Improvement in Detecting the Fate of Covid-19 Patients and Rule-based Analysis to Discover the Most Important Rules Governing their Fate

Sadegh Ilbeigipour
Tarbiat Modares University
Amir Albadvi (Albadvi@modares.ac.ir)
Tarbiat Modares University
Elham Akhondzadeh Noughabi
Tarbiat Modares University

Research Article

Keywords: Covid-19, Coronavirus, Machine learning, Classification, Association rules minig

Posted Date: June 10th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-515541/v1

Improvement in detecting the fate of Covid-19 patients and Rule-based analysis to discover the most important rules governing their fate

Sadegh Ilbeigipour[1], Amir Albadvi[1 *], Elham Akhondzadeh Noughabi[1]

[1]Department of Information Technology Engineering, Industrial and Systems Engineering Faculty, Tarbiat Modares University, Tehran, Iran. *email: Albadvi@modares.ac.ir

Abstract

The world today faces a new challenge that is unprecedented in the last 100 years. The emergence of a new coronavirus has led to a human catastrophe. The new coronavirus is the cause of the Covid-19 disease, which kills many people in the world every day. Scientists in various sciences have been looking for solutions to this problem so far. In addition to general vaccination, maintaining social distance and hygienic principles are the most well-known strategies to prevent Covid-19 infection. In this research, we have tried to examine the symptoms of Covid-19 cases through different supervised machine learning methods. We solved the class imbalance problem using the SMOTE up-sampling method and then developed some classification models to predict the recovery or death of patients. Besides, we implemented a rule-based technique to identify important symptoms that affect patients' fate and calculate the range of values in these features that lead to recovery or death of patients. Our results showed that the random forest model with 94% accuracy, 95.2% sensitivity, 92.7% specification, 93.2% precision, and 94.2% F-score outperforms state-of-the-art classification models. Finally, we identified the ten most significant rules in the data set. The rules state that different combinations of 6 features in certain ranges of their values lead to patients' recovery with 90% confidence. In conclusion, the classification results in this study show better performance than recent researches. Besides, help physicians consider other important factors in improving health services to different groups of Covid-19 patients.

Keywords: Covid-19, Coronavirus, Machine learning, Classification, Association rules minig.

Introduction

In late 2019, the world suffered a great menace that influenced all viewpoints of human life. A new type of coronavirus has appeared in Wuhan, China [1]. The virus causes lung infection and has a very high rate of transmission [2]. Finally, with the extent of the virus in many countries, the World Health Organization on March 11, 2020, announced the prevalence of novel coronavirus as a fatal pandemic [3].

There are various types of coronaviruses in the world. Acute Respiratory Syndrome and Middle East Respiratory Syndrome are the most famous of these viruses [4]. The recently recognized virus is called SARS-COV-2 and is the object of Covid-19 disease [4].

Several efficient Covid-19 vaccines have been produced by well-known companies around the world so far. But, masking, hygiene, and social distancing are still the three most effective actions in preventing Covid-19 [4]. However, computer science applications are an effective way to help physicians cope with coronavirus and improve hospital care. These applications generally include machine learning, deep learning, mathematical modeling, and Social Network Analysis (SNA) techniques and are used to detect the disease or forecast the pattern of disease outbreaks. For example, regression and SEIR analysis based

on time series data have been applied to predict the future prevalence of the disease [5]. Moreover, various machine learning models have been proposed to foretell the number of cases, recovery, or death of Covid-19 infection [6-15]. Furthermore, to forecast the spread rate of the new coronavirus, some techniques based on mathematical modeling have been developed [16-18]. Today, in addition to short-term estimating of the prevalence pattern and assessing the risk of infection with the Covid-19 [19,20], real-time applications are increasing in various fields of medicine, such as the diagnosis of cardiac arrhythmias [21]. Also, researchers have used social network analysis approaches to answer how the novel coronavirus will spread and then recognized high-risk areas [22, 23]. As a last attempt, researchers have employed different deep learning methods to distinguish positive cases based on chest x-ray and CT-Scan images [24-31]. In [32], the authors have developed a deep recurrent neural network to foresight the future impact of the virus [32]. Another study modeled deep parallel networks based on a deep learning technique. It discovers spatiotemporal characteristics using one-dimensional and two-dimensional deep convolutional networks [33]. In the next step, the derived features recognize the confirmed cases [33]. Finally, although deep artificial neural networks have a high capability in processing image and audio data, these methods greatly increase computational costs. [34].

In this study, supervised machine learning methods are implemented to predict the recovery or death of positive cases with Covid-19 and identify the factors that have the greatest impact on their fate. We trained Decision Tree, Random Forest, Support Vector Machine, and K Nearest Neighbor classification algorithms to diagnose death or recovery data classes.

In this research, we first report the study area and collected data. After that, we describe the data preprocessing and developed models in the processing subsections. We explain our findings in the results section. Then we discuss different techniques, their results, and research applications. Finally, in the conclusion section, we state the goal of the research, our findings, and our limitations, and several suggestions for future works.

Method and materials

Research data was approved by the Saveh Medical Center (SMC) in Iran and was provided by figshare repository with unique identifier "http://doi.org/10.6084/m9.fihshare.12446120.v1" and under "Attribution 4.0 (CC BY 4.0)" license.

The Institutional Review Board of Department of Information Technology Engineering, Industrial and Systems Engineering Faculty, Tarbiat Modares University confirmed our research and waived the requirement for informed consent from participants in this study. Besides, the SMC waived the requirement for informed consent from participants because the data samples lacked the participants' personal information, and our study did not violate participants' privacy. Finally, research methods were performed in accordance with the relevant policies and regulations.

Experiment data. We provide data through interviews, questionnaires, and medical records of patients with Covid-19 admitted to Saveh Hospital in Iran between February 2020 and April 2020. Data combine 1,142 samples with 39 features per case. The most important symptoms include age, sex, hospital unit, breathing conditions, fever, cough, different underlying diseases, the length of hospitalization, blood rate oxygen, intubation, and death or recovery class labels. Our data includes 1131, and 111 released and died cases, respectively. Thus, our data suffer from a class imbalanced problem, which can affect learning models.

Data visualization. A useful tool for revealing hidden statistical properties in a dataset is data visualization. We visualize the data with different visualization plots. Fig. 1 show the age distribution of patients relative

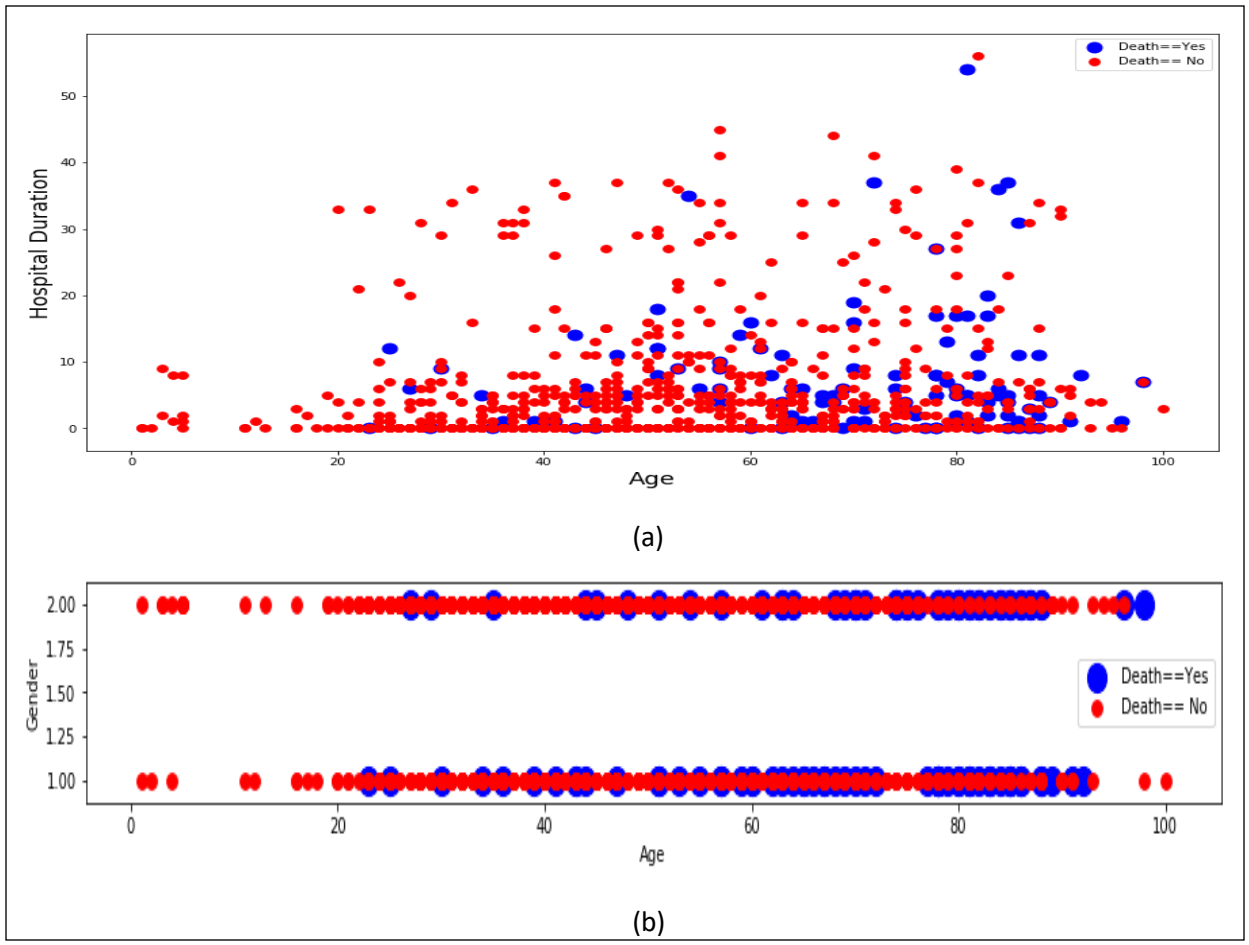

Figure 1. Scatter plot of the age of patients relative to their (a) sex and (b) hospital duration based on death or recovery class labels.

