

Respiratory failure in COVID-19 patients a comparative study of smokers to nonsmokers

Mohammad Kharabsheh^{1,2}, Shadi Banitaan², Hakam W. Alomari³, Mohammad Alshirah⁴,
Sukaina Alzyoud⁵

¹Department of Computer Information Systems, Faculty of Prince Al-Hussein bin Abdullah II of Information Technology, The Hashemite University, Zarqa, Jordan

²Department of Computer Science, College of Engineering and Science, University of Detroit Mercy, Detroit, United States of America

³Department of Computer Science and Software Engineering, College of Engineering and Computing, Miami University, Oxford, United States of America

⁴Departments of Computer Information Systems, Faculty of Information Technology, Al Albayt University, Mafrq, Jordan

⁵Department of Community and Mental Health Nursing, Faculty of Nursing, The Hashemite University, Zarqa, Jordan

Article Info

Article history:

Received Jan 23, 2022

Revised Jun 9, 2022

Accepted Jun 23, 2022

Keywords:

COVID-19

Decision support system

Machine learning

Respiratory

Smokers

ABSTRACT

For many decades, smoking tobacco has been a crucial concern due to respiratory failure. The potential relationship between smoking and COVID-19 has been recently investigated. In this paper, we study and investigate the role of the decision support system to predict the ratio of respiratory failure in smokers versus non-smokers among COVID-19 patients. We employed a classifier that predicts the ratio of respiratory failure as well as the ratio of the death toll between smokers and non-smokers using machine learning methods. The employed model demonstrate a prediction accuracy of 77% when applied on a sample from 23 countries that confirmed the highest number of COVID-19 patients. This was obtained from The World Bank Data-Health Nutrition and Population Statistics. As a result, a strong (significant) relationship between smoking tobacco and COVID-19 was illustrated by the employed model. Our approach achieves a good recall (78%). Thus, smokers are more susceptible to respiratory failure than non-smokers, as COVID-19 complications.

This is an open access article under the [CC BY-SA](#) license.



Corresponding Author:

Mohammad Kharabsheh

Department of Computer Science, College of Engineering and Science, University of Detroit Mercy

Detroit, MI 48221, United States of America

Email: mohkh86@hu.edu.jo, kharabmo@udmercy.edu

1. INTRODUCTION

Machine learning (ML) is defined as a branch of artificial intelligence (AI) that is interested in “teaching” computer how to process without the need to explicitly implement every possible scenario [1]. The main idea, in short, is to develop algorithms that are able to learn, by training, on a very large number of inputs, possibly with known results [2], [3]. Supervised learning is one of the major types of ML. This type fundamentally relies on estimating future instances based on known (current) instances. The major goal of supervised learning is to discover patterns of class labels that rely on predictor features. These patterns are utilized for the selection of class labels for testing instances. The selection process is based on the predictor features that are known [4], [5]. Feature selection (FS), also known as attribute selection, is an essential phase of building any predictive model [6], [7]. This is essential since the number of features could be large and others less informative. Without a doubt, the COVID-19 pandemic has changed the world’s view on life. This requires the world

to make great efforts, go hand in hand, and use the best available technologies to help predict infection and control the spread of this real threat [8], [9]. Coronavirus (SARS-CoV-2) still a widespread health contingency in the world [10]. Even though the evidence surrounding pharmaceutical therapy has been developed quickly, however the lack of well established preventive methods has made the efficient triage of COVID-19 patients a difficult process. While the forecast scores, such as the modified early warning score (MEWS) [11], are very helpful in determining the severity of illness of a COVID-19 patient, there is a lack of research studies that examine the capability of these scoring systems when predicting COVID-19 infected patients and mortality rates [12]. Also, the discriminatory capability of such rules-based results has been proven recently to lack quality [10].

From the epidemiologic predictions perspective, the expectation is that the hospitals will face an increased number of admissions of COVID-19 patients. And the expectation is that the patient triage will stay sustainable to smooth the efficient distribution of minimal resources [13]. Due to the early similarities of symptoms between patients who are in danger of decompensation and ones in need of mechanical ventilation, physicians become more aggressive in monitoring patients which consequently narrows down the procedures of more controlled climate for intubation. The longer the waiting time to decide about intubation puts patients at risk of danger complications, including, hypotension, peri-intubation hypoxia, cardiac arrest and arrhythmia [14].

The COVID-19 pandemic sets most of the health care systems in an inadequate situation and demands for modern methods to tackle this unmatched public health and clinical contingency. The clinical complications of COVID-19 range from asymptomatic issues to acute pneumonia, in which progression to respiratory failure is difficult to predict. Pneumonia, in any event, occurs in the second or third week of symptomatic infection, and it is described to have a death rate of 3% to 10%. Complications like pneumonia increase the need for mechanical intubation and run the risk of multi organ failure. Generally, patients should report the abrupt onset of dyspnoea during activity or rest [15], [16]. A respiratory rate greater than 30 breaths per minute, blood oxygen saturation less or equal to 93% and a partial pressure of arterial oxygen to fraction inspired oxygen ratio ($\text{PaO}_2/\text{FiO}_2$) is less than 300 mmHg are significant clinical signs of sharp respiratory failure syndrome acute respiratory distress syndrome (ARDS) leading to mild up to severe respiratory failure. Overall, there is a rising stage of doubt jointly in the progression of the patient's health care and in the speed at that patients improve respiratory fail demand mechanical ventilation. ML models, like those utilized to make the model, have shown potential to make predictive models that can be adopted to help and develop clinical ruling for a wide variety of results and have lately been used in echo to the COVID-19 emergency [17].

