can be obtained by using these methods. For this reason, if there is no MCAR feature in missing data mechanisms; it is more suitable to choose robust methods in missing data mechanisms without the MCAR feature (Mislevy, Little, and Rubin 1991).
3. *Missing not at random-MNAR:* If the probability of responses of missing data of variables depends on each variable, missing data is defined as a missing not at random. In other words, the lack of data is not random. Since missing data cannot be estimated using other variables, missing data in the NMAR mechanism depends on unobserved values. All imputation methods are biased in the case of the NMAR mechanism (Mislevy, Little, and Rubin 1991).

In some studies, when it is not possible to obtain full dataset due to the scope of research, missing observations can be ignored during analysis considering missing data is negligible. Missing data mechanisms developed against the negligibility of missing data are very important in determining appropriate analysis methods and interpreting results. Based on assumptions of random missing (MAR) and completely random missing (MCAR), it can be decided whether the missing data mechanism is "negligible." The missing data mechanism can be neglected if there is a relationship between estimation parameters under normal conditions and parameters estimated in presence of the MAR or MCAR assumptions and the presence of the missing data. If MAR assumption is not provided, the missing data mechanism is not negligible. The fact that there is a negligible missing data mechanism will enable missing data to be excluded from analysis; it will facilitate the work of the researcher, otherwise, the missing data mechanism will be modeled. When the dataset contains indispensable data, there is often insufficient information on which models will be more appropriate. As the results of the research will be quite sensitive to the selected model, it is stated that there can be an effective estimation with irrevocable missing data when there is very strong preliminary information about the nature of the missing data process (Allison 2002).

There is no exact cutoff point in literature where an acceptable rate for missing observations is specified. Some sources argue that a missing data rate of 5% or less is insignificant; there are also studies emphasizing that statistical analysis will have a biased estimate in cases where more than 10% of data is missing observation, or that structure of data will be greatly affected when more than 15% of missing observation is observed (Schafer 1999; Bennet 2001; Acuña and Rodriguez 2004).

Although the best method of dealing with missing observation is to plan study correctly and to collect data with care, this issue has become more sensitive especially in clinical studies. The recommended imputation methods to minimize the amount of missing data are included in two main titles, classical and advanced imputation methods. There are classical methods such as listwise deletion or pairwise deletion, hot-deck imputation, last observation carried forward, regression imputation, interpolation imputation method (Acock 2005). Also, the advanced imputation methods such as maximum likelihood method, multiple imputation method, single random imputation, and multiple random imputation methods have been developed (Sinharay, Stern, and Russell 2001).

2.2.2. Problem 2: Noisy data

Noisy data is defined as the data containing a large amount of meaningless information called noise. The term noise is also often referred to as corrupt data and is also defined as data that cannot be correctly understood and interpreted by machine learning methods. The noise problem frequently encountered in health data is included in the established classification model as a measurement error. As in other problems, in noisy data problems, the quality of data affects the performance of machine learning methods in terms of classification accuracy and interpretability of the results (Wang, Storey, and Firth 1995). Noise usually occurs during data collection and data preparation and depends on two main sources. These are implicit errors caused by measurement tools and defined by experts as random errors during data entry (Zhu and Wu 2004). While machine learning methods used in such cases are expected to have high classification performance, the quality of training datasets decreases, and at the same time, a classification algorithm is not robust to noise, which reduces the performance of methods (Wu and Zhu 2008). Components such as class labels and variable values directly affect the quality of a classification dataset. The quality of class labels indicates whether each individual's class has been correctly assigned. The quality of variable values, on the other hand, expresses the ability of individuals to accurately identify the diagnosis of disease. The presence of noisy data causes estimation errors in the classification problems of disease diagnosis. The boundaries of diagnostic classes lose their reliability due to noisy data and decision rules become difficult to develop. Based on these two sources of information, two different types of noisy data are defined (Zhu and Wu 2004).

1. *Class Noise:* This type of noise occurs when the individual is mislabeled. Class noise occurs due to various reasons during the tagging processes, such as subjectivity, data entry errors, or insufficient information used to tag each individual.
2. *Attribute Noise:* This noise type refers to distortions in one or more variable values or incorrect data entry (Frénay and Verleysen 2014).

Two approaches have been proposed in the literature to reduce the effects of noise (Zhu and Wu 2004). The first of these approaches is the stimulation of algorithms to properly handle noise. These methods are known as robust learners and are defined to be less affected by noisy data (Salzberg 1994). As a second approach, it is recommended to preprocess datasets that allow deleting or correct individuals with noisy data (Brodley and Friedl 1999). Although these two approaches proposed to reduce the effect of noise also perform well, they have some disadvantages. As the first approach, the stimulation of algorithms is based on a classification algorithm, the same results cannot be generalized to other machine learning algorithms because of extreme compatibility in the models. The second approach, the data preprocessing approach, is also faced with the situations mentioned in the first approach, and at the same time, these approaches are designed to detect only a certain type of noise; as a result, the data obtained is still not sufficient to achieve an excellent dataset (Wu and Zhu 2008). For these reasons, it is important to investigate other mechanisms that can lead to the reduction of effects caused by noise. This can be done without the need to adapt to each specific machine learning algorithm or without having to make assumptions about the current noise level and type in the data. Filtering methods are data preprocessing mechanisms used to detect and eliminate noisy individuals in the training set. Noise filters commonly used in literature, respectively; It is specified as an edited nearest-neighbor (ENN), all k-nearest neighbors (AllKNN), relative neighborhood graph edition (RNGE), modified edited nearest neighbor (MENN), nearest centroid neighbor edition (NCNEdit), cut edges weight statistic (CEWS), edited nearest neighbor estimating class probabilistic and threshold (ENNTh), classification filter (CF), multiedit, Tomek links (TL), iterative partitioning filter (IPF), cross-validated committees filter (CVCF), ensemble filter (EF), and iterative class noise filter based on the fusion of classifiers (INFFC) (Krawczyk and McInnes 2018; Morales et al. 2017). With the "NoiseFiltersR" package developed in the R programming language, there are class noise filters for classification data preprocessing (Morales et al. 2017).

2.2.3. Problem 3: Class imbalanced/unbalanced

In the class imbalance problem, unbalanced data occurs when a particular class (for instance disease class) is represented less proportionally than others. Thus, a majority-oriented learning bias is expected. Since bias is often defined as a misclassification problem resulting from a tendency to ignore minority class, this problem leads to a diagnostic error (Kübler, Liu, and Sayyed 2017; Zhang et al. 2017). The algorithms used to diagnose in the presence of unbalanced class problems are affected by this situation and may cause misdiagnosis. For classification algorithms to work with the data having such problem, various algorithms and/or approaches must be used to represent large amounts of raw data and information. Thus, researchers working in various fields such as patient safety/quality improvement, decision making, and problem-solving can safely use these algorithms and/or approaches to make medical diagnosis more reliable. In a study on medical diagnosis in literature, it has been stated that classification guides medical shaping and practice. Thus, it has been argued that understanding classification should be part of a quest to better understand diagnostic results (Jutel 2011). To make the correct diagnosis, it is necessary to use algorithms developed to deal with the unbalanced class problem and the existing machine learning algorithms in an integrated way. In one study, it was emphasized that due to the collection of large-scale industrial data in industrial systems, erroneous diagnoses and prognostics based on data constitute a field of study in these systems (Wu, Lin, and Ji 2018). Easy-SMT imbalanced learning algorithm has been proposed to provide solutions to imbalanced data problem obtained from these systems. This algorithm is defined as an integrated community-based method that includes the over-sampling method based on the SMOTE method to extend unbalanced minority classes and the EasyEnsemble method used for an ensemble-based balanced learning problem of an imbalanced class learning problem (Wu, Lin, and Ji 2018).

