Journal of Big Data

RESEARCH

Open Access

An enhanced random forest approach using CoClust clustering: MIMIC-III and SMS spam collection application

Zeynep Ilhan Taskin[1*], Kasirga Yildirak[2] and Cagdas Hakan Aladag[3]

*Correspondence:
zeynepilhan@ogu.edu.tr

[1] Department of Statistics, Faculty of Science and Literature, Eskisehir Osmangazi University, Eskisehir, Turkey
[2] Department of Actuarial Science, Faculty of Science, Hacettepe University, Ankara, Turkey
[3] Department of Statistics, Faculty of Science, Hacettepe University, Ankara, Turkey

Abstract

The random forest algorithm could be enhanced and produce better results with a well-designed and organized feature selection phase. The dependency structure between the variables is considered to be the most important criterion behind selecting the variables to be used in the algorithm during the feature selection phase. As the dependency structure is mostly nonlinear, making use of a tool that considers nonlinearity would be a more beneficial approach. Copula-Based Clustering technique (CoClust) clusters variables with copulas according to nonlinear dependency. We show that it is possible to achieve a remarkable improvement in CPU times and accuracy by adding the CoClust-based feature selection step to the random forest technique. We work with two different large datasets, namely, the MIMIC-III Sepsis Dataset and the SMS Spam Collection Dataset. The first dataset is large in terms of rows referring to individual IDs, while the latter is an example of longer column length data with many variables to be considered. In the proposed approach, first, random forest is employed without adding the CoClust step. Then, random forest is repeated in the clusters obtained with CoClust. The obtained results are compared in terms of CPU time, accuracy and ROC (receiver operating characteristic) curve. CoClust clustering results are compared with K-means and hierarchical clustering techniques. The Random Forest, Gradient Boosting and Logistic Regression results obtained with these clusters and the success of RF and CoClust working together are examined.

Keywords: CoClust, Random forest, Copula, Feature selection, MIMIC-III

Introduction

The random forest technique is an effective and popular method to solve classification and regression problems based on decision trees. It is a combination of tree predictors such that each tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest. Random forest (RF) has been used in biology and medicine, such as high-dimensional genetic or tissue microarray data and MIMIC-III [1–6]. It is specifically devised to operate quickly and efficiently over large datasets because of the simplification and it offers the highest prediction accuracy compared to other models in the setting of classification.

The main contribution of this study is to increase the speed and accuracy of RF by adding a new feature selection step. Especially when working with big data, it is very important to increase speed and accuracy by using a correct clustering method. Correct determination of the dependency between variables in the feature selection step is one of the most critical steps of the study. Although there is an expectation of linear dependence in the studies, nonlinear dependence is also frequently encountered. The efficient operation of the clustering method used in nonlinear dependence is one of the side benefits of the study. Working with non-linear dependency during the correct determination of the relationship between variables is one of the side benefits of the article. One of the popular methods used in analyzing nonlinear dependencies is copulas. The main advantage of the proposed approach using CoClust is to achieve high accuracy in big data in a short time.

Copula-Based Clustering technique called CoClust, which examines dependencies using copulas, is an alternative to classical clustering techniques. It overcomes linear dependency constraints. In this technique, the power and type of multivariate dependency between sets are modeled with a copula function and dependency parameter.

In the feature selection step, the determination of nonlinear dependency is emphasized, and copulas are preferred. CoClust gives effective results by clustering variables that show nonlinear dependency using copulas. We mainly work on the feature selection phase employing CoClust rather than regular feature selection methods and show that high efficiency in terms of CPU time and prediction is obtained from this version of RF because CoClust implies the noninclusion of the uncorrelated variables in clusters.

The data-oriented purpose of our work lies in the use of a more efficient prediction model for mortality prediction and spam SMS classification through copulas and CoClust. It also aimed to develop a different approach for mortality prediction in intensive care patients and spam SMS classification by examining the nonlinear dependency structure between variables through copulas.

This method proposed for the Random Forest method is also applied in other classification techniques such as Gradient Boosting and Logistic Regression, and the results are evaluated in terms of both other clustering methods and machine learning methods.

Another important aspect of the study is that it works with two large datasets, MIMIC-III (Medical Information Mart for Intensive Care) and SMS Spam Collection. The MIMIC-III is a large free access database including more than forty thousand patients who were treated at the intensive care units of Beth Israel Deaconess Medical Center between 2001 and 2012. MIMIC-III, the latest version of MIMIC, includes the hospital records of 46520 patients, 38645 of whom are adults and 7875 newborns.

Examining the proposed method in a dataset with a large number of variables is another important step of the study. In this context, the SMS Spam Collection dataset is used, which helps short message services classify messages as spam. While the number of text messages used is 5574, the number of variables is remarkable in this dataset. The dataset consists of 770 variables. For the feature selection step, testing the proposed method on a dataset consisting of many variables expands the vision for comparison.

In this context, a literature review of the techniques used is clarified in "Literature review" Section. CoClust, RF and the proposed approach for RF are explained in "The proposed approach for random forest" Section. "Datasets" Section presents the

experimentation of data sampling. In "Application" Section, the results obtained by applying the proposed approach are presented. "Discussion" Section and "Conclusion" Section focus on the discussion and conclusion of the application.

Literature review

RF is a flexible, easy-to-use machine learning algorithm that often produces a great result and mainly depends on the celebrated method so-called classification and regression trees (CART). Breiman [7] provided an early example of bagging with random selection to grow each tree without replacement. Dietterich [8] and Ho [9] make use of random subspace and random split selection.

Breiman [10] uses new training sets by randomizing the outputs in the original training set. He defines the RF as a combination of tree predictors such that each tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest. He also suggests that some or all of the input variables may be categorical, and since it is wanted to define additive combinations of variables, it is necessary to define how categorical variables will be treated so they can be combined with numerical variables. [11].

Mistry et al. [5] draw attention to classifiers that allow us to predict which tool will be most suitable for reducing the toxicity of a drug. They demonstrate the use of data mining and machine learning techniques by examining models using RF and decision trees. Accordingly, an accuracy of 80% is obtained from the RF models. Thus, RF gives efficient results in the field of health.

The use of RF in mortality predictions also has an important place in the literature. Levantesi and Nigri [3] propose a novel approach based on the combination of RF and two-dimensional P-spline. The two-dimensional P-spline is used to smooth and project the RF estimator in the forecasting phase. All the analyses were carried out on data from the Human Mortality Database and considering the Lee–Carter model.

RF could be used in biology and medicine, such as high-dimensional genetic or tissue microarray data [12, 13]. The RF technique has also been studied on MIMIC, which is an important database. Thus, remarkable studies have emerged for both RF and MIMIC databases.

The RF technique has also been studied on MIMIC, which is an important database. Thus, remarkable studies have emerged for both RF and MIMIC databases. Poucke et al. [6] concentrate on quantitative analysis of the predictive power of laboratory tests and early detection of mortality risk by using predictive models and feature selection techniques in the MIMIC-III database. RF and logistic regression were used on patients with renal failure admitted to ICUs at Boston's Beth Israel Deaconess Medical Center.

McWilliams et al. [4] object to developing an automated method for detecting patients who are ready for discharge from intensive care. Two cohorts derived from the GICU and MIMIC-III were analyzed with RF and a logistic classifier.

Another important step regarding RF is the feature selection phase. Although RF inherently enables feature selection, using different techniques in feature selection sheds light on the RF technique and literature.

Hapfelmeier and Ulm [14] claim that feature selection has been suggested for RF to improve data prediction and interpretation. Three approaches to selecting variables,

i.e., Multiple imputations, complete case analysis and the application of a self-contained measure are applied to half of the data. In the rest of the study, unbiased RF is preferred.

Uddin and Uddin [15] propose a feature selection method based on guided RF. The guided RF is used to select a small set of important variables. First, an ordinary RF is trained on the dataset to collect the feature importance scores, and then, the collected importance scores are injected to influence the feature selection process in the guided RF.

Gupta [16] uses three approaches (wrappers, filters, embedded methods) for feature selection, and then four machine learning models are used to solve classification problems. RF is one of these methods. The highest accuracy of 56.99% is achieved with the RF model.

Copulas are first used by Abe Sklar [17]. Sklar's theorem elucidates the role that copulas play in the relationship between multivariate distribution functions and their univariate margins [18]. It expresses that any multivariate joint distribution can be written on the basis of univariate marginal distribution functions and a copula that describes the dependence structure between the variables [19].

Mesiar and Sheikhi [20] emphasize the importance of nonlinear dependence in their studies and offer a solution to the problem through copulas. In this study, the simulated data are obtained through copulas and each of them is placed in a correlated cluster. However, in CoClust, if the variable is not related, it is excluded from the clusters, which is one of its most important distinguishing features.

Although copulas are used in many areas, the introduction to CoClust is with Di Lascio [21]. The study that introduced the CoClust technique into the field is a doctoral dissertation that develops the technique further on clinical microarray data analysis [22], Di Lascio [21]. In 2017, the development of forty European countries was examined by healthy nutrition rules with CoClust, and later in 2019, they examined the improved version of the technique (Di Lascio, Durante, and Pappada [23]; Di Lascio and Giannerini [24].

CoClust, based on copula functions, allows clustering of observations according to multivariate dependency structure without any assumption on marginals. The basic idea behind CoClust is that the row data matrix separates the K group at once, that is, it creates an advanced procedure that separates the p-dimensional vector for each set (Di Lascio [21]).

Di Lascio [21] also compared CoClust with another well-known clustering technique based on probability models and found that the latter is not able to model the true dependence relationship between observations.

There are also studies in the literature in which copulas and decision trees are used together. Khan et al. [25] bring a joint approach to copulas and decision trees. They appraised a novel nonparametric copula-based decision tree organization using a measure of dependence and applied their proposed method to credit card records for Taiwan and coronary heart disease records of Pakistan and acquired the desirable outcomes. As a result of the application, the desired results are obtained.

Eling and Toplek [26] and Messiar and Sheikhi [20] emphasize the importance of nonlinear dependence in their studies and offer a solution to the problem through

copulas. In this study, a solution proposal to this problem is presented by using the nonlinear dependency skill of CoClust.

Zhu et al. [27] aim to establish prediction scores on mechanically ventilated patients in ICU and they use the machine learning methods of k-nearest neighbors, logistic regression, bagging, decision tree, random forest, Extreme Gradient Boosting, and neural network for model establishment. The efficiency of the resulting models is measured via AUC and a value of AUC is reached 0.819 with RF.

Khope and Elias [28] examine the MIMIC-III data set over KNN, LR and ANN and compare the results obtained with the confusion matrix on accuracy.

Based on the literature review, it is decided to use CoClust and RF techniques together. Thus, by applying the feature selection step with CoClust, it is possible to work on the goal of achieving more perfect accuracy in a shorter time. The methods mentioned in the literature review are explained in "The proposed approach for random forest" Section.

The proposed approach for random forest

CoClust brings a different perspective to the literature by using copulas in the clustering technique. In this study, a novel approach is proposed by adding CoClust to RF as a feature selection step. In the proposed approach, clusters are formed by considering the dependency between variables with CoClust, and then the most efficient model is obtained with RF by using the relevant variables. Thus, it aims to bring a different approach to the feature selection phase of RF. These techniques utilized in the proposed method are explained in this section.

CoClust

CoClust was introduced by Di Lascio in [21] through her doctoral thesis, developed in 2017, and the final version of the technique was presented by Di Lascio and Giannerini [24].

CoClust includes copula families in the clustering algorithm. It refers to the clustering of multivariate dependent variables based on the likelihood copula function. CoClust assumes that the data are derived from the multivariable copula function, which is known to represent each cluster by the marginal function. The power and type of multivariate dependency between clusters are modeled by the copula function and the dependency parameter of the copula, respectively.

The copula function is defined as "functions that join or couple multivariate distribution functions to their one-dimensional marginal distribution functions" by Nelsen [18].