While current data science and machine learning technologies have proven to be very useful in diagnosing patients, tracking the spread of the virus and speeding up the process of finding an effective vaccine, health organizations and governments are still struggling to contain the spread of COVID-19 virus. So far, the two most effective techniques in combating the dissemination of COVID-19 are data science and machine learning. And those techniques are the ones that have aided China curb the spread of the virus in a short time [18], [19]. The use of ML to better understand risk factors in large and mixed groups of patients with Corona so that the use of algorithms in the objective evaluation of these factors can help determine the percentage of respiratory failure in smokers and non-smokers among patients with COVID-19.

Unfortunately, the pandemic in progression, there is limited research regarding the health status of the patients as well as their risk factors, such as smoking. Due to its detrimental impact on societies, smoking has been a significant concern for many generations. A study of the role of decision support system for COVID-19 patients to develop a prediction model to obtain a ratio of respiratory failure in smokers to non-smokers, who are suffering from COVID-19. In addition to predicting the death toll in those patients. As a result, to highlight the negative impact of tobacco as extensive evidence of a plethora of respiratory diseases. In conclusion, there are no studies in the literature that show that smoking increases the death tolls among COVID-19 patients or the severity of disease in those smokers.

The current study provides a new machine learning model that aims to predict the ratio of respiratory failure in smokers to non-smokers among patients with COVID-19, and the ratio of the death toll in smokers to non-smokers. This would help provide care to patients in a system where resources are limited by enabling risk recording, based on data from many health care delivery centers, including demographics, laboratory findings, and existing diseases. This study tries to address two major research questions:

- RQ1: Do machine learning approaches support predicting the ratio of respiratory failure between COVID-19 smokers patients and their opposites?

- RQ2: Could we predict a ratio of the death toll in smokers to non-smokers between COVID-19 patients through machine learning methods?

To answer these research questions, the study used supervised classifiers in which the introductory body is divided into two groups: i) a training group and ii) a testing group. The first group is the group that is used to train the advanced machine learner. On the other hand, learner performance is calculated by the second group. A 10-fold cross-validation [20] technique was used to obtain both training and test sets. Also, WEKA Toolkit [2], [21], [22] was used to perform supervised classification in our work. Finally, a set of different classification algorithms were used for the decision support systems provided to the healthcare industry that have been used to develop the employed models [23]-[26].

This paper is organized as follows. Section 2 reviews related works. Section 3 discusses the methodology that follows in this study. The study results are presented in section 4. Section 5 introduces the main threats to validity, followed by conclusions and some future research directions in section 6.

2. RELATED WORK

Burdick *et al.* [14] study aimed to improve machine learning based models for risk prediction critical illness outgrowth in COVID-19 patients. To evaluate how ML risk prediction models may help look after COVID-19 patients in a clinical setting. 197 patients were registered in the Respiratory decompensation and pattern for the triage of COVID-19 patients: a prospective study (READY) clinical trial. The study result showed that the algorithm had a higher diagnostic odds rate (DOR, 12.58) for predicting ventilation than a comparator before the usual time caution order, the modified early warning score (MEWS) [27]. The algorithm also carried out significantly higher sensitivity (0.90) than MEWS, which finished an allergy of 0.78, while preserving a higher specificity ($p < 0.05$).

Explaining research chronological, including research design, research procedure (in the form of algorithms, Pseudocode or other), how to test and data acquisition [7]-[17]. The description of the course of research should be supported references, so the explanation can be accepted scientifically [2], [6]. Figures 1-2 and Table 1 are presented center, as shown below and cited in the manuscript [7], [20]-[30]. The effects of electrical discharges to acidity of HVNE and NELV has been illustrated in Figure 2(a) and the effects of breakdown voltage of NE and NELV has been illustrated in Figure 2(b). Patanavanich and Glantz [28] study aimed to give out a meta-analysis of the association among smoking and the progression of the intended illness COVID-19. The results conclude that smoking is a danger factor for the progression of COVID-19, with smokers own higher odds of COVID-19 progress than those people who not smokers.

Ferrari *et al.* [17] study aimed to estimate a 48-hour prognosis of mild to acute respiratory failure, requiring mechanical ventilation, in hospitalized patients. The study represents a total of 198 patients giving a share in giving a rise to 1068 serviceable observations which let us build 3 predictive models founded respectively on 31-variables mark and symptoms, 39-variables laboratory biomarkers, and 91-variables as a structure of the two. The last model, the “boosted mixed model”, contains 20 variables chosen from model 3, carried out the best predictive execution (AUC=0.84) without doing to pot the FN rate. Its clinical performance was adopted in a narrative case report as an example. The study improved a machine model with 84% prognosis accuracy that is fit to help clinicians in the decision-making process and share to develop new analytics to develop care at high technology readiness levels.

