journals.uoz.edu.krd

Available online at sjuoz.uoz.edu.krd

Science Journal of University of Zakho

Vol. 12, No.2, pp.244–249 April-June, 2024

p-ISSN: 2663-628X
e-ISSN: 2663-6298

# CORRELATION OF COVID-19 TO LUNG INFECTION AND PREDICTION OF LUNG INFECTION IN COVID-19 PATIENTS IN IRAQ USING DATA MINING METHODS

Shivan Sarkawt Alomadi [a*], Jihan Abdulazeez Ahmed Rasool [a]

[a] Department of Computer Science, College of Science, University of Duhok, Kurdistan Region-Iraq – (shivansami@gmail.com, drjihanrasool@uod.ac)

Received: 18 Feb., 2024 / Accepted: 21 Apr., 2024 / Published: 23 June., 2024. https://doi.org/10.25271/sjuoz.2024.12.2.1273

## ABSTRACT:

The Covid-19 pandemic emerged as an unforeseen global crisis, exerting a profound influence on various aspects of human life. Hence, the need for collaborative efforts and scholarly investigations to address and alleviate the challenges arising from this crisis is crucial. One notable concern pertains to lung infections, which are recognized as a highly perilous consequence of the aforementioned virus. Thus, this study aims to investigate the potential correlation between Covid-19 and lung infections, and test the efficacy of various algorithms in predicting lung infections amongst Covid-19 patients. For this purpose, data has been procured from multiple health institutions in Iraq. Using this data, a robust correlation between Covid-19 and lung infection cases was found and the bagging, boosting, naïve Bayes, K-Nearest Neighbour, J48, random forest, PART, and logistic regression algorithms showcased a high accuracy in prediction lung infection in Covid-19 patients, with naïve Bayes achieving the highest accuracy of 93.41 percent.

**KEYWORDS:** Covid-19, lung infections, Iraq, correlation, data mining, machine learning, Bagging, Boosting, Naïve Bayes, K-Nearest Neighbour, J48 decision tree, Random Resonance Theory, Binary Logistic Regression.

## 1. INTRODUCTION

The Covid-19 pandemic was an unpredicted crisis that impacted the globe, influencing all factors of life. The first confirmed case of the Covid-19 virus was officially identified on December 20, 2019, at Wuhan Junyintan Hospital in Wuhan City, located in Hubei Province, China. The spreading of Covid-19 was rapid on a global scale (Harapan et al., 2020; Vieira et al., 2020; Yu et al., 2020). Because of the severity of the situation, the World Health Organization (WHO) declared formal pandemic state in March 2020 (Güner et al., 2020). Initial preventive measures for the population were behavioural advises, namely, keeping away from crowded spaces with inadequate ventilation, wearing face masks, and abiding to social distancing. (MacIntyre and Wang, 2020). Because of the rapid increase of infected persons, a vast amount of Covid-19 data became available. The available data were used to study the virus and its impacts. That was mostly done utilizing data mining methods and techniques.

Data mining is an approach that is employed to extract valuable, innovative, and patterns that are less prominent and information from datasets (Hussain et al., 2019). Data mining is a combination of procedures and algorithms acquired from the realm of data management, artificial intelligence, statistics, and machine learning. aimed at uncovering concealed patterns and/or relationships, methodologies of data mining are utilized across various domains including, but not limited to, science, marketing, insurance, finance, banking, retail, fraud detection, and engineering (Harding et al., 2006).

With the rapid spread of Covid-19 infections since early 2020, many researchers have turned to Data Mining techniques to acquire knowledge on Covid-19 from the available patient data. Skopljanac et al. (2021) revealed that a significant proportion of deaths associated with Covid-19 were attributed to pneumonia infections. Wang et al. (2021) investigated the relation between the severity of lung infection and clinical laboratory indicators in Covid-19 patients. They employed Spearman correlation test to analyse the data of 31 Covid-19 patients. Their results revealed a statistically significant correlation between the percentage of infection (POI) in lung and the peripheral blood lymphocyte and neutrophil levels. They also determined that the point of interest (POI) of the lung was not significantly correlation to either peripheral blood white blood cell (WBC) count, monocyte percentage (M%), or haemoglobin (HGB) content.

Yağmur et al. (2021) conducted a study to investigate the correlation of olfactory dysfunction (OD) with lung infection and severity in Covid-19 patients. They collected the data of 180 Covid-19 patients and used a combination of the Kolmogorov–Smirnov test and Multivariate Logistic Regression analysis. They found a significant correlation between OD and lung infection in Covid-19 patients, they concluded that patients who reported OD had milder cases of infection and a lower rate of lung involvement, and were also less likely to require admittance to the intensive care unit (ICU).