The copula function was first handled by Abe Sklar in [17] as a function that depends on univariate marginals to multivariate distributions within the scope of probable metric spaces [29].

Consider for a moment a pair of random variables X and Y, with distribution functions $F(x) = P(X \leq x)$ and $G(y) = P(Y \leq y)$, respectively, and a joint distribution function $H(x, y) = P(X \leq x, Y \leq y)$. For each pair of real numbers (x, y), we can associate three numbers: $F(x)$, $G(y)$, and $H(x, y)$. Each of these numbers lies in the interval [0,1]. In other words, each pair (x, y) of real numbers leads to a point $(F(x), G(y))$ in the unit square $[0,1] \times [0,1]$, and this ordered pair in turn corresponds to a number $H(x, y)$ in [0,1]. This correspondence, which assigns the value of the joint distribution function

to each ordered pair of values of the individual distribution functions, is indeed a function. Such functions are copulas [18].

Let H be a joint distribution function with margins F and G. A copula C is defined in Eq. 1 for all $x, y \in \overline{R}$[18].

$$H(x, y) = C(F(x), G(y)) \tag{1}$$

According to Sklar's theorem, any joint probability function *f(.)* can be split into the margins and a copula. For continuous random variables, the copula density *c(·)* is related to the density *f (·)* of the *distribution F(·)* through the well-known canonical representation and can be presented in Eq. 2 (Di Lascio, Durante, and Pappada 2017).

$$f(x_1, \ldots, x_K) = c(F_1(x_1), \ldots, F_K(x_K)) \prod_{k=1}^{K} f_k(x_k) \tag{2}$$

Such separation determines the modeling flexibility given by copulas since it is possible to decompose the estimation problem in two steps: in the first step, margins are estimated; and in the second step, the copula model is estimated. The most commonly used estimation method is the two-stage inference for margins method [30], which employs the log-likelihood estimation method to estimate both the parameter(s) of each margin and the copula parameter θ. This method can be used in a semiparametric approach (Genest, Ghoudi, and Rivest [31]) that does not require distributional assumptions on the margins. The log-likelihood copula function is used to estimate θ in Eq. 3 (Di Lascio, Durante, and Pappada 2017).

$$\hat{\theta} = argmax_{\theta} \sum_{i=1}^{n} logc\left\{\hat{F}_1(X_{1i}), \ldots, \hat{F}_k(X_{Ki}); \theta\right\} \tag{3}$$

The concept of CoClust refers to the aggregation of multivariate dependent variables based on a log-likelihood function of the copula model. To realize this clustering, CoClust assumes that the parameters of the data are derived by the multivariate copula function, which represents clusters, and each cluster is known to be represented by the univariate density function. The power and type of multivariate dependency between clusters are modeled by a copula function and dependency parameter of the copula, respectively.

The beginning of the algorithm is an *(n x p)* data matrix X. It is expressed by Eq. 4.

$$X = \begin{bmatrix} x_{11} & \ldots & x_{1j} & \ldots & x_{1p} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{i1} & \ldots & x_{ij} & \ldots & x_{ip} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{i'1} & \ldots & x_{i'j} & \ldots & x_{i'p} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{n1} & \ldots & x_{nj} & \ldots & x_{np} \end{bmatrix} \tag{4}$$

The purpose of clustering is to group the *(n x p)*-dimensional dataset into a K cluster.

Values in a row (or column) vector are independent functions of the same density function, so the observations in each set are from the same distribution. Here, the algorithm is described as applying the data matrix to the rows (Di Lascio [21]).

The main steps of the CoClust algorithm required for clustering the n row data matrix are explained as follows (Di Lascio and Giannerini [24]).

1. for $k=2...., K_{max}$, where $K_{max} \leq n$ is the maximum number of clusters to be tried:

   a. select a subset of n_k k-plets of rows/profiles in the data matrix on the basis of the following multivariate measure of association based on pairwise Spearman's ρ correlation coefficient in Eq. 5.

$$H(\Lambda_2|\Lambda_1) = \max_{i' \in \Lambda_2} \left\{ \underset{i \in \Lambda_1}{\psi} \left(\rho(x_i, x_{i'})\right) \right\} \tag{5}$$

   In Equation 5, Λ is a set of row index profiles such that $\Lambda = \Lambda_1 \cup \Lambda_2$, Λ_1 is the subset of profiles already selected to compose a k-plet, Λ_2 is the set of remaining candidates to complete a k-plet, x_i is the ith profile, ψ is a selected function among the mean, the median or the maximum;

   b. fit the copula model on the n_k k-plets of profiles/rows through the maximum pseudolikelihood estimation.

2. select the subset of n_k k-plets of rows/profiles, say n_K K-plets, that maximizes the log-likelihood copula function; hence, the number of clusters K, i.e., the dimension of the copula, is automatically chosen;
3. select a K-plet using the measure in Eq. (5) and estimate K! copulas by using the observations already clustered and a permutation of those candidates to the allocation;
4. allocate the permutation of the selected K-plet to the clustering by assigning each observation to the corresponding cluster if it increases the log-likelihood of the copula fit; otherwise, drop the entire K-plet of rows/profiles;
5. repeat steps 3. and 4. until all the observations are evaluated (either allocated or discarded).

Since nonnested models are tested at every step of the algorithm, that is, working with copula models using univariate dependency parameters, the defined log-likelihood-based criterion is equivalent to the Bayesian information criterion and the Akaike information criterion (Di Lascio [21]).

The Bayesian information criterion is defined as Eq. 6 for the K-dimensional copula model m (Di Lascio [21]).

$$BIC_{K,m} = -2\log \prod_{i=1}^{n} c_m \left\{ \hat{F}_1(X_{1i}), \ldots, \hat{F}_K(X_{Ki}); \hat{\theta} \right\} + s\log((n/K)p) \tag{6}$$

Accordingly, the model of the copula that minimizes the BIC value is selected. Similarly, the Akaike information criterion (AIC) is expressed in Eq. 7 and is used to select the model of the copula (Di Lascio [21])

$$AIC_{K,m} = -2\log \prod_{i=1}^{n} c_m\left\{\hat{F}_1(X_{1i}), \ldots, \hat{F}_K(X_{Ki}); \hat{\theta}\right\} + 2s \tag{7}$$

In the technique, clusters, each containing a maximum number of (n/K)p independent observations, are obtained. The configuration of multivariate relationships here is not based on intracluster relationships in classical clustering methods. Although each cluster is independent identical distributions obtained from the same marginal distribution, intercluster observations share the same multivariate dependency structure (Di Lascio [21]). Thus, each cluster is generated by a (marginal) univariate density function, and the interpretation of the clustering is based on within-group independence and among-group dependence (Di Lascio, Durante, and Pappada [32]). In classical clustering methods, elements that are correlated with each other are in the same cluster and are expressed in this way in tables. However, the situation is the opposite in CoClust tables. Di Lascio and Disegna [33] explain that the CoClust aims to describe the within-cluster independence and the between-cluster dependence instead of the within-cluster homogeneity and the between-cluster separation, as the more traditional clustering approaches. Therefore, in order not to create confusion for the reader, the expression of the sets in the tables is done as in the classical methods.

The most important advantage of the technique is that there is no need to set a priori the exact number of clusters K, nor is a starting classification required because the algorithm automatically selects the best number of clusters K within a given range of possibilities on the basis of the log-likelihood in Eq. 3 (Di Lascio and Giannerini [23]).

The other important feature of this technique is that it clusters only the variables that it identifies to be related, which means not all variables present are placed in clusters. Variables regarded as uncorrelated are kept outside of the clusters. In this respect, it differs from the tail dependency technique.

In the literature, many different copula models are available, but Nelsen [18] demonstrated that the elliptical and Archimedean families are the most useful in empirical modeling. The elliptical family includes the Gaussian copula and the t-copula. Both copulas are symmetric, and they can take into account both positive and negative dependence since $-1 \leq \theta \leq 1$. On the other hand, the Archimedean family enables us to describe both left and right asymmetry as well as weak symmetry among the margins by employing Clayton's, Gumbel's and Frank's models. Clayton's copula has the parameter $\theta \in (0, \infty)$, and as θ approaches zero, the margins become independent. The dependence parameter θ of a Gumbel model is restricted to the interval $[1, +\infty]$. where the value 1 means independence. Finally, the dependence parameter θ of a Frank copula may assume any real value, and as θ approaches zero, the marginal distributions become independent (Di Lascio, Durante, and Pappada [32]).

Di Lascio [21] tests the CoClust algorithm on simulated data drawn from Gaussian and Frank copulas in different situations and dependence settings. They found that

the algorithm is able to recover the true underlying dependence relationship between observations grouped in different clusters irrespective of the kind of margins, the value of the dependence parameter and the copula model.

The CoClust algorithm has been successfully applied to various datasets. Di Lascio et al. [32] attempted to determine the type of organs from tumors and cancer cell lines. Regarding biomedical applications, Di Lascio and Giannerini [24] applied the CoClust algorithm to formulate the possible functional relationship between genes with hypotheses. Di Lascio et al. [32] study can be given as an example for applications in other fields. The aim of this study is to analyze changes in EU country diets under the guidance of health diets and common European policies, and Di Lascio et al. [32] use them to investigate the geographic distribution of precipitation measurements.

The other important feature of this technique is that it clusters only the variables that it identifies to be related, which means not all variables present are placed in clusters. Variables regarded as uncorrelated are kept outside of the clusters. In this respect, it differs from the tail dependency technique.

Random forest

RF is a technique based on decision trees that uses rules to split data in a binary method (Ji, Yang, and Tang [34]). In the literature, when solving classification problems, the Gini index, deviance and the towing rule are used for the best split [35].

Breiman [11] defines that RF is a combination of tree predictors such that each tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest. Decision trees come together to form an RF. Each tree is a randomly selected subset from the dataset.

An RF is an ensemble classifier consisting of many decision trees, where the final predicted class for a test example is obtained by combining the predictions of all individual trees [11]. Each node is partitioned based on a single feature, and each branch ends in a terminal node. Terminal nodes provide a prediction for the class of a test example based on the path taken through the tree. The color of a terminal node indicates its class prediction. The final predicted class for a test example is obtained by combining the predictions of all individual trees [36].

In other words, many classification and regression trees are generated and then the results are aggregated. Each tree is independently constructed using a bagging sample of the training data (Ji, Yang, and Tang [34]).

Additionally, the technique is not affected by the interactions of correlated variables because each tree comprises random samples [37].

In the training phase, X represents the object in the training dataset (an $N\,x\,M$ matrix, where N is the number of training data and M is the number of variables); L represents the labels of the training set (an $N \times 1$ matrix); n_{tree} represents the number of trees in the forests; θ_k represents each random tree in the random forests (k = 1, 2, …, n_{tree}); M_{try} represents the number of features randomly selected to split (Ji, Yang, and Tang [34]).

RF is an integrated classifier composed of multiple decision tree classifiers, which can be described as in Eq. 8.

$$h(X, \theta_k); k = 1, 2, \ldots, n_{tree} \tag{8}$$

At the end of the algorithm, the predictive capability of the RF model should be assessed. Various statistical parameters or cross-validation procedures are used to validate the performance of the proposed models [38, 39].

The RF method has two important products: out-of-bag estimates of the generalization error and variable importance measures [11]. Two algorithms for calculating variable importance measures differ somewhat from the four heuristics originally suggested for variable importance measures [11]. The first heuristic is based on the Gini criterion, and the second calculates variable importance as the mean decrease in accuracy using out-of-bag observations [40, 11]. The OOB observations can also be used to calculate variable importance, and Gini impurity represents the probability that a randomly selected sample from a node will be incorrectly classified according to the distribution of samples in the node [40].