With such studies in literature, it is attempted to provide a more accurate definition of the disease. For example, the classification of chronic diseases is of great importance in making a medical diagnosis. The incidence of diabetes, one of the leading diseases of chronic diseases, is increasing faster than expected (Fatima and Pasha 2017). In a study using RUS (*random under-sampling*) and NCL (*neighborhood cleaning rule*) methods, which are among sub-sampling methods used in literature to examine the effect of class imbalance problem and to improve classification performance in unbalanced dataset for diagnosis of diabetes; original dataset is classified using decision tree algorithm (Folorunso and Adeyemo 2013). In this study, Kappa statistics, Matthew's correlation coefficient, accuracy, and error square mean square root criteria were compared with RUS and NCL method, and performances of datasets were compared and the NCL sub-sampling method has the best performance. In another study, after obtaining balanced classes using the SMOTE approach to deal with the unbalanced class problem, the diagnosis of diabetes was classified by using a decision tree, naive Bayes, logistic regression, logistic model tree, and random forest methods (Alghamdi et al. 2017). Kappa statistics, sensitivity, selectivity, precision, accuracy, and F-measure and performances of methods were compared, and random forest and Naive Bayes algorithms were found the best two algorithms according to the performance criteria. Similarly, in a previous study aiming to estimate disease risks arising from excessively unbalanced data, the performance of support vector machine, bagging, boosting, and random forest algorithms were compared in the classification of diagnosis of a total of eight chronic diseases (Khalilia, Chakraborty, and Popescu 2011). Since these datasets have extremely unbalanced classes, the problem has been solved by integrating the repeated random sub-sampling method. The algorithm with the best performance was a random forest algorithm in comparing performance with the ROC curve criterion. In all these studies, it is aimed to conduct a study to compare classification performances of existing machine methods, first after dealing with the imbalanced data problem, instead of applying existing machine learning algorithms directly to the original dataset. It was attempted to deal with this problem with resampling methods commonly used in literature. Resampling methods are statistical procedures that reuse data in the sample. Although there are many methods in literature to overcome the problem of class imbalance, resampling methods are specified as the most commonly used techniques to overcome this problem (Hu et al. 2009). These methods are briefly summarized below (Liu 2004):

- *Undersampling:* The class of individuals where the number of individuals in the dataset is high (majority class) is getting closer to the class (minority class) where the number of individuals is low. This reduction in the number of individuals in the dataset significantly reduces training time and provides a significant saving in terms of memory.
- *Oversampling:* Unlike the undersampling, the class of individuals where the number of individuals in the dataset is low (minority class) is getting closer to the class (majority class) where the number of individuals is high. Since the oversampling method greatly increases the size of the dataset, longer training time is observed and memory usage are generally very high.

2.3. Data preprocessing

Before using machine learning classification algorithms, possible problems such as missing data, noisy data, and class imbalanced were emphasized;

- The MICE approach, which is assumed that missing data is randomly lost and multiple imputation methods are generally MAR, was used for missing data preprocessing. Compared to a single imputation, multiple imputations focus on uncertainty in missing data.
- For noisy data problems, an R package called NoiseFiltersR was used for class noise data preprocessing in classification problems (Morales et al. 2017).
- Class imbalance SMOTE function is based on the basic logic of generating new artificial observations for minority class using the nearest neighbor algorithm. SMOTE also creates a more balanced dataset by sampling majority class observations at a low number.

2.4. Tree based ensemble learning algorithms

Ensemble learning methods named multiple classifier learning systems are used to train multiple classifiers simultaneously to solve the same problem (Yang et al. 2010). Contrary to learning approaches that try to model with a single classifier from training data, ensemble learning methods attempt to build and combine models with multiple classifiers.

Tree-based ensemble learning is a sensitive method that uses multiple classifiers rather than using a single classifier like other ensemble learning methods. The basic logic here is to combine estimates of a very different number of trees. Bagging and boosting methods are mentioned as two important basic ensemble learning methods. While the most preferred bagging algorithms in the health domain are random forest (rf) and weighted subspace random forest (wsrf); commonly used boosting algorithms are additive logistic regression (logitboost) and gradient boosting machines (gbm) algorithms (Witten, Frank, and Hall 2011). In a tree-based ensemble learning method, subsets of training dataset different from the original dataset are obtained. Models are then estimated for these training datasets and results are combined. Finally, the final model is set up and test data and models are tested (Dietterich 2000).

2.4.1. Random forest

The random forest algorithm is a bagging method that allows a combination of the bagging method proposed by Leo Breiman in 2000 and the random subspace method proposed by Ho in 1998 (Ho 1998; Breiman 2001). The random forest was proposed by Breiman in 2001 on grounds which provides more randomization than boosting algorithms (Breiman 2001). A random forest algorithm is also a simultaneous approach as it is based on the bagging algorithm. In this algorithm, decision trees, which are expressed as individual classifiers in the forest, are obtained with the CART algorithm (Sutton 2005). These individual classifiers come together to form a decision forest ensemble. Each decision tree that creates a decision forest creates different samples with a bootstrap sampling method selected by randomly replacing it from the original dataset. After the decision forest ensemble is created, results are combined by combining the forecast results, and the final forecast is obtained.

2.4.2. Weighted subspace random forest

The weighted subspace random forest (wsrf) algorithm is the bagging method proposed by Xu et al. (2012). This algorithm uses decision trees and variables for nodes, just like the rf algorithm. In particular, the materiality of variables is of great importance for the performance of the wsrf algorithm. The reason for this is that decision trees determine which variable to start split. If decision trees are not established with highly important variables, weak decision tree models will be established in the forest. Thus, it will cause the overall accuracy of the model with the wsrf algorithm to decrease (Kotsiantis 2011). In their study, Amaratunga, Cabrera, and Lee (2008) proposed a variable weighting method for subspace sampling to solve this problem. With this method, the weight of a variable is calculated from the t-test used for variance analysis, and the correlation between the variable and the class. This weight is used as the probability of a variable selected for a subspace. The higher correlation between variable and class, the higher the weight variable will have. Thus, there will be an increase in performance of decision trees established with a variable with high weight (Amaratunga, Cabrera, and Lee 2008). In wsrf model, it causes an increase in the average success of trees that make up the forest, since informative variables are more likely to be selected when growing decision trees in a random forest. Xu et al. (2012) proposes wsrf algorithm based on a weighted subspace random forest algorithm to classify very high-dimensional data using information gain ratio to calculate variable weights so that method proposed by Amaratunga, Cabrera, and Lee (2008) can be applied in a multi-class situation (Amaratunga, Cabrera, and Lee 2008; Xu et al. 2012).

2.4.3. Additive logistic regression

Additive logistic regression algorithm is a boosting algorithm proposed by Friedman in 1998 (Friedman, Hastie, and Tibshirani 2000). Additive regression can be applied in the classification model as well as in the linear regression model. The Logitboost classifier is a meta-estimator that builds an additive model that can minimize the logistic loss function. As with all boosting methods, the basis of the logitboost method is based on the AdaBoost algorithm. In other words, the

Table 1. Error matrix for calculating performance measurement metrics.

		Real status	
		Positive	Negative
Predicted status	Positive	TP	FN
	Negative	FN	TN

logitboost algorithm is a generalized version of the AdaBoost algorithm. In classification, the AdaBoost algorithm can be built only for binary classes, while the logitboost algorithm builds models for both binary and multiple classes.

2.4.4. Gradient boosting machines

Gradient boosting machines algorithm is a boosting algorithm proposed by Friedman (2001). The main purpose of the Gbm algorithm is to adapt the loss function obtained in the previous iteration to a negative gradient vector instead of re-weighting basic classifiers, as in the AdaBoost algorithm. The gradient has partial derivative information of a multivariable function. In this study, while the logistic regression loss function is used for datasets whose dependent variable is binary by default, it is preferred to use the multi-class logistic regression loss function for datasets with multiple classes.

2.5. Performance metrics used in classification

While comparing performances of ensemble learning algorithms to be used in this study; accuracy-ACC, sensitivity/recall-SEN, specificity-SPE, precision-PRE, Kappa statistic-KAP, Youden index-YI, F-measure–F, and ROC measurement metrics were used. These metrics were, calculated according to Table 1, called the error matrix or confusion matrix (Chawla et al. 2002). Performance metrics are between 0 and 1. Measurement metric values being very close to 1 mean that classification success is high. According to Table 1, TP refers to true positive, FP refers to false positive, FN refers to false negative, and TN refers to true negative.

- **Accuracy:** The accuracy of a test is the ability to accurately distinguish patients and healthy individuals (Swets 1988). While calculating the accuracy of the diagnostic test (Formula 1), correct positive and correct negative ratios are calculated for all patients and healthy individuals. The higher classification accuracy, the better system performs.