Francone et al. (2020) investigated the association between computed tomography (CT) score and the severity of lung infection in individuals diagnosed with Covid-19. The investigation involved the acquisition of data of 130 Covid-19 patients. To explore various aspects of the data, they conducted a series of statistical analyses on the dataset employing Mann-Whitney test, Kruskal-Wallis test, Chi-square test, and Kaplan-Meier method. The researchers determined that a robust correlation exists between the CT score and the severity of infection in individuals diagnosed with Covid-19, stating that the CT score could be used as an early predictor for the outcome of the infection. Dawoud et al. (2020) also performed a study testing the correlation of CT score to infection severity in Covid-19 patients. The study collected data of 212 individuals diagnosed with Covid-19. They employed Shapiro-Wilk test and the one-way ANOVA method for data analysis. Similar to the findings of Francone et al. (2020), they predicted that a strong correlation exists between CT scores and the severity of infection in Covid-19 patients, stating as well that chest CT examinations could be used as a method for diagnosing Covid-19 pneumonia.

* Corresponding author

Annoni et al. (2021) conducted a study aimed at estimating the correlation between percentage of lung parenchyma volume to the total lung volume and the patient's clinical course in Covid-19 infected patients. The records of 76 Covid-19 patients were utilized. They utilized Mann-Whitney test, Chi-square test, and Multivariable Logistic Regression using MedCalc software. Their results indicated a notable correlation between the percentage of lung parenchyma volume and the clinical outcome of Covid-19 patients.

In a different case, Tung-Chen et al. (2020) explored the correlation between chest CT score and lung ultrasonography (LUS) in patients diagnosed with Covid-19. The data of 51 Covid-19 patients were analysed using Man-Whitney, Chi-Square, and Cohen's kappa tests. They found that there was no missing diagnosis in LUS exams when compared to CT scan findings, concluding that there was a highly significant correlation in the results of LUS exams and those of CT scan.

In a more recent study, Mahammedi et al. (2021) attempted to find the correlation between brain and lung imaging findings in individuals diagnosed with Covid-19. The data of 135 individuals diagnosed with Covid-19 were involved in the study. They employed Kolmogorov-Smirnov test, Wilcoxon rank sum test, Chi-square test, and Fisher exact test. The results of the study revealed that there is a significant correlation between the brain CT findings and lung CT findings. The conclusion was that the lung CT score could be used as a predictive indicator for acute abnormalities observed in neuroimaging among Covid-19 patients.

In a study conducted by Chen et al. (2020), the correlation between cytokine profiles and lung injury in individuals diagnosed with Covid-19 was evaluated. They utilized a combination of student's t test, Kruskal-Wallis test, Chi-square test, and Spearman rank test. The dataset was consisting of the medical records of 106 Covid-19. The results indicated a significant correlation between the presence of lung injury and the elevation of circulating cytokines.

Based on the aforementioned research conducted on Covid-19 via data mining methodology, a substantial body of evidence from diverse sources indicates a potential correlation between Covid-19 infections and pulmonary infections, albeit with different outcomes (Duzgun et al., 2020). In those findings, the occurrence and intensity of the lung infections are greatly influenced by the surrounding environment, as the prevalence of infection associated with the same diseases can differ across different environments (Olivieri & Scoditti, 2005).

Thus, the objective of this study is to gather data from Iraqi patients who have experienced the post Covid-19 pandemic period. This was followed by conducting an analysis and examination of the potential correlation between Covid-19 and lung infections in the environment of Iraq. Then, data mining and machine learning algorithms will be used in the identification of pulmonary infection in individuals afflicted with Covid-19.

The structuring of this paper is as follows: section 2 presents the methodology used in this study and provides a theoretical background for the techniques; section 3 displays the results and their discussion; and section 4 is the conclusion of the paper.

## 2. METHODOLOGY AND THEORETICAL BACKGROUND

The study employed a systematic methodology encompassing data collection, evaluation of Covid-19-induced lung infection through various tests, and the utilization of algorithms for the prediction of lung infection in patients affected by Covid-19. The following is a concise summary of each individual step.

### 2.1 Data Collection

Following the approval obtained from the Research Ethics Committee of the Directorate General of Health in Duhok, located in the Kurdistan Region of Iraq, permission was granted for the collection of patient data for the purpose of this study. A substantial dataset was acquired from an internal physician clinic situated within the Kurdistan Private Hospital in Duhok, Kurdistan Region-Iraq. The dataset consists of records pertaining to 1024 patients from the year 2021. The selected attributes include patient gender, age, Covid-19 infection status, and lung infection status. In addition, a second dataset was obtained through personal communications with Talib (2021). With approval from the Public Health Unit /Al-Karkh, affiliated to the Iraqi Ministry of Health, the data was collected from the patients of a wide number of hospitals in the year 2020. This dataset comprises records of 501 patients. Each record includes: age, fever, dry cough, chronic diseases, nasal congestion, hearing problems, loss of smell, diarrhoea, body-ache and fatigue, sore throat, travel history, breathing problems, Covid-19 status, and lung infection status.