To evaluate the classification ability and the performance of the model, parameters such as error (Er) and accuracy (Ac) are calculated, which are given in Eqs. 9 and 10. In the equations, TP, FP, TN and FN denote true positives, false positives, true negatives and false negatives, respectively. The relationship of these four factors can be best shown by the confusion matrix [41].

$$Er = \frac{FP + FN}{TP + TN + FP + FN} \tag{9}$$

$$Ac = \frac{TP + TN}{TP + TN + FP + FN} \tag{10}$$

Note that sensitivity is also called the true positive rate, defined as the ability and proportion of a classifier to correctly predict positively labeled molecules, while specificity is also called the true negative rate, defined as the capability and percentage of negatively labeled instances identified as negative [39, 42, 43]. Accuracy is the percentage coverage of correct predictions, generally applied to judge the predictive power of models [44].

There are other metrics such as F1 and Recall calculated with the confusion matrix. Recall score represents the model's ability to correctly predict the positives out of actual positives. Recall is also known as sensitivity or the true positive rate and is given in Eq. 11. F1 score represents the model score as a function of precision and recall score. F-score is a machine learning model performance metric that gives equal weight to both the Precision and Recall for measuring its performance in terms of accuracy, making it an alternative to accuracy metrics. F1 score is also given in Eq. 12.

$$Recall = \frac{TP}{TP + FN} \tag{11}$$

$$F1 = \frac{2TP}{2TP + FP + FN} \tag{12}$$

Accuracy is a machine learning classification model performance metric that is defined as the ratio of true positives and true negatives to all positive and negative observations. In other words, accuracy tells us how often we can expect our machine learning model

will correctly predict an outcome out of the total number of times it made predictions. For this reason, accuracy criterion is preferred in the interpretation of the models.

The second performance criterion aims at measuring the extent to which the RF model can distinguish between the classes, i.e., the ability of the RF model to rank the events with "$y=1$" relative to those with "$y=0$". This can be evaluated using the receiver operating characteristic (ROC) curve [45]. The closer the curve is to the left-hand corner of the ROC space, the better the classification. The area between the first bisector and the ROC curve (denoted as AUC) allows the performance of the RF model to be quantified [46]. In other words, the AUC is a combined measure of sensitivity (true positive rate) and specificity (true negative rate) at various probability threshold settings. Since both the x and y axes have values between 0 and 1, it can take any value between 0 and 1. The closer the AUC is to 1, the better the overall diagnostic performance of the test, so it is expected to be as close to 1 as possible.

Random forest with CoClust

In this subsection, the proposed method combining RF and CoClust is introduced. The main purpose of this study is to achieve efficient results while reducing CPU time by adding a feature selection step to the RF technique. Thus, RF was developed, which is currently a powerful and effective method. In addition to this advantage, obtaining high-efficiency models in short CPU time and using this efficient result in estimating mortality and spam messages is another gain of the study.

CoClust works through copula families using the nonlinear dependency structure. It gives effective results by clustering variables that show nonlinear dependency using copulas. Nonlinear dependency is included by using CoClust in the feature selection step. This is one of the important advantages of the study. Therefore, using a new method such as CoClust, the traditional point of view has been viewed from an innovative perspective.

In addition to other feature selection techniques, the reasons for choosing CoClust are listed below.

- It does not require a starting classification to be chosen;
- It does not require the number of clusters to be set a priori;
- It is able to capture multivariate and nonlinear dependence relationships underlying the observed data;
- It does not require the marginal probability distributions to be set as Gaussian;
- It is able to discard irrelevant observations [32].

CoClust makes a difference, as it gives results that are completely compatible with the data structure without interfering with the data and results. When all the mentioned features are examined, it is clearly seen that the researcher cannot directly intervene in the process. It is very important in terms of the reliability and objectivity of the method.

The additional gains of this study are to bring a new approach to RF with the proposed method, to predict mortality by working with a large dataset and to correctly classify spam messages with fewer variables. Using the correct variables together is very important for predicting mortality in ICU patients. By observing the effect of the variables

to be determined in the MIMIC-III dataset on mortality prediction with the proposed method, an important improvement will be made for ICU patients.

Based on this, first, the RF technique is applied to the MIMIC-III and SMS Spam Collection datasets without adding any steps. At this stage, forests of different sizes are created, and CPU time, accuracy and ROC curve results are recorded. In this step, all results obtained from RF are presented, and the efficiency of the method is emphasized again.

Then, feature selection is carried out with CoClust in the MIMIC-III and SMS Spam Collection datasets, and models are created with the RF method by using variables that passed the selection stage. In the CoClust step, Archimedean and Gaussian all copula families are used. Copula families resulting in clustering are used. As an advantage of the method, the choice of the number of clusters is left to CoClust without restricting the number of clusters. After clustering with CoClust, RF application is applied to the variables in the clusters. The study was repeated in forests of the same size for each cluster, and CPU time, accuracy and ROC curve values were recorded for the forests belonging to the clusters. At the end of modeling, the most efficient models are selected and evaluated according to the saved results.

The application of CoClust and RF techniques, whose theoretical background is explained, to the MIMIC-III and SMS Spam Collection datasets is explained in "Datasets" Section.

Datasets

The MIMIC-III and SMS Spam Collection datasets are used to determine the efficiency of the proposed method. In the following sections, datasets are introduced.

MIMIC-III dataset

The MIMIC used in the study is a large free access database including more than forty thousand patients who were treated at the intensive care units of Beth Israel Deaconess Medical Center between 2001 and 2012. This database includes demographic information, laboratory test results, the procedures applied, medications, caregiver notes, imaging reports, hourly records of vital signs and death variables [47].

MIMIC-III, the latest version of MIMIC, includes the hospital records of 46520 patients, 38645 of whom are adults and 7875 newborns. The latest data cover the period between June 2001 and October 2012. Although the database has not been identified, it contains detailed information about the clinical care of patients. MIMIC-III database is closed. The academic paper about database is here: https://physionet.org/content/mimiciii/1.4/. To get data from the database, proceed from this link: https://physionet.org/settings/credentialing/. In order to use the restricted-access clinical databases hosted on PhysioNet, users must have a credentialed PhysioNet account. If user doesn't have credentialed account, s/he must apply for access from this link: https://physionet.org/credential-application/. In order to become a credentialed PhysioNet user and access the restricted-access clinical databases like MIMIC-III, you must complete a suitable training program in human research subject protections and HIPAA regulations. After these steps, personal information is completed. In our team, these processes are carried out by Prof.Dr. Kasırga Yıldırak completed.

From the database, fourty physiological and demographic variables were obtained. These variables are used in scores (SOFA, SAPS II, APACHE) used in intensive care patients. Vital signals such as blood pressure, temperature, and respiration have been proven to have a strong relationship with mortality [48]. In the literature, a variable pool is created by adding vital variables such as albumin, hemoglobin and glucose, which are associated with mortality [22, 48].

A death variable was used for the mortality model. Here, the death variable is the "*no death*" category, which is coded with reference category *0*.

The variables that are used are given in Table 1.

Respiration, coagulation, liver, renal, central nervous system and cardiovascular function were categorical variables. Vincent et al. [49] express the categorization conditions of these variables, as shown in Table 2.

According to Johnson et al. [22], 25800 patients, who were only adults, were studied after removing the data caused by registration errors and patients who stayed in intensive care for less than 4 h.

While approximately 60% of these patients were women (15536 people), 40% were men (10264 people). These results can be obtained from Fig. 1.

Frequencies and percentages for categorical variables are shown in Table 3.

Descriptive statistics of the age variable, vital variables, laboratory results and blood values of the patients are shown in Table 4.

The death rate was 31.3% (8075 people), and the nondeath rate was 68.7% (17725 people). The prediction is made by recording the patient data.

Table 1 Variables used

Variable	Measurement unit	Variable	Measurement unit
Age	Continuous	Potassium	mEq/L
Gender	Categorical	Partial Thromboplastin Time	Second
Heart Rate	Discrete	International Normalized Ratio	Ratio
Systolic Blood Pressure	mm/Hg	Prothrombin Time	Second
Diastolic Blood Pressure	mm/Hg	Sodium	mEq/L
Mean Arterial Pressure	mm/Hg	Blood Urea Nitrogen	mg/dL
Respiration	mmHg	White Blood Cells	$10^3/mm^3$
Temperature	°C	Norepinephrine	pg/mL
SpO_2	mmHg	Epinephrine	pg/mL
Glucose	mg/dL	Dopamine	pg/mL
Albumin	g/L	Dobutamine	mL/h
Immature Neutrophil Cells	With > 10%	Urine Output	mL/kg/hr
Bicarbonate	mEq/L	PaO_2/FiO_2	mmHg
Bilirubin	mg/dL	Glasgow Coma Scale	Score
Creatinine	mg/dL	Mechanical Respiration	Categorical
Chloride	mEq/L	Coagulation	$\times 10^3/\mu L$
Hematocrit	%	Liver	Categorical
Hemoglobin	g/dL	Cardiovascular	Categorical from mmHg
Lactate	mEq/L	Central Nervous System	Score
Platelet	$10^9/L$	Renal	Categorical from mg/dL

Table 2 Categorical variables

	0	1	2	3	4
Respiration	≥ 400	≥ 300	≥ 200	≥ 100	> 0
Coagulation	≥ 150	< 150	< 100	< 50	< 20
Liver	< 1,2	≥ 1,2	≥ 2,0	≥ 6,0	> 12
Renal	< 1,2	≥ 1.2	≥ 2.0	≥ 3.5	≥ 5.0
Central Nervous System	Out of range	≤ 14	≤ 12	≤ 9	< 6
Cardiovascular	Out of range	MAP* < 70	Dopamine ≤ 5 or dobutamine (any dose)	Dopamine > 5 or Epinephrine ≤ 0,1 or Norepinephrine ≤ 0,1	Dopamine > 15 or Epinephrine > 0,1 or Norepinephrine > 0,1

* MAP: Mean Arterial Pressure

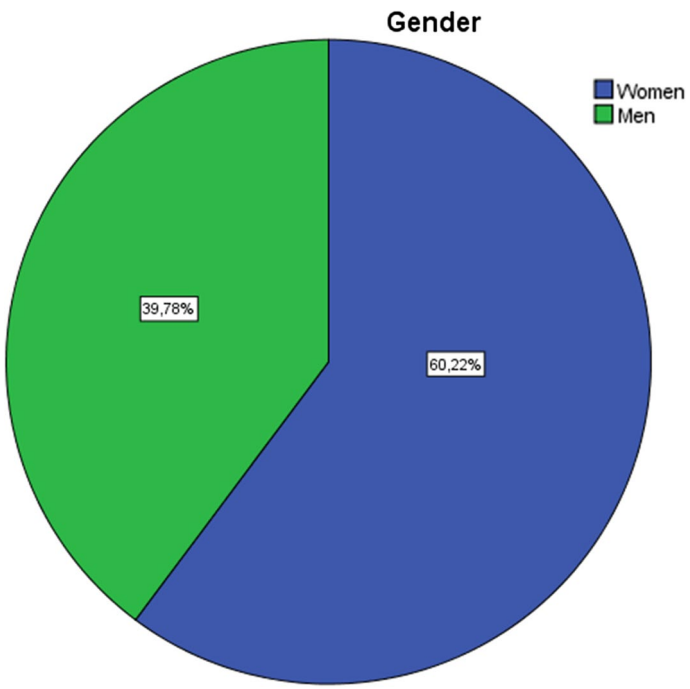

Fig. 1 Gender percentages

SMS spam collection dataset

The SMS Spam Collection is a public set of SMS-labeled messages that was created by Tiago A. Almeida and José María Gómez Hidalgo. 425 SMS from the Grumbletext Web site, 3375 SMS randomly chosen ham messages of the NUS SMS Corpus (NSC), 450 SMS ham messages collected from Caroline Tag's PhD Thesis and 1324 SMS from SMS Spam Corpus v.0.1 Big have incorporated in SMS Spam Collection Dataset. The dataset consists of 5574 English, real and nonencoded messages and spam sms [50], Almeida, Hidalgo, and Yamakami [51]; Hidalgo, Almeida, and Yamakami [52]).