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

- **Sensitivity:** The sensitivity of a test (Formula 2) is the conclusion of the patient as a result of the test when an individual knows that he is a patient (Maimon and Rokach 2005). True positive rate is calculated to differentiate patient individuals correctly. The sensitivity value is also equal to the power of the test (1 − β). Sensitivity is calculated with a similar formula to the recall value.

$$SEN = REC = \frac{TP}{TP + FN} \tag{2}$$

- **Specificity:** The specificity of a test (Formula 3) is defined as the probability that diagnostic test results will be negative when it is known that an individual is not a patient (Maimon and Rokach 2005). The true negative rate is calculated to differentiate healthy individuals correctly.

$$SPE = \frac{TN}{TN + FP} \tag{3}$$

- **Precision**: The precision of a test (Formula 4) is calculated when a diagnostic test's positive predictive value or positive predictions are positive (Buckland and Gey 1994). In other words, it is defined as the probability of an individual who is positive as a result of the diagnostic test.

$$PRE = \frac{TP}{TP + FP} \tag{4}$$

- **Kappa statistic**: Kappa statistic (Formula 5), often used to test interobserver reliability, is a conformity metric that was first proposed by Cohen (Cohen 1960; McHugh 2012). Kappa statistic takes values ranging from -1 to 1. The value of Kappa greater than 0.75 indicates a perfect fit; a value less than 0.40 indicates a poor fit. Kappa values between 0.40 and 0.75 are defined as acceptable fit.κ is shown by p_o when observed fit rate and are p_e taken as expected fit rate.

$$\kappa = \frac{p_o - \ p_e}{1 - \ p_e} \tag{5}$$

- **Youden index:** This metric, like the ROC curve, is a criterion (Formula 6) used to measure the effectiveness of the diagnostic test and to determine the most appropriate threshold value (Youden 1950; Fluss, Faraggi, and Reiser 2005). Despite the ROC curve, which is the most commonly used measure of diagnostic accuracy, Youden index is also preferred. Youden index ranges from 0 to 1. In case of complete separation of patients and healthy individuals, $J = 1$; in case of full overlap, $J = 0$.

$$J = max_C\left\{sensitivity(c) + specificity(c) - 1\right\} \tag{6}$$

- **F-measure:** The metric F-measure (Formula 7) is calculated by taking a harmonic average of precision and precision values (Buckland and Gey 1994). In this case, it is not easy to understand as accuracy because it calculates by considering both false positive and false negative values at the same time, but if there is an unbalanced class distribution, it is recommended to evaluate the F-measure value with accuracy. If positive and false negatives have similar costs, it is more beneficial to use the correct value. Similarly, if the costs of false positives and false negatives are different, it is better to prefer to use the F-measure value (Watson and Petrie 2010).

$$F = \ 2* \frac{\left(\frac{TP}{TP+FN}\right)*\left(\frac{TP}{TP+FP}\right)}{\left(\frac{TP}{TP+FN}\right) + \left(\frac{TP}{TP+FP}\right)} = \ 2* \frac{REC*PRE}{REC + PRE} \tag{7}$$

- **ROC curve:** The ROC curve, also referred to as Area Under the Curve (AUC), is a method used in cases such as evaluating diagnostic tests and prediction models, determining appropriate threshold and discrimination of test, and comparing the performance of two or more diagnostic tests (Hajian-Tilaki 2013). The area under the curve shows how much differentiating power it has in diagnosing patients and healthy individuals. The ROC curve gives a graph of $1 - specificity(c)$ values versus $sensitivity(c)$ for all possible threshold values (c) in determining the appropriate threshold value. The area under the curve is between 0 and 1 and is not affected by the prevalence of mass. While the area under the curve that a diagnostic test can randomly take is 0.5; the accuracy of an excellent diagnostic test is expressed as 1 (Bradley 1997; Maimon and Rokach 2005).

Table 2. General information on datasets.

Data set	Sample size	Variable number	Continuous variable number	Categorical variable number	Class levels	Percentage of missing data	Percentage of class noise	Class unbalance rate
Cleveland	303	13	5	8	5	1,980	19,802	1,179
Heart	270	13	6	7	2	–	4,074	1,250
Hepatitis	155	19	6	13	2	48,387	4,516	3,844
Lymphography	148	18	3	15	4	–	–	1,209
Mammographic	961	5	1	4	2	13,632	9,886	1,159
Newthyroid	215	5	5	–	3	–	2,791	2,308
Pima	768	8	8	–	2	–	7,943	1,866
Thyroid	7,200	21	6	15	3	–	0,056	12,483

3. Results

Before proceeding to data preprocessing stages, the percentage of missing data, percentage of class noise, and class imbalance rate of eight different datasets were given, and performances of classification models were calculated with original datasets. Then, the performances of processed datasets were recalculated and results were compared. Since the class noise package used in data preprocessing stages is sensitive to missing data, firstly, necessary assignments were made to datasets with missing data by multiple imputation method. As a second step, data with class noise had been cleared to avoid balancing with noisy classes. At the last step, the class imbalance problem was eliminated for all datasets, balanced classes were obtained and new datasets were recorded. Later, the algorithm that has the most successful classification performance metrics were determined with processed datasets. The running times of algorithms were recorded during all these processes. The first stage of the running time of algorithms was an analysis of original data, the second stage was data preprocessing and the final stage was an analysis of processed data.

Table 2 contains general information about original datasets. Sample sizes, variable numbers, missing observation percentages, class noise percentages, and class imbalance ratios of eight different datasets; the numbers of both continuous and categorical variables were not similar. While the dataset with the lowest sample width was the Lymphography data, the data with the highest sample width was the Thyroid data. In addition, the Mammographic and the Newthyroid datasets had the least number of variables, while the Thyroid dataset had the highest number of variables. It is noteworthy that Cleveland, Heart, Hepatitis, and Lymphography datasets had low sample widths and large numbers of variables. In this case, there was a size problem between sample width and a number of variables. While Newthyroid and Pima datasets had only continuous variables, other datasets included both types of variables. Another important point here is that the numbers of continuous and categorical variables of datasets were differ from each other. The categorical variable numbers of Hepatitis, Lymphography, and Thyroid datasets were considerably higher than the number of continuous variables.

When original datasets were considered in terms of class levels of a dependent variable, Heart, Hepatitis, Mammographic and Pima data had two categories; while multi-class datasets Newthyroid and Thyroid had three categories, Lymphography had four categories, and Cleveland had five categories. All datasets were examined in terms of possible problems that may be encountered during the preprocessing stage, Cleveland, Hepatitis, and Mammographic datasets had missing data problem. The dataset with the least missing data was Cleveland (1.980%), while data with the most missing observation was Hepatitis (48.387%). The least class noise dataset was Thyroid dataset (0.056%), while the most class noise dataset is Cleveland dataset. There was a class imbalance problem in all datasets. While the dataset with the lowest class imbalance was Mammographic data (1.159), the data with the highest class imbalance was Thyroid data (12.483). It should be noted here that Cleveland, Hepatitis, and Mammographic data had a feature of having all three problems. Heart, Newthyroid, Pima, and Thyroid data were datasets with

Table 3. Model performance measurement metrics values for original data sets.

Data set	Algorithm	ACC	SEN	SPE	PRE	Kappa	F	YI	ROC
Cleveland	rf	0,568	0,244	0,854	0,238	0,216	0,241	0,099	0,766
	wsrf	0,579	0,264	0,857	0,289	0,235	0,277	0,121	**0,802**
	logitboost	0,575	**0,351**	0,866	0,355	0,276	**0,353**	**0,217**	0,782
	gbm	**0,591**	0,293	**0,878**	**0,426**	**0,305**	0,347	0,169	0,777
Heart	rf	**0,926**	**0,933**	**0,917**	**0,933**	**0,850**	**0,933**	**0,850**	**0,961**
	wsrf	0,877	0,889	0,861	0,889	0,750	0,889	0,750	0,949
	logitboost	0,817	0,892	0,735	0,786	0,631	0,835	0,627	0,909
	gbm	0,914	**0,933**	0,889	0,913	0,825	0,923	0,822	0,949
Hepatitis	rf	0,800	0,444	0,889	0,500	0,348	0,471	0,333	**0,813**
	wsrf	**0,867**	0,556	**0,944**	**0,714**	**0,546**	0,625	0,500	0,796
	logitboost	0,800	0,556	0,861	0,500	0,400	0,526	0,417	0,741
	gbm	0,844	**0,667**	0,889	0,600	0,533	**0,632**	**0,556**	0,756
Lymphography	rf	0,767	0,385	0,881	0,382	0,529	0,384	0,266	**0,857**
	wsrf	0,721	0,365	0,861	0,357	0,442	0,361	0,226	0,843
	logitboost	**0,786**	**0,401**	**0,895**	**0,391**	**0,577**	**0,396**	**0,296**	0,528
	gbm	0,767	0,389	0,892	0,389	0,547	0,389	0,281	0,836
Mammographic	rf	0,843	0,864	0,819	0,847	0,684	0,855	0,683	0,889
	wsrf	0,822	0,844	0,797	0,828	0,642	0,836	0,641	0,868
	logitboost	**0,878**	**0,909**	**0,844**	**0,866**	**0,755**	**0,887**	**0,753**	0,865
	gbm	0,843	0,870	0,812	0,843	0,684	0,856	0,682	**0,895**
Newthyroid	rf	**0,922**	**0,826**	**0,912**	**0,967**	**0,813**	**0,891**	**0,738**	0,881
	wsrf	0,859	**0,826**	0,899	0,811	0,701	0,819	0,725	0,857
	logitboost	0,891	0,811	0,900	0,882	0,749	0,845	0,711	**0,894**
	gbm	0,859	0,767	0,876	0,858	0,677	0,810	0,643	0,844
Pima	rf	0,761	0,562	0,867	0,692	0,449	0,621	0,429	0,838
	wsrf	0,752	0,575	0,847	0,667	0,436	0,617	0,422	0,832
	logitboost	0,748	**0,588**	0,828	0,635	0,424	0,611	0,416	0,742
	gbm	**0,778**	0,538	**0,907**	**0,754**	**0,476**	**0,628**	**0,445**	**0,851**
Thyroid	rf	0,995	0,986	0,992	0,959	0,964	0,972	0,978	0,921
	wsrf	0,997	**0,996**	**0,997**	0,963	0,977	0,979	**0,993**	0,833
	logitboost	**0,998**	0,972	0,988	0,958	0,959	0,965	0,961	0,815
	gbm	**0,998**	0,993	**0,997**	**0,981**	**0,984**	**0,987**	0,989	**0,969**

