Available online at www.medicinescience.org

REVIEW ARTICLE

Medicine Science International Medical Journal

Medicine Science 2022;11(2):924-33

An investigation of ensemble learning methods in classification problems and an application on non-small-cell lung cancer data

Mehmet Kivrak[1], Cemil Colak[2]

[1]Recep Tayyip Erdogan University Faculty of Medicine, Department of Biostatistics and Medical Informatics, Rize, Turkey
[2]Inonu University, Faculty of Medicine, Department of Biostatistics and Medical Informatics, Malatya, Turkey

Received 14 October 2021; Accepted 13 February 2022
Available online 20.03.2022 with doi: 10.5455/medscience.2021.10.339

Abstract

This study aims to classify NSCLC death status and consists of patient records of 24 variables created by the open-source dataset of the cancer data site. Besides, basic classifiers such as SMO (Sequential Minimal Optimization), K-NN (K-Nearest Neighbor), random forest, and XGBoost (Extreme Gradient Boosting), which are machine learning methods, and their performances, and voting, bagging, boosting, and stacking methods from ensemble learning methods were used. Performance evaluation of models was compared in terms of accuracy, specificity, sensitivity, precision, and Roc curve. The basic classifier performances of random forest, SMO, K-NN, and XGBoost classifiers, their performances in the bagging ensemble learning method, and their performances in the boosting ensemble learning method are evaluated. In addition, Model 1 (random forest + SMO), Model 2 (XGBoost + K-NN), Model 3 (random forest + K-NN), Model 4 (XGBoost+SMO), Model 5 (SMO+K-NN + random forest), Model 6 (SMO+K-NN+XGBoost) and Model 7 (SMO+K-NN + random forest + XGBoost) the performances of in different metrics were expressed. The boosting ensemble learning method, which provides the maximum classification performance with XGBoost, achieved a 0.982 accuracy value, 0.971 sensitivity value, 0.989 precision value, 0.989 specificity value, and 0.998 ROC curve. It is recommended to use ensemble learning methods for classification problems in patients with a high prevalence of cancer to achieve successful results.

Keywords: NSCLS, machine learning, ensemble learning

Introduction

Lung cancer is one of the most common types of cancer in our country as well as all over the world [1]. Smoking and passive smoking, air pollution, gender, occupational reasons, genetic factors, chronic lung disease, and radiotherapy are the main risk factors. The most common type of NSCLC [2]. In NSCLC cases, factors such as age, gender, weight loss, performance status, tumor stage are the main prognostic factors. In NSCLC cases, factors such as age, gender, weight loss, performance status, tumor stage are the main prognostic factors. The 5-year survival rate is known as 67 % for stage IA, 55 % for stage IIA, and 23 % for stage IIIA, especially in patients who undergo surgery. In stage IIIB cases, this rate is only 3-7 % [3]. Positron emission tomography (PET), magnetic resonance (MR), and spiral thorax tomography (Thorax CT) are the imaging methods used for tumor diagnosis and staging in lung cancer [2]. In the pathological diagnosis of cancer, methods for primary tumor such as bronchoscopy, transthoracic needle aspiration, and sputum cytology are used, as well as methods for metastasis such as thoracentesis, closed pleural biopsy, thoracoscopy, lymph node, and skin biopsy [4]. Tumor characteristics, lymph node, and metastasis (TNM) systems are used in NSCLC staging. T defines primary tumor, N regional lymph nodes, and M distant metastasis [5]. Delays in diagnosis or treatment are an important issue for patients with NSCLC. These delays are known to have a serious impact on tumor stage and prognosis [3].

For NSCLC patients, the right treatment methods are determined, and a personalized treatment process is created through decision support systems developed through multi-factor decision-based radiation oncology, artificial intelligence, and machine learning algorithms. Thus, excessive or less than necessary treatment is avoided [6].

***Corresponding Author:** Mehmet Kivrak, Recep Tayyip Erdogan University Faculty of Medicine, Department of Biostatistics and Medical Informatics, Rize, Turkey E-mail: mehmetkivrak83@gmail.com

This study aims to classify NSCLC death status and consists of patient records of 24 variables created by the open-source dataset of the cancer data site. Besides, basic classifiers such as SMO, K-NN, random forest, and XGBoost, which are machine learning methods, and their performances, and voting, bagging, boosting, and stacking methods from ensemble learning methods were used. Performance evaluation of models was compared in terms of accuracy, specificity, sensitivity, precision, and Roc curve. In trying to determine the most successful model in estimating the death status of NSCLS disease, the results were presented comparatively.

Materials and Methods

Dataset

The dataset used for the analysis was obtained from https://www.cancerdata.org/resource/doi:10.17195/candat.2016.04.1 [6]. The data set was examined in terms of prognostic and clinical values of blood biomarkers related to hypoxia, inflammation, immune response, and tumor burden in NSCLC. The data set includes a total of 181 inoperable stage I-IIIB NSCLC patient records. Approximately 55.2 % of the data set consisted of patients receiving radiotherapy or chemotherapy, while 44.8 % consisted of patients receiving radical treatment. Ethics committee approval was not required as the data was obtained from open-source access. A detailed explanation of the variables is given in Table 1.