The dataset contains the same information as the original dataset plus the additional DistilBERT classification embeddings. It contains 5574 rows and 770 columns. The *spam* column describes whether the message is spam or not. The *original message*

Table 3 Frequencies and percentages of categorical variables

	%	0	1	2	3	4	Total
Respiration	Frequency	18373	989	2795	2743	900	25800
		71.2	3.8	10.8	10.6	3.5	100%
Coagulation	Frequency	17997	4229	2442	900	232	25800
		69.8	16.4	9.5	3.5	0.9	100%
Liver	Frequency	21641	1391	1670	577	521	25800
		83.9	5.4	6.5	2.2	2.0	100%
Renal	Frequency	16760	4243	1569	1298	1930	25800
		65.0	16.4	6.1	5.0	7.5	100%
Central Nervous System	Frequency	17532	5259	1297	1031	681	25800
		68.0	20.4	5.0	4.0	2.6	100%
Cardiovascular	Frequency	6720	15782	295	1317	1686	25800
		26.0	61.2	1.1	5.1	6.5	100%

column expresses unprocessed messages. The other 768 columns contain the DistilBERT classification embeddings for the message after it is processed. The dataset and detailed information about the dataset can be found on the UCI Repository website (http://archive.ics.uci.edu/ml/datasets/SMS+Spam+Collection). SMS SPAM database is open. The database can be accessed here: https://archive.ics.uci.edu/ml/machine-learning-databases/00228/.

The variables used in the dataset are named v1, v2, etc. All variables except the *spam* variable are continuous variables. The number of spam messages is 13.42% (748 messages), and the number of nonspam messages is 86.58% (4826 messages).

The problem of SMS spam is evaluated in legal, economic and technical aspects as in e-mails. Unlike e-mails, text messages usually consist of a few words and are filtered by word bag-of-words-based spam filters. By evaluating feature-based and compression-model-based spam filters, it has been determined that compression model filters perform well and bag-of-words-based filters are open to improvement. It is also found that content filtering for short messages is surprisingly effective [53].

The success of Bayesian filtering techniques, which are very effective in e-mails, in English and Spanish text messages has been examined. They tested a number of message representation techniques and machine learning algorithms in terms of effectiveness. The results showed that Bayesian filtering techniques can be used effectively in SMS spam detection [54].

Application

In this section, we present the results of CoClust, K-means, and hierarchical clustering to build models to evaluate the death rates of ICU patients and the variables that directly affect spam message classification. We compare the results by classifying the obtained clusters with Random Forest, Gradient Boosting (GB) and Logistic Regression (LR) methods.

The R implementation of CoClust and RF are used in the application. The Random Forest, rpart, prediction, caret, cluster, copula, CoClust, and copBasic packages are used

Table 4 Descriptive statistics of vital variables

	Minimum	Maximum	Mean	St. Deviation
Age	18.02	64.99	49.637	11.802
Heart rate	37	280	106.38	20.158
Systolic blood pressure	0	181	94.09	17.816
Diastolic blood pressure	0	114	47.25	12.141
Mean arterial pressure	0.20	125.00	61.272	14.164
Respiration	8	69	27.08	6.77
Temperature	15	39.72	36.203	0.785
Albumin	1.00	5.70	3.218	0.459
SpO_2	29	100	96.07	4.096
Glucose	21	2440	179.02	110.173
Immature neutrophil cells	1	68	9.00	3.158
Bicarbonate	5.00	51.50	23.784	4.388
Bilirubin	0.10	80.90	2.211	3.595
Creatinine	0.10	33.30	1.444	1.795
Chloride	54.80	154	104.423	5.596
Hematocrit	10.35	68.95	33.222	5.784
Hemoglobin	3.35	28.05	11.225	2.055
Lactate	0.30	26.75	2.439	1.451
Platelet	5	1297	209.82	116.616
Potassium	1.90	15.45	4.196	0.626
Partial thromboplastin time	14	150	36.42	16.874
International normalized ratio	0.70	32.40	1.453	0.836
Prothrombin time	8.45	150	15.441	5.759
Sodium	103.5	167.50	138.186	4.225
Blood urea nitrogen	1.00	251	22.142	19.416
White blood cells	0.10	204.65	11.759	6.738
Norepinephrine	0.00	50	0.332	0.473
Epinephrine	0.006	2.00	0.063	0.022
Dopamine	0.00	325	14.499	24.941
Dobutamine	0.50	20.455	5.991	0.355
Urine Output	0.00	51520	2203.91	3786.925
PaO_2/FiO_2	23.00	1542.50	246.375	82.241
Glasgow coma scale	3	15	13.93	2.421

while applying CoClust and RF. All computational runs were performed on a device with an Intel Core i7-6700 HQ CPU @ 2.60 GHz.

Study design

First, the RF, GB and LR techniques are applied to the MIMIC-III and SMS Spam Collection datasets without adding any steps. At this stage, forests consisting of 100, 200, 500, 1000, 2000, 5000 and 10000 trees are created, and CPU time, accuracy and ROC curve results are recorded for RF and GB methods. Then, the datasets are divided into clusters with CoClust, K-means and hierarchical clustering techniques. After that clustering results are evaluated in the RF, GB and LR applications. The dependency structure between variables is analyzed with copulas so that the variables to be used in prediction would be related to each other. RF application consisting of 100, 200, 500, 1000, 2000,

5000 and 10000 trees is applied to the clusters obtained. The CPU time, accuracy and ROC curve results obtained by applying RF to these clusters separately are compared with the results obtained from RF application only. In light of the results obtained, a model proposal for mortality prediction is investigated. In addition, it aims to quickly classify spam messages correctly with fewer variables.

Application of the proposed approach

In this section, first, RF, GB and LR applications are carried out without any clustering application. The CPU time, accuracy and ROC curve results to be obtained from these forests are recorded. In the next step, the datasets are clustered with the CoClust, K-means and hierachical clustering techniques. In the clusters obtained, RF and GB applications are carried out with forests of 100, 200, 500, 1000, 2000, 5000 and 10000 trees. The CPU time, accuracy and ROC curve results of the forests obtained for clusters are also recorded and compared with the previous results. The efficiency of the proposed method is questioned as a result of this comparison. The model that gives the most efficient result in the shortest CPU time is selected and suggested for mortality prediction.

The use of RF combined with CoClust is referred to as the proposed RF, while the application without CoClust is called traditional RF within the scope of the study.

Traditional RF and other classification methods without applying CoClust

In accordance with the purpose of the study, RF, GB and LR applications are performed without adding a feature selection step to the datasets obtained. The results obtained by creating forests consisting of 100, 200, 500, 1000, 2000, 5000 and 10000 trees are examined.

The CPU time, accuracy, OOB error rate and ROC curve results from the RF application for the datasets are given in Table 5. When the results are examined, it can be said that an application with 1000 trees for both datasets gives the most efficient result for both datasets according to CPU time, ROC, accuracy and OOB error rate results.

The CPU time, accuracy, OOB error rate and ROC curve results from the GB application for the datasets are given in Table 6. When the results are examined, it can be said that an application with 1000 trees for both datasets gives the most efficient result for both datasets according to CPU time, ROC, accuracy and OOB error rate results.

Table 5 RF results for datasets

n_{tree}	MIMIC-III				SMS spam collection			
	Accuracy	OOB error rate (%)	AUROC	CPU time	Accuracy	OOB error rate (%)	AUROC	CPU time
100	0.7663	23.17	0.904	27.67 secs	0.9085	9.89	0.968	53.9 secs
200	0.7641	23.29	0.907	51.81 secs	0.9097	9.78	0.969	1.49 min
500	0.7661	23.39	0.908	2.08 min	0.9183	9.09	0.970	3.99 min
1000	0.7671	23.09	0.908	4.14 min	0.9189	8.99	0.971	7.77 min
2000	0.7665	23.25	0.908	8.52 min	0.9188	9.07	0.971	15.34 min
5000	0.7661	23.39	0.909	23.94 min	0.9195	9.05	0.971	39.92 min
10000	0.7665	23.25	0.909	49.57 min	0.9289	8.60	0.971	1.38 h

Table 6 GB results for datasets

n_{tree}	MIMIC-III				SMS spam collection			
	Accuracy	OOB error rate (%)	AUROC	CPU time	Accuracy	OOB error rate (%)	AUROC	CPU time
100	0.7401	25.99	0.844	48.77 secs	0.8577	14.23	0.878	1.39 min
200	0.7471	25.29	0.845	59.71 secs	0.8602	13.98	0.881	2.06 min
500	0.7526	24.74	0.849	3.29 min	0.8609	13.91	0.881	4.28 min
1000	0.754	24.60	0.850	5.59 min	0.8576	14.24	0.877	8.37 min
2000	0.754	24.60	0.850	10.75 min	0.8571	14.29	0.877	17.44 min
5000	0.7989	20.11	0.875	25.69 min	0.8552	14.48	0.877	43.22 min
10,000	0.8101	18.99	0.878	56.69 min	0.8687	13.13	0.882	1.49 h

Table 7 LR results for MIMIC-III

Accuracy	MIMIC_III AUROC	CPU time
0.742	0.833	15.26 secs

When the Logistic Regression results are examined, this model obtained for the MIMIC-III data set is also found to be significant ($p < \alpha = 0.05$). The model is also found to be suitable ($p > \alpha = 0.05$). The Nagelkerke R2 value is determined as 0.679. The CPU time, accuracy and ROC curve results from the LR application for MIMIC-III are given in Table 7.

When the LR results for the SMS Spam data set are examined, this model is not found to be significant ($p > \alpha = 0.05$). The model is not found to be suitable ($p < \alpha = 0.05$).

Feature selection by CoClust

After editing the datasets, the dependency structures between variables are examined with CoClust. In the clustering step, all Gaussian and Archimedean copula families are used. Gumbel, Clayton and Frank families give results for the SMS Spam collection, while Clayton and Frank copula families give clustering results for MIMIC-III. Variables to be used in the RF technique are selected with CoClust.

Seven clusters obtained by clustering with the Clayton and Frank copulas for MIMIC-III are presented in Table 8, similar to classical clustering methods.

Similar clusters are observed in both the Clayton copula and Frank copula for MIMIC-III. In this context, the similarity of the five clusters has been determined. Common variables are not observed in the clusters determined to be different.

For the SMS Spam dataset, three clusters obtained with the Frank and Clayton copulas and five clusters obtained with the Gumbel copula are presented in Table 9.

The same clusters are obtained by Clayton and Frank copulas for SMS Spam Collection. Three of the five clusters obtained with the Gumbel copula family are identical to the clusters obtained from the Clayton and Frank copulas.

Table 8 CoClust clustering results for MIMIC-III

	Clayton			
Cluster 1	Hemoglobin	Hematocrit	Albumin	Coagulation
Cluster 2	Liver	Prothrombin Time	Int. Norm. Ratio	Part. Throm. Time
Cluster 3	Cardiovascular	Heart Rate	Glasgow Coma Scale	Cent. Nervous Sys
Cluster 4	Mean Arterial Pressure	Systolic Blood Pressure	Diastolic Blood Pressure	Mech. Respiration
Cluster 5	Bicarbonate	Sodium	Chloride	Platelet
Cluster 6	Respiration	PaO_2/FiO_2	SpO_2	Age
Cluster 7	White Blood Cells	Glucose	Temperature	Epinephrine
	Frank			
Cluster 1	Hemoglobin	Hematocrit	Albumin	Coagulation
Cluster 2	Diastolic Blood Pressure	Systolic Blood Pressure	Mean Arterial Pressure	Mech. Respiration
Cluster 3	Potassium	Blood Urea Nitrogen	Creatinine	Renal
Cluster 4	Platelet	Sodium	Chloride	Bicarbonate
Cluster 5	Age	PaO_2/FiO_2	SpO_2	Respiration
Cluster 6	Temperature	Glucose	White Blood Cells	Epinephrine
Cluster 7	Immature Neutr. Cells	Dopamine	Dobutamine	Norepinephrine

Table 9 CoClust clustering results for SMS Spam dataset

Frank and Clayton			
Cluster 1	v754	v275	v669
Cluster 2	v48	v316 v383	v590
Cluster 3	v182		v207
Gumbel			
Cluster 1	v754	v275	v669
Cluster 2	v48	v383	v590
Cluster 3	v182	v316	v207
Cluster 4	v231	v284	v604
Cluster 5	v472	v620	v8

Feature selection by other methods

After CoClust application, the relationship between variables are examined by K-means and hierarchical clustering methods. For MIMIC-III dataset, four clusters obtained with K-means clustering are presented in Table 10.