both class noise and class imbalance. Lymphography data, which was a dataset that contains only a single data preprocessing problem, had a class imbalance problem. In terms of all these situations, when original datasets were considered, advantages and differences of datasets were revealed.

3.1. Analysis of original datasets

First of all, different models were established to classify with original datasets and the performance measurement metrics of algorithms were calculated. To test the validity of the models in Table 3, all datasets were divided into two as 30% test and 70% training datasets. The results in bold in Table 3 represented the best values of the model performance measurement metrics for the original datasets.

When all algorithm models of Cleveland data were evaluated in terms of performance measurement metrics, modeling performance achievements yielded very low results. Small sample size, 1.98% missing observation, 19.802% class noise, and 1.179 class imbalance were among the reasons affecting the success of modeling performance. Although measurement metrics of all four algorithms gave values as close to each other as possible, these values were not sufficient when evaluated in terms of model successes. Meanwhile, boosting algorithms produced similar results in all other measurement metrics except for the ROC curve value (0.802%). Therefore, algorithms with the best classification success were logitboost and gbm algorithms.

Although Heart data showed high values in all measurement metrics in general, Kappa statistic (0.631%) and Youden index (0.627%) of logitboost algorithm gave slightly lower results. On the other hand, rf and gbm algorithms had the same sensitivity value (0.933%). At the same time, the Heart data had a class noise of 4.074% and a class unbalance ratio of 1.25. The best classification algorithm of heart data was the rf algorithm. The reasons why the Hepatitis data measurement metrics generally gave low results included 48.387% missing observation, 4.516% class noise, and 3.844 class imbalance. In general, it was not correct to generalize by looking at a single measurement value, as it will be understood that all measurement values except accuracy and specificity yielded very low results. According to these results, since wsrf and gbm algorithms had very close values, both algorithms were chosen as the most successful classification algorithms. On the other hand, the rf algorithm had the highest ROC curve value (0.813%).

Lymphography data, like Hepatitis data, had very low classification successes, and its class imbalance rate was 1,209. Although the algorithm with the best classification success in terms of all measurement values was the logitboost algorithm model, the rf algorithm had the highest ROC curve value (0.857%). In the analyzes of the Mammographic data, although all models generally had high classification success; classification successes of the models were not sufficient since the dataset had 13.632% missing observations, 9.886% classroom noise, and 1.159 class imbalance. Although the model with the most successful classification algorithm for Mammographic data was the logitboost algorithm, gbm algorithm had the highest ROC curve value (0.895%). Although Newthyroid data had high classification success in all algorithms in general, it had 2.791% class noise and 2.308 class imbalance.

The dataset gave approximately similar results in all algorithms; however, the best classification success belongs to the rf algorithm, and the algorithm with the highest value in terms of ROC curve value (0.894%) was the logitboost algorithm. Pima data gave very low values in all other measurement metrics except specificity and ROC curve values. Possible reasons for this may be that Pima data had 7.943% class noise and 1.866 class imbalance. Given all these situations and other algorithms gave approximately similar results, the best classification success belongs to the gbm algorithm. Thyroid data had 0.056% class noise and 12,483 class imbalance, although it gave very high results in terms of all algorithms. The algorithm with the best successful classification performance for thyroid data was the gbm algorithm. As stated in Figure 2, all performance measurement metrics for original data had values between 0.20% and 0.99%. This range was expected to narrow further after data preprocessing. Thus, improvements in model performances were observed more clearly.

3.2. Analysis of preprocessed datasets

The results indicated in bold in Table 4 represented the best values of model performance measurement metrics for preprocessed datasets. When measurement metrics of the Cleveland data for all algorithm models were evaluated, modeling performance success was quite high compared to original data. The algorithm with the best classification success for Cleveland data was the gbm algorithm. However, only in terms of precision criteria, the best value with a rate of 0.927% was obtained in the rf algorithm. The Heart data gave high values in all measurement metrics in general. Only in terms of sensitivity criteria, the wsrf and the gbm algorithms had the same value as 0.979%. Accordingly, the most successful classification algorithm of the Heart data was the gbm algorithm. Generally, measurement metrics of Hepatitis data provided high results. In particular, identical results (0.975%) were obtained in all four algorithms in terms of specificity criteria. According to these results, the most successful algorithm of Hepatitis data was the logitboost algorithm. The algorithm with the best classification success in all measurement values for Lymphography data was the gbm algorithm. Mammographic data had a fairly high classification success for all models. Although it received the best value in gbm algorithm with only 0.95% in specificity criterion and 0.999% in ROC curve criterion, the data achieved the best classification success in the logitboost algorithm. New thyroid data showed high classification success in all

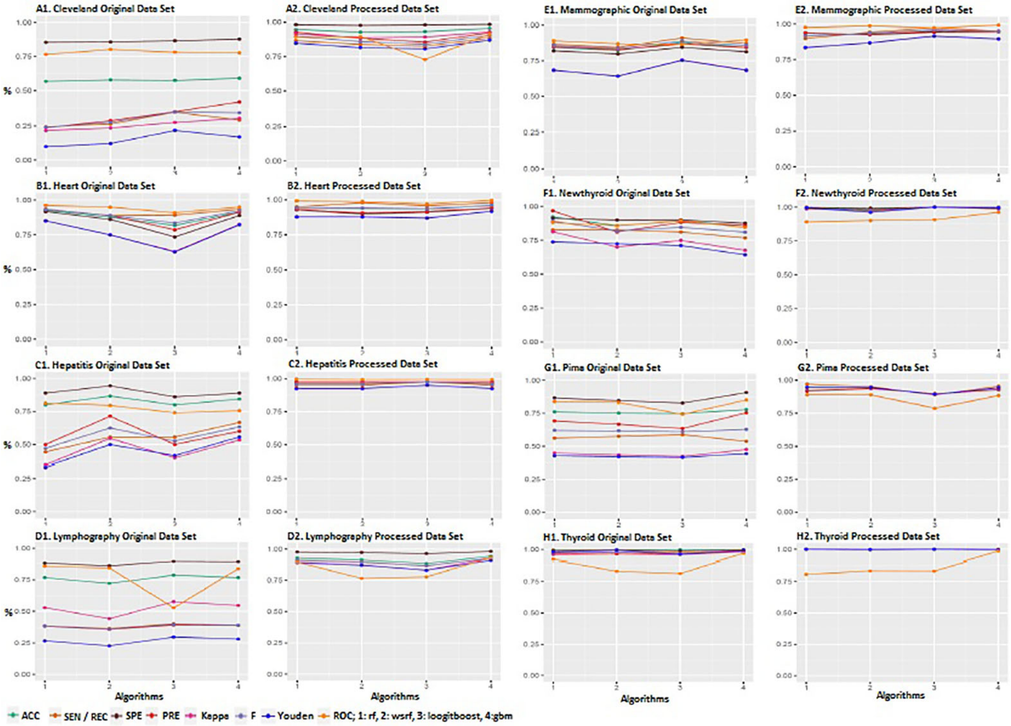

Figure 2. Interaction effect of group × intervention on weekly alcohol consumption over time.

classification algorithms. The algorithm for the best classification success of Newthyroid data was logitboost algorithm. In spite of this algorithm shows 100% success in all performance measurement metrics, it only had a value of 0.907% ROC curve.