Table 1. A Detailed Explanation of the Variables

Variables	Explanation
Status	Target (0: Death,1: Alive)
Age	age
Gender	Gender (1=male, 0=female)
Survival	Expected Life Time
Stage	Disease Level (I-IIIB)
Histology	Disease History (Adeno / NOS / SCC / Undefined)
WHO-PS	World Health Organization Performance Status
FEV1s %	Breathing Tests: Difficult Vital Capacity Maneuver
Lymph nodes	Positive lymph nodes count identified by PET scan
RT Protocol	Standard beam radiation therapy
Total dose(1st)	A total dose of 1 administered in Chemo-Radiotherapy
Total dose(2nd)	A total dose of 2 administered in Chemo-Radiotherapy
GTV	Some of the primary tumor and metastatic lymph nodes
OPN	Osteopontin
CA-9	Carbonic Anhydrase-9
IL-6	Interleukin-6
IL-8	Interleukin-8
CRP	C-Reactive Protein
CEA	Tumor burden carcinoembryonic antigen
Cyfra 21-1	Cytokeratin 21-1
α2M	Alpha-2-Macroglobulin
sIL2R	Soluble in serum IL2 Respector
TLR-4	Toll-like Respector 4
VEGF	Vascular Endothelial Growth Factor

Knowledge Discovery in Databases (KDD)

In the process of KDD; data selection (NSCLS dataset), data preprocessing (extreme and missing value analyses), data transformation (normalization, etc.), data mining and evaluation, and interpretation of the results were performed.

Basic Classification Method

The most used data mining methods on the analyzed datasets have been applied for the classification of NSCLC death status. Performance data obtained by using random forest, SMO, K-NN, XGBoost, and ensemble learning classification methods were comparatively presented to the data sets.

Random Forest

The random forest algorithm is a forest classifier consisting of many decision trees, and classification or regression trees can be established and clustered with this method. If the classification variable is categorical, classification-based trees are established, if continuous, regression-based trees are established. In this method, trees are created with the classification and regression trees (CART) algorithm, and the trees are not pruned [7]. The CART algorithm uses the concept of knowledge gain and entropy to optimally separate nodes. When there are k probabilities for X variable (attribute) P_1,P_2,P_3,P_k respectively, entropy for variable X is given in the equation below [8].

$$Entropy = H(X) = -\sum_{j=1}^{k} p_j log_2(p_j) \quad (1)$$

When the target attribute of the sub-clusters $T_1,T_2,T_3,..T_{k-}$ in the training set is subdivided into sub-compartments, the weighted average of the information required to determine the class of each T is given as the weighted sum of entropies.

$$H_S(T) = \sum_{i=1}^{k} p_i H_S(T_i) \quad (2)$$

Information gain is calculated to perform the separation process. The random forest algorithm realizes the optimal separation process by determining the most information gain in each decision node. Information gain is given in the equation below [9,10].

$$IG\ (S) = H(T) - H_S(T) \quad (3)$$

Sequential Minimal Optimization (SMO)

SMO classifier is essentially a kind of SVM algorithm. It makes model estimation by running the support vector machine at each stage of the smallest optimization problem with two Lagrange multipliers [11]. Sequential Minimum Optimization (SMO) improves the training of the SVM classifier using polynomial nuclei. This generally replaces all missing values and converts the nominal properties to binary values [12]. To find a decision boundary between the two classes, SVM tries to maximize the gap between classes, choosing linear separations in a property area. Classification of the k-core function points in space x_i is y_i, which varies between -1 and +1. If x'is a point with an unknown classification, the prediction classification y'is as in the equation below.

$$y' = Sign(\sum_{i=1}^{n} \alpha_i y_i K(X_i, X') + d \quad (4)$$

In the equation, K; core function, n; support vector number, α; adjustable weight and d are defined as bias. The classification process is linear in the number of support vectors [13].

K-Nearest Neighbor (K-NN)

The K-NN classifier is a generally used machine learning method. KNN, like the classifiers of the supervised learning paradigm, requires P≤X training data, in which h (x) contains data elements xЄX where known. The classifier is predicted to wager the brand-new sample classification tag utilizing the P understanding. K-NN is greatly utilized in machine learning approaches because of its good efficiency as well as its easy use [14].

Extreme Gradient Boosting (XGBoost)

Developed with the aid of Chen and Guestrin, the XGBoost classifier is a scalable machine learning process in an end-to-end tree approach used in classification and regression issues.

$$(\emptyset) = \sum ii\ (yyii, yyii) + \sum kk\ \acute{\Omega}(ffkk) \quad (5)$$

The hyperparameters of the XGBoost mannequin are the maximum depth, number of divisions, and learning pace set through the grid search optimization algorithm [15].

Ensemble Learning

Ensemble learning methods essentially aim to achieve the most accurate result by combining different methods. It can also be applied successfully in various machine learning systems such as feature extraction, error correction, unstable data, learning to deviate in non-stationary distributions, and confidence estimation. "bagging, boosting, voting and stacking" are the most used algorithms for the training of ensemble classifiers. The most common unification rule used to combine individual classifiers is majority voting. The choice of the W_c class with the majority vote is as inequality [16].

$$\sum_{t=1}^{T} d_{t,c} = max_c \sum_{t=1}^{T} d_{t,c} \quad (6)$$

Boosting

Boosting algorithm is an approach to obtaining strong classifiers from weak classifiers with low training error. Boosting combines a community of weak classifiers using the simple majority vote method [17] (Figure 1).

Bagging

Bagging is short for "bootstrap aggregation". It is an algorithm created by combining the classifiers obtained by the resampling method. In this method, new classifiers are created by taking samples from the data set to replace the ensemble learning algorithm is developed from these classifiers [18] (Figure 2).