The study that conducted by Lyu *et al.* [29] used qualitative and quantitative CT indicators of the chest to assess the clinical severity of COVID-19 pneumonia and characterize the topography of critical cases. 51 patients with COVID-19 pneumonia were registered and they were divided into three groups, one for normal cases (group A, n=12), severe cases (group B, n=15) and critical cases (group C, n=24), retroactively. Qualitative and quantitative indicators of chest CT were recorded and compared using fisher’s exact test, one-way ANOVA test, Kruskal-Wallis H test, and receiver operating characteristics analysis. The results showed that and depending on the severity of the disease, the number of affected lung segments and lobes, the frequency of consolidation, the insane paving pattern and the bronchopulmonary gram increase in more severe cases. Qualitative indicators, including total lung severity score and the overall result of mad paving and uniformity, could distinguish groups B and C of A (69% sensitivity, 83% specificity, and 73% accuracy) but were similar between group B and group C. The quantitative and qualitative indicators pooled among these three groups were of high sensitivity (B+C versus A, 90%; C versus B, 92%), qualitative (100%, 87%) and accuracy (92%, 90%). Critical cases had a higher overall severity score (> 10) and a higher overall score for insane paving and

consolidation (> 4) than normal cases.

Alshirah and Al-Fawa'reh [30] study aimed to become aware of phishing URLs using machine learning lexical feature-based analysis during adopting the method detects phishing URLs across analyzing URLs to take out lexical characteristics features. Subsequently, apply a machine learning method based on the take out features. The dataset was gathered from different sources, and it contains four dissimilar attack scenarios: (Defacement, spam, phishing, and malware). In spite of this, in this research, the emphasis was on Phishing URLs. The dataset was operated as input for numerous machine learning and statistical uncovering models “(Random forest (RF), decision tree classifier (DT), gaussian naive bayes (GNB), k-nearest neighbor (KNN), logistic regression, support vector classifier (SVC), quadratic discriminant analysis (QDA), perceptron, synthetic minority oversampling technique (SMOTE))”. These models were employed to predict Phishing URLs based on lexical characteristics features. The outcomes point to a comparatively good accuracy rate. The Random forest model has shaped the best accuracy (98%) likened to the other detection models. In addition, the RF takes shaped the best precision and recall (98%), correspondingly.

3. METHOD

The methodology that we followed in our study is presented in this section and shown in Figure 1. Firstly, we discuss the dataset that is used for our evaluation, specifically the collecting and processing processes. Then, we introduce the factors that are used in the learning of the classifier. Lastly, we present the developed model and the metrics that are used in the evaluation experiments.

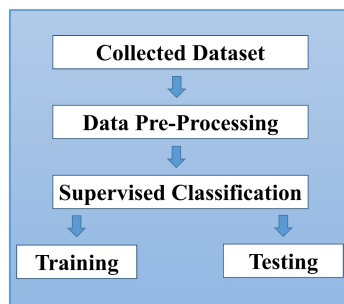


Figure 1. The proposed methodology

3.1. Studied dataset

Our study was conducted among a sample of 23 countries that confirmed the highest number of COVID-19 patients during January and February of 2020 from the world bank data: health nutrition and population statistics. In this research, we consider the variables classification factors that shows in Table 1. In this paper, was pre-processed the dataset using Microsoft Excel. This data was used as an input for various prediction models based on statistical model (logistic regression (LR)) and machine learning model (support vector machine (SVM), and multi-layer perceptron (MLP)). These models were utilized to predict potential patients of COVID-19 based on their signs and symptoms.

Table 1. Summary of classification factors

No	Classification factors
1	Country
2	TotalSmokingRate
3	MaleSmokingRate
4	FemaleSmokingRate
5	Pop2020
6	COVID19_confirmed
7	COVID19_recovered
8	COVID19_deaths
9	Confirmed to pop
10	Recovered to confirmed
11	Deaths to confirmed

3.2. Creating the corpus

The key step in performing the classification purpose is generating the corpus that characterizes the input of the classifiers, Figure 2, shows the major steps of our prediction model. For this work, the corpus includes the extracted values relevant to every classification factor for each instance of our studied dataset. Table 2, summarizes the corpus information. We used the level of failure in smokers to non-smokers from A to F.

Table 2. Summary of corpus information

Number of instances	A	B	C	D	E	F
250	10	65	119	22	22	12

3.3. Classification algorithms

In our experiments, we employed the supervised classifiers in which the included corpus is distributed into two groups as a training set and a testing set. The first group is the one that is used to train the developed machine learner. On the other hand, the performance of the learner is computed through the second group. We used the widely popular 10-fold cross-validation [31] technique to obtain both the training and testing sets to get unbiased results, which offered better model performance in my dataset. The WEKA toolkit is employed [2] to perform the supervised classification. There are various classification algorithms that widely used of decision support systems presented for the healthcare domain and have been used to develop the employed models [32], [33], these algorithms are as follows:

- SVM: seek to figure out a decision boundary between classes, expanding the margin of the separating line; while one of the drawbacks of this approach is that it can be only applied for binary classification [34]. SVM can construct the optimal separating lone, which increase the distance between the contiguous sample data [35]. SVM: this algorithm rises the dimensionality of training instances to achieve differentiable points in one of the dimensions. This algorithm is very popular since it is efficient in high dimensional spaces and thus provides more accurate results [36].
- Random tree: random tree is an ensemble training method for classification. This method is a set of separate decision trees in which each tree is produced from different samples and subsets of the training data. Random Tree is a supervised learning algorithm that produces many individual learners. It generates a random set of data for creating a decision tree. Random trees deal with both classification and regression problems. Random tree is a set of tree predictors (forest). The classifier gets the input feature vector, classifies it with every tree in the predictors. Random Tree is an active data mining algorithm that is used with large amounts of data. The technique employs several classification trees to a data set and next generates the prediction from all of the correlated trees [37], [38].
- Decision tree: decision tree in particular J.48 algorithm is commonly used to classify different data sets and perform accurate results of the classification. J48 algorithm is one of the best machine learning algorithms to investigate the data category continuously. it engages more memory space and reduces the performance and accuracy in classified data. This algorithm creates a binary tree for classification problems. The approach splits the data into range using the values of attributes for that item that are recognized in the training set [39].
- Naïve Bayes: Naive Bayes allocates the highly expected class when given characteristics are independent of any particular class. Naive Bayes is effective in many fields such as text categorization, and therapeutic diagnosis. This method assumes that all classification factors are independent. It shows great performance in terms of accuracy when it was applied in medical domain studies [40].
- SMO: sequential minimal optimization is used for solving the quadratic equation programming problem that occurs throughout the training of support-vector machines. SMO is commonly used for training SVMs because of high-speed training. This approach that trains a support vector learner using polynomial. It converts attributes from nominal to binary values [41].
- Logistic regression: logistic regression is a predictive analysis which estimates the probability of one dependent variable based on one or more independent variables. Logistic Regression is a linear model for categorization rather than regression. This approach uses regression models for classification tasks that models the posterior class probabilities for each of the needed n-classes from the dataset [42].
- K-Star: the main idea of the k-star is to take advantage of instance-based classifier and dataset features

reduction, the model has the ability to recognize features with high detection rate and low false negative. Selecting a good quality subset of features demonstrates to be significant in enhancing the performance of the system. Features are filtered to generate the most important feature subset before the start of the training process. K-Star represents a nearest neighbor method uses the distance calculations from the training set, such as the mahalanobis metric, to classify the instances of the testing set [43].

- Decision table: decision table is a method for prediction from decision trees and it is an ordered set of If-Then rules that have the possibility to be more efficient and therefore more reasonable than the decision trees. Selection to explore decision tables because it is an easier, less intensive algorithm than the decision tree-based approach. Decision table creates a decision table classifier and estimates feature subsets using best-first search and can utilize cross-validation for evaluation. The table for a given dataset is generated using grouping-and-counting in order to apply classification over unknown sample [44].
- K-NN: k-nearest neighbors' algorithm is a non-parametric technique used for categorization and regression. Nearest neighbor is a commonly used text classifier since of its ease and effectiveness. Its learning phase comprises storing all learning examples as classifier; therefore, it can be called as lazy learner because it's suspended the decision on how to generalize the learning data until each new instance is encountered. This technique is based on discovering the unidentified instances using the formerly known instances (e.g., nearest neighbor) and hence classify other instances using the voting approach [45].
- IBk: IBk is nearest-neighbor algorithm that uses the distance metrics created from the training set as closest associated vectors that would be used to classify data instances of the testing set [46].



Figure 2. The major steps of our prediction model

4. RESULTS AND DISCUSSION

To evaluate the effectiveness of our proposed classification model, we choose to use the following metrics:

- Precision: the ratio of retrieved instances that are truly relevant. It is calculated as $(P = \text{true positives} / (\text{true positives} + \text{false positives}))$ [47].
- Recall: the ratio of relevant instances that are retrieved by the classifier and hence it is computed as $(R = \text{true positives} / (\text{true positives} + \text{false negatives}))$ [47].
- F-Measure: a metric that depends on both recall and precision of a model and thus calculated by a combination of these two metrics as $((2 * \text{Recall} * \text{Precision}) / (\text{Recall} + \text{Precision}))$. The value of this metric is between

0 and 1 [47].

- Accuracy: A metric that represents the number of predictions that are incorrectly categorized positive and incorrectly categorized negative and is calculated as $(R=(\text{true positives}+\text{true negative})/(\text{true positives}+\text{true negative}+\text{false negatives}+\text{false positive}))$ [47].

We present our obtained results from the undertaken classification experiments and hence we answer the research questions mentioned early. The evaluation results show that our approach achieves a recall of 78% as we shown in Figure 3.

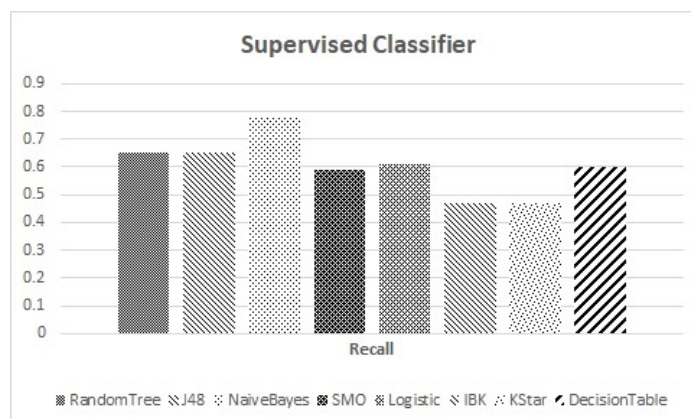


Figure 3. Evaluation results of our approach (recall)

RQ1: Do machine learning approaches support predicting the ratio of respiratory failure between COVID-19 smokers patients and their opposites? To answer the above question, the recall, precision, and F-measure metrics are used to evaluate the effectiveness of employed models. Using the factors that are given in Table 3, our proposed classifiers are trained using a combination of all these factors. A comparison between several classification techniques has been performed. A comparison with a baseline approach was also conducted. The performance results of the employed models (classifiers) are shown in Table 3. Our results show that there is an improvement in the prediction process in terms of all evaluation measures. For example, a comparison between our Naïve Bayes classifier and the baseline model shows a 0.78 in terms of recall and 0.77 in terms of precision improvement ratio. Which means it is possible to build machine learning models that have a highly accurate prediction capability of the ratio of respiratory failure in smokers versus non-smokers.