Pima data had the best classification value with 0.98% ROC curve value only in the gbm algorithm. Besides, 0.944% accuracy, 0.889% Kappa statistics, and 0.889% Youden index values were obtained in the rf and the wsrf algorithms. According to all these results, both of the rf and the wsrf algorithms were the most successful classification algorithms for Pima data. Thyroid data, as in the original dataset, again presented quite high results in terms of all algorithms. Although thyroid data had 100% success in the rf and the logitboost algorithms, had the best classification success in the gbm algorithm with only 0,987% in ROC curve value. According to these results, the most successful classification algorithms for the Thyroid data were the rf and the logitboost algorithms. As stated in Figure 2, all performance measurement metrics for processed data took values between 0.70% and 0.100%. Thus, it was observed that the range narrowed after data preprocessing, and improvements in model performance were observed. In Table 5, the differences between the classification accuracy values of the algorithms before and after data preprocessing were found to be statistically significant with 95% confidence, and the accuracy values of the algorithms with data preprocessing were found to be statistically significantly higher ($p < 0.05$).

According to these results, boosting algorithms yielded more successful results than bagging algorithms. One of the reasons for this is that bagging algorithms were more difficult to train in these datasets and the modeling success was lower than boosting algorithms. Another reason is that boosting algorithms update models with observation focus each time. Thanks to these updates, misclassified observations increase the chances of being classified correctly. It is also understood that data preprocessing had a statistically significant effect on model achievements.

Table 4. Model performance measurement metrics values for processed data sets.

Data set	Algorithm	ACC	SEN	SPE	PRE	Kappa	F	YI	ROC
Cleveland	rf	0,946	0,865	0,982	**0,927**	0,915	0,895	0,847	0,895
	wsrf	0,927	0,838	0,977	0,882	0,886	0,859	0,815	0,895
	logitboost	0,931	0,825	0,981	0,855	0,892	0,839	0,806	0,728
	gbm	**0,954**	**0,884**	**0,985**	0,922	**0,928**	**0,903**	**0,869**	**0,914**
Heart	rf	0,939	0,949	0,929	0,930	0,878	0,939	0,878	0,992
	wsrf	0,939	**0,979**	0,898	0,906	0,878	0,941	0,878	0,988
	logitboost	0,934	0,957	0,912	0,917	0,869	0,936	0,869	0,970
	gbm	**0,959**	**0,979**	**0,939**	**0,941**	**0,918**	**0,960**	**0,918**	**0,997**
Hepatitis	rf	0,963	0,950	**0,975**	0,974	0,925	0,962	0,925	**0,997**
	wsrf	0,963	0,950	**0,975**	0,974	0,925	0,962	0,925	0,994
	logitboost	**0,975**	**0,975**	**0,975**	**0,975**	**0,950**	**0,975**	**0,950**	0,995
	gbm	0,963	0,950	**0,975**	0,974	0,925	0,962	0,925	0,994
Lymphography	rf	0,928	0,911	0,977	0,909	0,893	0,910	0,889	0,894
	wsrf	0,913	0,896	0,973	0,893	0,871	0,894	0,869	0,764
	logitboost	0,884	0,865	0,964	0,865	0,829	0,865	0,829	0,776
	gbm	**0,942**	**0,927**	**0,982**	**0,927**	**0,914**	**0,927**	**0,909**	**0,930**
Mammographic	rf	0,917	0,897	0,937	0,934	0,834	0,915	0,834	0,975
	wsrf	0,933	0,942	0,924	0,926	0,867	0,934	0,867	0,987
	logitboost	**0,958**	**0,972**	0,942	**0,948**	**0,915**	**0,960**	**0,914**	0,972
	gbm	0,947	0,950	**0,945**	0,945	0,894	0,947	0,894	**0,992**
Newthyroid	rf	0,995	0,996	0,998	0,987	0,992	0,992	0,994	0,891
	wsrf	0,985	0,969	0,992	0,979	0,974	0,974	0,961	0,901
	logitboost	**1,000**	**1,000**	**1,000**	**1,000**	**1,000**	**1,000**	**1,000**	**0,907**
	gbm	0,995	0,996	0,998	0,987	0,992	0,992	0,994	0,961
Pima	rf	**0,944**	**0,971**	0,918	0,922	**0,889**	**0,946**	**0,889**	0,986
	wsrf	**0,944**	0,952	**0,937**	**0,938**	**0,889**	0,945	**0,889**	0,979
	logitboost	0,894	0,889	0,899	0,898	0,788	0,894	0,788	0,927
	gbm	0,942	0,956	0,927	0,929	0,884	0,943	0,884	**0,980**
Thyroid	rf	**1,000**	**1,000**	**1,000**	**1,000**	**1,000**	**1,000**	**1,000**	0,809
	wsrf	0,998	0,998	0,999	0,998	0,997	0,998	0,997	0,835
	logitboost	**1,000**	**1,000**	**1,000**	**1,000**	**1,000**	**1,000**	**1,000**	0,833
	gbm	0,998	0,998	0,999	0,998	0,997	0,998	0,997	**0,987**

3.3. Running time of algorithms

Running time of all calculations for the algorithms were recorded. According to the results in Table 6, boosting algorithms run longer than the bagging algorithms. As the width of the dataset increases, the running times of algorithms increase in the same way. Processed data run longer than original data. While the gbm algorithm worked with the longest run time; the rf algorithm had the shortest runtime. Since the Thyroid data was the dataset with the largest sample size, it had the longest run time (168.77 min) in both original and processed data. On the other hand, the new thyroid data had the shortest running time (2.36 min) in both original and processed data. In the data preprocessing step, filtering took the longest time for the class noise problem.

4. Discussion

Based on hypotheses, results discussed in this study were compared with similar and different studies in the literature. In the currently available literature, studies on this subject have been evaluated only in terms of general accuracy values of the models. 10-fold cross-validation method was used for model validity in all studies in literature and including in our study. Also, datasets we used in our study were chosen differently in terms of sample sizes, class levels, missing data percentages, class noise percentages, and class imbalance rates.

In the previous study of Moon et al. (2007), they developed a classification approach with an ensemble learning-based random partitions and compared the performance of this approach with random forest, Adaboost, logitboost, decision forest, support vector machine, diagonal linear discriminant analysis, shrunken centroids, CART, neutral interaction selection and predictive

Table 5. Comparison of accuracy values of algorithms.

Algorithm	Status	Mean ± SS	t	*p**
rf	Without preprosessing	0,822 ± 0,132	−3,075	0,018
	With preprosessing	0,954 ± 0,030		
wsrf	Without preprosessing	0,809 ± 0,125	−3,801	0,007
	With preprosessing	0,950 ± 0,029		
logitboost	Without preprosessing	0,811 ± 0,123	−3,733	0,007
	With preprosessing	0,947 ± 0,044		
gbm	Without preprosessing	0,824 ± 0,119	−3,615	0,009
	With preprosessing	0,962 ± 0,022		

*Paired sample T-test.

classification rule and fast, neutral and efficient statistics tree algorithms. Genomic data of Lymphoma, Lung Cancer, and Breast Cancer diseases were used in the study. The dimensions of these data are considerably larger than the datasets we use in our study. Lymphoma data is a dataset consisting of 47 samples with 4,026 genes. For datasets with missing observation problems, imputations were made using the 10-nearest neighbor method means. Lung cancer data is a dataset consisting of 181 samples with 12,533 genes. Since noisy data was filtered from this data, analysis was performed with 5,000 genes. Breast cancer data is a dataset consisting of 78 samples with 25,000 genes. Since there are 5,000 significant genes in this dataset, the remaining genes are removed from the dataset. As a result of all these data preprocessing, when comparing model performances for all three datasets, the algorithm of classification from random sections has been the most successful algorithm. Different from our study, the diagnosis of the disease was classified using genomic data (Moon et al. 2007).