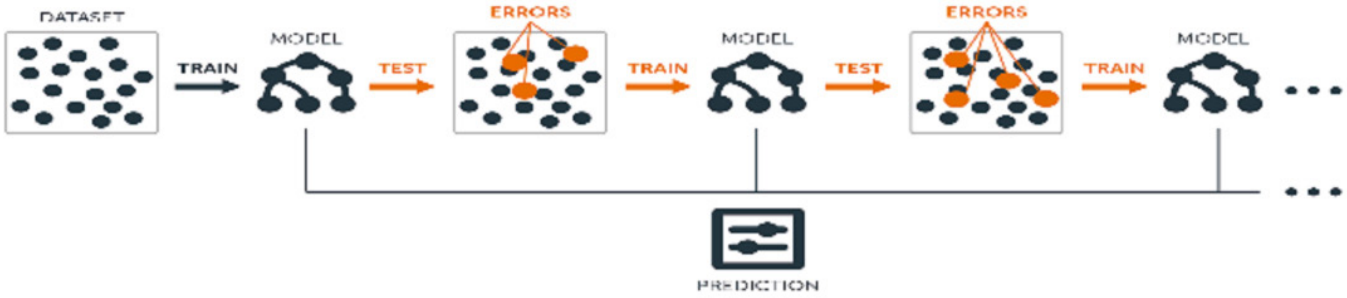

Figure 1. Boosting Algorithm

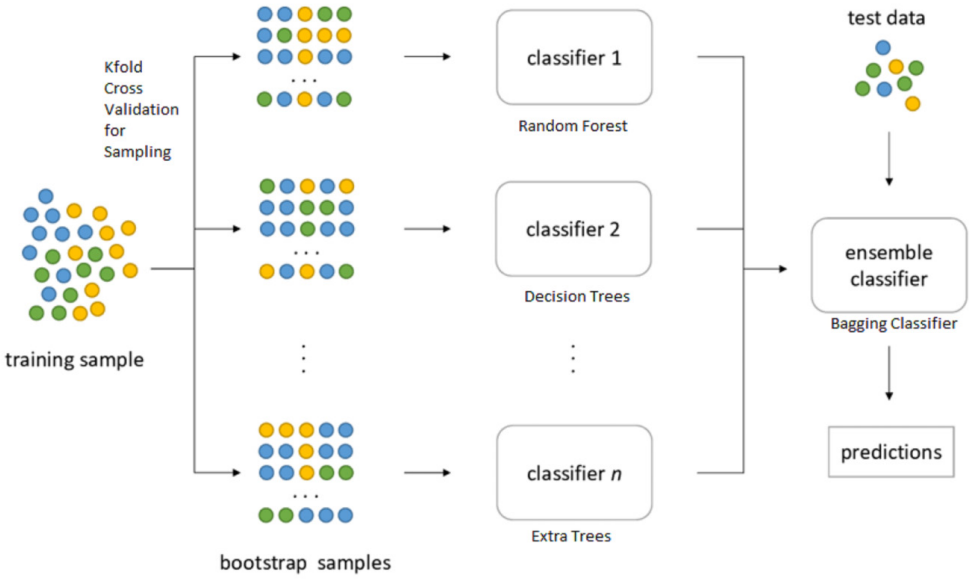

Figure 2. Bagging Algorithm

Voting

The Voting approach is the most greatly used to become a member of the system for nominal results [19]. It is among the simplest ways to combine predictions from more than one desktop studying algorithm. With this system, a combo approach that includes exclusive agencies trained and evaluated in parallel is carried out to improve from the extraordinary facets of every classifier [7] (Figure 3).