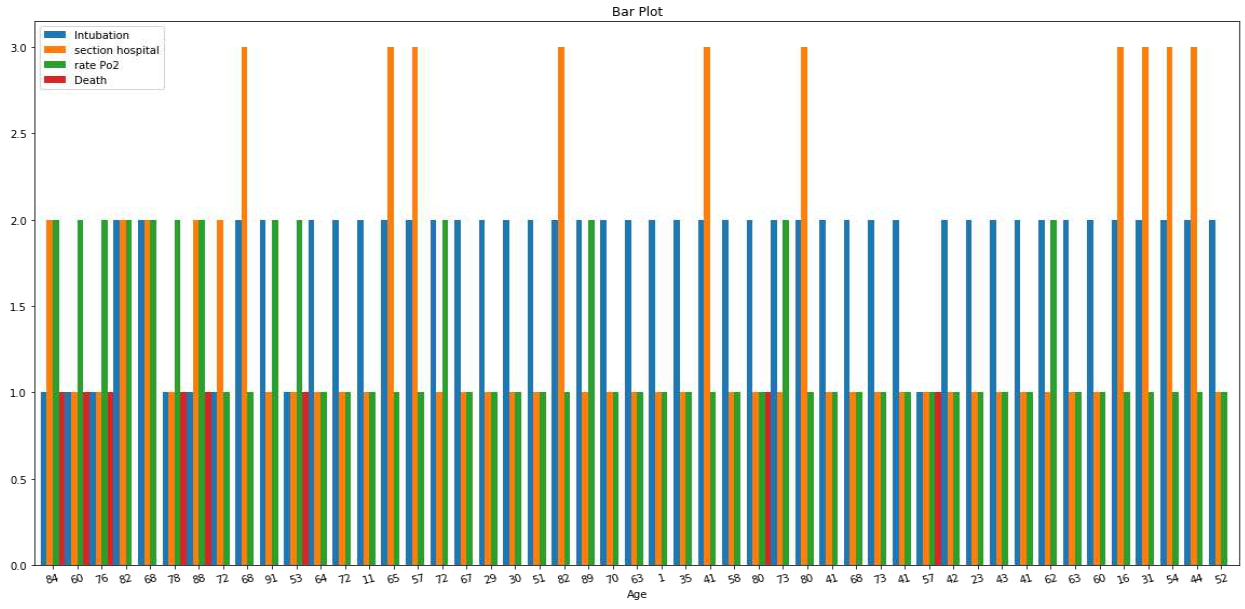

Figure 2. Bar chart of Intubation, hospital unit, rate Po2, and the class features of patient based on their age.

to hospital duration and the sex of patients based on their fate (death or recovery), respectively. As can be seen, most of the patients were hospitalized for 1 to 10 days, and the number of patients who died (blue spots) was more than 60 years old (Fig. 1.a). On the other hand, according to Fig. 1.b, the number of male patients (value 2) in this study is more than female patients (value 1). Furthermore, the mortality rate is almost the same among both sexes.

A useful way to explore the different characteristics of patients relative to each other is to present data with a bar plot. Fig. 2 shows the intubation, hospital section, blood oxygen level, and death or recovery of some patients toward their age. The bar diagram explains how to allocate a variety of features to a specific case. Besides, it is a suitable way to compare the values of one sample with another sample based on a particular feature.

Visualization also provides a way to show the rate of change in the values of different features relative to each other. The line chart is the tool used for this purpose. The line chart in Fig. 3 shows the rate of change in intubation, hospital ward, blood oxygen level (ratePo2), and class label (death) characteristics for 50 Covid-19 cases.

Finally, we divide the data into separate sections using a facet chart and display it as a single plot. The facet chart in this study (Fig. 4) divides the data set into two subsets based on the class labels and shows the age distribution of patients in each subset. Accordingly, the highest age recurrence is 40 years in the recovered patients set and 80 years in the deceased cases set.

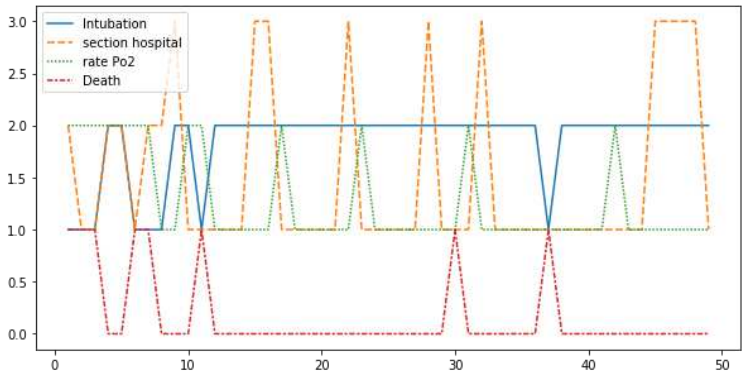

Figure 3. Line chart of intubation, hospital ward, blood oxygen level, and class label variables for 50 Covid-19 cases.

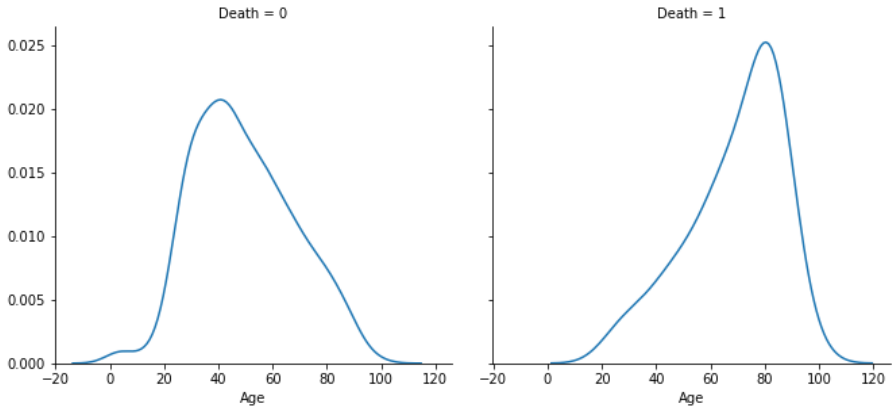

Figure 4. Facet chart of the age distribution of patients based on different class labels.

Data pre-processing. In data analysis applications, data should be preprocessed before training the learning models. In this research, the preprocessing stage includes removing outliers values, replacing null values, solving the class imbalance problem, and feature selection. The output of the preprocessing step is a set of data erased from redundancy. It has a significant impact on improving the performance of learning models. [34]. Additionally, preprocessing steps may require data sampling and normalization based on the learning method.

In this study, null values in all variables are replaced with the mean value of that variable. It assures that the learning model does not tend to a specific value in a variable. Moreover, the Kmeans clustering method was applied to discover outliers in the data. Our data lacked out-of-range values due to the accurate methods we use in collecting data for this research.

Feature selection. The purpose of feature selection is to find an optimal subset of characteristics by eliminating unrelated variables in the research data. So it leads to advancing the model performance results and reduces computational complexity [35]. The most well-known methods for picking out features are Filter, Wrappers, and Embedded procedures [35]. The filter method does attribute ranking independent of the learning algorithm. Feature ranking describes the degree of influence of a feature in separating between data classes [35]. The wrapper-based techniques are executed along with the learning algorithm [35]. These methods evaluate the model metrics to choose the best subset of attributes by applying the learning algorithm on different subsets of data regularly. Finally, a hybrid of filter and wrapper approaches makes the embedded techniques [35].

In this study, a filter-based method is used to select the ten most important variables in predicting the death or recovery classes of Covid-19 patients (Fig. 5). We developed the extra tree algorithm that is an ensemble and majority-vote-based classifier and employs a set of decision trees to distinguish class labels [36].
Fig. 5 presents the most relevant features to train learning models. According to this figure, intubation, age, and the number of hospitalization days are the most effective properties to determine class labels, respectively.

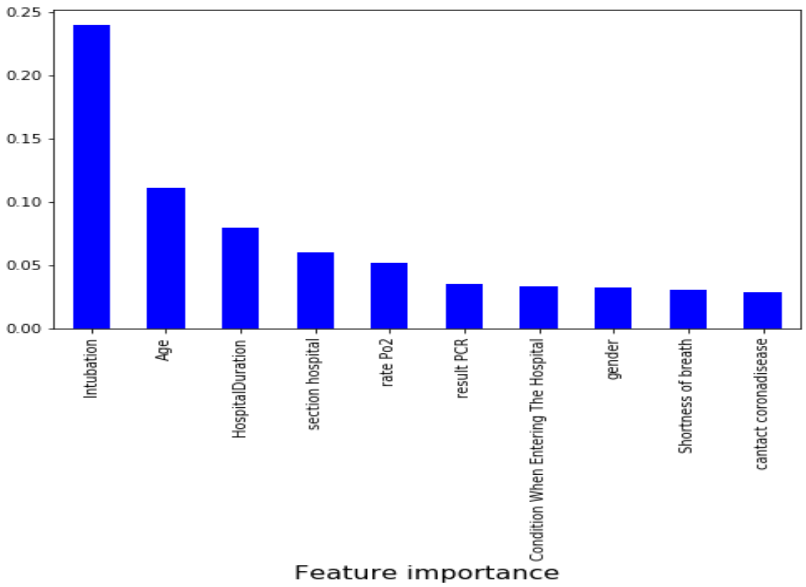

Figure 5. Top 10 influence attributes selected using the filter-based technique.