For MIMIC-III dataset, five clusters obtained with hierarchical clustering technique are presented in Table 11.

For the SMS Spam dataset, four clusters obtained with K-means are presented in Table 12.

For the SMS Spam dataset, four clusters obtained with hierarchical clustering technique are presented in Table 13.

Similar clusters are observed for both K-means and hierarchical clustering for the SMS Spam dataset. In both techniques, there are about 700 variables in Cluster 1.

Table 10 K-means clustering results for MIMIC-III dataset

K-means Clustering			
Cluster 1	**Cluster 2**	**Cluster 3**	**Cluster 4**
Hematocrit	Heart Rate	Age	Immature Neutrophil Cells
Hemoglobin	Systolic Blood Pressure	Gender	Dobutamine
International Normalized Ratio	Diastolic Blood Pressure	Temperature	
Norepinephrine	Mean Arterial Pressure	Bicarbonate	
Epinephrine	Respiration	Chloride	
Liver	SpO_2	Lactate	
	Glucose	Sodium	
	Albumin	Dopamine	
	Bilirubin	Mechanical Respiration	
	Creatinine	Coagulation	
	Platelet	Cardiovascular	
	Potassium	Central Nervous System	
	Partial Thromboplastin Time		
	Prothrombin Time		
	Blood Urea Nitrogen		
	White Blood Cells		
	Urine Output		
	PaO_2/FiO_2		
	Glasgow Coma Scale		
	Renal		

The results of the proposed approach with applying CoClust and RF with applying the other clustering methods

In this section, the feature selection step has been added to the RF technique by considering the dependency between variables in the datasets. All Gaussian and Archimedean copula families are used, but the clustering result cannot be obtained from all copula families in CoClust application. Although clustering is performed with the Clayton, Gumbel and Frank copula families for SMS SPAM, it is only possible with the Clayton and Frank copula families for the MIMIC-III dataset in CoClust application. Therefore, it is continued by using only clusters from these families.

First, RF application is performed in clusters obtained from the MIMIC-III dataset. The application is first examined in clusters obtained from the Clayton copula. The third and seventh clusters from the Frank copula are different from the Clayton copula clusters. Later, RF performance is measured in these clusters. After examining the results of the RF application of the clusters obtained with CoClust, the RF results of the clusters obtained with other clustering techniques are examined.

After the MIMIC-III application, RF application is made to the clusters obtained from the SMS SPAM Collection. Since the clusters obtained from the Gumbel copula also include clusters obtained from other copula families, the results are observed by applying RF to the clusters obtained from the Gumbel copula family. After examining

Table 11 Hierarchical clustering results for MIMIC-III dataset

Hiyerarchical clustering				
Cluster 1	**Cluster 2**	**Cluster 3**	**Cluster 4**	**Cluster 5**
Age	Heart Rate	Glucose	Platelet	Urine Output
Gender	Systolic Blood Pressure		PaO_2/FiO_2	
Diastolic Blood Pressure	SpO_2			
Mean Arterial Pressure	Chloride			
Respiration	Sodium			
Temperature				
Albumin				
Immature Neutrophil Cells				
Bicarbonate				
Bilirubin				
Creatinine				
Hematocrit				
Hemoglobin				
Lactate				
Potassium				
Partial Thromboplastin Time				
International Normalized Ratio				
Prothrombin Time				
Blood Urea Nitrogen				
White Blood Cells				
Dobutamine				
Glasgow Coma Scale				
Mechanical Respiration				
Coagulation				
Liver				
Cardiovascular				
Central Nervous System				
Renal				

Table 12 K-means clustering results for SMS Spam dataset

K-means clustering			
Cluster 1	**Cluster 2**	**Cluster 3**	**Cluster 4**
And other variables	v46	v72	v313
	v92	v205	v503
	v109	v214	v752
	v142	v496	v755
	v273	v588	v767
	v308		
	v368		
	v667		

the results of the RF application of the clusters obtained with CoClust, the RF results of the clusters obtained with other clustering techniques are examined.

The CPU time results of RF with CoClust clusters are shown in Table 14.

Table 13 Hierarchical clustering results for SMS Spam dataset

Hierarchical clustering			
Cluster 1	**Cluster 2**	**Cluster 3**	**Cluster 4**
And other variables	v46	v205	v752
	v109	v588	
	v142		
	v273		
	v308		
	v667		

Table 14 The CPU time results of the proposed approach

CPU time						
n_{tree}	CoClust		K-Means		Hierarchical	
	MIMIC-III	SMS Spam Collection	MIMIC-III	SMS Spam Collection	MIMIC-III	SMS Spam Collection
100	4.01 secs	1.36 secs	5.28 secs	1.21 min	5.07 secs	1.01 min
200	6.79 secs	1.836 secs	8.14 secs	2.01 min	8.56 secs	2.00 min
500	16.00 secs	3.17 secs	19.25 secs	3.48 min	19.12 secs	3.59 min
1000	30.45 secs	5.16 secs	1.27 min	7.46 min	1.36 min	7.58 min
2000	1.02 min	8.48 secs	3.17 min	17.02 min	3.55 min	16.24 min
5000	2.52 min	20.55 secs	6.01 min	45.26 min	5.06 min	45.02 min
10000	5.07 min	40.17 secs	11.02 min	1.48 h	10.22 min	1.32 h

In the previous section, it is determined that the 1000-tree forests give the most efficient result in the RF applications without the feature selection step, and the applications are completed in 4.14 and 7.77 min. At this stage, the 1000-tree forests are formed in 30.45 and 5.16 s with CoClust clustering.

On the other hand, 1.27 and 1.36 min are recorded in K-means and hierarchical clustering techniques for MIMIC-III. For SMS Spam Collection dataset, almost no progress could be made with 7.46 and 7.58 min. Here we see a significant efficient result of CoClust clustering only related variables. Both K-means and hierarchical clustering techniques cluster all data, unlike CoClust. Here we see an important contribution from CoClust. These are remarkable improvements.

In the MIMIC-III dataset, there are similar clusters in two families except for the second and third clusters from the Clayton copula and the third and seventh clusters from the Frank copula. The accuracy, error rate, and ROC curve values of 1000-tree forests give the same result for similar clusters found in both the Clayton copula and Frank copula. The accuracy, error and AUROC results of the clusters belonging to the MIMIC-III dataset are given in Table 15.

When Table 12 is examined, it can be said that when the highest accuracy and the lowest error rate are selected, Cluster 3 from the Clayton copula and Cluster 7 from the Frank copula are the most efficient results. Efficient results can be obtained in the prediction of mortality by using the variables in these clusters. The ROC curve results of RF are also an important step for validation.

Table 15 Results of RF with CoClust for MIMIC-III

		Accuracy	OOB error rate (%)	AUROC
Clayton Copula	Cluster 1	0.8453	15.47	0.911
	Cluster 2	0.8667	13.33	0.949
	Cluster 3	0.9343	6.57	0.987
	Cluster 4	0.8355	16.45	0.915
	Cluster 5	0.7963	20.37	0.883
	Cluster 6	0.8254	17.46	0.894
	Cluster 7	0.8269	17.31	0.895
Frank Copula	Cluster 1	0.8453	15.47	0.911
	Cluster 2	0.8355	16.45	0.915
	Cluster 3	0.7940	20.60	0.900
	Cluster 4	0.7963	20.37	0.883
	Cluster 5	0.8254	17.46	0.894
	Cluster 6	0.8269	17.31	0.895
	Cluster 7	0.9436	5.64	0.992

The accuracy, error and AUROC results of the K-means clusters belonging to the MIMIC-III dataset are given in Table 16. When the results are examined, Cluster 2 and 3 give the most efficient results when the highest accuracy and lowest error rate are selected.

The accuracy, error and AUROC results of the hierarchical clusters belonging to the MIMIC-III dataset are given in Table 17. When the results are examined, Cluster 1 gives the most efficient results when the highest accuracy and lowest error rate are selected.

In the SMS Spam Collection, the first three of the clusters obtained from the Gumbel copula family are exactly the same as the clusters obtained from the Frank and Clayton copula families. The accuracy, error rate, and ROC curve values of 1000-tree forests give the same result for similar clusters. The accuracy, error and AUROC results of the clusters belonging to the SMS Spam Collection dataset are given in Table 18.

Table 16 Results of RF with K-means clustering for MIMIC-III

		Accuracy	OOB error rate (%)	AUROC
K-means	Cluster 1	0.7589	24.11	0.849
	Cluster 2	0.7601	23.99	0.850
	Cluster 3	0.7621	23.79	0.850
	Cluster 4	0.7598	24.02	0.849

Table 17 Results of RF with hierarchical clustering for MIMIC-III

		Accuracy	OOB error rate (%)	AUROC
Hierarchical	Cluster 1	0.7562	24.38	0.849
	Cluster 2	0.7079	29.21	0.810
	Cluster 3	0.7032	29.68	0.810
	Cluster 4	0.7056	29.44	0.810
	Cluster 5	0.7069	29.31	0.810

Table 18 Results of RF with CoClust for SMS Spam Collection

		Accuracy	OOB error rate (%)	AUROC
Frank and Clayton Copula	Cluster 1	0.9803	2.60	0.974
	Cluster 2	0.982	2.45	0.972
	Cluster 3	0.984	2.27	0.963
Gumbel Copula	Cluster 1	0.9803	2.60	0.974
	Cluster 2	0.982	2.45	0.972
	Cluster 3	0.984	2.27	0.963
	Cluster 4	0.9838	2.27	0.982
	Cluster 5	0.9838	2.24	0.987

When Table 15 is examined, it can be said that when the highest accuracy and the lowest error rate are selected, Cluster 3 from the Gumbel copula (also Frank and Clayton copula) and Cluster 5 from the Gumbel copula are the most efficient results. Efficient results can be obtained in the prediction of spam messages by using the variables in these clusters. The ROC curve results of RF are also an important step for validation.

The accuracy, error and AUROC results of the K-means clusters belonging to SMS Spam Collection are given in Table 19. When the table is examined, Cluster 1 gives the most efficient results when the highest accuracy and lowest error rate are selected.

The accuracy, error and AUROC results of the hierarchical clusters belonging to SMS Spam Collection are given in Table 20. When the results are examined, Clusters 1 gives the most efficient results when the highest accuracy and lowest error rate are selected.

As a result, while a significant efficiency is achieved in CPU time with clusters obtained from CoClust, it cannot be said about the same for K-means and hierarchical clustering results. All models from CoClust clusters work quite well. These results obtained by using CoClust and RF together are remarkable.