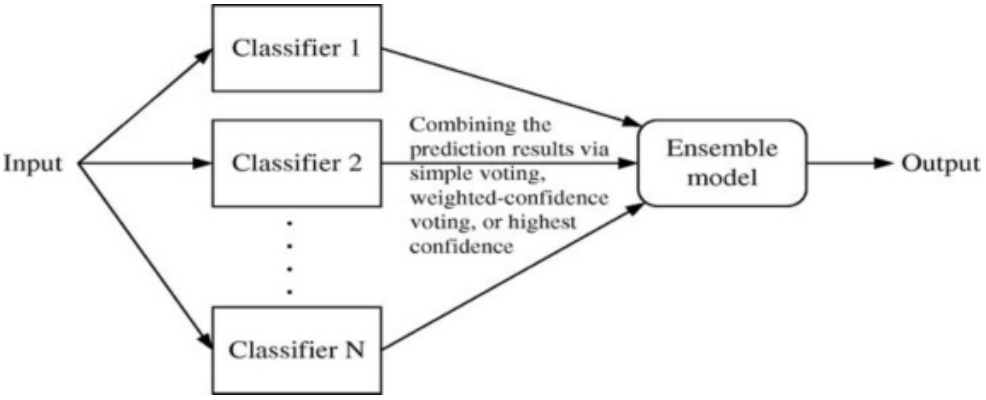

Figure 3. Voting Algorithm

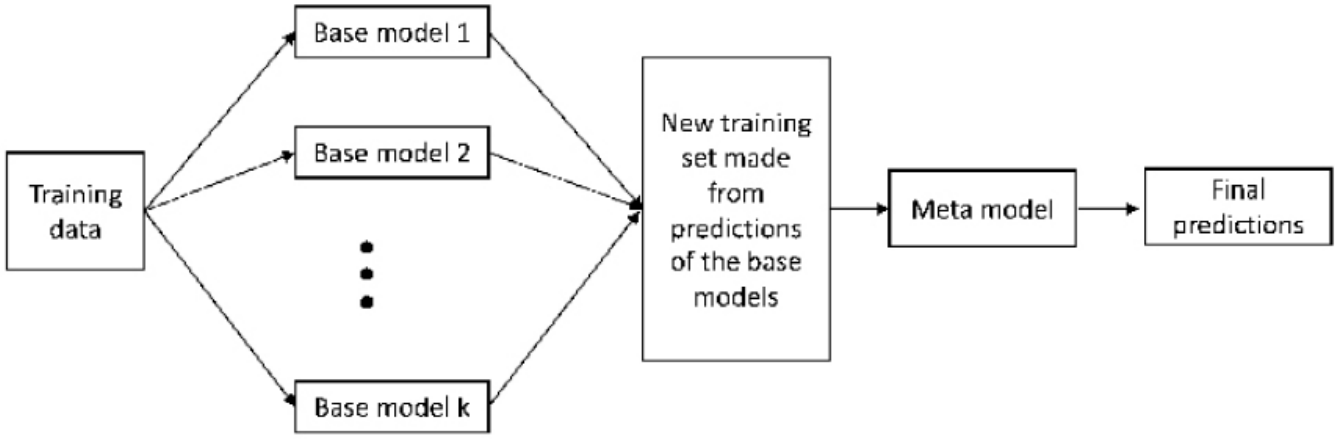

Figure 4. Stacking Algorithm

Performance Metrics

Accuracy (AC) is defined as the division of values incompatible eyes by the total number of observations and is indicated by equation 7.

$$AC = \frac{TP+TN}{TP+TN+FN+FP} \quad (7)$$

Sensitivity is the capability of the test to differentiate patients from actual sufferers and is indicated by using equation 8.

$$Sensitivity = \frac{TP}{TP+FP} \quad (8)$$

Specificity is the ability of the experiment to distinguish robots from actual robots and is indicated by using equations 9 [9,10].

Stacking

It is a procedure that accepts classifier estimates as input for the meta classifier to provide high accuracy efficiency with the estimates of exceptional varieties of classifiers. In this procedure, even as estimates are obtained with exceptional classifiers from the training information set, then the estimates got are combined within the meta classifier and the ensemble finding out mannequin is estimated [20] (Figure 4).

$$Specificity = \frac{TN}{TN+FN} \quad (9)$$

Precision gives the likelihood of those who are sick and is indicated by equation 10.

$$Precision = \frac{TP}{TP+TN} \quad (10)$$

The ROC Curve process is used to evaluate the efficiency of diagnostic tests used to diagnose a disorder and to check the reduce-off features. The ROC curve has sensitivity values on the vertical axis and 1-specificity values on the horizontal axis [21].

Results

Statistical Analysis

Quantitative data had been summarized because of the arithmetic

means with standard deviation, median with min and max values, and qualitative knowledge as numbers with the aid of percentage. After the suitability of the data to more than one normal distribution, the change between the companies is commonly distributed organizations used to be examined by t-experiment in independent samples and the Mann-Whitney u-scan for variables that didn't exhibit average distribution. The frequency distribution relationship between categorical variables was evaluated using the chi-square test and fisher's exact chi-square test. For statistical analysis, IBM SPSS version 22 [22], RStudio version 1.1.463 [23], and Rapid Miner Studio version 8.1.001 [24] were used.

Descriptive Statistics

The dataset consisted of 76 (46%) patients who died of lung cancer and 91 (54 %) patients with lung cancer. State distributions are given in Figure 5. Looking at the distribution of lung cancer by gender in Table 2, 54% of women died due to cancer, while 61% survived. While 44% of men died of cancer, 56% survived. There was no gender difference in death and survival rates due to lung cancer (p-value >0.05). Looking at the distribution of lung cancer according to the lymph nodes variable, 57% of the very low ones died while 43% survived, 56% of the low ones died while 44% survived, 53% of the moderate ones died and 47% survived. 59% of the high died while 41% survived, 30% of the very high died and 70% survived. There was no difference in mortality and survival rates due to lung cancer according to the lymph nodes variable (p-value >0.05). According to the WHO-PS variable of lung cancer, 65% of the active ones died while 35% survived, 41% of the partially limited ones died, 59% survived, 30% of the partial actives died and 70% survived., 25% of the limited died, 75% survived, and 100% of the inactive survived. There was no difference in mortality and survival rates due to lung cancer according to the WHO-PS variable (p-value >0.05) (Figure 5) (Table 2).