### 2.2 Correlation Testing

Before testing the predictive algorithms, the correlation between Covid-19 infections and lung infections must be determined. Towards this purpose, the Chi-Squared Test of Independence is used to test the significance of the correlation (McHugh, 2013), followed by the Cramer's V test to determine the strength of the association between the two variables (Wu et al., 2012).

### 2.3 Machine Learning Algorithms

After the correlation between Covid-19 and lung infections was evaluated, data mining and machine learning algorithms were used in predicting lung infections in Covid-19 patients. The algorithms are briefly introduced in the following.

**2.3.1 Decision Tree:** A classifier can be constructed by employing a recursive partition of the instance space, as described by Maimon and Rokach (2005). The decision tree is composed of nodes that form a rooted tree structure. The rooted tree is characterized by a node referred to as the "root," which does not have any incoming edges, unlike other nodes in the network which have a single incoming edge. In addition, a node within a graph or tree structure that has outgoing edges is referred to as a test or internal node. The remaining nodes are known as leaves, which are also known as decision or terminal nodes. In a decision tree, every internal node partitions the instance space into multiple sub-spaces based on a specific discrete function that considers the values of the input attributes. Each test in the experimental procedure is designed to specifically target and evaluate a singular attribute. This is done in order to effectively and accurately divide the instance space based on the value of the attribute, following the most straightforward and commonly observed scenario. The condition refers to a range within the context of numeric attributes.

The J48 algorithm is classified as one of several Univariate Decision Tree algorithms. A Univariate Decision Tree refers to a type of decision tree in which the process of splitting is conducted by utilizing only one attribute at the internal nodes (Bhargava et al., 2013). Bagging and Boosting: When multiple models' outputs are aggregated, a unified prediction is generated by consolidating all of the individual outputs. When it comes to classification, a straightforward approach involves conducting a voting process. Similarly, in the context of numerical prediction, a basic method is to calculate the average. In both scenarios, the vote and the average can be either conducted using standard methods or employing a weighted approach.

Both Bagging and Boosting use this strategy, although they differ in their respective implementations. The difference lies in the calculation of the variables' weights. In Bagging, attributes are assigned equal weight, whereas Boosting incorporates weighting to assign greater importance to attributes that have a stronger impact. This can be equated to a managerial decision where one expert's advice is given more weight over another based on how accurate their past predictions have been (Witten et al., 2011).

**2.3.2 Random Forest:** The random forest algorithm proposed by Breiman (2001) has seen great success as a general-purpose classification and regression technique. The method has demonstrated good performance in situations where the quantity of variables is significantly greater than the quantity of observations. The proposed method involves the integration of multiple randomized decision trees, followed by the aggregation of their individual predictions through averaging. Moreover, the algorithm is highly adaptable, making it suitable for addressing complex problems. And it is easily customizable for a diverse range of ad hoc learning tasks, and it generates metrics of diverse significance (Biau & Scornet, 2016).

**2.3.3 Naïve Bayes:** Naive Bayes, derived from Bayes' theory, is a statistical classification algorithm that can be employed to estimate the probability of a particular class (Abdulrahman & Rasool, 2020). The theory shares similar classification capabilities with decision trees and neural networks. This theory offers a systematic approach to assess the probability of a specific event by considering the probabilities of associated events. Nevertheless, the model's name suggests that it is naive in its assumption of the independence of all features within the dataset. It is crucial to acknowledge that the presence of co-dependent features can decrease the accuracy of the model. However, Naive Bayes has been proven to achieve high levels of accuracy and speed when employed in the context of databases containing large volumes of data (Santoso et al., 2020).

**2.3.4 K-Nearest Neighbour:** The K-Nearest Neighbour (KNN) classifier, a supervised and non-parametric data mining technique, is employed for both classification and regression tasks (Altman, 1992). The K nearest training datasets in the feature space are the input variables for both tasks. The K-Nearest Neighbours (KNN) algorithm relies on labelled input data in order to understand the underlying function and generate the desired output for unlabelled input. In the context of K-nearest neighbours (KNN) classification, the outcome is a class membership for each data instance based on the class that receives the greatest percentage of support from its K-nearest neighbours. On the other hand, the outcome of a K-nearest neighbours (KNN) regression is the attribute value of a given data instance, which is computed as the mean of the attribute values of its K-nearest neighbours.

**2.3.5 Projective Adaptive Resonance Theory (PART):** The PART neural network, developed by Cao and Wu (2002), is an enhanced version of the adaptive resonance theory (ART) neural network. It is specifically designed to identify projected clusters within datasets with high dimensions. Unlike traditional approaches, the PART neural network takes into account both the data points and the dimensions, and can deal with the lack of flexibility in the cluster (Chen & Chuang, 2008).