The second observation is that Naïve Bayes and SMO are more precise than the rest of the machine learning classifiers in terms of accuracy and F-measure. For example, Naïve Bayes calculate a probability for each class based on the probability distribution in the training dataset. As a result, the probability and prior are able to be updated dynamically to achieve flexibility and robustness to classification errors with each training example. On the other hand, the SMO learner achieves better F-measure because of increasing the dimensionality of data until the data points are differentiable in some dimension. Additionally, the space usage needed for SMO is linear in the size of the training set; therefore, it allows SMO to handle very large training sets with higher accuracy.

Table 3. Obtained classification results of the ratio of respiratory failure in smokers to non-smokers.

Learner	Accuracy	Recall	Precision	F-measure
RandomTree	0.56	0.65	0.53	0.58
J48	0.59	0.65	0.61	0.63
NaiveBayes	0.77	0.78	0.77	0.79
SMO	0.63	0.59	0.59	0.59
Logistic	0.48	0.61	0.47	0.53
IBK	0.49	0.47	0.43	0.45
KStar	0.52	0.47	0.71	0.57
DecisionTable	0.61	0.6	0.64	0.62

RQ2: Could we predict a ratio of the death toll in smokers to non-smokers through machine learning methods? To answer the second research question, we need to evaluate the usefulness of each feature separately as a predictor of mortality among smokers and non-smokers patients who infected with COVID-19 through machine learning methods. To do this, we developed the employed models (classifiers) using decision trees that were trained using all of the classification factors previously discussed. Using decision trees, we can classify traits based on their utility in our prediction experiments by performing a Top Node analysis linked to the decision tree approach. This node analysis approach calculates the presence of each factor under consideration by examining the structure and levels of the developed decision tree. Then, the tree level where the attribute occurs and the counted number of the attribute are used to determine the utility rank of that attribute, the most influencing factor will be the generated decision tree root node, while the factor's effectiveness decreases as we move toward the tree's leaves. Thus, in our study, we developed a decision tree using the C4.5 algorithms, which was trained using all the factors researched in this work. C4.5 is a greedy technique that adds decision nodes at each level of the generated tree by following a training set team-and-con algorithm. At each stage of the running algorithm, the information observed from each attribute is computed, and next to the attribute with the highest ranking, the steps of running the greedy algorithm are set to a certain threshold value, which is used to determine the number of records in the terminal nodes while building the tree. The performance results obtained from our decision tree classifier are We present our obtained results from the undertaken classification experiments and hence we answer the research questions mentioned early. The evaluation results show that our approach achieves a recall of 78% as we shown in Figure: i) recall: 0.52, ii) precision: 0.49 and iii) f-measure: 0.50.

In addition, our analysis results are mentioned in Table 4. Specifically, for each influential factor, the table provides the level at which it appears in the created tree (e.g., The first column) and the occurrence frequency associated with the factor (such as the second column). As we can see, the percentage of smokers represents the root node of our resulting tree, and is therefore the most influential factor in our experiments. That is, the death rate among smokers infected with COVID-19 will be the most expected rate. As for the gender of the smoker, the percentage of females represents the highest rate of confirmation of infection with COVID-19, and the frequency associated with factor (1) in the first column, and finally, the proportion of those recovering from those infected with COVID-19 is considered the most influential factor in the incidence associated with factor (2) for the first column.

Table 4. Outcomes of top node analysis

Level	Occurrence Count	Attribute
0	6	TotalSmokingRate
	2	MaleSmokingRate
1	7	FemaleSmokingRate
	11	Covid19_Confirmed
2	13	Covid19_Recovered
	2	Covid19_Deaths
	3	Confirmed To Pop

5. THREATS TO VALIDITY

As with any case study that based on a sample of smokers and non-smokers, we have some potential threats that prevent us from generalizing our findings to different data sets in various settings. The data set could not be illustrative for all the samples, so we could not generalize our results to a variety of data sets. Moreover, there may be other features that were not present that were used in this study (for example, smoking sessions, the psychological state of the patient with COVID-19, and the age of the person). These factors may positively influence the results we obtained. Our developed classifiers are based on successful machine learning techniques that are widely used in the literature. However, there is a classification of some flaws in each method that may negatively affect the validity of our experiments. Therefore, developing classifiers using other machine learners will be our future consideration.