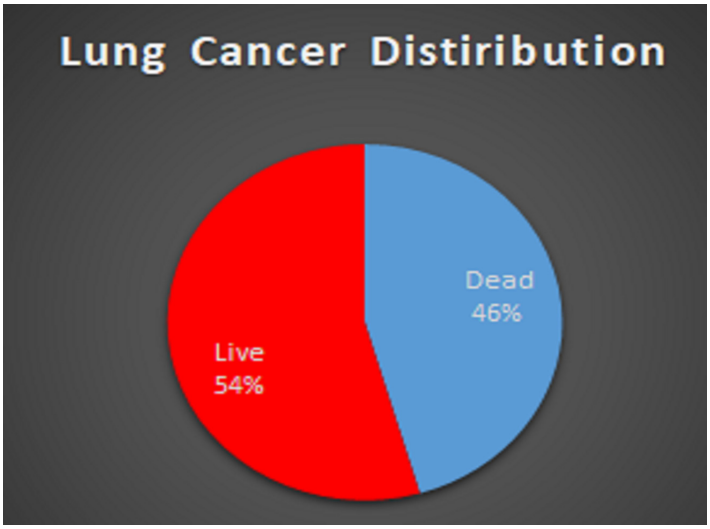

Figure 5. Distributions of Lung Cancer

Table 2. Lung Cancer by Categorical Variables

Gender (Count (Percent))	Dead	Alive	Total	Chi-square	p-value
Female	52 (54 %)	61 (46 %)	113 (100 %)	0.036	0.849
Male	24 (44 %)	30 (56 %)	54 (100 %)		
Total	76 (46 %)	91 (54 %)	167 (100 %)		
Lymphnodes (Count (Percent))					
Very Low	18 (57 %)	24 (43 %)	42 (100 %)	0.018	0.671
Low	15 (56 %)	12 (44 %)	27 (100 %)		
Midle	17 (53 %)	15 (41 %)	32 (100 %)		
High	13 (59 %)	9 (41 %)	22 (100 %)		
Very High	13 (30 %)	31 (70 %)	44 (100 %)		
Total	76 (46 %)	91 (54 %)	167 (100 %)		
WHO-PS (Count (Percent))					
Active	30 (65 %)	16 (35 %)	46 (100 %)	0.021	0.796
Partial Restricted	39 (41 %)	57 (59 %)	96 (100 %)		
Partial Active	6 (30 %)	14 (70 %)	20 (100 %)		
Restricted	1 (25 %)	3 (75 %)	4 (100 %)		
Not Active	0 (0.0 %)	1 (100 %)	1 (100 %)		
Total	76 (46 %)	91 (54 %)	167 (100 %)		

According to the comparison results in Table 3, the difference in terms of OPN and SIL2R variables of patients who died and survived due to lung cancer was statistically significant (p-value <0.05). OPN and SIL2R variants were seen at higher values in survivors. In terms of other variables, there was no statistically significant difference between patients who died and survived due to lung cancer (p-value >0.05) (Table 3).

Continuous Variables

According to the comparison results in Table 4, the differences in age and FEV1s variables of patients who died and survived due to lung cancer were not statistically significant (p-value >0.05). In terms of the total dose (2) variable, the difference between patients who died and survived due to lung cancer was statistically significant (p-value<0.05).

Table 3. Lung Cancer Status By Non-Normally Distributed

Variables($\bar{X}$±SS/Min-Max)	Dead	Alive	U-statistics	p-value
Age	68.3±5.6/58-78	51.8±3.28/43-81	2.78	0.493
Fev1s	73.1±197.3/48-985	74.7±316.6/58-1546	3.56	0.514
TotalDose	11.1±18.7/0.5-181	17.5±14.2/2-65.1	0.104	0.035
***:Mann Whitney U Test**				

Table 4. Lung Cancer Statusby Normally Distributed Continuous Variables

Variables($\bar{X}$±SS/Min-Max)	Dead	Alive	t	p-value
OPN	102.8±51.6/16-489	128.2±56.3/43-298	3.042	0.003
CA9	291.7±197.3/48-985	359.5±316.6/58-1546	1.622	0.107
IL8	17.1±18.7/0.5-181	18.6±14.2/2-65.1	0.589	0.556
CEA	23.5±85.3/0.7-635	18.2±40.95/1.5-279	-0.517	0.606
SIL2R	5378.4±2220.8/1121-10600	6768.6±3353.2/2278-20000	3.091	0.002
TLR4	7.12±3.7/1.9-17.7	7.26±4.63/1.4-29.9	0.21	0.834
VEGF	101.5±82.3/17.8-504.8	117.2±98.6/20.9-465.4	1.01	0.272

Data Mining

In this study, basic classifiers such as SMO, K-NN, random forest, and XGBoost, which are machine learning methods, and their performances, and the performances of different classifiers using voting, bagging, boosting, and stacking methods from ensemble learning methods were examined.

Model Development

The 10-fold cross-validation procedure used to be used in the performance development of all classifier methods to verify the excellent of the items. Cross-validation is the re-sampling method used to evaluate machine learning models in a data pattern. The method has a single parameter named k that expresses the number of groups to split a given data sample. In 10-fold cross-validation, the items are proficient and established ten distinctive occasions, after which, imply performance metrics (i.E., accuracy, precision, etc) are estimated on the finish of the process [25].

Evaluation of the Models

In the classification of death status in NSCLC patients, dissimilar classifiers and different ensemble learning methods were applied to determine the model that provides the best performance. The performance metrics are given in Table 5. In the table, basic classifier performances of random forest, SMO, K-NN, and XGBoost classifiers, their performances in the bagging ensemble learning method, and their performances in the boosting ensemble learning method are evaluated. In addition, Model 1 (random forest + SMO), Model 2 (XGBoost + K-NN), Model 3 (random forest + K-NN), Model 4 (XGBoost + SMO), Model 5 (SMO + K-NN + random forest), Model 6 (SMO + K-NN + XGBoost) and Model 7 (SMO + K-NN + random forest + XGBoost) the performances of in different metrics were expressed.

General Assessment

In line with the general assessment, the boosting ensemble learning process supplied the highest efficiency in the metrics of accuracy, sensitivity, precision, specificity, and the field beneath the ROC curve. The boosting approach, which supplies the easiest classification performance with XGBoost, finished a 0.982 accuracy price, 0.971 sensitivity price, 0.989 precision value, 0.989 specificity value, and 0.998 ROC curve. The pseudo-code of the XGBoost classifier is shown in figure 6 classification performances of all classifiers and ensemble methods are expressed in figure 7. In Table 6, the importance levels of the variables of the XGBoost classifier in the boosting algorithm, which gives the best results in the ensemble learning method, and the XGBoost classifier, which gives the best results in individual classifiers, are shown in the classification of NSCLC mortality. In the XGBoost base classifier, the variables GTV (29.5%), CEA (V10.4%), WHOPS (10.0%), age (8.5%), and OPN (7.2%) provide the highest importance, while lymph nodes (2.5%) and SIL2R (2.9%).) were the variables with the lowest significance. In the XGBoost classifier in the Boosting ensemble learning algorithm, the variables GTV (11.2%), CEA (6.2%), WHOPS (5.2%), Total Dose 2 (5.2%) have the highest importance, while FEV1s (2.0%) and lymph nodes (2.1%) have the most. There were variables with low significance.