**2.3.6 Binary Logistic Regression:** unlike discriminant analysis, binary logistic regression (BLR) does not require the assumptions of linearity, normality, or homoscedasticity to hold. As a result, according to Abdulhafedh (2017), BLR is utilized more commonly than discriminant analysis. Logistic regression is a statistical technique used to establish a model that describes the associations between a group of independent variables, also known as predictors, and a dependent variable with multiple outcomes, often more than two. BLR focuses only on analysing dependent variables with dichotomous characteristics, meaning they only have two possible outcomes. Thus, BLR is a suitable method for analysing an unordered (cannot be reasonably ordered) categorical dependent variable.

## 3. RESULTS AND DISCUSSION

For statistical analysis, only the first dataset was utilized, however in the classification testing phase, both datasets were utilized with machine learning algorithms. The first dataset consisted of 1024 patient records. The average age of the patients was found to be 43.97, with a standard deviation of ± 17.334. The age range of the patients spanned from 10 years old to 110 years old. Moreover, a total of 616 individuals identified as female, while 408 individuals identified as male. Among these individuals, 377 were diagnosed with Covid-19, while 647 tested negative for the virus. Additionally, 245 patients exhibited lung infections, while 779 patients had no lung-related ailments, as indicated in Table 1.

Table 1: Patient's data. The figures in the parenthesis are the ratio to the number of patients (1024).

| Number of patients | | 1024 |
|---|---|---|
| Age range (years) | | 10 – 110 |
| Mean age (year) | | 43.97 |
| Number of Patients | Male | 408(39.84%) |
| | Female | 616(60.16%) |
| Covid-19 infections | P* | 377(36.82%) |
| | N | 647(63.18%) |
| Lung infections | P | 245(23.93%) |
| | N | 779(76.07%) |

* P stand for "Positive" and N for "Negative"

The age distribution of the patients ranges from 10 to 110 years, as depicted in Fig. 1.

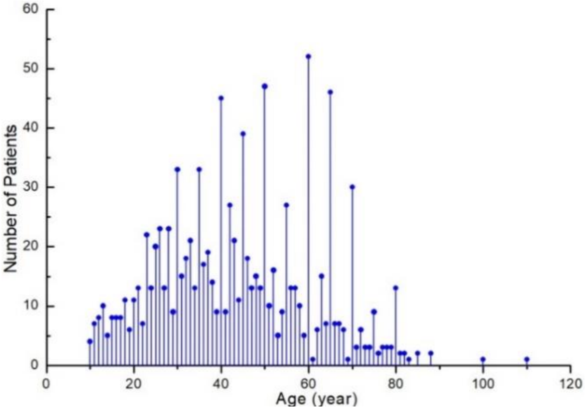

Figure 1: The distribution of the ages of the patients.

According to the data presented in Table 1, it can be observed that the male patient population constitutes approximately two-thirds of the female patient population. Additionally, among the total patient cohort, 39.84% were diagnosed with Covid-19 infection, 23.93% were diagnosed with lung infection, while the remaining individuals did not exhibit symptoms of either condition. Evidently, there existed patients who tested positive for Covid-19 but exhibited negative indications of lung infection, as well as patients who displayed positive indications of lung infection but tested negative for Covid-19. Thus, the patients were classified into four categories based on their infection status: (i) Patients who tested positive for both Covid-19 and lung infection, referred to as PP, (ii)

Patients who tested positive for Covid-19 but negative for lung infection, referred to as PN, (iii) Patients who tested negative for Covid-19 but positive for lung infection, referred to as NP, and (iv) Patients who tested negative for both Covid-19 and lung infection, referred to as NN. Fig. 2 illustrates the distribution of patients across the four categories. The data presented in the figure indicates that a majority of patients who tested positive for Covid-19 also exhibited signs of lung infection. Specifically, the proportion of patients with both positive Covid-19 and positive lung infection (PP) was found to be greater than half of the total number of patients who were tested for both conditions (PP > (PP+PN)/2). A significantly smaller proportion of patients exhibited positive lung infection but were not infected with Covid-19, as indicated by the NP bar in Fig. 2.

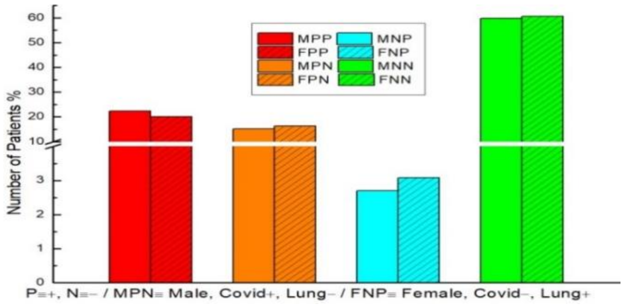

Figure 2: Number of patients with Covid-19 and lung infections, respectively, both positive (PP), positive and negative (PN), negative and positive (NP), and both negative (NN).