Balancing class labels. The class imbalance problem is a common problem in supervised feature learning. Class imbalance indicates that the number of data samples with different labels is not balanced. It can greatly affect the prediction results. If the number of a particular label in a data set is very different from

other labels, the data suffer from the class imbalance problem. Therefore, it is necessary to solve this problem, otherwise, the classification results will not be sufficiently reliable.

Conventional methods for solving the class imbalance problem are based on Over-sampling or Under-sampling approaches. The over-sampling methods balance class labels by replicating minority class samples. The under-sampling methods, on the other hand, balance class labels by eliminating some majority class samples.

In this study, the number of data classes collected for recovered and dead patients is 1031 and 111 samples, respectively. Therefore, data classes are imbalanced in our research. So, we have used the Synthetic Minority Oversampling Technique (SMOTE) [37] to solve the class imbalance problem. The SMOTE algorithm is a method based on a data Over-sampling approach that incorporates the minority class samples to create new cases instead of duplicating them. SMOTE first randomly selects a sample from the minority class set and finds its best K nearest neighbor by the Euclidean distance measure. Then randomly selects some cases from its neighbors and generates a new synthetic sample by the selected samples [37]. This process is repeated for all the minority class samples until the number of the minority class cases equals the majority class cases. Equation 1 shows a way of combining samples in the SMOTE algorithm [37]. While X_n is the new sample, X is the sample selected from the minority class sample, and X_k is the sample selected from the neighbors set. Plus, the rand function generates a random number between zero and one.

$$X_n = X + \text{rand}(0, 1) * |X - X_k| \quad (1)$$

By solving the class imbalance problem in this study, the number of dead samples (minority class) increased from 111 to 1131 cases, equal to recovered samples (majority class). Table 1 shows the number of patients belonging to each class label in this research after applying the SMOTE algorithm and dividing the data into a train and test set.

Table 1. Number of samples in different classes in the train and test data set.

Data set	**Recovery class samples**	**Death class samples**	**Total sample**
Train (70%)	729	714	1443
Test (30%)	302	317	619
All	1031	1031	2062

Processing. In this stage, we train learning models using preprocessed data. In the literature, researchers have proposed many machine learning approaches for different artificial intelligence applications. Supervised, unsupervised, and semi-supervised feature learning approaches are the well-known machine learning categories [34].

The supervised machine learning methods can utilize what they have seen in the past to foretell the future. In these methods, the data set is divided into two train and test data sets. In the training phase, the model learns how to separate class labels (supervised learning), and in the testing phase, we evaluate how well the model can recognize the class of the samples [34]. On the other side, unsupervised methods do not need class labels, and we do not provide any guidance to the learning model [34]. Unsupervised techniques are mostly applied to recognize similar groups of the studied data.

Finally, a combination of supervised and unsupervised strategies forms semi-supervised learning methods [34]. Nevertheless, the learning models deal with a set of data that some of its tuples lack the class label [34].

In this study, we trained a set of supervised classification models to predict the death or recovery of the Covid-19 patients. Next, we implement a supervised rule-based technique to identify a set of significant factors in determining the fate of patients.

Classification. Several classification models were developed for supervised features learning to diagnose the recovery or death of patients infected with the coronavirus in this study. These learning models include Decision Tree, Support Vector Machine (SVM), K Nearest Neighbor, and Random Forest. In the results section, the performance metrics of the models are detailed.

Association rule mining. Association rules extraction is a data mining operation that finds connections between the features of a data set [34]. In other words, association analysis is the study of features or characteristics that are related to each other and tries to extract rules from these features. This method seeks to discover rules to quantify the relationship between two or more properties [34]. Association rules are defined as if and then with two indices of Support and Confidence [34].

In this study, the aim is to determine rules for predicting the novel coronavirus behavior in different patients. In the next section, we describe the concepts needed to derive the rules that may not be familiar to the readers.

Important definition. Suppose that $I= [I_1, I_2,\ldots, I_m]$ is the set of total features available in the data set. Each subset of I is called a transaction, denoted by T, and D is the set of transactions in T. Then an association rule is shown as follows:

$X \rightarrow Y$ {Support , Confidence} So that, $X \subset I$, $Y \subset I$, $X \cap Y = \emptyset$.

Support: Indicates the percentage or number of D transactions that include both X and Y.
Confidence: Expresses the dependence of a particular feature on another feature and is calculated as Equation 2.
Strong rules: Strong rules are rules that have greater support and confidence than the determined threshold. In association rules analysis, the goal is to find and extract these strong rules.

$$\text{Confidence}(X, Y) = \frac{\text{support}(X \cup Y)}{\text{support}(X)} \qquad (2)$$

Lift index: Lift index is a measure for evaluating association rules and shows the attractiveness of a rule [34], which is calculated as Equation 3.

$$\text{Lift}(X \rightarrow Y) = \frac{\text{support}(X \cup Y)}{\text{support}(X).\text{support}(Y)} \qquad (3)$$

According to Equation 3, if the lift index for a rule is less (or greater) than one, then there is a negative (or positive) dependency between X and Y in the $X \rightarrow Y$ condition.

Many algorithms have been proposed to extract strong rules in the literature. One of the most widely used methods in this field is the Apriori algorithm, which has lower computational complexity than other methods [34]. We have used the Apriori algorithm with 30% and 90% support and confidence thresholds, respectively, to extract strong rules from the symptoms of Covid-19 disease in determining the fate of patients. The extracted rules that meet all the defined conditions include 269 rules. Fig. 6 and 7 show the scatter diagram and the grouped matrix of the extracted rules, respectively. The higher color intensity in

Fig. 6 indicates a higher degree of confidence (close to one). But, the higher color intensity in Fig. 7 represents the lift value, and the node size describes the support value. The matrix elements in Fig. 7 show the symptoms of Covid-19 cases with specific intervals that lead to recovery or death of patients. Although all 269 rules are important, the rules filtered in Fig. 6 (top left corner) are more important than the rest of the rules because they have higher lift and confidence indices. In the next section, we will examine the ten most important of these rules in more detail and state what important factors together determine the fate of patients with high confidence.

Analysis Environment. Our system is supported by CPU 2.3Ghz (five-core), 6 gigabytes of RAM, and one terabyte of disk space to implement algorithms in this study. We implemented visualization, preprocessing, and classification steps with the Python programming language version 3.5. Also, we use R language version 3.5.1 to discover association rules due to its capability to display results.

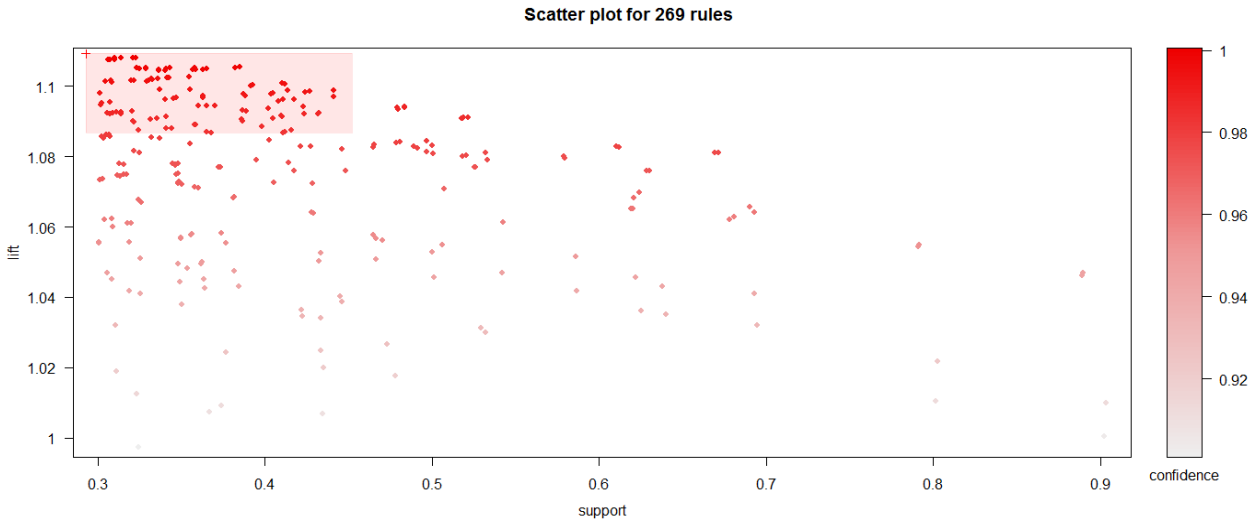

Figure 6. Scatter plot of the association rules extracted with the expected conditions (support = 0.3, confidence = 0.9, lift> 1).