In the study conducted by Verma and Hassan (2011), diagnosis of disease was defined in Mammography, Wisconsin Breast Cancer, and Pima Diabetes data using unsupervised ensemble learning approaches. The hybrid ensemble learning methods used in the study gave 100% accuracy for Mammography data as in our study. This result shows that it is possible to achieve this accuracy in training and test data if excellent modeling success is obtained between data (Verma and Hassan 2011; UCI 2019). Lavanya and Rani (2012) used the CART bagging algorithm, a variable selection approach that works integrated with the CART algorithm, and a hybrid approach which can preprocess data. They compared model performances using Breast Cancer, Wisconsin Breast Cancer (Original) and Wisconsin Breast Cancer (Diagnostic) datasets from the UCI database. The sample sizes of these datasets consist of 286, 699, and 569 observations, respectively. In the data preprocessing phase, the most significant independent variables were selected and missing observations were deleted from datasets. In the study, models' running time were recorded considering only accuracy criteria. According to these results, the hybrid approach has the best accuracy values for all three datasets (ACC: 74.47%; 97.85%; 95.96%), while the longest working time was recorded as 28.25 min for Breast Cancer data. In this study, it can be concluded that how difficult it is to learn from data as a reason for Breast Cancer data to have the longest running time although it has the least sample size. In this case, it is not always correct to make a judgment such as obtaining the fastest working time in the smallest sample. The most obvious situation we can see is that there is no significant difference between the running time of the original data we use in our study and the running time of processed data. There is an obvious difference between working times of Thyroid data only. The reason for this is that the sample size is reduced by half (Lavanya and Rani 2012; UCI 2019).

Kumar, Kongara, and Ramachandra (2013) used Wisconsin Breast Cancer and Pima Diabetes datasets from the UCI database. Accuracy, sensitivity, specificity, error rate metrics performance were calculated for these datasets using bagging, boosting, AdaBoost, Multiboosting, and random forest algorithms. They also recorded algorithms' running times and compared them. These variables were deleted from the dataset without any modifications to independent variables. While the random forest was the most successful algorithm for the Breast Cancer dataset (ACC: 94.49%), the MultiBoost algorithm was found to be the most successful algorithm for the diabetes dataset

Table 6. Running times of algorithms.

	Step1: Original data				Step2: Data preprocessing			Step3: Processed data				
Data set	rf	wsrf	logit boost	Gbm	Missing data	Class noise	Class imbalance	rf	wsrf	logit boost	gbm	For three stages total time (min)
Cleveland	0,17	0,49	2,53	4,57	0,06	0,15	0,01	0,34	1,24	4,53	13,28	27,37
Heart	0,12	0,36	0,34	0,59	–	0,12	0,01	0,33	1,09	0,53	2,19	5,68
Hepatitis	0,05	0,11	0,53	0,48	0,10	0,12	0,01	0,09	0,23	0,52	1,23	3,47
Lymphography	0,12	0,28	4,25	2,56	–	–	0,01	0,11	0,25	4,25	4,57	16,40
Mammography	0,31	1,32	0,27	1,10	0,26	0,32	0,01	1,17	2,56	0,54	2,48	10,34
Newthyroid	0,05	0,13	0,18	0,28	–	0,07	0,01	0,11	0,14	0,27	1,12	2,36
Pima	0,27	1,24	0,24	1,03	–	0,18	0,01	0,39	1,47	0,33	1,39	6,55
Thyroid	4,10	3,02	35,12	25,36	–	42,48	0,01	1,37	2,13	12,14	43,04	168,77

(ACC: 76.31%). In addition, bagging algorithms are the longest running algorithms. These times are about 0.17 min for Wisconsin Breast Cancer data and about 0.50 min for Pima Diabetes data (Kumar, Kongara, and Ramachandra 2013; UCI 2019). In a study conducted by Sáez, Krawczyk, and Woźniak (2016), the class noise problem was discussed in detail. In this study, it was stated that class noise in the classification of medical data consisted of many sources such as human errors, machine errors, digitization, and archiving errors. In solving this problem, algorithm level and data level approaches are proposed (Sáez, Krawczyk, and Woźniak 2016).

Similar to our study, Alanis-Tamez, Villuendas-Rey, and Yáñez-Márquez (2017) used disease diagnosis datasets from the KEEL database. In this study, although attention was drawn to missing data and class imbalance problems, no data preprocessing was performed. With this original data; classification models were obtained using logistic, multi-class logistic, k-nearest neighbors, support vector machines, and C4.5 decision tree weak classifiers. In addition, unlike other studies, model verification was performed with a 5-fold cross validation method (Alcalá-Fdez et al. 2009; Alanis-Tamez, Villuendas-Rey, and Yáñez-Márquez 2017). When all these results were tested with the Friedman's test statistics logistic model was found to be the best classification algorithm ($p < 0.05$).

Gunčar et al. (2018) developed two different ensemble-based smart blood analytics algorithms for the classification of hematology disease diagnosis. These two approaches are called SBA-HEM61 and SBA-HEM181. These approaches can only estimate hematological data using blood test result parameters. These data were taken from a database of 8233 patients consisting of 43 different hematological categories. In this study, in the data preprocessing process, excessive deviations were removed from datasets, and imputation was made with the MICE algorithm for missing observation problem as in our study. According to all these results, the modeling general accuracies of both algorithms increased from 59% and 57%, respectively, to 88% and 86% (Gunčar et al. 2018). Although it was used to balance class distributions with over-sampling and under-sampling methods recommended in the literature for the problem of class imbalance, it was emphasized by Fernández et al. (2013) that there are some serious handicaps in the use of methods. Classification performance in test datasets can often yield much worse results, even if the accuracy of the training dataset is high since over-sampling problem occurs in the dataset due to over-sampling (Fernández et al. 2013). In this study, taking into consideration this situation, SMOTE approach, which is more resistant to excessive adaptation problem, was preferred. Thus, the ability to provide a balanced distribution without losing information about the majority class of over-sampling was used. In their study, Zhang et al. (2019) aimed to evaluate the performance of the Gbm algorithm by simulating a binary class disease diagnosis dataset. Accordingly, the gbm algorithm used when modeling complex relationships of data obtained by simulation has been shown to give better results than the logistic regression model approach based on the generalized linear model. The performance of the logistic regression algorithm with Gbm was tested with DeLong's test, which allows measuring the relationship between ROC values (Zhang et al. 2019). According to this result, it was concluded that the gbm algorithm was statistically more significant than the logistic regression algorithm (0.98; 95% CI: 0.972–0.997).

There is no study in the literature evaluating rf, wsrf, logitboost and gbm algorithms together. When examining all these previous studies in the literature, there was no similar study comparing performances of ensemble learning algorithms for both original and processed datasets in cases where problems such as missing data, class noise, and class imbalance exist, as in our study. These studies also show that there can be many problems with the data and modeling performances decrease when any learning algorithm is applied to data without solving these problems. After problems are solved, it is more accurate to try to learn from these models.

5. Conclusions

According to findings obtained from the study, it was observed that performance classification performances of original datasets were not significantly successful. One possible reason of this might be that the sample sizes of some datasets are quite low compared to the number of variables, and sample representation power is insufficient. Other important reasons are issues such as missing data, class noise, and class imbalance. When processed datasets that are cleared of these problems are re-modeled with the same parameters, very high classification successes have been achieved in all datasets. When boosting and bagging algorithms were evaluated in terms of running time, it was found that algorithms with the longest working time were boosting algorithms. The main reason for this is that, while learning boosting algorithms from data, weak ones of sequentially developed decision tree models are combined and turned into strong ones. Thus, weights of observations trained in weak classifiers are updated. As this process continues with each renewal, running time of boosting algorithms are prolonged. This is not the case with bagging algorithms. Accordingly, running time of bagging algorithms do not depend on whether the size of the datasets is small or large, and they work quickly in any case. It can be suggested to use bagging algorithms especially in large datasets, but high modeling accuracy depends on the representation power of data and the learning skill of the algorithm. These periods can be extended if the dataset is quite large. Boosting algorithms are suggested to be used due to long working times. It is not easy to determine the best approach in comparing machine learning algorithms. It is possible to link this to a single cause or multiple causes. Increasing sample size is not always an advantage. The reason for this is to reveal data that will enable the algorithm to learn correctly. In general, in such studies using machine learning algorithms, the success of one or more algorithms may not give the same successful results in similar studies. Especially if machine learning algorithms will be used in big data, data preprocessing should not be ignored. In addition to missing data, class noise, and class imbalance problems mentioned in this study, analyzes should be made after problems specific to the dataset are solved. Thus, an increase in the classification success of data is expected.