Each of the four categories were divided into two groups, males (M) and females (F). The patient population was divided into eight groups, consisting of four male groups (MPP, MPN, MNP, and MNN) and four female groups (FPP, FPN, FNP, and FNN). Instead of representing the absolute number of patients in each of the eight groups, the ratios of patients within each group were depicted, as illustrated in Fig. 3. The ratios pertaining to the four male groups are in relation to the total number of male patients, and likewise for the four female groups.

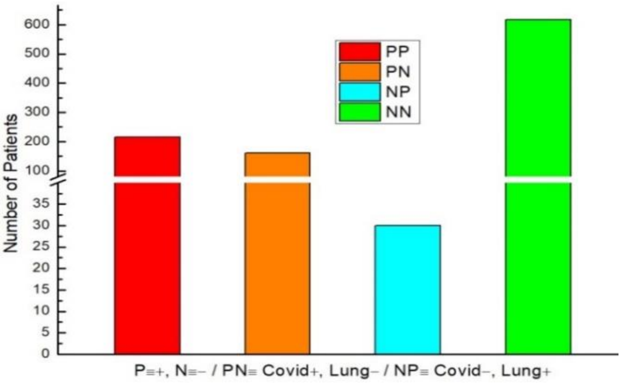

Figure 3: Ratio of male and female patients with Covid-19 and lung infections, respectively, both positive (MPP and FPP), positive and negative (MPN and FPN), negative and positive (MNP and FNP), and both negative (MNN and FNN).

According to Fig. 3, it can be observed that a majority of male and female patients who tested positive for Covid-19 also exhibited positive lung infections. Specifically, the proportion of male patients with positive lung infections (MPP) was found to be greater than half of the sum of male patients with positive lung infections (MPP) and male patients with negative lung infections (MPN); MPP > (MPP+MPN)/2. Additionally, the proportion of female patients with positive lung infections (FPP) was also greater than half of the sum of female patients with positive lung infections (FPP) and female patients with negative lung infections (FPN); FPP > (FPP+FPN)/2. Furthermore, it is worth noting that there were significantly fewer male and female patients who tested negative for Covid-19 but still had a positive lung infection. This observation can be seen in the MNP and FNP bars depicted in Fig. 3. The figure also displays a comparison between prevalence of positive Covid-19 and positive lung infection in male and that in female patients. It is observed that the ratio of male patients with both positive Covid-19 and positive lung infection is higher than that of female patients, as indicated by the MPP and FPP bars. In contrast, the proportion of female patients exceeds that of male patients in the remaining three categories. Nevertheless, irrespective of the presence of lung infection, the proportions of male and female patients infected with Covid-19 exhibit a nearly equal distribution, as depicted by the bars MPP, FPP, MPN, and FPN in Fig. 3.

Subsequently, the correlation analysis of the first dataset was conducted using IBM SPSS Statistics, software version 26 (Hinton et al., 2014). The software was utilized to conduct the Chi-Squared test of independence and Cramer's V test of strength. The purpose of this analysis was to investigate the potential association between Covid-19 infections and lung infections. The statistical analysis revealed a significant relationship between the two variables, X2 (1, N = 1024) = 359.235, p = .001, V = .592. There is a high likelihood of patients diagnosed with Covid-19 experiencing concurrent lung infections. This trend is also evident in Fig. 4, where it is observed that 57% of patients who tested positive for Covid-19 exhibited lung infections, whereas only 4.6% of patients who tested negative for Covid-19 displayed lung infections.

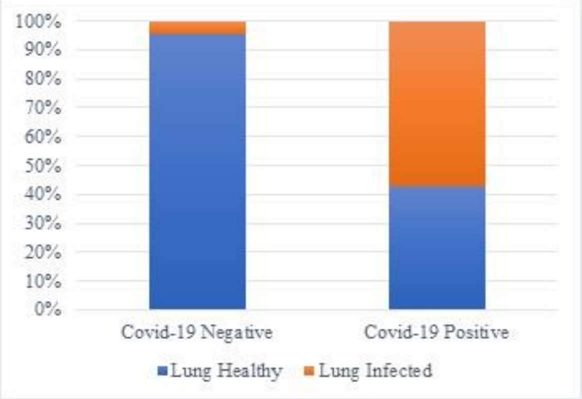

Figure 4: Distribution of lung infection in Covid-19 negative and positive patients.