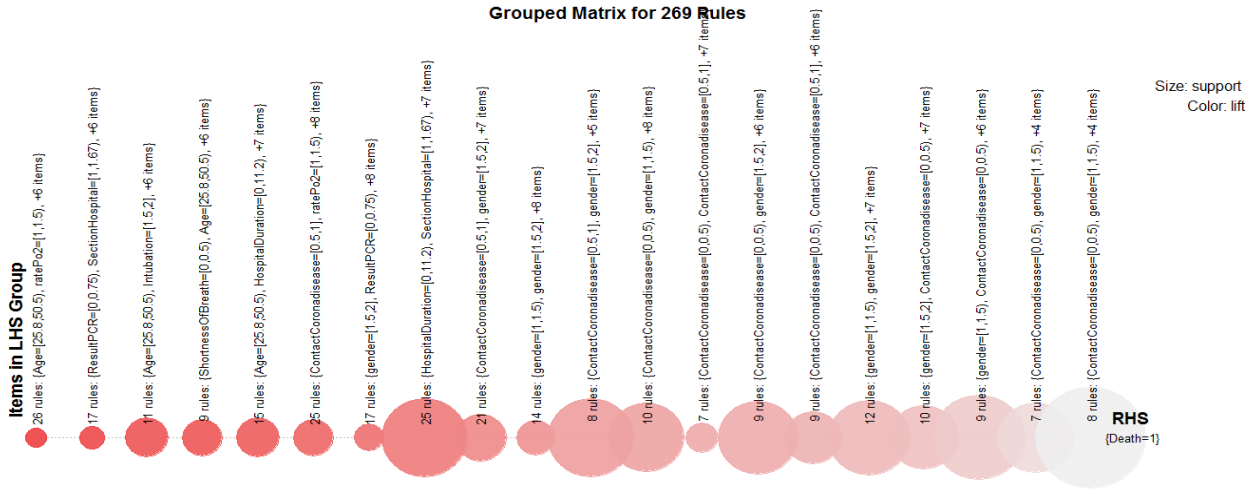

Figure 7. Grouped matrix diagram of extracted association rules with expected conditions (support 0.3, confidence 0.9, lift> 1).

Result

In this section, we present the research results for different methods separately. First, the classification performance metrics to predict the recovery or death of Covid-19 patients are computed in detail for each classification algorithm and compared with the results of previous research. Second, a set of symptoms that play a significant role in determining the fate of patients are identified using the association rules mining technique.

Classification performance. We trained the decision tree, K nearest neighbor, support vector machine, and random forest classifiers to predict the death or recovery of Covid-19 patients with 70% of the data and tested with the remaining 30%. Classification performance metrics include accuracy, precision, sensitivity, specificity, and F1-score measures, which are calculated by Equations (4-8), respectively. In these equations, true positive (TP) represents positive samples, and the model predicts them as positive correctly. False Positive (FP) are negative samples, and the model has evaluated them as positive mistakenly. Also, true negatives (TN) are the number of negative samples, and the model predicts them as negative correctly. Lastly, false negatives (FN) are examples that the model should predict positive but considered negative wrongly [34]. Table 2 shows the classification performance results for the different classification models. Based on the data provided in Table 2, the random forest classification model has a better performance than other models in determining the fate of Covid-19 patients.

Finally, Table 3 compares the classification performance results in this study with the results of previous researches. Our results show a remarkable improvement over previous studies.

$$\text{Accuracy: } \frac{TP + TN}{TP + TN + FP + FN} \tag{4}$$

$$\text{Precision: } \frac{TP}{TP + FP} \tag{5}$$

$$\text{Sensitivity: } \frac{TP}{TP + FN} \tag{6}$$

$$\text{Specificity: } \frac{TN}{TN + FP} \tag{7}$$

$$\text{F1} - \text{score: } \frac{2(Precision * Sensitivity)}{(Precision + Sensitivity)} \tag{8}$$

Table 2. Classification performance results to diagnosis the death or recovery of Covid-19 patients.

Model	**Acc(%)**	**Pr(%)**	**Se(%)**	**Sp(%)**	**F1_score(%)**
Decision tree	93.37	92.85	94.32	92.38	93.58
SVM	86.26	86.94	86.11	86.42	86.52
Knn	85.13	80.48	93.69	76.15	86.58
Random forest	94.02	93.20	95.26	92.71	94.22

Association rules results. We have already said that association rules techniques are used to identify characteristics that, with a certain confidence and support, lead to the occurrence of a class label. This technique is also applied in various applications such as customer shopping patterns evaluation and banking. In this study, we used association rules mining by the Apriori algorithm to identify a set of symptoms that

Table.3: Classification performance of the proposed method and comparison with state-of-the-art methods.

Research	Acc(%)	Pr(%)	Se(%)	Sp(%)	F1_score(%)
An, Chansik, et al. [38]	91.9	25.6	92.0	91.8	40.0
Chen, et al. [39]	-	-	91.4	76.0	-
Chowdhury, et al. [40]	-	-	92.0	92.0	-
Iwendi, et al. [41]	94.0	100	75.0	-	86.0
Yan, et al. [42]	-	95.0	92.0	-	93.0
Mohammad and Mahdi [43]	89.9	93.6	87.7	93.2	90.5
Proposed method	94.0	93.2	95.2	92.7	94.2

lead to recovery or death of patients with high confidence (0.9) and support (0.3). In each specific application, the rules generated are usually high. Therefore, we need to extract only the most essential rules that satisfy the value of the support and confidence threshold, and the correlation between their items and their results is positive. We used the lift index to extract the ten most momentous rules in the data set by evaluating the relationship between the item(s) and the label of the rules. We visualized these rules with a parallel coordinates plot in Fig. 8. The vertical axis of the diagram in Fig. 8 shows the set of items, and the horizontal axis represents the position of the items in different rules. According to the result, all the rules lead to the recovery of patients (death = 1). In our data set, the class label (death) column with numbers 1 and 2 describes the recovered and dead patients, respectively. Other items include hospital section, hospital duration, patient's respiratory condition, the patient's condition when visiting the hospital, blood oxygen level, need for a ventilator, and patient's age, respectively.

Before extracting the rules, the values of each attribute are discretized into different intervals. Each rule specifies the extent to which features can together determine the fate of patients. In addition, it recognizes variables in which range of their values have the most occurring in the data set. The range [1,1.67) for the hospital section indicates the high range of its values and, as mentioned previously, refers to the hospital wards where patients with usual symptoms are admitted. Also, the rules record that the number of hospitalization days with a range between 0 and 11 days has an effective role in determining the recovery of patients. In this study, the set of values for lack of the patient's shortness of breath and the patient's shortness of breath is 0 and 1, respectively. Shortness of breath (interval [0,0.5)) in this variable indicates lack of shortness of breath or mild shortness of breath in patients. The next feature is the patient's condition when visiting the hospital. This feature with a value of zero describes the patients with normal conditions, and a value of one expresses the patients with the severe condition. So, the rules indicate that patients recover in a critical condition (interval [0.5,1]). This unusual situation in the rules occurs because almost all patients in our data set are hospitalized in unfavorable conditions. Blood oxygen level in patients is another feature that plays a significant role in their recovery. The higher the rate of pulmonary infection caused by the coronavirus, the lower the blood oxygen level in patients. In coronavirus, the values of 1 and 2 determine the oxygen level above 93% and less than 93%, respectively. As Fig. 8 shows, the oxygen level with intervals [1, 1.5) plays a substantial role in various rules. Moreover, the items that have been reviewed so far, the intubation feature is another principle feature that occupies a place in the set of rules. Intubation indicates whether a patient has needed a ventilator device or not. Patients who did not require intubation are marked with the number 2, otherwise, the value of the intubation variable is 1. Naturally, the rules

indicate that intervals [1.5,2] are necessary for patients to recover because they do not need artificial respiration. Finally, the age property of patients as the last item is essential in determining the recovery of patients. Patients between the ages of 25 and 50 are more likely to recover.

According to Fig. 8, a combination of different items participates in various rules. For example, one rule states that a patient will recover with a hybrid of age, shortness of breath, hospital duration, and oxygen rate within defined intervals with 90% confidence and 30% support. Another rule replaces the intubation feature with the length of hospitalization feature and infers the same result.

Finally, it is principal to note that the extracted rules are one-sided. It means that although a combination of properties can result in the recovery of patients with high confidence, a patient may have recovered but not meet any of the conditions set out in Fig. 8. In other words, the rules do not guarantee that a patient who is not covered by any of the rules will die.

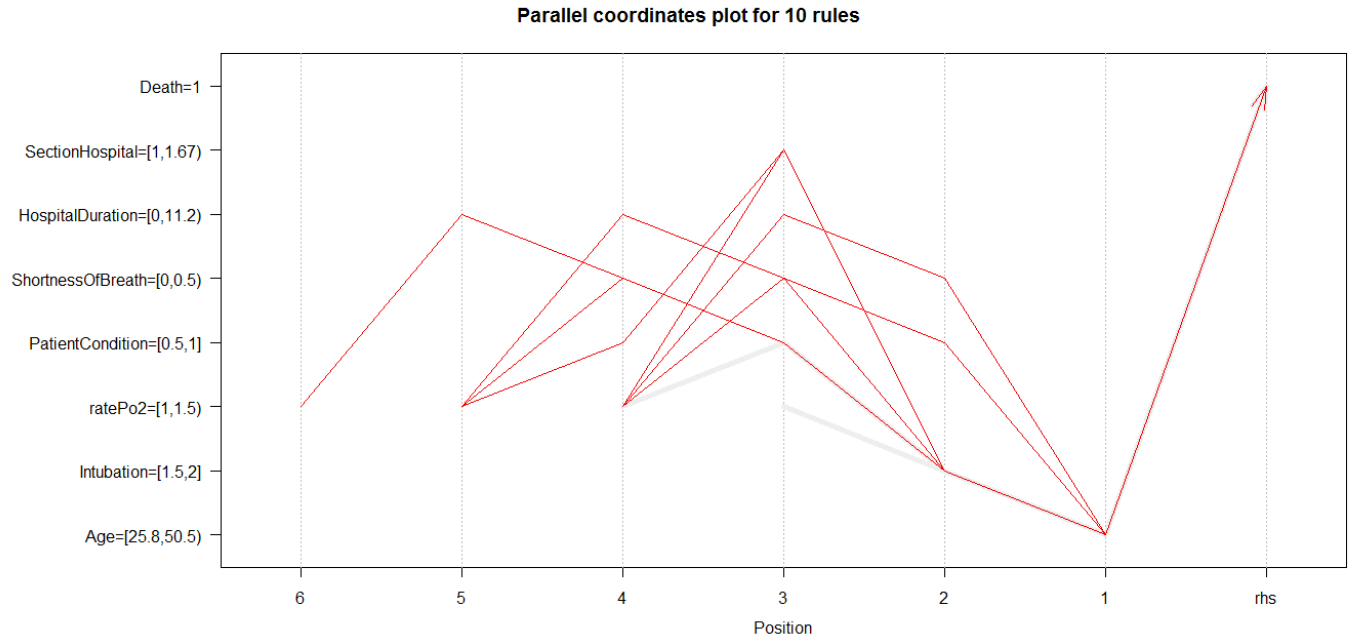

Figure 8. The ten most essential association rules discovered in this research.