Table 5. Model Performance Metrics

Basic **Classifiers**	**Accuracy (%)**	**Sensitivity (%)**	**Specificity (%)**	**Precision (%)**	**AUC (%)**
Random Forest	88.9	81.6	94.4	85.6	98.9
SMO	91.5	89.5	93.4	91.9	96.9
K-NN	59.8	54.8	63.8	54.7	84.6
XGBoost	97.5	97.1	97.8	97.6	98.6
Ensemble (Bagging)					
Random Forest	98.2	97.1	98.9	98.9	99.4
SMO	94.0	90.9	96.7	95.9	98.0
K-NN	59.8	54.8	63.8	54.7	71.6
XGBoost	97.5	97.1	97.8	97.6	99.4
Ensemble (Boosting)					
Random Forest	98.1	97.1	98.9	98.6	99.8
SMO	92.2	89.6	94.6	93.6	97.4
K-NN	59.8	54.8	63.8	54.7	84.6
XGBoost	**98.2**	**97.1**	**98.9**	**98.9**	**99.8**
Ensemble (Voting)					
Model 1	94.5	88.0	100.0	100.0	99.7
Model 2	78.4	53.4	98.9	96.7	98.4
Model 3	78.4	53.4	98.9	97.5	98.6
Model 4	93.9	88.0	98.9	98.8	99.7
Model 5	95.1	94.6	95.6	94.6	98.8
Model 6	95.7	94.6	96.7	96.1	98.1
Model 7	96.3	93.2	98.9	98.8	99.0
Ensemble (Stacking)					
Model 1	94.6	93.4	95.6	95.0	97.5
Model 2	75.6	68.8	81.1	74.3	86.7
Model 3	59.8	54.8	63.8	54.7	75.7
Model 4	96.4	95.9	96.7	96.5	96.2
Model 5	59.8	54.8	63.8	54.7	80.4
Model 6	75.6	68.8	81.1	74.3	89.0
Model 7	75.6	68.8	81.1	74.3	88.0

Algorithm 1 GBDT Training

Input: a training set $\{X_i, y_i\}_{i=1}^{N}$

Output: a model $M_T(X)$, which is based on T decision trees with their corresponding weights, and T augmentation functions

1: **function** TRAIN($\{X_i, y_i\}_{i=1}^{N}$)
2: Initialize $M_0(X)$ as $\underset{\rho}{argmin} \sum_{i=1}^{N} \mathcal{L}(y_i, \rho)$ ▷ $\mathcal{L}$ is the loss
3: Initialize $\{\tilde{y}_i\}_{i=1}^{N}$ as $\{y_i\}_{i=1}^{N}$
4: **for** $t \leftarrow 1$ to T **do**
5: Update the targets $\{\tilde{y}_i\}_{i=1}^{N}$ using the last gradient of $\mathcal{L}$: $-\left[\frac{\partial \mathcal{L}(y_i, M(X_i))}{\partial M(X_i)}\right]_{M(X)=M_{t-1}(X)}$
6: Train a decision tree $\mathcal{D}_t$, using $\{X_i, \tilde{y}_i\}_{i=1}^{N}$
7: Set ρ_t, the weight of the new model, to be $\underset{\rho_t}{argmin} \sum_{i=1}^{N} \mathcal{L}(y_i, M_{t-1}(X_i) + \rho_t \cdot \mathcal{D}_t(X_i))$
8: $M_t(X) \leftarrow M_{t-1}(X) + \rho_t \cdot \mathcal{D}_t(X))$
9: **return** $M_T(X)$