Following the statistical analysis, machine learning was utilized on both datasets to test the ability of multiple algorithms in predicting lung infections in Covid-19 patients. The Bagging, Boosting, Naïve Bayes, KNN, J48, Random Forest, PART, and BLR classification algorithms were tested using the WEKA Data Mining Software version 3.8.6 (Hall et al., 2009). The Cross-Validation training and testing method was employed to test each algorithm, at 10 folds. For the first dataset, each of the algorithms demonstrated an accuracy ranging from 83.3% to 84.86%. The specific accuracy and precision of each algorithm is presented in Table 2. It is noteworthy to mention that the findings indicate that both the PART and J48 Decision Tree algorithms demonstrate that female Covid-19 patients over 29 years old and male Covid-19 patients over 32 years old were identified as the most probable individuals susceptible to lung infection.

Table 2: Results of lung infection prediction with Machine Learning algorithms on the first dataset.

| Algorithm | Accuracy (%) | Precision (%) | RMSE |
|---|---|---|---|
| Bagging | 84.28 | 84.7 | 0.3315 |
| Boosting | 84.67 | 84.9 | 0.3325 |
| Naïve Bayes | 84.86 | 85.1 | 0.3269 |
| KNN | 83.40 | 84.1 | 0.3661 |
| J48 | 83.89 | 85.2 | 0.3364 |
| Random Forest | 83.30 | 83.3 | 0.3518 |
| PART | 84.08 | 85.5 | 0.3364 |
| BLR | 84.86 | 85.0 | 0.3278 |

Meanwhile, for the second dataset, the lowest accuracy presented by the tested algorithms was 92.42%, which was the result of the KNN, J48 Decision Tree, Random Forest, and PART algorithms, while the highest accuracy was 93.41%, which was achieved by the Naïve Bayes algorithm. The specific accuracy and precision of the results of the testing on the second dataset can be seen in Table 3. The improvement of the accuracy of the algorithms when tested on the second dataset compared to the first dataset can be accredited to the higher dimensionality of the second dataset. The larger number of attributes provides the algorithms more information on the patient’s condition, allowing the algorithms to make better predictions.

Table 3: Results of lung infection prediction with Machine Learning algorithms on the second dataset.

| Algorithm | Accuracy (%) | Precision (%) | RMSE |
|---|---|---|---|
| Bagging | 92.61 | 92.8 | 0.2316 |
| Boosting | 92.81 | 93.0 | 0.235 |
| Naïve Bayes | 93.41 | 93.5 | 0.2388 |
| KNN | 92.42 | 92.4 | 0.2708 |
| J48 | 92.42 | 92.6 | 0.2506 |
| Random Forest | 92.42 | 92.4 | 0.2432 |
| PART | 92.42 | 92.5 | 0.2443 |
| BLR | 93.01 | 93.1 | 0.2315 |

## CONCLUSIONS

The global impact of the Covid-19 pandemic has been substantial, surpassing initial expectations. The provision of information aimed at alleviating the difficulties faced by individuals affected by the virus is of great significance, particularly in developing nations where data scarcity is prevalent, such as Iraq. This study aimed to examine the correlation between Covid-19 and lung infections as well as predicting the occurrence of lung infection in Covid-19 patients. The results confirmed a robust correlation between these two variables. Moreover, the experimentation involving various Data Mining and Machine Learning algorithms also demonstrated the ability to reasonably predict the incidence of lung infections in patients affected by Covid-19, with the accuracy of the predictions rising with the increase in dimensionality of the data. Maximum accuracy of 93.41% is achieved in this paper. This can allow medical personnel to implement preventative measures for lung infection in Covid-19 patients before the onset of the infection, whether to prevent the infection from occurring, or to at least limit the amount of damage in could inflict on the patient.

## ACKNOWLEDGEMENTS

The authors express their gratitude to the Research Ethics committee of the Directorate General of Health Duhok, Duhok, Kurdistan Region-Iraq for their kind approval in granting access to medical data. Additionally, the authors extend their appreciation to Dr. Ramadhan Salah-Aldeen Kocher of Kurdistan Private Hospital, Duhok, Kurdistan Region-Iraq for generously providing his patient records for this study.

## REFERENCES

Abdulhafedh, A. (2017). Incorporating the Multinomial Logistic Regression in Vehicle Crash Severity Modeling: A Detailed Overview. *Journal of Transportation Technologies*, 07(03), 279–303. https://doi.org/10.4236/jtts.2017.73019

Abdulrahman, M.S., and Rasool, J.A. (2020). Using Data Mining Algorithms to Predict Recommendations on Products. *International Journal of Advanced Science and Technology*, 29(3), 4370 - 4381. Retrieved from http://sersc.org/journals/index.php/IJAST/article/view/5263

Altman, N. S. (1992). An Introduction to Kernel and Nearest-Neighbor Nonparametric Regression. *The American Statistician*, 46(3), 175–185. https://doi.org/10.1080/00031305.1992.10475879