Discussion

In this study, we employed machine learning techniques to evaluate the clinical symptoms of two groups of recovered and deceased Covid-19 patients. Machine learning techniques in this research mainly include supervised methods. Classifying patients based on recovery or death classes and discovering association rules to identify a set of vital symptoms in determining the fate of patients are two supervised methods used. One problem with data classification is the class imbalance problem. The total number of patients in this study is 1142 patients, of which 1031 patients have recovered, and 111 patients have died. So our data suffer from the class imbalance problem. We used the SMOTE algorithm to solve the problem by increasing the dead class. We then developed the four decision tree classification, nearest neighborhood, support vector machine, and random forest models with 70% of the train data and tested them with the remaining 30% samples. Our results showed that the random forest classifier with 94.02% accuracy, 93.20% precision,

95.26% sensitivity, 92.71% specification, and 94.22% F1-score has better performance in diagnosing recovery or death of patients than other classification models (Table 2).

Random forest is an ensemble classification method based on the majority vote and uses a combination of decision trees to classify data classes. Researches show that this classifier performs more beneficial than other classification models due to its ensemble nature, although it has a higher computational complexity. It could justify the higher performance of this model in this research.

Also, we present a rule-based approach to determining the set of factors that together affect the fate of patients. We set a 30% support and 90% confidence for extracting significant rules. The purpose of adopting a high threshold for support and confidence measures is to find the highest quality rules. In the next step, we calculated the quality for all generated rules by the lift metric and represented the ten highest quality rules through a parallel coordination diagram (Fig. 8). All extracted rules are related to the recovered patients and do not provide knowledge for deceased patients. Although it does not impair the accuracy of our results, it is because more than 80% of the Covid-19 cases from recovered cases in our dataset, and the number of patients who die is much lower. However, the high number of recovered cases confirms the produced rules because they follow more patterns in the data set. Accordingly, the rules tell us what symptoms and in what range of their values with high confidence will result in the recovery of patients. The items that make up the ten most important rules are age, intubation, hospital section, breathing condition, the patient condition when visiting the hospital, hospital duration, and blood oxygen rate. Different combinations of these symptoms lead to the production of various rules (Fig. 8).

Finally, our results confirm the research presented in the past. Most studies have identified patients' age as an essential factor in the recovery or death of coronavirus cases. According to the latest research, more than 70% of deaths in Iran are among people over 60 years old. Besides, our results present new information to physicians. For example, specialists should consider the value of clinical variables such as hospital duration, patient's respiratory condition, and patient condition at the time of hospitalization as other important factors to improve treatment services at different stages of patients' treatment, in addition to the symptoms identified in the past, such as age and underlying diseases.

Conclusion

We investigated different aspects of the new coronavirus using two supervised machine learning methods. After data preprocessing and fixing the class imbalance problem, we developed an efficient random forest model to diagnose the recovery or death of Covid-19 patients with performance metrics of accuracy, sensitivity, specifications, and F-1 score of 94.0%, 95.2%, 92.7%, and 94.2% respectively. Our results show better performance in predicting the fate of Covid-19 cases than previous studies. Next, we developed a rule-based method to extract the essential association rules governing the Covid-19 cases. The rules identify different combinations of 6 features and the range of their values that determine the recovery of Covid-19 patients with a 90% confidence.

Our study proves past research and reports new facts. Moreover, our results help health specialists consider other factors to enhance healthcare services for the Covid-19 patients. Specialists can handle different groups of patients by measuring the characteristics that have been identified as efficient in this research. It conclusively leads to decreasing in Covid-19 fatality.

This research faced some constraints. First, we considered ten features as the best attributes to minimize computational costs and improve reliability. Therefore, more variables will be checked by boosting this threshold. Second, the methods based on deep learning cannot implement on our data. We could provide a Big Data framework by adding a variety (big data feature) of data types such as CT-scan images and ECG signals of patients to execute deep neural networks to solve this problem. On the other hand, it significantly raises the computational costs of the operations.

References

1. N. Chen et al., "Epidemiological and clinical characteristics of 99 cases of 2019 novel coronavirus pneumonia in Wuhan, China: a descriptive study," The lancet, vol. 395, no. 10223, pp. 507-513, 2020.
2. J. Yoosefi Lebni et al., "How the COVID-19 pandemic effected economic, social, political, and cultural factors: A lesson from Iran," International Journal of Social Psychiatry, p. 0020764020939984, 2020.

3. "World Health Organization (2020) Coronavirus Disease (COVID-19) Pandemic. Geneva: World Health

Organization.."

4. H. A. Rothan and S. N. Byrareddy, "The epidemiology and pathogenesis of coronavirus disease (COVID-19) outbreak," Journal of autoimmunity, vol. 109, p. 102433, 2020.

5. R. Gupta, G. Pandey, P. Chaudhary, and S. K. Pal, "Machine learning models for government to predict COVID-19 outbreak," Digital Government: Research and Practice, vol. 1, no. 4, pp. 1-6, 2020.

6. S. Kushwaha et al., "Significant applications of machine learning for COVID-19 pandemic," Journal of Industrial Integration and Management, vol. 5, no. 4, 2020.

7. S. Tuli, S. Tuli, R. Tuli, and S. S. Gill, "Predicting the growth and trend of COVID-19 pandemic using machine learning and cloud computing," Internet of Things, vol. 11, p. 100222, 2020.

8. C. M. Ye?ilkanat, "Spatio-temporal estimation of the daily cases of COVID-19 in worldwide using random forest machine learning algorithm," Chaos, Solitons & Fractals, vol. 140, p. 110210, 2020.

9. P. Wang, X. Zheng, J. Li, and B. Zhu, "Prediction of epidemic trends in COVID-19 with logistic model and machine learning technics," Chaos, Solitons & Fractals, vol. 139, p. 110058, 2020.

10. Z. Malki, E.-S. Atlam, A. E. Hassanien, G. Dagnew, M. A. Elhosseini, and I. Gad, "Association between weather data and COVID-19 pandemic predicting mortality rate: Machine learning approaches," Chaos, Solitons & Fractals, vol. 138, p. 110137, 2020.

11. K. Santosh, "AI-driven tools for coronavirus outbreak: need of active learning and cross-population train/test models on multitudinal/multimodal data," Journal of medical systems, vol. 44, no. 5, pp. 1-5, 2020.

12. M. Yadav, M. Perumal, and M. Srinivas, "Analysis on novel coronavirus (COVID-19) using machine learning methods," Chaos, Solitons & Fractals, vol. 139, p. 110050, 2020.

13. A. Alimadadi, S. Aryal, I. Manandhar, P. B. Munroe, B. Joe, and X. Cheng, "Artificial intelligence and machine learning to fight COVID-19," ed: American Physiological Society Bethesda, MD, 2020.

14. A. M. Leeuwenberg and E. Schuit, "Prediction models for COVID-19 clinical decision making," The Lancet Digital Health, vol. 2, no. 10, pp. e496-e497, 2020.

15. K. Santosh, "COVID-19 prediction models and unexploited data," Journal of medical systems, vol. 44, no. 9, pp. 1-4, 2020.

16. F. Ndaïrou, I. Area, J. J. Nieto, and D. F. Torres, "Mathematical modeling of COVID-19 transmission dynamics with a case study of Wuhan," Chaos, Solitons & Fractals, vol. 135, p. 109846, 2020.

17. H. M. Yang, L. L. Junior, F. F. M. Castro, and A. C. Yang, "Mathematical model describing CoViD-19 in S?o Paulo, Brazil–evaluating isolation as control mechanism and forecasting epidemiological scenarios of release," Epidemiology & Infection, vol. 148, 2020.

18. K. N. Nabi, "Forecasting COVID-19 pandemic: A data-driven analysis," Chaos, Solitons & Fractals, vol. 139, p. 110046, 2020.

19. T. Chakraborty and I. Ghosh, "Real-time forecasts and risk assessment of novel coronavirus (COVID-19) cases: A data-driven analysis," Chaos, Solitons & Fractals, vol. 135, p. 109850, 2020.

20. K. Roosa et al., "Real-time forecasts of the COVID-19 epidemic in China from February 5th to February 24th, 2020," Infectious Disease Modelling, vol. 5, pp. 256-263, 2020.

21. Ilbeigipour, Sadegh, Amir Albadvi, and Elham Akhondzadeh Noughabi. "Real-Time Heart Arrhythmia Detection Using Apache Spark Structured Streaming." *Journal of Healthcare Engineering* 2021 (2021).

22. S. Saraswathi, A. Mukhopadhyay, H. Shah, and T. Ranganath, "Social network analysis of COVID-19 transmission in Karnataka, India," Epidemiology & Infection, vol. 148, 2020.

23. P. Pascual-Ferr, N. Alperstein, and D. J. Barnett, "Social Network Analysis of COVID-19 Public Discourse on Twitter: Implications for Risk Communication," Disaster medicine and public health preparedness, pp. 1-9, 2020.