In this study, it is suggested that researchers should not ignore the data preprocessing process by emphasizing the importance of data preprocessing in ensemble learning classification algorithms which are among machine learning methods. Thus, researchers will contribute significantly to the success of algorithms using data preprocessing methods. In many studies conducted in recent years, attention has been drawn to the data preprocessing process in machine learning methods. In similar studies to be carried out in the future, the structure of data must be carefully assimilated and understood before using machine learning methods. Then, working disciplines of algorithms used should be learned. If, as in this study, it will not be worked with many datasets and modeling algorithms, simulations can be made by adjusting the parameters of models for different situations. Thus, the most successful algorithms can be obtained according to the best parameter values of models. Similarly, another way to increase modeling performance is to choose the most important variables for the model. For this, variable selection methods can be used before modeling. Apart from data preprocessing problems discussed in this study, it is planned to cover issues such as parameter tuning and variable selection in the possible future studies.

Funding

This study is supported by Ege University Scientific Research Projects Coordination Unit. Project Number: TKB-2019-20199. The initial results of the study was presented at the 5th *International Researchers, Statisticians and Young Statisticians Congress (IRSYSC2019)* at October 2019 in Aydın, Turkey.

ORCID

Yüksel Özkan http://orcid.org/0000-0003-0534-1173
Mert Demirarslan http://orcid.org/0000-0001-8848-7340
Aslı Suner http://orcid.org/0000-0002-6872-9901

References

Acock, A. C. 2005. Working with missing values. *Journal of Marriage and Family* 67 (4):1012–28.

Acuña, E, and C. Rodriguez. 2004. The treatment of missing values and its effect on classifier accuracy. In *Classification, clustering, and data mining applications. studies in classification, data analysis, and knowledge organisation*, ed. D. Banks, F. R. McMorris, P. Arabie, W. Gaul. Berlin, Heidelberg: Springer. doi:10.1007/978-3-642-17103-1_60.

Alanis-Tamez, M. D., Y. Villuendas-Rey, and C. Yáñez-Márquez. 2017. Computational intelligence algorithms applied to the pre-diagnosis of chronic diseases. *Research in Computing Science* 138 (1):41–50. Accessed July 6, 2019. https://www.semanticscholar.org/paper/Computational-Intelligence-Algorithms-Applied-to-of-Alanis-Tamez-Villuendas-Rey/055677db623b94f2cc940b94e3c732c111a9677d. doi:10.13053/rcs-138-1-4.

Alcalá-Fdez, J., L. Sánchez, S. García, M. J. del Jesus, S. Ventura, J. M. Garrell, J. Otero, C. Romero, J. Bacardit, V. M. Rivas, et al. 2009. KEEL: A software tool to assess evolutionary algorithms for data mining problems. *Soft Computing* 13 (3):307–18. doi:10.1007/s00500-008-0323-y.

Alghamdi, M., M. Al-Mallah, S. Keteyian, C. Brawner, J. Ehrman, and S. Sakr. 2017. Predicting diabetes mellitus using SMOTE and ensemble machine learning approach: The Henry Ford ExercIse Testing (FIT) project. *PLoS One* 12 (7):e0179805. doi:10.1371/journal.pone.0179805.

Allison, P. 2002. *Missing data*. 1st ed. San Francisco, CA, USA: SAGE Publications, Inc. doi:10.4135/9781412985079.

Amaratunga, D., J. Cabrera, and Y. S. Lee. 2008. Enriched random forests. *Bioinformatics* 24 (18):2010–4. doi:10.1093/bioinformatics/btn356.

Azencott, C. 2018. Machine learning and genomics: Precision medicine vs. patient privacy. *Philosophical Transactions of the Royal Society A Mathematical Physical and Engineering Sciences* 376(20170350):1–13. doi:10.1098/rsta.2017.0350

Bennet, D. A. 2001. How can I deal with missing data in my study? *Australian and New Zealand Journal of Public Health* 25 (5):464–9.

Bradley, A. P. 1997. The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition* 30 (7):1145–59. doi:10.1016/S0031-3203(96)00142-2.

Breiman, L. 2001. Random forests. *Machine Learning* 45 (1):5–32. doi:10.1023/A:1010933404324.

Brodley, C. E, and M. A. Friedl. 1999. Identifying mislabeled training data. *Journal of Artificial Intelligence Research* 11:131–67. doi:10.1613/jair.606.

Buckland, M, and F. Gey. 1994. The relationship between Recall and Precision. *Journal of the American Society for Information Science* 45 (1):12–9. doi:10.1002/(SICI)1097-4571(199401)45:1 < 12::AID-ASI2 > 3.0.CO;2-L.

Chawla, N. V., K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. 2002. SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research* 16:321–57. doi:10.1613/jair.953.

Chen, J. H, and S. M. Asch. 2017. Machine learning and prediction in medicine – Beyond the peak of inflated expectations. *The New England Journal of Medicine* 376 (26):2507–9. doi:10.1056/NEJMp1702071.

Cohen, J. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement* 20 (1):37–46. doi:10.1177/001316446002000104.

Dietterich, T. G. 2000. Ensemble methods in machine learning. In International Workshop on Multiple Classifier Systems, Springer, Berlin, 1–15. doi:10.1007/3-540-45014-9_1.

Dong, Y, and C. Y. J. Peng. 2013. Principled missing data methods for researchers. *SpringerPlus* 2 (1):222. doi:10.1186/2193-1801-2-222.

Fatima, M, and M. Pasha. 2017. Survey of machine learning algorithms for disease diagnostic. *Journal of Intelligent Learning Systems and Applications* 9 (1):1–16. doi:10.4236/jilsa.2017.91001.

Fernández, A., V. López, M. Galar, M. J. del Jesus, and F. Herrera. 2013. Analysing the classification of imbalanced data-sets with multiple classes: Binarization techniques and ad-hoc approaches. *Knowledge-Based Systems* 42: 97–110. doi:10.1016/j.knosys.2013.01.018.

Fluss, R., D. Faraggi, and B. Reiser. 2005. Estimation of the Youden index and its associated cutoff point. *Biometrical Journal. Biometrische Zeitschrift* 47 (4):458–72. doi:10.1002/bimj.200410135.

Folorunso, S. O, and A. B. Adeyemo. 2013. Alleviating classification problem of imbalanced dataset. *African Journal of Computing & ICT* 6 (2):137–44.

Frénay, B, and M. Verleysen. 2014. Classification in the presence of label noise: A survey. *IEEE Transactions on Neural Networks and Learning Systems* 25 (5):845–69. doi:10.1109/TNNLS.2013.2292894.

Friedman, J. H. 2001. Greedy function approximation: A gradient boosting machine. *The Annals of Statistics* 29 (5):1189–232. doi:10.1214/aos/1013203451.

Friedman, J., T. Hastie, and R. Tibshirani. 2000. Additive logistic regression: A statistical view of boosting. *The Annals of Statistics* 28 (2):337–407. doi:10.1214/aos/1016218223.

Gammerman, A. 2010. *Modern machine learning techniques and their applications to medical diagnostics*. Berlin: Springer. doi:10.1007/978-3-642-16239-8_2.

Gunčar, G., M. Kukar, M. Notar, M. Brvar, P. Černelč, M. Notar, and M. Notar. 2018. An application of machine learning to haematological diagnosis. *Scientific Reports* 8 (411):1–12. doi:10.1038/s41598-017-18564-8.

Hajian-Tilaki, K. 2013. Receiver operating characteristic (ROC) curve analysis for medical diagnostic test evaluation. *Caspian Journal of Internal Medicine* 4 (2):627–35.

Ho, T. K. 1998. The random subspace method for constructing decision forests. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 20 (8):832–44. doi:10.1109/34.709601.

Hu, S., Y. Liang, L. Ma, and Y. He. 2009. MSMOTE: Improving classification performance when training data is imbalanced. 2nd International Workshop on Computer Science and Engineering, WCSE 2009, Qingdao, China, vol. 2, 13–7. doi:10.1109/WCSE.2009.756.

Jutel, A. 2011. Classification, disease, and diagnosis. *Perspectives in Biology and Medicine* 54 (2):189–205. doi:10.1353/pbm.2011.0015.

Kang, H. 2013. The prevention and handling of the missing data. *Korean Journal of Anesthesiology* 64 (5):402–6. doi:10.4097/kjae.2013.64.5.402.

Khalilia, M., S. Chakraborty, and M. Popescu. 2011. Predicting disease risks from highly imbalanced data using random forest. *BMC Medical Informatics and Decision Making* 11 (1):51. doi:10.1186/1472-6947-11-51.