Annoni, A. D., Conte, E., Mancini, M. E., Gigante, C., Agalbato, C., Formenti, A., Muscogiuri, G., Mushtaq, S., Guglielmo, M., Baggiano, A., Bonomi, A., Pepi, M., Pontone, G., and Andreini, D. (2021). Quantitative Evaluation of COVID-19 Pneumonia Lung Extension by Specific Software and Correlation with Patient Clinical Outcome. *Diagnostics*, 11(2), 265. https://doi.org/10.3390/diagnostics11020265

Bhargava, N., Sharma, G., Bhargava, R., and Mathuria, M. (2013). Decision Tree Analysis on J48 Algorithm for Data Mining. *Computer Science and Software Engineering*. https://www.academia.edu/4375403/Decision_Tree_Analysis_on_J48_Algorithm_for_Data_Mining

Biau, G., and Scornet, E. (2016). A random forest guided tour. *TEST*, 25(2), 197–227. https://doi.org/10.1007/s11749-016-0481-7

Breiman, L. (2001). Random Forests. *Machine Learning*, 45(1), 5–32. https://doi.org/10.1023/a:1010933404324

Cao, Y., and Wu, J. (2002). Projective ART for clustering data sets in high dimensional spaces. *Neural Networks*, 15(1), 105–120. https://doi.org/10.1016/s0893-6080(01)00108-3

Chen, L.-D., Zhang, Z.-Y., Wei, X.-J., Cai, Y.-Q., Yao, W.-Z., Wang, M.-H., Huang, Q.-F., and Zhang, X.-B. (2020). Association between cytokine profiles and lung injury in COVID-19 pneumonia. *Respiratory Research*, 21(1). https://doi.org/10.1186/s12931-020-01465-2

Chen, R.-C., and Chuang, C.-H. (2008). Automating construction of a domain ontology using a projective adaptive resonance theory neural network and Bayesian network. *Expert Systems*, 25(4), 414–430. https://doi.org/10.1111/j.1468-0394.2008.00476.x

Dawoud, M. M., Dawoud, T. M., Ali, N. Y. A., and Nagy, H. A. (2020). Chest CT in COVID-19 pneumonia: a correlation of lung abnormalities with duration and severity of symptoms. *Egyptian Journal of Radiology and Nuclear Medicine*, 51(1). https://doi.org/10.1186/s43055-020-00359-z

Duzgun, S. A., Durhan, G., Demirkazik, F. B., Akpinar, M. G., and Ariyurek, O. M. (2020). COVID-19 pneumonia: the great radiological mimicker. *Insights into Imaging,* 11(1). https://doi.org/10.1186/s13244-020-00933-z

Francone, M., Iafrate, F., Masci, G. M., Coco, S., Cilia, F., Manganaro, L., Panebianco, V., Andreoli, C., Colaiacomo, M. C., Zingaropoli, M. A., Ciardi, M. R., Mastroianni, C. M., Pugliese, F., Alessandri, F., Turriziani, O., Ricci, P., and Catalano, C. (2020). Chest CT score in COVID-19 patients: correlation with disease severity and short-term prognosis. *European Radiology*,30(12),6808–6817. https://doi.org/10.1007/s00330-020-07033-y

GÜNER, R., HASANOĞLU, İ., and AKTAŞ, F. (2020). COVID-19: Prevention and control measures in community. *TURKISH JOURNAL OF MEDICAL SCIENCES*, 50(SI-1), 571–577. https://doi.org/10.3906/sag-2004-146

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. H. (2009). The WEKA data mining software. ACM *SIGKDD Explorations Newsletter*, 11(1), 10–18. https://doi.org/10.1145/1656274.1656278

Harapan, H., Itoh, N., Yufika, A., Winardi, W., Keam, S., Te, H., Megawati, D., Hayati, Z., Wagner, A. L., and Mudatsir, M. (2020). Coronavirus disease 2019 (COVID-19): A literature review. *Journal of Infection and Public Health*, 13(5), 667–673. https://doi.org/10.1016/j.jiph.2020.03.019

Harding, J. H., Shahbaz, M., Srinivas, and Kusiak, A. (2006). Data Mining in Manufacturing: A Review. *Journal of Manufacturing Science and Engineering-Transactions of The Asme*, 128(4), 969–976. https://doi.org/10.1115/1.2194554

Hinton, P., McMurray, I., and Brownlow, C. (2014). *SPSS Explained*. Taylor and Francis.