6. CONCLUSION AND FUTURE WORK





In this study, we've investigated whether machine learning approaches can help predict the ratio of respiratory failure in smokers to non-smokers among COVID-19 patients and the ratio of the death toll in smokers to non-smokers. Employed the supervised classifiers in which the inputted corpus is distributed into two groups: a training set and a test set. The first group is the one that is used to train the developed machine learner. On the other hand, the performance of the learner is computed through the second group. Here, we used the widely popular 10-fold cross-validation technique to obtain both the training and test sets. We employed the WEKA toolkit to perform the supervised classification in our work. We also discussed the various classification algorithms that were widely used in the literature of decision support systems presented for the healthcare domain and have been used to develop the employed models (classifiers). Our results show that the best equitable recall of 65% and the worst recall value is 47%. Add to that, the employed model achieved the best precision value of 71% and the worst value of 43%. By performing a top node analysis, we found that the attribute smokers are the most influential attribute in predicting the death rate among smokers infected with COVID-19 will be the most expected rate. As for the gender of the smoker, the percentage of females represents the highest rate of confirmation of infection with COVID-19 and the frequency associated with the factor (1). Finally, the proportion of those recovering from those infected with COVID-19 is considered the most influential factor in the incidence associated with the factor (2). We aim to explore more classification factors and study predicting the ratio of respiratory failure in smokers to non-smokers between COVID-19 patients using the machine learning approach and predicting nicotine dependencies in future studies in order to achieve better prediction performance. We plan to enrich our study by investigating more varies datasets from different countries and environments.

REFERENCES




- [1] F. Thung, S. Wang, D. Lo, and L. Jiang, "An empirical study of bugs in machine learning systems," in *2012 IEEE 23rd International Symposium on Software Reliability Engineering*, 2012, pp. 271–280, doi 10.1109/ISSRE.2012.22.
- [2] F. Eibe, M. A. Hall, and I. H. Witten, "The weka workbench. online appendix for data mining: practical machine learning tools and techniques," in *Morgan Kaufmann*, 2016.
- [3] M. Bkassiny, Y. Li, and S. K. Jayaweera, "A survey on machine-learning techniques in cognitive radios," *IEEE Communications Surveys & Tutorials*, vol. 15, no. 3, pp. 1136–1159, 2012, doi:10.1109/surv.2012.100412.00017.
- [4] A. Al-Nusirat, F. Hanandeh, M. Kharabsheh, M. Al-Ayyoub, and N. Al-dhufairi, "Dynamic detection of software defects using supervised learning techniques," *International Journal of Communication Networks and Information Security*, vol. 11, no. 1, pp. 185–191, 2019.
- [5] J. Alzyoud, M. Kharabsheh, S. Alzyoud and E. Alzbon, "Use of healthcare informatics applications and data for research purposes by students: opportunities and challenges in Jordan," 2019 IEEE 89th Vehicular Technology Conference (VTC2019-Spring), 2019, pp. 1-6, doi: 10.1109/VTCSpring.2019.8746618.
- [6] M. Kharabsheh, O. Meqdadi, M. Alabed, S. Veeranki, A. Abbadi, and S. Alzyoud, "A machine learning approach for predicting nicotine dependence," *International Journal of Advanced Computer Science and Applications*, vol. 10, no. 3, pp. 179–184, 2019.
- [7] M. Kharabsheh, A. Qawasmeh, O. D. Megdadi, N. Jawabreh, R. H. Mudallal, and S. A. Alzyoud, "A critical analysis of the relationship between depression and smoking using machine learning," *International Journal of Scientific & Technology Research*, vol. 8, pp. 22–26, 2019.
- [8] A. Kumar, P. K. Gupta, and A. Srivastava, "A review of modern technologies for tackling covid-19 pandemic," *Diabetes & Metabolic Syndrome: Clinical Research & Reviews*, vol. 14, no. 4, pp. 569–573, 2020.
- [9] R. Vaishya, M. Javaid, I. H. Khan, and A. Haleem, "Artificial intelligence (ai) applications for covid-19 pandemic," *Diabetes & Metabolic Syndrome: Clinical Research & Reviews*, vol. 14, no. 4, pp. 337–339, 2020.
- [10] A. Lorusso, P. Calistri, A. Petrini, G. Savini, and N. Decaro, "Novel coronavirus (sars-cov-2) epidemic: a veterinary perspective," *Veterinaria italiana*, vol. 56, no. 1, pp. 5–10, 2020.
- [11] C. P. Subbe, M. Kruger, P. Rutherford, and L. Gemmel, "Validation of a modified early warning score in medical admissions," *Qjm*, vol. 94, no. 10, pp. 521–526, 2001.
- [12] L. Y. Hsu, P. Y. Chia, and J. Lim, "The novel coronavirus (sars-cov-2) pandemic," *Ann Acad Med Singap*, vol. 49, no. 3, pp. 105–7, 2020.
- [13] B. M. Robinson, "Malignant pleural mesothelioma: an epidemiological perspective," *Annals of cardiothoracic surgery*, vol. 1, no. 4, p. 491, 2012.
- [14] H. Burdick *et al.*, "Prediction of respiratory decompensation in covid-19 patients using machine learning: The ready trial," *Computers in biology and medicine*, vol. 124, p. 103949, 2020.
- [15] C. Solinas, L. Perra, M. Aiello, E. Migliori, and N. Petrosillo, "A critical evaluation of glucocorticoids in the management of severe covid-19," *Cytokine & growth factor reviews*, vol. 54, pp. 8–23, 2020.
- [16] K. Swedberg *et al.*, "Guidelines for the diagnosis and treatment of chronic heart failure: executive summary (update 2005) the task force for the diagnosis and treatment of chronic heart failure of the european society of cardiology," *European heart journal*, vol. 26, no. 11, pp. 1115–1140, 2005.
- [17] D. Ferrari *et al.*, "Machine learning in predicting respiratory failure in patients with covid-19 pneumonia—challenges, strengths, and opportunities in a global health emergency," *PloS one*, vol. 15, no. 11, p. e0239172, 2020.