24. E. Luz, P. L. Silva, R. Silva, and G. Moreira, "Towards an efficient deep learning model for covid-19 patterns detection in x-ray images," arXiv preprint arXiv:2004.05717, 2020.

25. A. Haghanifar, M. M. Majdabadi, and S. Ko, "Covid-cxnet: Detecting covid-19 in frontal chest x-ray images using deep learning," arXiv preprint arXiv:2006.13807, 2020.

26. F. A. Saiz and I. Barandiaran, "COVID-19 detection in chest X-ray images using a deep learning approach," International Journal of Interactive Multimedia and Artificial Intelligence, InPress (InPress), vol. 1, 2020.

27. K. Kamal, Z. Yin, M. Wu, and Z. Wu, "Evaluation of deep learning-based approaches for COVID-19 classification based on chest X-ray images," Signal, Image and Video Processing, pp. 1-8, 2021.

28. H. Panwar, P. Gupta, M. K. Siddiqui, R. Morales-Menendez, P. Bhardwaj, and V. Singh, "A deep learning and grad-CAM based color visualization approach for fast detection of COVID-19 cases using chest X-ray and CT-Scan images," Chaos, Solitons & Fractals, vol. 140, p. 110190, 2020.

29. M. Hajij, G. Zamzmi, and F. Batayneh, "TDA-Net: Fusion of Persistent Homology and Deep Learning Features for COVID-19 Detection in Chest X-Ray Images," arXiv preprint arXiv:2101.08398, 2021.

30. E. Hussain, M. Hasan, M. A. Rahman, I. Lee, T. Tamanna, and M. Z. Parvez, "CoroDet: A deep learning based classification for COVID-19 detection using chest X-ray images," Chaos, Solitons & Fractals, vol. 142, p. 110495, 2021.

31. H. Panwar, P. Gupta, M. K. Siddiqui, R. Morales-Menendez, and V. Singh, "Application of deep learning for fast detection of COVID-19 in X-Rays using nCOVnet," Chaos, Solitons & Fractals, vol. 138, p. 109944, 2020.

32. S. Shastri, K. Singh, S. Kumar, P. Kour, and V. Mansotra, "Time series forecasting of Covid-19 using deep learning models: India-USA comparative case study," Chaos, Solitons & Fractals, vol. 140, p. 110227, 2020.

33. C.-J. Huang, Y. Shen, P.-H. Kuo, and Y.-H. Chen, "Novel spatiotemporal feature extraction parallel deep neural network for forecasting confirmed cases of coronavirus disease 2019," Socio-Economic Planning Sciences, p. 100976, 2020.

34. J. Han, M. Kamber, and J. Pei, "Data mining concepts and techniques third edition," The Morgan Kaufmann Series in Data Management Systems, vol. 5, no. 4, pp. 83-124, 2011.

35. G. Chandrashekar and F. Sahin, "A survey on feature selection methods," Computers & Electrical Engineering, vol. 40, no. 1, pp. 16-28, 2014.

36. P. Geurts, D. Ernst, and L. Wehenkel, "Extremely randomized trees," Machine learning, vol. 63, no. 1, pp. 3-42, 2006.

37. N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: synthetic minority over-sampling technique," Journal of artificial intelligence research, vol. 16, pp. 321-357, 2002.

38. An, Chansik, et al. "Machine learning prediction for mortality of patients diagnosed with COVID-19: a nationwide Korean cohort study." *Scientific reports* 10.1 (2020): 1-11.

39. Chen, Xingdong, and Zhenqiu Liu. "Early prediction of mortality risk among severe COVID-19 patients using machine learning." *MedRxiv* (2020).

40. Chowdhury, Muhammad EH, et al. "An early warning tool for predicting mortality risk of COVID-19 patients using machine learning." *Cognitive Computation* (2021): 1-16.

41. Iwendi, Celestine, et al. "COVID-19 patient health prediction using boosted random forest algorithm." *Frontiers in public health* 8 (2020): 357.

42. Yan, Li, et al. "A machine learning-based model for survival prediction in patients with severe COVID-19 infection." *MedRxiv* (2020).

43. Pourhomayoun, Mohammad, and Mahdi Shakibi. "Predicting mortality risk in patients with COVID-19 using machine learning to help medical decision-making." *Smart Health* 20 (2021): 100178.

Data Availability

The data utilized for finding the outcomes of this research have been taken through questionnaires and patients' medical records in Saveh Hospital, Iran.

Competing interests

The authors declare no competing interests

Funding Statement

The authors received no financial support for the research and/or authorship of this article.

Acknowledgments

None

Author contributions

S.I. collected research data, developed machine learning methods, provided the ideas of study, performed the statistical analysis, and wrote the article. A.A. and E.A. conceived the study, reviewed the various sections of the paper, conceptualized the results, and approved the final version of the manuscript.

Figures

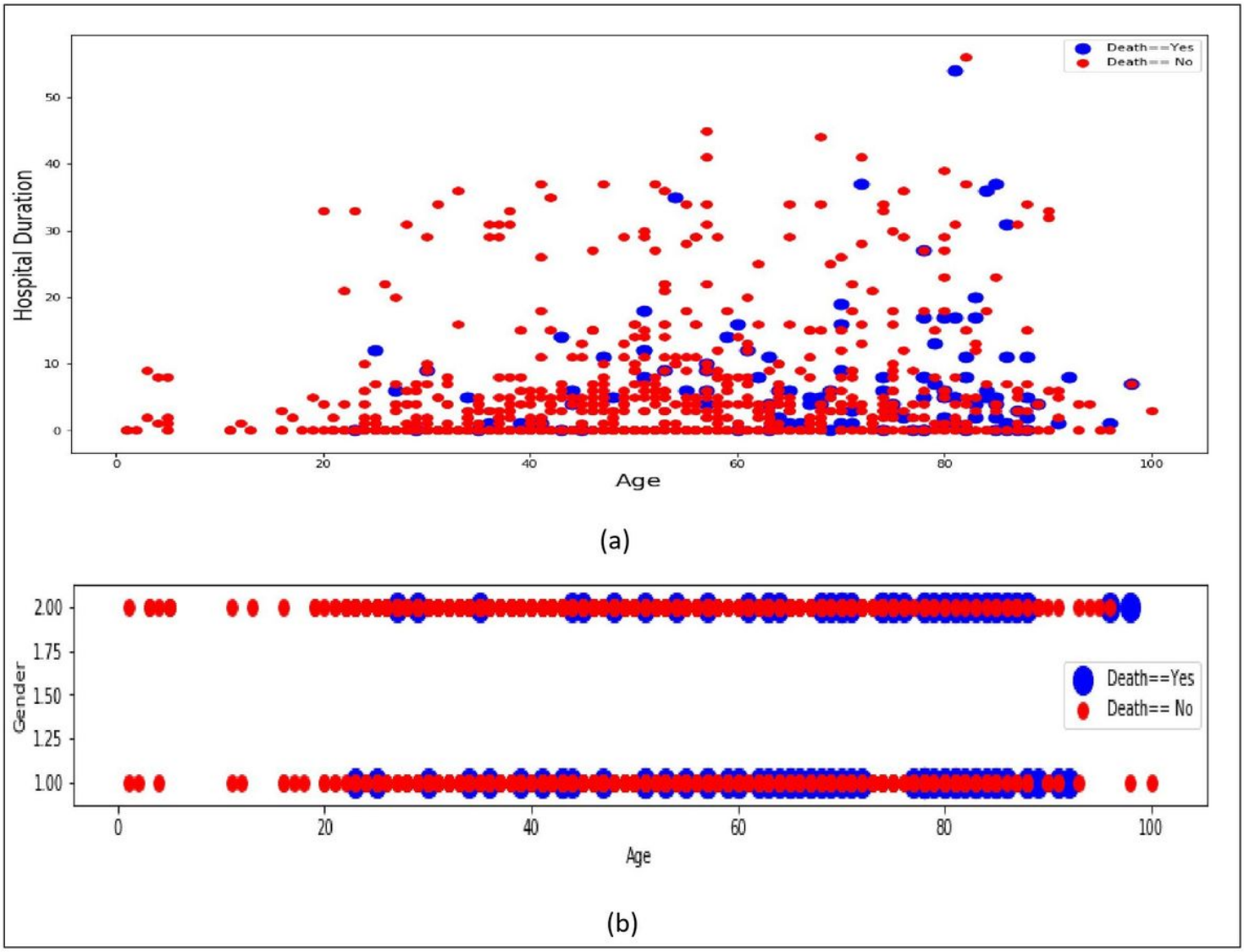

Figure 1

Scatter plot of the age of patients relative to their (a) sex and (b) hospital duration based on death or recovery class labels.