Hussain, S., Muhammad, L. J., Ishaq, F. S., Yakubu, A., and Mohammed, I. A. (2019). Performance Evaluation of Various Data Mining Algorithms on Road Traffic Accident Dataset. *Information and Communication Technology for Intelligent Systems*, 67–78. https://doi.org/10.1007/978-981-13-1742-2_7

MacIntyre, C. R., and Wang, Q. (2020). Physical distancing, face masks, and eye protection for prevention of COVID-19. *The Lancet*, 395(10242), 1950–1951. https://doi.org/10.1016/s0140-6736(20)31183-1

Mahammedi, A., Ramos, A., Bargalló, N., Gaskill, M., Kapur, S., Saba, L., Carrete, H., Sengupta, S., Salvador, E., Hilario, A., Revilla, Y., Sanchez, M., Perez-Nuñez, M., Bachir, S., Zhang, B., Oleaga, L., Sergio, J., Koren, L., Martin-Medina, P., … Vagal, A. (2021). Brain and Lung Imaging Correlation in Patients with COVID-19: Could the Severity of Lung Disease Reflect the Prevalence of Acute Abnormalities on Neuroimaging? A Global Multicenter Observational Study. *American Journal of Neuroradiology*, 42(6), 1008–1016. https://doi.org/10.3174/ajnr.a7072

Maimon, O., and Rokach, L. (2005). Data mining and knowledge discovery handbook. *Choice Reviews Online*, 48(10), 48–5729. https://doi.org/10.5860/choice.48-5729

McHugh, M. M. (2013). The Chi-square test of independence. *Biochemia Medica*, 143–149. https://doi.org/10.11613/bm.2013.018

Olivieri, D., and Scoditti, E. (2005). Impact of environmental factors on lung defences. *European Respiratory Review*, 14(95), 51–56. https://doi.org/10.1183/09059180.05.00009502

Rasool, J. A. A. (2018). Analysis the Relationship between Social Media and Education System in Kurdistan region of Iraq Using Chi-Square Test. *Academic Journal of Nawroz University*, 7(4), 133–138. https://doi.org/10.25007/ajnu.v7n4a282

Santoso, P., Fauziah, F., and Nurhayati, N. (2020). Application Of Data Mining Classification For COVID-19 Infected Status Using Algortima Naïve Method. *Jurnal Mantik*, 4(1), 267–275. http://iocscience.org/ejournal/index.php/mantik/article/view/740

Skopljanac, I., Ivelja, M. P., Barcot, O., Brdar, I., Dolic, K., Polasek, O., and Radic, M. (2021). Role of Lung Ultrasound in Predicting Clinical Severity and Fatality in COVID-19 Pneumonia. *Journal of Personalized Medicine*, 11(8), 757. https://doi.org/10.3390/jpm11080757

Talib, H. J. (2021). *Predicting the Correct Procedure of COVID-19 Patients in Hospitals Using Machine Learning*. M.Sc. Thesis, Applied science private university, Amman-Jordan.

Tung-Chen, Y., de Gracia, M., Díez-Tascón, A., Alonso-González, R., Agudo-Fernández, S., Parra-Gordo, M. L., Ossaba-Vélez, S., Rodríguez-Fuertes, P., and Llamas-Fuentes, R. (2020). Correlation between Chest Computed Tomography and Lung Ultrasonography in Patients with Coronavirus Disease 2019 (COVID-19). *Ultrasound in Medicine and Biology*, 46(11), 2918–2926. https://doi.org/10.1016/j.ultrasmedbio.2020.07.003

Vieira, J. M., Ricardo, O. M. de P., Hannas, C. M., Kanadani, T. C. M., Prata, T. dos S., and Kanadani, F. N. (2020). What do we know about COVID-19? A review article. *Revista Da Associação Médica Brasileira*, 66(4), 534–540. https://doi.org/10.1590/1806-9282.66.4.534

Wang, X., Che, Q., Ji, X., Meng, X., Zhang, L., Jia, R., Lyu, H., Bai, W., Tan, L., and Gao, Y. (2021). Correlation between lung infection severity and clinical laboratory indicators in patients with COVID-19: a cross-sectional study based on machine learning. *BMC Infectious Diseases*, 21(1). https://doi.org/10.1186/s12879-021-05839-9

Witten, I., Frank, E., and Hall, M. (2011). *Data Mining: Practical Machine Learning Tools and Techniques (The Morgan Kaufmann Series in Data Management Systems)* (3rd ed.). Morgan Kaufmann.

Wu, B., Wang, X., Shen, H., and Zhou, X. (2012). Feature selection based on max–min-associated indices for classification of remotely sensed imagery. *International Journal of Remote Sensing*, 33(17), 5492–5512. https://doi.org/10.1080/01431161.2012.663111

Yağmur, A. R., Akbal Çufalı, Ş., Aypak, A., Köksal, M., Güneş, Y. C., and Özcan, K. M. (2021). Correlation of olfactory dysfunction with lung involvement and severity of COVID-19. *Irish Journal of Medical Science* (1971 -), 191(4), 1843–1848. https://doi.org/10.1007/s11845-021-02732-x

Yu, F., Du, L., Ojcius, D. M., Pan, C., and Jiang, S. (2020). Measures for diagnosing and treating infections by a novel coronavirus responsible for a pneumonia outbreak originating in Wuhan, China. *Microbes and Infection*, 22(2), 74–79. https://doi.org/10.1016/j.micinf.2020.01.003