- [18] S. Latif *et al.*, “Leveraging data science to combat covid-19: A comprehensive review,” *IEEE Transactions on Artificial Intelligence*, vol. 1, no. 1, pp. 85–103, 2020.
- [19] L. Surya, “How government can use ai and ml to identify spreading infectious diseases,” *International Journal of Creative Research Thoughts (IJCRT)*, ISSN, vol. 6, no. 1, pp. 2320–2882, 2018.
- [20] C. Chien and G. J. Pottie, “A universal hybrid decision tree classifier design for human activity classification,” *2012 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, 2012, pp. 1065–1068, doi: 10.1109/EMBC.2012.6346118.
- [21] W.-L. Zuo, Z.-Y. Wang, T. Liu, and H.-L. Chen, “Effective detection of parkinson’s disease using an adaptive fuzzy k-nearest neighbor approach,” *Biomedical Signal Processing and Control*, vol. 8, no. 4, pp. 364–373, 2013.
- [22] C.-H. Jen, C.-C. Wang, B. C. Jiang, Y.-H. Chu, and M.-S. Chen, “Application of classification techniques on development an early-warning system for chronic illnesses,” *Expert Systems with Applications*, vol. 39, no. 10, pp. 8852–8858, 2012.
- [23] C. Vaghela, N. Bhatt, and D. Mistry, “A survey on various classification techniques for clinical decision support system,” *International Journal of Computer Applications*, vol. 116, no. 23, 2015.
- [24] B. A. Thakkar, M. I. Hasan, and M. A. Desai, “Health care decision support system for swine flu prediction using naïve bayes classifier,” *2010 International Conference on Advances in Recent Technologies in Communication and Computing*. 2010, pp. 101–105, doi: 10.1109/ARTCom.2010.98.
- [25] M. W. Moreira, J. J. Rodrigues, V. Korotaev, J. Al-Muhtadi, and N. Kumar, “A comprehensive review on smart decision support systems for health care,” *IEEE Systems Journal*, vol. 13, no. 3, pp. 3536–3545, 2019, doi: 10.1109/JSYST.2018.2890121.
- [26] B. Ozaydin, J. M. Hardin, and D. C. Chhieng, “Data mining and clinical decision support systems,” in *Clinical Decision Support Systems*. Springer, 2016, pp. 45–68, doi: 10.1007/978-0-387-38319-4_3.
- [27] J. Gardner-Thorpe, N. Love, J. Wrightson, S. Walsh, and N. Keeling, “The value of modified early warning score (mews) in surgical in-patients: a prospective observational study,” *The Annals of The Royal College of Surgeons of England*, vol. 88, no. 6, pp. 571–575, 2006.
- [28] R. Patanavanich and S. A. Glantz, “Smoking is associated with covid-19 progression: a meta-analysis,” *Nicotine and Tobacco Research*, vol. 22, no. 9, pp. 1653–1656, 2020.
- [29] P. Lyu, X. Liu, R. Zhang, L. Shi, and J. Gao, “The performance of chest ct in evaluating the clinical severity of covid-19 pneumonia: identifying critical cases based on ct characteristics,” *Investigative radiology*, vol. 55, no. 7, pp. 412–421, 2020.
- [30] M. Alshira’h and M. Al-Fawa’reh, “Detecting phishing urls using machine learning lexical feature-based analysis,” *Int. J. Adv. Trends Comput. Sci. Eng.*, vol. 9, no. 4, pp. 5828–5837, 2020.
- [31] B. Efron, “Estimating the error rate of a prediction rule: improvement on cross-validation,” *Journal of the American statistical association*, vol. 78, no. 382, pp. 316–331, 1983.
- [32] M. Shouman, T. Turner, and R. Stocker, “Applying k-nearest neighbour in diagnosing heart disease patients,” *International Journal of Information and Education Technology*, vol. 2, no. 3, pp. 220–223, 2012.
- [33] I. H. Witten, E. Frank, M. A. Hall and C. Pal, “Data Mining: Practical machine learning tools and techniques,” in *The Morgan Kaufmann Series in Data Management Systems*, 2011.
- [34] W. Noble, “What is a support vector machine? nature biotechnology,” *Nature Biotechnology*, vol. 24, pp. 1565-1567, 2006.
- [35] K. P. Bennett and C. Campbell, “Support vector machines: hype or hallelujah?” *ACM SIGKDD explorations newsletter*, vol. 2, no. 2, pp. 1–13, 2000.
- [36] P. Ahmad, S. Qamar, and S. Q. A. Rizvi, “Techniques of data mining in healthcare: a review,” *International Journal of Computer Applications*, vol. 120, no. 15, 2015.
- [37] B. Pfahringer, “Random model trees: an effective and scalable regression method,” University of Waikato, Department of Computer Science, Hamilton, New Zealand, 2010. [Online]. Available: <https://hdl.handle.net/10289/4056>.
- [38] K. Wisaeng, “A comparison of different classification techniques for bank direct marketing,” *International Journal of Soft Computing and Engineering (IJSCE)*, vol. 3, no. 4, pp. 116–119, 2013.
- [39] N. Saravana and D. V. Gayathri, “Performance and classification evaluation of j48 algorithm and kendall’s based j48 algorithm (knj48),” *Int. J. Comput. Trends Technol. (IJCTT)–Volume*, vol. 59, 2018.
- [40] I. Rish *et al.*, “An empirical study of the naive bayes classifier,” in *IJCAI 2001 workshop on empirical methods in artificial intelligence*, vol. 3, no. 22, 2001, pp. 41–46.
- [41] J. Platt, “Sequential minimal optimization: A fast algorithm for training support vector machines,” <https://www.microsoft.com/en-us/research/publication/sequential-minimal-optimization-a-fast-algorithm-for-training-support-vector-machines/>, accessed (May 26, 2022).
- [42] E. W. Steyerberg, F. E. Harrell Jr, and P. H. Goodman, “Neural networks, logistic regression, and calibration,” *Medical Decision Making*, vol. 18, no. 3, pp. 349–350, 1998.
- [43] D. Y. Mahmood and M. A. Hussein, “Intrusion detection system based on k-star classifier and feature set reduction,” *International Organization of Scientific Research Journal of Computer Engineering (IOSR-JCE) Vol*, vol. 15, no. 5, pp. 107–112, 2013.
- [44] G. Banerji and K. Saxena, “An efficient classification algorithm for real estate domain,” *India: International Journal of Modern Engineering Research (IJMER) www.ijmer.com*, vol. 2, no. 4, 2012.
- [45] F. Sebastiani, “Machine learning in automated text categorization,” *ACM computing surveys (CSUR)*, vol. 34, no. 1, pp. 1–47, 2002.
- [46] D. W. Aha, D. Kibler, and M. K. Albert, “Instance-based learning algorithms,” *Machine learning*, vol. 6, no. 1, pp. 37–66, 1991.
- [47] D. M. Powers, “Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation,” *arXiv preprint arXiv:2010.16061*, 2020.