Figure 6. Pseudo Code of XGBoost

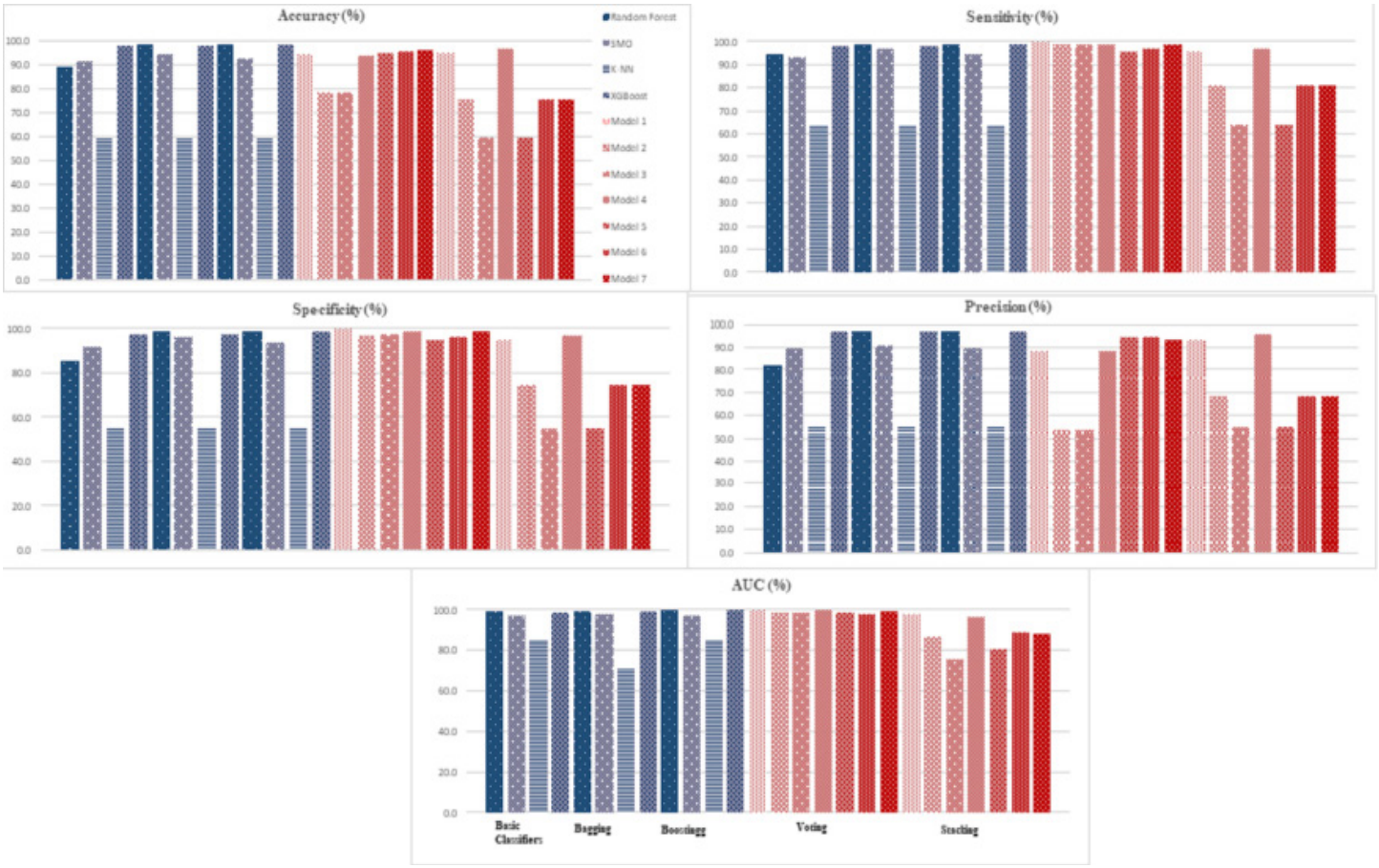

Figure 7. Models Performance Metrics

Table 6. Variable Significance of The Best Models

	Model			
	XGBoost (Basic)		Boosting /XGBoost (Ensemble)	
Variable	Relative Importance	Percent (%)	Relative Importance	Percent (%)
GTV	29.5	29.4	11.2	11.2
CEA	10.4	10.4	6.2	6.2
WHOPS	10.0	9.9	5.2	5.3
Age	8.5	8.5	2.3	2.3
OPN	7.2	7.2	4.7	4.7
Total_Dose (2)	6.6	6.6	5.2	5.1
FEV1s	6.4	6.4	2.0	2.0
CA9	4.9	5.0	3.2	3.2
VEGF	4.6	4.5	4.1	4.1
IL8	3.6	3.6	2.6	2.6
TLR4	3.3	3.3	3.3	3.3
SIL2R	2.9	2.8	4.9	5.0
Lymphnodes	2.5	2.5	2.1	2.1

Discussion

Lung cancer is one of the most common types of cancer and begins to grow out of control by occurring when lung cells become abnormal [26]. Cancer cells form a tumor as their numbers increase and spread to other parts of the body [27]. NSCLC accounts for more than 80% of all lung cancer cases [28].

In recent years, the use of machine learning and artificial intelligence methods has become a wide application area, especially in the early diagnosis and treatment of cancer diseases. Machine learning and artificial intelligence methods have been used frequently to predict cancer diseases and have achieved high classification performance [29,30]. These methods, which are used in many health fields today, attach importance to the automatic acquisition of hidden patterns in databases and the efficiency of the data analysis process [31].

The ensemble learning method is an algorithm that combines several basic models to produce an optimal prediction model and trains data by modeling multiple classifiers rather than training the training dataset through classifiers [32].

This study aims to investigate and improve the usability of artificial intelligence-based ensemble learning methods in medicine. Using the data set including various clinical variables to classify NSCLC death status, NSCLC variable; The NSCLC death status classification performance of ensemble learning methods will be examined, and the best model will be determined, with the response /output /target and measurements in the data set and other factors being explanatory/predictive/independent variables.

In a study in which the same data set was used in the literature, Nishio et al. Utilized SVM and XGBoost classifiers in their studies on the computer-aided diagnosis of lung nodules using XGBoost and Bayes optimization. According to the performance metrics, the XGBoost classifier provided the highest success [33]. Faisal et al, of their work on the comparison of machine learning classifiers and communities for early-stage prediction of lung cancer, a variety of classifier performances, together with gradient-boosted tree (GBT), SVM, C4.5, decision tree, multi-layer perceptron (MLP), and NB. And also, with famous communities such as random wooded areas (bagging) and Majority voting. According to the performance evaluations, it was observed that GBT performed better than all other individual and group classifiers [34]. As a result, it is seen that the classification and estimation studies in NSCLC cases according to the ensemble learning method are not at a sufficient level. Exceptionally in this very usual form of cancer, using ensemble learning ways will furnish excessive classification efficiency thanks to the truth that the data units include a tremendous quantity of variables and the sample in the training data set will also be discovered with many samples.

As a result

The ensemble learning XGBoost classifier has provided the highest classification performance boosting. Boosting method successfully predicted death status in NSCLC patients. The boosting method evaluated 98.2 % accuracy in classification and 99.8 % area values under the ROC curve.

While the boosting method provided the best classifier performance, the bagging method also yielded successful results. The random forest classifier provided a high-performance metric, especially in the bagging method. It is recommended to use ensemble learning methods for classification problems in patients with a high prevalence of cancer to achieve successful results.

It is suggested that new ensemble learning methods such as stacking will be more successful especially with the correct selection of basic classifiers.