When the clustering techniques are examined in terms of accuracy, the most efficient result is obtained from the clustering with CoClust. As a result of clustering with CoClust, while accuracy and ROC values increased for both data sets, OOB error rates decrease. On the other hand, there is no positive or negative improvement in accuracy

Table 19 Results of RF with K-means clustering for SMS Spam Collection

		Accuracy	OOB error rate (%)	AUROC
K-means	Cluster 1	0.8887	11.13	0.889
	Cluster 2	0.8787	12.13	0.873
	Cluster 3	0.8711	12.89	0.871
	Cluster 4	0.8715	12.85	0.872

Table 20 Results of RF with hierarchical clustering for SMS Spam Collection

		Accuracy	OOB error rate (%)	AUROC
Hierarchical	Cluster 1	0.8815	11.85	0.878
	Cluster 2	0.8789	12.11	0.873
	Cluster 3	0.8800	12.00	0.878
	Cluster 4	0.8788	12.12	0.873

and other evaluation criteria for the MIMIC-III data set in RF applied with K-means and hierarchical clustering results. The decrease in accuracy and ROC values for SMS Spam Collection is remarkable. When both CPU time development and model selection criteria are examined, it is seen that only clustering with CoClust yields efficient results.

At the stage of choosing the best model for the MIMIC-III dataset, one model from both the Frank copula and Clayton copula is chosen. When the selected clusters are examined, there are "*Cardiovascular, Heart Rate, Glasgow Coma Scale, Central Nervous System*" variables in the third cluster and "*Immature Neutrophil Cells, Dopamine, Dobutamine, Norepinephrine*" variables in the seventh cluster.

The SMS Spam Collection dataset is a fairly large dataset working with 770 variables. Although a large number of variables is very useful in classification, it has been seen that effective and efficient classification results can be obtained by using only three variables with the application. Successful classification models can be obtained by using the *v182, v316, and v207* variables in the third cluster and the *v472, v620, and v8* variables in the fifth cluster.

Gradient boosting and logistic regression results for clustering methods

In this section, firstly, the results obtained with applying CoClust, K-means and hierarchical clustering techniques for Gradient Boosting are examined.

The CPU time results of GB with CoClust clusters are shown in Table 21. When Table 6 and Table 21 are evaluated together, the most efficient result is obtained with CoClust, even though the CPU time for all clustering techniques decreases compared to the GB application without clustering methods.

The accuracy, error and AUROC results of the CoClust clusters belonging to the MIMIC-III dataset are given in Table 22.

When Table 22 is examined, it can be said that when the highest accuracy and the lowest error rate are selected, Cluster 1 from the Clayton copula and the Frank copula are the most efficient results. Efficient results can be obtained in the prediction of mortality by using the variables in these clusters.

The accuracy, error and AUROC results of the K-means clusters belonging to the MIMIC-III dataset are given in Table 23. When Table 23 is examined, Cluster 2 gives the most efficient results when the highest accuracy and lowest error rate are selected.

Table 21 The CPU time results of GB application

CPU time						
n_{tree}	CoClust		K-Means		Hierarchical	
	MIMIC-III	SMS Spam Collection	MIMIC-III	SMS Spam Collection	MIMIC-III	SMS Spam Collection
100	5.43 secs	2.04 secs	6.05 secs	1.31 min	5.27 secs	1.21 min
200	7.88 secs	3.36 secs	9.16 secs	1.59 min	9.06 secs	2.27 min
500	20.03 secs	5.16 secs	20.25 secs	3.58 min	19.36 secs	3.45 min
1000	1.04 min	7.46 secs	2.58 min	9.25 min	2.58 min	9.05 min
2000	3.45 min	8.48 secs	4.55 min	16.37 min	4.55 min	16.47 min
5000	5.36 min	20.55 secs	6.09 min	47.39 min	6.09 min	45.28 min
10000	11.26 min	40.17 secs	11.22 min	1.17 h	11.22 min	1.35 h

Table 22 Results of GB with CoClust for MIMIC-III

		Accuracy	OOB error rate (%)	AUROC
Clayton Copula	Cluster 1	0.8125	18.75	0.883
	Cluster 2	0.7625	23.75	0.852
	Cluster 3	0.7611	23.89	0.852
	Cluster 4	0.7614	23.86	0.852
	Cluster 5	0.7584	24.16	0.849
	Cluster 6	0.7579	24.21	0.849
	Cluster 7	0.7600	24.00	0.863
Frank Copula	Cluster 1	0.8125	18.75	0.883
	Cluster 2	0.7614	23.86	0.852
	Cluster 3	0.7589	24.11	0.849
	Cluster 4	0.7584	24.16	0.849
	Cluster 5	0.7579	24.21	0.849
	Cluster 6	0.7600	24.00	0.863
	Cluster 7	0.7528	24.72	0.849

Table 23 Results of GB with K-means clustering for MIMIC-III

		Accuracy	OOB error rate (%)	AUROC
K-means	Cluster 1	0.7479	25.21	0.839
	Cluster 2	0.7591	24.09	0.850
	Cluster 3	0.7541	24.59	0.849
	Cluster 4	0.7588	24.12	0.850

Table 24 Results of GB with hierarchical clustering for MIMIC-III

		Accuracy	OOB error rate (%)	AUROC
Hierarchical	Cluster 1	0.711	28.89	0.811
	Cluster 2	0.7054	29.46	0.809
	Cluster 3	0.6987	30.13	0.809
	Cluster 4	0.7048	29.52	0.809
	Cluster 5	0.6994	30.06	0.809

The accuracy, error and AUROC results of the hierarchical clusters belonging to the MIMIC-III dataset are given in Table 24. When the results are examined, Cluster 1 gives the most efficient results when the highest accuracy and lowest error rate are selected.

The accuracy, error and AUROC results of the CoClust clusters belonging to the SMS Spam Collection are given in Table 25.

The accuracy, error and AUROC results of the K-means clusters belonging to SMS Spam Collection are given in Table 26. According to the results, Cluster 1 gives the most efficient results when the highest accuracy and lowest error rate are selected.

The accuracy, error and AUROC results of the hierarchical clusters belonging to SMS Spam Collection are given in Table 27. When Table 27 is examined, Cluster 1 gives the most efficient results when the highest accuracy and lowest error rate are selected.

Table 25 Results of GB with CoClust for SMS Spam Collection

		Accuracy	OOB error rate (%)	AUROC
Frank and Clayton Copula	Cluster 1	0.8903	10.97	0.884
	Cluster 2	0.892	10.80	0.885
	Cluster 3	0.894	10.60	0.884
Gumbel Copula	Cluster 1	0.8903	10.97	0.884
	Cluster 2	0.892	10.80	0.885
	Cluster 3	0.894	10.60	0.884
	Cluster 4	0.883	11.70	0.879
	Cluster 5	0.883	11.70	0.879

Table 26 Results of RF with K-means clustering for SMS Spam Collection

		Accuracy	OOB error rate (%)	AUROC
K-means	Cluster 1	0.8517	14.83	0.877
	Cluster 2	0.8499	15.01	0.874
	Cluster 3	0.8498	15.02	0.874
	Cluster 4	0.8475	15.25	0.872

Table 27 Results of RF with hierarchical clustering for SMS Spam Collection

		Accuracy	OOB error rate (%)	AUROC
Hierarchical	Cluster 1	0.8101	18.99	0.883
	Cluster 2	0.7989	20.11	0.879
	Cluster 3	0.8005	19.95	0.881
	Cluster 4	0.7980	20.20	0.879

The accuracy, error and AUROC results of the CoClust clusters belonging to the MIMIC-III dataset are given in Table 28.

The accuracy, error and AUROC results of the K-means clusters belonging to the MIMIC-III dataset are given in Table 29. When the results are examined, Cluster 2 gives the most efficient results when the highest accuracy and lowest error rate are selected.

The accuracy, error and AUROC results of the hierarchical clusters belonging to the MIMIC-III dataset are given in Table 30. According to results, Cluster 1 gives the most efficient results when the highest accuracy and lowest error rate are selected.

The accuracy, error and AUROC results of the CoClust clusters belonging to the SMS Spam Collection are given in Table 31.

The accuracy, error and AUROC results of the K-means clusters belonging to SMS Spam Collection are given in Table 32. The model belonging to the first cluster obtained by K-means clustering is not found significant and appropriate. Cluster 2 gives the most efficient results when the highest accuracy and lowest error rate are selected.

The accuracy, error and AUROC results of the hierarchical clusters belonging to SMS Spam Collection are given in Table 33. The model belonging to the first cluster obtained by hierarchical clustering is not found significant and appropriate. When the results are

Table 28 Results of LR with CoClust for MIMIC-III

		Accuracy	AUROC
Clayton Copula	Cluster 1	0.704	0.801
	Cluster 2	0.702	0.800
	Cluster 3	0.693	0.798
	Cluster 4	0.689	0.789
	Cluster 5	0.690	0.799
	Cluster 6	0.692	0.799
	Cluster 7	0.688	0.788
Frank Copula	Cluster 1	0.704	0.801
	Cluster 2	0.689	0.799
	Cluster 3	0.709	0.800
	Cluster 4	0.690	0.799
	Cluster 5	0.692	0.799
	Cluster 6	0.688	0.789
	Cluster 7	0.687	0.788

Table 29 Results of LR with K-means clustering for MIMIC-III

		Accuracy	AUROC
K-means	Cluster 1	0.713	0.811
	Cluster 2	0.725	0.826
	Cluster 3	0.717	0.812
	Cluster 4	0.687	0.809

Table 30 Results of LR with hierarchical clustering for MIMIC-III

		Accuracy	AUROC
Hierarchical	Cluster 1	0.737	0.845
	Cluster 2	0.698	0.809
	Cluster 3	0.687	0.809
	Cluster 4	0.687	0.809
	Cluster 5	0.689	0.809

Table 31 Results of LR with CoClust for SMS Spam Collection

		Accuracy	AUROC
Frank and Clayton Copula	Cluster 1	0.806	0,801
	Cluster 2	0.874	0,881
	Cluster 3	0.894	0.884
Gumbel Copula	Cluster 1	0.806	0,801
	Cluster 2	0.874	0,881
	Cluster 3	0.894	0.884
	Cluster 4	0.832	0.829
	Cluster 5	0.861	0.879

Table 32 Results of LR with K-means clustering for SMS Spam Collection

		Accuracy	AUROC
K-means	Cluster 1	–	–
	Cluster 2	0.866	0.881
	Cluster 3	0.859	0.878
	Cluster 4	0.843	0.869

Table 33 Results of LR with hierarchical clustering for SMS Spam Collection

		Accuracy	AUROC
Hierarchical	Cluster 1	–	–
	Cluster 2	0.866	0.881
	Cluster 3	0.866	0.881
	Cluster 4	0.866	0.881

examined, all clusters give the most efficient results when the highest accuracy and lowest error rate are selected.

The RF results of the obtained models and their comparison with each other are examined in the following section.

Comparison of the results of the proposed approach with the results of other methods

To show the applicability of the new approach, the results obtained from the application of the proposed approach are compared with the results of the normal Random Forest, Gradient Boosting and Logistic Regression. The results obtained are compared primarily in terms of accuracy, OOB error rate and AUROC values, and the progress achieved is examined. The other comparison is made in terms of CPU time.

Firstly, RF results with CoClust are examined. The improvement achieved in terms of accuracy, error rate and ROC curve in is quite remarkable. In accuracy, an increase of up to 0.12 is observed in the MIMIC-III dataset, while an increase of 0.06 is observed in the SMS Spam collection dataset.

The CPU time results and the improvement observed in the results of RF with CoClust are given in Table 34. While the efficiency obtained in CPU time is approximately 85% and 97% in an application with 100-tree, it reaches 90% and 99% in an application with 10000-tree.