BIOGRAPHIES OF AUTHORS

Mohammad Kharabsheh, Ph.D.,     Visiting Research Scholar at the University of Detroit Mercy. He joined The Hashemite University in the Department of Computer Information System in the Prince al Hussein bin Abdullah II faculty of IT. He received his BSc. in Computer Science from Al-Yarmouk University and his MSc Master degree in Computer Science from Jordan University of Science and Technology, Irbid, Jordan (2007). Dr Kharabsheh holds the PhD in Computer Science from University of Leicester, United Kingdom. His research interests include Artificial Intelligence Applications and Machine Learning, Database Management Systems, in addition to the Health Informatics. He can be contacted at email: mohkh86@hu.edu.jo.






Shadi Banitaan, Ph.D.,    (Member, IEEE) received the B.S. degree in computer science and the M.S. degree in computer and information sciences from Yarmouk University and the Ph.D. degree in computer science from North Dakota State University. He worked as an Instructor with the University of Nizwa, from 2004 to 2009. In 2013, he joined the University of Detroit Mercy, where he is currently an Associate Professor and the Director of computer science and software engineering. His research interests include software engineering, machine learning, and data mining. He is a member of the Association for Computing Machinery (ACM) and the IEEE Computer Society. He can be contacted at email: banitash@udmercy.edu.






Hakam W. Alomari, Ph.D.,    is an Assistant Professor in the Department of Computer Science and Software Engineering. He joined Miami University in 2015. Dr. Alomari received his Bachelor degree in Computer Science from the Yarmouk University in 2004, a Master's degree in Computer Science from Jordan University of Science and Technology in 2006, and a Ph.D. in Computer Science from Kent State University, Ohio, USA in 2012. Dr. Alomari's research focuses on developing and constructing methods for lightweight static program analysis. The objective is to develop new analysis methods that are highly scalable for application on very large software systems. He can be contacted at email: alomarhw@miamioh.edu.



Mohammad Alshirah, Ph.D.,    is an Associate Professor at the Information Systems Department at Al al-Bayt University. He received his BSc. in Software Engineering from the Hashemite University, Jordan in 2007 and received his MSc in Computer Science from Al al-Bayt University, Jordan in 2010. He obtained his PhD in Software Engineering at the Department of Computer Science of the University of Leicester in the United Kingdom in 2016. He has a variety of academic and professional qualifications and experience. His research interests include, but are not limited to, Machine Learning, Software Usability and User Experience. He can be contacted at email: alshirah@aabu.edu.jo.



Sukaina Alzyoud    is a professor and the REACH program Manager. She is also the focal point for U.S.-Jordan UCN at Hashemite University. She has served in various administrative and academic roles. Most recently Dr. Alzyoud was the Dean of Academic Development and International Outreach [DADIO] at Hashemite University. She is also considered an active researcher serving as a PI on several projects in the fields of public health and higher education. She secured more than \$300K in funding for her projects from national and international organizations. In collaboration with her research team, she has published several articles. In recent years Dr. Alzyoud has founded and co-founded several initiatives and research groups in higher education and public health. She can be contacted at email: sukaina-alzyoud@hu.edu.jo.