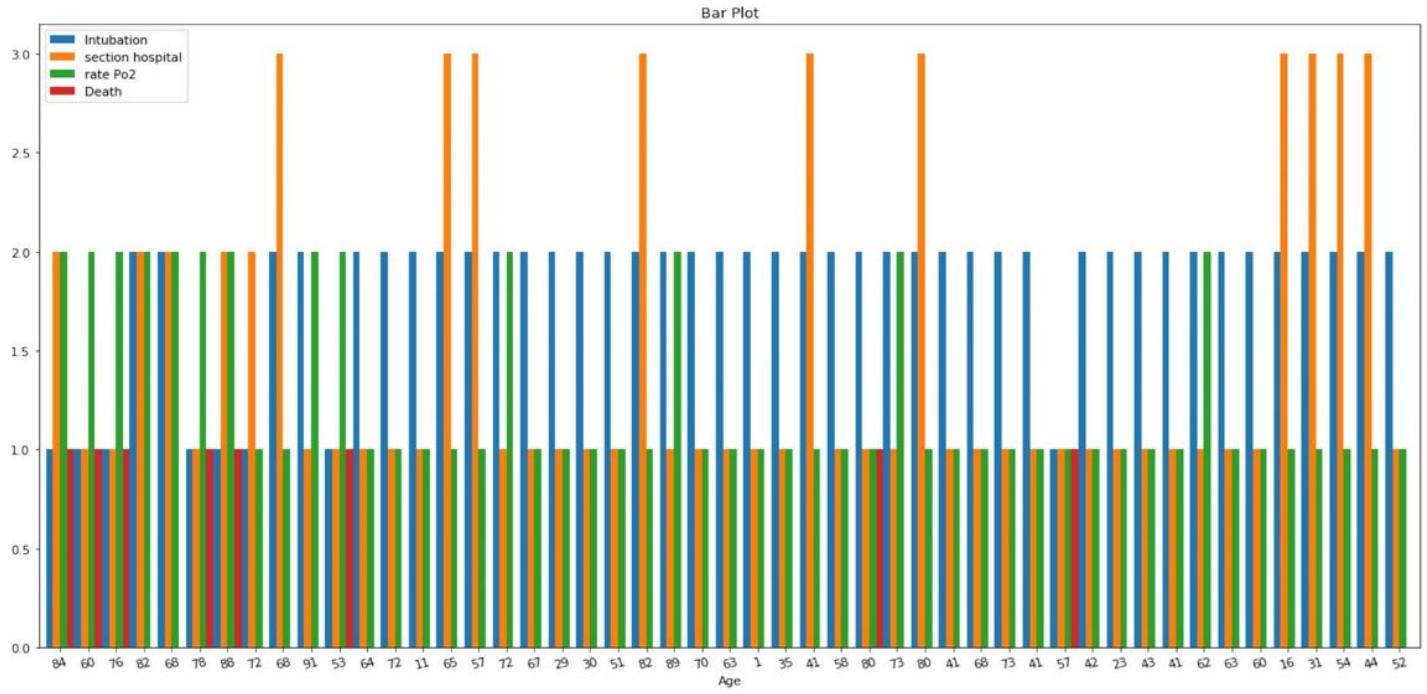

Figure 2

Bar chart of sex, hospital duration, CT-scan result, and the class of patient features based on their age.

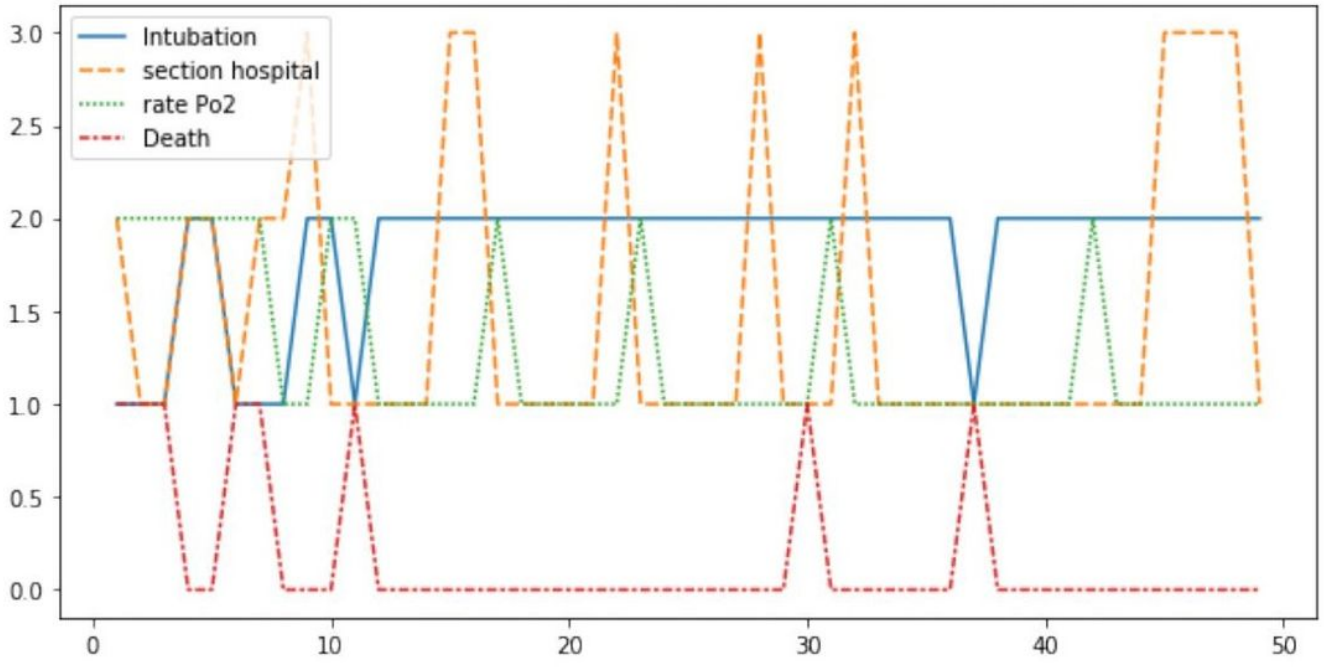

Figure 3

Line chart of intubation, hospital ward, blood oxygen level, and class label variables for 50 Covid-19 cases.

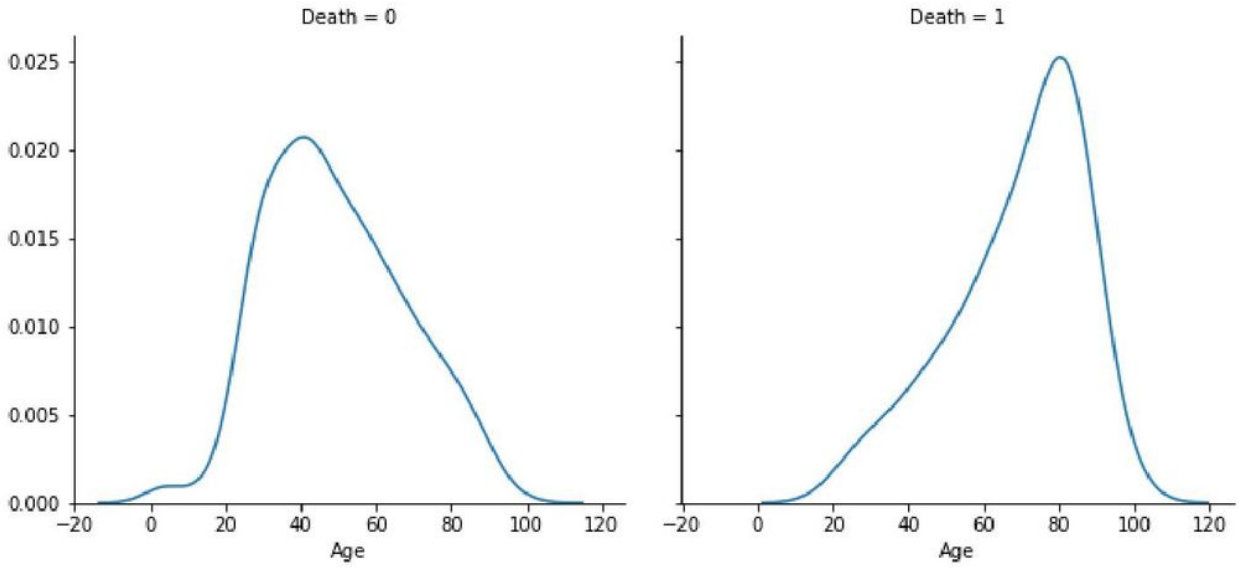

Figure 4

Facet chart of the age distribution of patients based on different class labels.

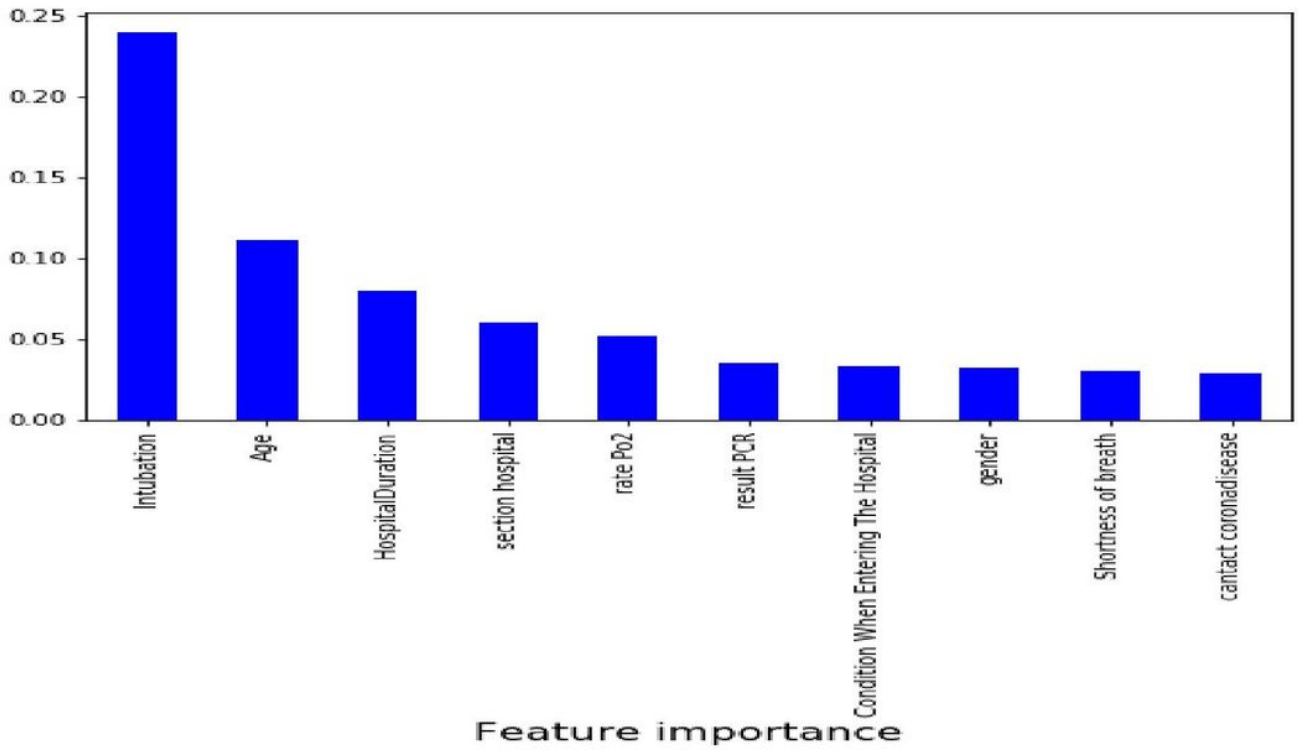

Figure 5

Top 10 influence attributes selected using the filter-based technique.