When the CPU time improvement for both datasets is examined in the results of Random Forest with CoClust, it is 87.79% for the first dataset and 98.63% for the second dataset in all forests. The closest result is obtained in applications with 1000-tree in both datasets. This result will significantly contribute to the time constraint problem, especially in big data. When analyzed with artificial intelligence, as the number of parameters in machine learning increases, the speed decreases. Since this increases the time, the researcher goes to reduce the number of trees. However, here, we see that the desired result can be achieved by increasing the accuracy without reducing the number of trees. Moreover, it is possible to eliminate the time constraint. Instead of working with

Table 34 CPU time comparison of CoClust and Random Forest applications

	n_{tree}	Traditional RF	The Proposed RF	% of improvement
MIMIC-III	100	27.67 secs	4.01 secs	85.51
	200	51.81 secs	6.79 secs	86.89
	500	2.08 min	16.00 secs	87.18
	1000	4.14 min	30.45 secs	87.74
	2000	8.52 min	1.02 min	88.03
	5000	23.94 min	2.52 min	89.47
	10,000	49.57 min	5.07 min	89.77
SMS Spam Collection	100	53.9 secs	1.36 secs	97.48
	200	1.49 min	1.836 secs	97.95
	500	3.99 min	3.17 secs	98.68
	1000	7.77 min	5.16 secs	98.89
	2000	15.34 min	8.48 secs	99.08
	5000	39.92 min	20.55 secs	99.14
	10,000	1.38 h	40.17 secs	99.19

100 trees in traditional RF, it will be able to work with 10000 trees at the same time using the proposed method.

The change in CPU time of CoClust and RF application in MIMIC-III dataset is given in Fig. 2 below.

The change in CPU time of CoClust and RF application in SMS Spam Collection is given in Fig. 3 below.

Adding the feature selection step considerably reduces the duration of the application. Achieving results in a shorter CPU time with accuracy and error results in the application is just as important. Especially as the number of trees increases, the decrease in CPU time becomes even more striking.

In addition to obtaining a result in a short CPU time, improvement in accuracy and error results is also very important in terms of analysis. As a result of RF applied to clusters, accuracy results up to 0.9436 and 0.9840 are obtained. These striking results can be

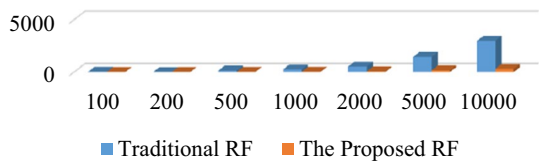

Fig. 2 CPU time results of RF with CoClust applications in MIMIC-III

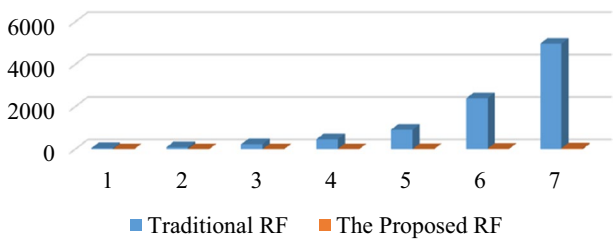

Fig. 3 CPU time results of RF with CoClust applications in SMS Spam Collection

seen in Tables 11 and 12. The results of 0.992 and 0.987 obtained in ROC curve values are especially important for mortality prediction and spam message classification.

When GB results are compared according to clustering techniques, the highest accuracy and ROC values for both datasets are obtained from CoClust results. While the accuracy reaches 0.81 for MIMIC-III, it reaches 0.89 for SMS Spam Collection, which has a large number of variables. On the other hand, RF results for this technique are found to be more efficient than GB results. Accordingly, the efficiency of the proposed method is remarkable.

However, it is clear that this fruitful result do not come from simply adding the variable selection step. When K-means and hierarchical clustering techniques are examined, the efficiency of the results obtained from CoClust is remarkable.

When we compare it according to LR results, although K-means clustering for MIMIC-III gives more efficient results in terms of accuracy and ROC results than CoClust, we cannot say the same for SMS Spam Collection. For the second data set, it could not give a meaningful and appropriate model in Cluster 1. It is not enough for a technique to give positive results in one place.

When we examine the LR results for CoClust clusters, a decrease is observed in accuracy and ROC values for MIMIC-III. On the other hand, an increase is observed in the criteria sought for SMS Spam Collection, which includes a large number of variables. All the models obtained are found to be significant and appropriate.

Discussion

The main goal of the study is to increase the prediction power while reducing the application CPU time by adding a novel feature selection step to RF. As seen in the results obtained, the study has reached its aim. CoClust turns out to be a highly effective method.

When the results of RF application without adding CoClust are examined for both datasets, the most efficient result is obtained from a forest of 1000 trees according to accuracy, error rate and ROC curve values. In the 1000-tree application, a CPU time of 4.14 and 7.77 min was reached. However, when the feature selection step is added with CoClust, the time required for 1000-tree decreases to 30.45 s for the first dataset and to 5.16 s for the other. This result is important and remarkable.

In the MIMIC-III dataset, an 85% reduction in CPU time is observed for an application with 100 trees, while the reduction reaches approximately 90% when the number of trees reaches 10000. In the SMS Spam Collection dataset, this decrease reaches up to 99%, making a large difference. Since the modeling is carried out on fewer variables, the analysis CPU time is considerably shortened. This is a very important development.

The accuracy, OOB error rate and AUROC results have also been carefully studied, as they are not enough to reach a solution in a short CPU time. There is also a visible improvement in accuracy, OOB error rate and AUROC, which are the most important result information. A model proposal for mortality prediction can also be made here. On the other hand, it is a very important development to accurately classify spam messages with three variables.

In the MIMIC-III dataset, CoClust selects 4 variables out of 40 variables and forms clusters. In the study performed on a dataset of 25800 people, the average processing

CPU time was determined to be 4.01 s for 100 trees and 30.45 s for 1000 trees. The lowest AUROC value is determined to be 0.883. In the SMS Spam collection dataset, it successfully realizes the classification estimation by decreasing from 770 to 3 variables. In this dataset, an application with 100 trees needs 1.36 s, while it takes approximately 40 s to complete an application with 10000 trees. While completing the classification in such a short time, the ROC curve value reaches 99%. It is observed that the lowest ROC curve value obtained in our study is quite good compared to the ROC value obtained in the study of Zhu et al. [27].

In this study, CPU time improvement was between 85.51 and 99.19% in all forests. For an application with 10000 trees in MIMIC-III, this efficiency reaches 90%, while in the other dataset, it reaches 99%. This is a very serious development in today's big data age because, when analyzing with artificial intelligence, as the number of parameters in machine learning increases but speed decreases. Since this increases the time, the researcher goes to reduce the number of trees. However, we see here that the number of iterations can be increased without compromising accuracy by being afraid of time constraints. With this proposed method, while bringing a new perspective to traditional RF, researchers are provided with the opportunity to reach higher accuracy in the same CPU time.

The fact that CoClust works efficiently in nonlinear dependencies and in the field of health has also contributed greatly to the RF step. Thus, a model proposal could be made for mortality prediction. A mortality prediction to be carried out through variables in the third cluster of Clayton's copula and the seventh cluster of Frank copula yields efficient results.

When K-means and hierarchical clustering techniques and CoClust cluster results are compared, the most efficient results in terms of both accuracy and CPU time are obtained from CoClust clusters. CoClust creates balanced clusters because it does not include the uncorrelated variable in clustering. This is one of the important differences between other methods. On the other hand, the clusters obtained by these clustering techniques are examined with GB and LR classification methods as well as RF. Again, the most efficient results wae obtained in the classification made with RF.

The results obtained from the clustering stage with CoClust, which is one of the important steps of the study, are carefully examined. This step is very important both for the development of CoClust and the next modeling phase with RF. The most efficient results are obtained in the third cluster from the Clayton copula and the seventh cluster from the Frank copula.

Frank Copula is symmetrical, and Clayton Copula is an asymmetrical copula family. The Clayton copula family is an asymmetrical Archimedean copula that examines dependency in the lower (left) tail, and the Frank copula family is a symmetrical Archimedean copula. It shows that the CoClust technique differs from other techniques by offering solutions with both asymmetrical and symmetrical approaches. For this reason, CoClust is thought to be working more efficiently when performing dependency research in tails [55].

When the validity of the models obtained by modeling the variables in the clusters is examined, high validity results are obtained in all of them. The importance of putting correlated variables into modeling is remarkable. The results of the models obtained are

satisfactory because of the importance of working with high accuracy in mortality risk studies.

According to accuracy, OOB est. of error rate and ROC curves, two clusters are selected from the Clayton and Frank copula families. Cardiovascular, heart rate, Glasgow Coma Scale, and central nervous system variables are included in Cluster 3 from the Clayton copula. Immature neutrophil cells, dopamine, dobutamine, and norepinephrine variables are included in Cluster 7 from the Frank copula.

It is noteworthy that the variables in the selected clusters are also emphasized in the literature. The relationship between heart rate variability and patient coma status and the Glasgow Coma Scale value was revealed. A notable reduction in heart rate is found in patients according to the Glasgow Coma Scale [56]. The purpose of Cooke et al. [57]'s study was to assess heart rate variability and its association with mortality in prehospital trauma patients. They also used the Glasgow Coma Scale values in this study, and the relationship of these variables with mortality in trauma patients was examined.

Wan-Ting et al. [58], Hekmat et al. [59] and Hasanin et al. [60] examine the relationship between heart rate and cardiovascular and Glasgow Coma Scale variables and mortality risk in adult severe trauma and cardiac patients.

Baser et al. [61] investigated the relationship between neutrophil cells and mortality risk prediction and emphasized that vasopressors are used in patients who survive.

On the other hand, with the rapid increase in big data in the field of technology, data management becomes more difficult. In this context, it is very important to classify more accurately with fewer variables. With only 3 variables instead of 770 variables in the SMS Spam Collection dataset, it is very valuable to reach spam classification in a very short time.

The results obtained in areas such as technology and health, where it is very important to make the right decision quickly, are striking. Just as fast and accurate estimation of mortality is important in healthcare, fast and accurate classification of spam messages is very important in the age of technology. The high validity results obtained from each of the models clearly show the importance of using correlated variables in modeling.

Conclusion

According to the results obtained, the use of RF and CoClust together improves CPU time and prediction. In addition, CoClust produces groups of uncorrelated variables where interpretation becomes easier for practitioners, especially for medicine data with highly correlated factors.

It has been shown that the proposed methodology works well for different types of big data. This fact can be easily generalized to the case that significant improvements in artificial intelligence should be possible. For researchers, it has been possible to eliminate the constraint of decreasing speed with significantly increasing learning**.**

CoClust's ability to select variables should continue to be rigorously examined in future studies. Examining different data types and their behaviors in machine learning techniques and making them applicable in practice will both facilitate researchers in the age of big data and lead to other variable selection methods.

Acknowledgements
Not applicable.

Author contributions
The three authors contributed to the planning, design and writing of the manuscript. The conception of the definitions, the results and corresponding proofs were regularly discussed by the three authors. The first draft of the manuscript was written by ZI, and all authors read and commented on each version and future directions to take. The final version was read and approved by all authors. ZI is also correspondence author. All authors read and approved the final manuscript.

Funding
Not applicable.