Conclusion

Conflict of interests
The authors declare that they have no known competing financialinterestsor personal relationships that could have appeared to influence the work reported in this paper.

Financial Disclosure
The authors declare that they have received no financial support for the study.

Ethical approval
Ethics committee approval was not obtained as an open-source data set was used.

References

1. Gunbatar H, Sertogullarindan B, et al. Evaluation of cases with lung cancer; 3-year analysis. Van Med J. 2012;13-20.
2. TT. Association and Annual Congress. Lung and pleural malignancies working group, Turkey's lung cancer map project. Turkey's Lung Cancer Incidence. 2005;13.
3. Yilmaz A, Damadoglu E, Salturk C, et al. Delays in the diagnosis and treatment of primary lung cancer: are longer delays associated with the advanced pathological stage? Ups J Med Sci. 2008;113.3:287-96.
4. Schreiber G, McCrory DC. Performance characteristics of different modalities for diagnosis of suspected lung cancer: summary of published evidence. Chest. 2003;123;115-28.
5. Yoh W.M. TNM classification for lung cancer. Ann Thorac Cardiovasc Surg. 2003;9:343-50.

6. Carvalho S, Troost EG, Bons J, et al. Prognostic value of blood-biomarkers related to hypoxia, inflammation, immune response, and tumor load in non-small cell lung cancer–A survival model with external validation. Radiother Oncol. 2016;119:487-94.
7. Akman M, Genc Y, Ankarali H. Random Forests Methods and an Application in Health Science. Turkiye Klinikleri J Biostat. 2011;36-48.
8. Rao S, Gupta P. Implementing Improved Algorithm Over Apriori Data Mining Association Rule Algorithm. Int J Comput Sci Inf Technol. 2012;3:489-93.
9. Alatas B. Fuzzy Logic and Genetic Algorithm Approach to the Discovery of Quantitative Association Rules. FiUniv Unv. Graduate School of Natural and Applied Sciences. 2003; Elazig.
10. Karabatak M, Ince MC. Student Success Analysis with Apriori Algorithm. Elektrik Elektronik Bilgisayar Mühendisliği Sempozyumu (ELECO), Bursa. 2004.
11. Platt J. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. Advances in large margin classifiers,, 1999;10.3:61-74.
12. Manimaran J, Velmurugan T. Analysing the quality of association rules by computing an interestingness measure. Indian J Sci Technol. 2015;1-12.
13. Kumar S, Joshi N. Rule power factor: a new interest measure in associative classification. Procedia Comput Sci. 2016;12-8.
14. Yildiz O. Melanoma detection from dermoscopy images with deep learning methods: A comprehensive study. J Fac Eng Archit Gazi Univ. 2019;2241-60.
15. Chen MS, Han J, Yu PS. Data mining: an overview from a database perspective. IEEE Trans Knowl Data Eng. 1996;866-83.
16. Berzal F, Blanco I, Sánchez D, Vila M.A. Measuring the accuracy and interest of association rules: A new framework. Intell Data Anal. 2002;221-35.
17. Polikar R. Ensemble learning. in Ensemble machine learning: Springer. 2012;1-34.
18. Breiman L. Random forests. Mach Learn. 2001;45:5-32.
19. Zhou ZH. Ensemble methods: foundations and algorithms. CRC Press. 2012.
20. Wolpert DH. Stacked generalization. Neural Networks. 1992;241-59.
21. Alpar R. Applied Statistics and Validity-Reliability with Examples from Sports, Health and Education Sciences. 2016;513-57.
22. Yasar S, Arslan A, Colak C, Yologlu S. A Developed Interactive Web Application for Statistical Analysis: Statistical Analysis Software. MBSJ Health Sci. 2020;227-39.
23. Campbell M. RStudio Projects. in Learn RStudio IDE: Springer. 2019;39-48.
24. Hofmann M, Klinkenberg R. RapidMiner: Data mining use cases and business analytics applications. CRC Press. 2016.
25. Yadav S, Shukla S. Analysis of k-fold cross-validation over hold-out validation on colossal datasets for quality classification. in 2016 IEEE 6th IACC. 2016;78-83.
26. Jemal A, Thomas A, Murray T, Thun M. Cancer statistics. Ca-Cancer J Clin. 2002;52:23-47.
27. Cokkinides V, Albano J, Samuels A, et al. American cancer society: Cancer facts and figures. Atlanta: American Cancer Society. 2005.
28. Capewell S. Patients presenting with lung cancer in South East Scotland. Thorax. 1987;42:853-7.
29. Kourou K, Exarchos TP, et al. Machine learning applications in cancer prognosis and prediction. Comput Struct Biotechnol J. 2015;13:8-17.
30. Asri H, Mousannif H, Al Moatassime H, Noel T. Using machine learning algorithms for breast cancer risk prediction and diagnosis. Procedia Comput Sci. 2016;83:1064-9.
31. Karacan H, Yesilbudak M. User-Centered Interactive Data Mining: A Literature Review. J Information Technologies2010;3.
32. Li K, Liu Z, Han Y. Study of selective ensemble learning methods based on support vector machine. Phys Procedia. 2012;33:1518-25.
33. Nishio M, Nishizawa M, Sugiyama O, et al. Computer-aided diagnosis of lung nodule using gradient tree boosting and Bayesian optimization. PloS One. 2018;13:e0195875.
34. Faisal MI, Bashir S, Khan ZS, Khan FH. An evaluation of machine learning classifiers and ensembles for early-stage prediction of lung cancer. Paper presented at 2018 3rd International Conference on Emerging Trends in Engineering, Sciences and Technology (ICEEST)] IEEE. 2018.