Kononenko, I. 2001. Machine learning for medical diagnosis: History, state of the art and perspective. *Artificial Intelligence in Medicine* 23 (1):89–109. doi:10.1016/S0933-3657(01)00077-X.

Kotsiantis, S. 2011. Combining bagging, boosting, rotation forest and random subspace methods. *Artificial Intelligence Review* 35 (3):223–40. doi:10.1007/s10462-010-9192-8.

Krawczyk, B, and B. T. McInnes. 2018. Local ensemble learning from imbalanced and noisy data for word sense disambiguation. *Pattern Recognition* 78:103–19. doi:10.1016/j.patcog.2017.10.028.

Kübler, S., C. Liu, and Z. A. Sayyed. 2017. To use or not to use: Feature selection for sentiment analysis of highly imbalanced data. *Natural Language Engineering* 24:3–37. doi:10.1017/S1351324917000298.

Kumar, G. R., V. S. Kongara, and G. A. Ramachandra. 2013. An efficient ensemble based classification techniques for medical diagnosis. *International Journal of Latest Technology in Engineering, Management & Applied Science* 2 (8): 5–9. Accessed July 5, 2019. https://www.academia.edu/35248870/An_Efficient_Ensemble_Based_Classification_Techniques_for_Medical_Diagnosis.

Lavanya, D, and U. Rani. 2012. Ensemble decision tree classifier for breast cancer data. *International Journal of Information Technology Convergence and Services* 2 (1):17–24. doi:10.5121/ijitcs.2012.2103.

Little, R, and D. Rubin. 1987. *Statistical analysis with missing data*. 1st ed. New Jersey, USA: John Wiley & Sons, Inc. doi:10.4135/9781412985079.

Little, R. J. A., and D. B. Rubin. 1991. Statistical Analysis with Missing Data. *Journal of Educational Statistics*, 16 (2):150–5.

Liu, A. Y. C. 2004. The effect of oversampling and undersampling on classifying imbalanced text datasets (Doctoral dissertation, University of Texas at Austin).

Maimon, O, and L. Rokach. 2005. *Data mining and knowledge discovery handbook*. Berlin: Springer-Verlag.

McHugh, M. L. 2012. Interrater reliability: The kappa statistic. *Biochemia Medica* 22 (3):276–82. doi:10.11613/BM.2012.031.

Mislevy, R., R. Little, and D. Rubin. 1991. Statistical analysis with missing data. *Journal of Educational Statistics* 16 (2):150–5. doi:10.2307/1165119.

Moon, H., H. Ahn, R. L. Kodell, S. Baek, C. J. Lin, and J. J. Chen. 2007. Ensemble methods for classification of patients for personalized medicine with high-dimensional data. *Artificial Intelligence in Medicine* 41 (3):197–207. doi:10.1016/J.ARTMED.2007.07.003.

Morales, P., J. Luengo, L. P. F. Garcia, A. C. Lorena, A. C. P. L. F. Carvalho, and F. Herrera. 2017. The NoiseFiltersR package: Label noise preprocessing in R. *The R Journal* 9 (1):219–28. doi:10.32614/RJ-2017-027.

Obermeyer, Z, and E. J. Emanuel. 2016. Predicting the future – Big data, machine learning, and clinical medicine. *The New England Journal of Medicine* 375 (13):1216–9. doi:10.1056/NEJMp1606181.

Sáez, J. A., B. Krawczyk, and M. Woźniak. 2016. On the influence of class noise in medical data classification: Treatment using noise filtering methods. *Applied Artificial Intelligence* 30 (6):590–609. doi:10.1080/08839514.2016.1193719.

Salcedo-Bernal, A., M. P. Villamil-Giraldo, and A. D. Moreno-Barbosa. 2016. Clinical data analysis: An opportunity to compare machine learning methods. *Procedia Computer Science* 100:731–8. doi:10.1016/j.procs.2016.09.218.

Salzberg, S. L. 1994. C4.5: Programs for machine learning. *Machine Learning* 16 (3):235–40. doi:10.1007/BF00993309.

Schafer, J. L. 1999. Multiple imputation: A primer. *Statistical Methods in Medical Research* 8 (1):3–15. doi:10.1177/096228029900800102.

Shah, P., F. Kendall, S. Khozin, R. Goosen, J. Hu, J. Laramie, M. Ringel, and N. Schork. 2019. Artificial intelligence and machine learning in clinical development: A translational perspective. *npj Digital Medicine* 2 (1):1–5. doi:10.1038/s41746-019-0148-3.

Sinharay, S., H. S. Stern, and D. Russell. 2001. The use of multiple imputation for the analysis of missing data. *Psychological Methods* 6 (4):317–29. doi:10.1037/1082-989X.6.4.317.

Sutton, C. D. 2005. Classification and regression trees, bagging, and boosting. In *Handbook of statistics*, ed. R. C. Rao and A. S. R. S. Rao, 24th ed., 303–30. San Francisco, CA, USA: Elsevier. doi:10.1016/S0169-7161(04)24011-1.

Swets, J. 1988. Measuring the accuracy of diagnostic systems. *Science* 240 (4857):1285–93. doi:10.1126/science.3287615.

Thottakkara, P., T. Ozrazgat-Baslanti, B. Hupf, P. Rashidi, P. Pardalos, P. Momcilovic, and A. Bihorac. 2016. Application of machine learning techniques to high-dimensional clinical data to forecast postoperative complications. *PLoS One* 11 (5):e0155705. doi:10.1371/journal.pone.0155705.

UCI. 2019. UC Irvine Machine Learning Repository. Center for Machine Learning and Intelligent Systems. Accessed July 27, 2019. https://archive.ics.uci.edu/ml/index.php.

Verma, B, and Z. S. Hassan. 2011. Hybrid ensemble approach for classification. *Applied Intelligence* 34 (2):258–78. doi:10.1007/s10489-009-0194-7.

Wang, R. Y., V. C. Storey, and C. P. Firth. 1995. A framework for analysis of data quality research. *IEEE Transactions on Knowledge and Data Engineering* 7 (4):623–40. doi:10.1109/69.404034.

Watson, P. F, and A. Petrie. 2010. Method agreement analysis: A review of correct methodology. *Theriogenology* 73 (9):1167–79. doi:10.1016/J.THERIOGENOLOGY.2010.01.003.

Witten, I. H., E. Frank, and A. Hall. 2011. *Data mining: Practical machine learning tools and techniques*, Data Mining (Third Edit, Vol. 277). San Francisco, CA, USA: Elsevier. doi:10.1002/1521-3773(20010316)40:6<9823::AID-ANIE9823>3.3.CO;2-C.

Wu, X, and X. Zhu. 2008. Mining with noise knowledge: Error-aware data mining. *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans* 38 (4):917–32. doi:10.1109/TSMCA.2008.923034.

Wu, Z., W. Lin, and Y. Ji. 2018. An integrated ensemble learning model for imbalanced fault diagnostics and prognostics. *IEEE Access*. 6:8394–402. doi:10.1109/ACCESS.2018.2807121.

Xu, B., J. Z. Huang, G. Williams, Q. Wang, and Y. Ye. 2012. Classifying very high-dimensional data with random forests built from small subspaces. *International Journal of Data Warehousing and Mining* 8 (2):44–63. doi:10.4018/jdwm.2012040103.

Yang, P., J. Y. H. Yang, B. Zhou, and A. Zomaya. 2010. A review of ensemble methods in bioinformatics. *Current Bioinformatics* 5 (4):296–308. doi:10.2174/157489310794072508.

Youden, W. J. 1950. Index for rating diagnostic tests. *Cancer* 3 (1):32–5. doi:10.1002/1097-0142(1950)3:1<32::AID-CNCR2820030106>3.0.CO;2-3.

Zhang, Y., B. Liu, J. Cai, and S. Zhang. 2017. Ensemble weighted extreme learning machine for imbalanced data classification based on differential evolution. *Neural Computing and Applications* 28 (s1):259–67. doi:10.1007/s00521-016-2342-4.

Zhang, Z., Y. Zhao, A. Canes, D. Steinberg, O. Lyashevska, and written on behalf of AME Big-Data Clinical Trial Collaborative Group. 2019. Predictive analytics with gradient boosting in clinical medicine. *Annals of Translational Medicine* 7 (7):152–9. doi:10.21037/atm.2019.03.29.

Zhu, X, and X. Wu. 2004. Class noise vs. attribute noise: A quantitative study. *Artificial Intelligence Review* 22 (3):177–210. doi:10.1007/s10462-004-0751-8.