Availability of data and materials
Since MIMIC-III is provided individually by MIT, the data set will not be shared. But here is the reference link: https://physionet.org/content/mimiciii/1.4/. The SMS Spam dataset is available here: https://archive.ics.uci.edu/ml/machine-learning-databases/00228/.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Received: 27 April 2022 Accepted: 21 March 2023
Published online: 30 March 2023

References

1. Darwiche Aiman A. 2018. "Machine learning methods for septic shock prediction." PhD Thesis, Nova Southeastern University. Retrieved from NSUWorks, College of Engineering and Computing. (1051) https://nsuworks.nova.edu/gscis_etd/1051
2. Lee J. Patient-specific predictive modeling using random forests: an observational study for the critically Ill. JMIR Med Informat. 2017. https://doi.org/10.2196/medinform.6690.
3. Levantesi S, Nigri A. A random forest algorithm to improve the Lee-carter mortality forecasting: impact on q-forward. Soft Comput. 2020;24(12):8553–67. https://doi.org/10.1007/s00500-019-04427-z.
4. McWilliams CJ, et al. Towards a decision support tool for ıntensive care discharge: machine learning algorithm development using electronic healthcare data from MIMIC-III and Bristol, UK. BMJ Open. 2019. https://doi.org/10.1136/bmjopen-2018-025925.
5. Mistry P, Neagu D, Trundle PR, Vessey JD. Using random forest and decision tree models for a new vehicle prediction approach in computational toxicology. Soft Comput. 2016;20(8):2967–79. https://doi.org/10.1007/s00500-015-1925-9.
6. Van Poucke S, Kovacevic A, Vukicevic M. Early prediction of patient mortality based on routine laboratory tests and predictive models in critically Ill patients. In Data Mining InTech. 2018. https://doi.org/10.5772/intechopen.76988.
7. Breiman L. Bagging predictors. Mach Learn. 1996;24:123–40. https://doi.org/10.1007/BF00058655.
8. Dietterich TG. An experimental comparison of three methods for constructing ensembles of decision trees: bagging, boosting, and randomization. Mach Learn. 2000;40:139–57. https://doi.org/10.1023/A:1007607513941.
9. Ho K. The random subspace method for constructing decision forests. IEEE Trans Pattern Anal Mach Intell. 1998;20(8):832–44. https://doi.org/10.1109/34.709601.
10. Breiman L. "Using Adaptive Bagging To Debias Regressions." Technical Report 547. Berkeley: University of California at Berkeley; 1999.
11. Breiman L. Random forests. Mach Learn. 2001;45:5–32. https://doi.org/10.1023/A:1010933404324.
12. Shi T, et al. Tumor classification by tissue microarray profiling: random forest clustering applied to renal cell carcinoma. Mod Pathol. 2005;18(4):547–57. https://doi.org/10.1038/modpathol.3800322.
13. Shi T, Horvath S. Unsupervised learning with random forest predictors. J Comput Graph Stat. 2006;15(1):118–38.
14. Hapfelmeier A, Ulm K. Variable selection by random forests using data with missing values. Comput Stat Data Anal. 2014;80:129–39. https://doi.org/10.1016/j.csda.2014.06.017.
15. Uddin Taufeeq, Azher Uddin. 2015. "A guided random forest based feature selection for activity recognition." In 2nd Int'l Conf. On electrical engineering and ınfonnation & communication technology (ICEEICT). https://doi.org/10.1109/ICEEICT.2015.7307376
16. Gupta Chelsi. 2019. "Feature selection and analysis for standard machine learning of audio beehive samples." Msc Thesis, Utah State University. https://digitalcommons.usu.edu/etd/7564.
17. Sklar A. Fonctions de repartition á n dimensions et leurs marges. Publications de l'Institut Statistiquede l'Université de Paris. 1959;8:229–31.
18. Nelsen RB. An ıntroduction to copulas. 2nd ed. Berlin: Springer Science & Business Media; 2006.
19. Jaworski Piotr, Fabrizio Durante, Wolfgang Hardle, Tomasz Rychlik. 2009. "Copula Theory And Its Applications." Proceedings of the Workshop Held in Warsaw, 25–26. https://doi.org/10.1007/978-3-642-12465-5
20. Mesiar R, Sheikhi A. Nonlinear random forest classification, a copula-based approach. Appl Sci. 2021;11:7140. https://doi.org/10.3390/app11157140.

21. Di Lascio, Francesca Marta Lilja. 2008. "Analyzing the dependence structure of microarray data: a copula-based approach." PhD Thesis, University of Bologna.
22. Johnson AEW, Mark RG. Real-time mortality prediction in the ıntensive care unit. AMIA Annu Symp Proc. 2018;2017:994–1003.
23. Lascio Di, Lilja FM, Giannerini S. A copula-based algorithm for discovering patterns of dependent observations. J Classif. 2012;29(1):50–75. https://doi.org/10.1007/s00357-012-9099-y.
24. Lascio Di, Lilja FM, Giannerini S. Clustering dependent observations with copula functions. Stat Pap. 2019;60(1):35–51. https://doi.org/10.1007/s00362-016-0822-3.
25. Khan YA, Shan QS, Liu Q, Abbas SZ. A nonparametric copula-based decision tree for two random variables using MIC as a classification index. Soft Comput. 2021;25(15):9677–92. https://doi.org/10.1007/s00500-020-05399-1.
26. Eling M, Toplek D. Modeling and management of nonlinear dependencies-copulas in dynamic financial analysis. J Risk Insur. 2009;76:651–81. https://doi.org/10.1111/j.1539-6975.2009.01318.x.
27. Zhu Y, et al. Machine learning prediction models for mechanically ventilated patients analyses of the MIMIC-III database. Front Med. 2021;8:662340. https://doi.org/10.3389/fmed.2021.662340.
28. Khope SR, Elias S. Critical correlation of predictors for an efficient risk prediction framework of ICU patient using correlation and transformation of MIMIC-III dataset. Data Sci Eng. 2022;7:71–86. https://doi.org/10.1007/s41019-022-00176-6.
29. Frees EW, Valdez EA. Understanding relationships using copulas. North Am Actuar J. 1998;2(3):1–25. https://doi.org/10.1080/10920277.1998.10595667.
30. Joe H, Xu JJ. The estimation method of ınference functions for margins for multivariate models. Vancouver: University of British Columbia; 1996.
31. Genest C, Ghoudi K, Rivest L-P. A semiparametric estimation procedure of dependence parameters in multivariate families of distributions. Biometrika. 1995;82(3):543–52.
32. Lascio Di, Lilja FM, Durante F, Pappada R. Copulas and dependence models with applications. in copulas and dependence models with applications. Berlin: Springer International Publishing; 2017;49–65.
33. Lascio Di FML, Disegna M. A copula-based clustering algorithm to analyse EU country diets, Knowledge-Based Systems. 2017;132:72–84. https://doi.org/10.1016/j.knosys.2017.06.004
34. Xue Ji, Yang B, Tang Q. Seabed sediment classification using multibeam backscatter data based on the selecting optimal random forest model. Appl Acoust. 2020;167:108387. https://doi.org/10.1016/j.apacoust.2020.107387.
35. Rivest RL, Hellman ME, Anderson JC. Responses to NIST's proposal. Commun ACM. 1992;35(7):41–54. https://doi.org/10.1145/129902.129905.
36. Gray KR, et al. Random forest-based similarity measures for multi-modal classification of Alzheimer's disease. Neuroimage. 2013;65:167–75. https://doi.org/10.1016/j.neuroimage.2012.09.065.
37. Qiu Z, Qin C, Jiu M, Wang X. A simple iterative method to optimize protein-ligand-binding residue prediction. J Theor Biol. 2013;317:219–23. https://doi.org/10.1016/j.jtbi.2012.10.028.
38. Friedman Jerome, Trevor Hastie, Robert Tibshirani. 2008. The elements of statistical learning preface to the second edition.
39. Sonam G, Jamal S, Open source drug discovery consortium, and Vinod Scaria. "Cheminformatics models for inhibitors of Schistosoma Mansoni Thioredoxin glutathione reductase." Sci World J. 2014. https://doi.org/10.1155/2014/957107.
40. Archer KJ, Kimes RV. Empirical characterization of random forest variable importance measures. Comput Stat Data Anal. 2008;52(4):2249–60. https://doi.org/10.1016/j.csda.2007.08.015.
41. Li BK, et al. Modeling, predicting and virtual screening of selective inhibitors of MMP-3 and MMP-9 over MMP-1 using random forest classification. Chemom Intell Lab Syst. 2015;147:30–40. https://doi.org/10.1016/j.chemolab.2015.07.014.
42. Jamal S, Scaria V. Cheminformatic models based on machine learning for pyruvate kinase ınhibitors of leishmania mexicana. BMC Bioinformatics. 2013;14(1):329. https://doi.org/10.1186/1471-2105-14-329.
43. Kovalishyn V, et al. Predictive QSAR modeling of phosphodiesterase 4 inhibitors. J Mol Graph Model. 2012;32:32–8. https://doi.org/10.1016/j.jmgm.2011.10.001.
44. Chang KY, Yang J-R. Analysis and prediction of highly effective antiviral peptides based on random forests. PLoS ONE. 2013;8(8):e70166.
45. Metz CE. Basic principles of ROC analysis. Seminars in nuclear medicine. 1978;8(4):283–298. https://doi.org/10.1016/s0001-2998(78)80014-2
46. Rohmer J, et al. Casting light on forcing and breaching scenarios that lead to marine inundation: combining numerical simulations with a random-forest classification approach. Environ Model Softw. 2018;104:64–80. https://doi.org/10.1016/j.envsoft.2018.03.003.
47. Johnson AEW, et al. MIMIC-III, a freely accessible critical care database. Sci Data. 2016;3:1.
48. Zhang Q, Xiao M, Singh VP. Uncertainty evaluation of copula analysis of hydrological droughts in the east river Basin, China. Global Planet Change. 2015;129:1–9. https://doi.org/10.1016/j.gloplacha.2015.03.001.
49. Vincent J-L, et al. The SOFA (sepsis-related organ failure assessment) score to describe organ dysfunction/failure. Intensive Care Med. 1996;22(7):707–10.
50. Almeida TA, Hidalgo JMG, Hidalgo JMG, Silva TP. Towards SMS spam filtering: results under a new dataset. Int J Informat Secur Sci. 2013;2(1):1–18.
51. TA Almeida, JMG Hidalgo, A Yamakami. 2011. "Contributions to the study of SMS spam filtering: new collection and results." In proceedings of the 2011 ACM symposium on document engineering, Association for Computing Machinery. 259-262. https://doi.org/10.1145/2034691.2034742
52. Hidalgo JMG, Tiago AA, Akebo Y. 2012. "On the Validity of a New SMS Spam Collection." In Proceedings—2012 11th International Conference on Machine Learning and Applications, ICMLA. 240–245. https://doi.org/10.1109/ICMLA.2012.211
53. Cormack GV, María J, Sánz EP, Hidalgo G. Spam filtering for short messages. Int Conf Informat Knowl Manag Proc. 2007. https://doi.org/10.1145/1321440.1321486.

54. Hidalgo, José María Gómez, Guillermo Cajigas Bringas, Enrique Puertas Sánz, and Francisco Carrero García. 2006. "Content Based SMS Spam Filtering." In Proceedings of the 2006 ACM symposium on document engineering, DocEng. 2006, 107–114. https://doi.org/10.1145/1166160.1166191
55. İlhan, Zeynep. 2019. "Kopula Temelli Değişken Kümeleme Tekniklerinin İncelenmesi ve Mortalite Tahmini Uygulaması." PhD Thesis, Eskisehir Osmangazi University.
56. Machado-Ferrer Y, et al. Heart rate variability for assessing comatose patients with different Glasgow coma scale scores. Clin Neurophysiol. 2013;124(3):589–97. https://doi.org/10.1016/j.clinph.2012.09.008.
57. Cooke WH, et al. Heart rate variability and its association with mortality inprehospital trauma patients. J Trauma Injury Infect Crit Care. 2006;60(2):363–70. https://doi.org/10.1097/01.ta.0000196623.48952.0e.
58. Wan-Ting C, et al. Reverse shock index multiplied by Glasgow coma scale (RSIG) predicts mortality in severe trauma patients with head injury. Sci Rep. 2020;10(1):2095. https://doi.org/10.1038/s41598-020-59044-w.
59. Hekmat K, et al. Daily assessment of organ dysfunction and survival in intensive care unit cardiac surgical patients. Ann Thorac Surg. 2005;79(5):1555–62. https://doi.org/10.1016/j.athoracsur.2004.10.017.
60. Hasanin A, et al. Incidence and outcome of cardiac injury in patients with severe head trauma. Scand J Trauma Resusc Emerg Med. 2016;24(1):1–6. https://doi.org/10.1186/s13049-016-0246-z.
61. Kazım B, et al. Changes in neutrophil-to-lymphocyte ratios in postcardiac arrest patients treated with targeted temperature management. Anatol J Cardiol. 2017;18(3):215–22. https://doi.org/10.14744/anatoljcardiol.2017.7716.

Publisher's Note