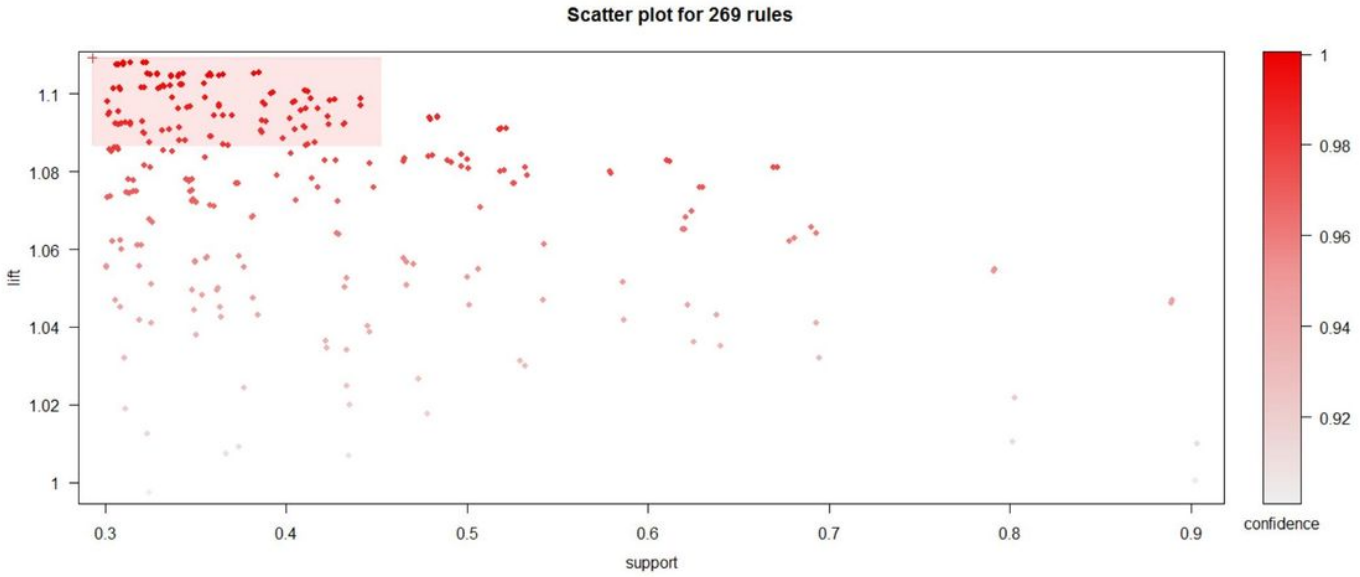

Figure 6

Scatter plot of the association rules extracted with the expected conditions (support = 0.3, confidence = 0.9, lift> 1).

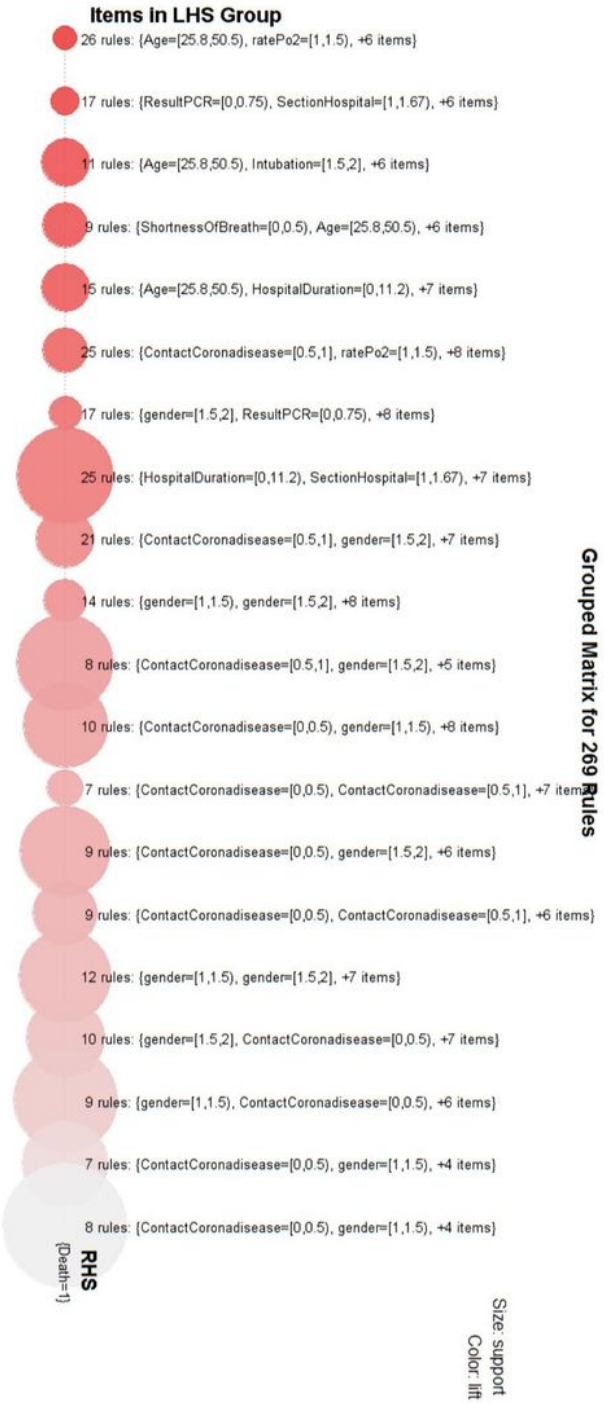

Figure 7

Grouped matrix diagram of extracted association rules with expected conditions (support 0.3, confidence 0.9, lift> 1).

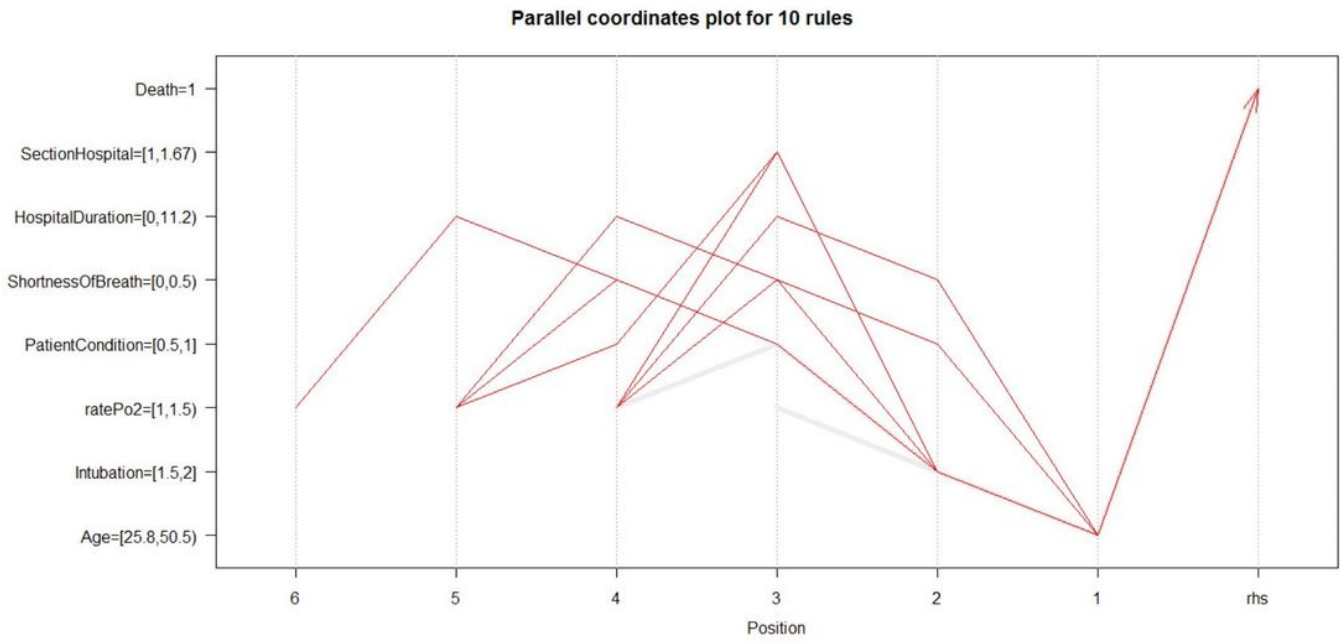

Figure 8

The ten most essential association rules discovered in this research.