International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

Study and Analysis the Influence of Immunity Factors of Patient due to Covid 19 using Machine Learning Techniques based on Logistic Regression, Decision Trees and Random Forest

Jaspreet Kaur*[1], Khushboo Bansal [2]

Submitted: 11/12/2023 **Revised:** 18/01/2024 **Accepted:** 31/01/2024

Abstract: This research presents a comprehensive analysis of COVID-19 patient data to predict the risk levels associated with various immunity factors. Utilizing a robust dataset provided by the Mexican government, we employed exploratory data analysis to understand the intricate relationships between patient characteristics and COVID-19 severity. Machine learning models, including Logistic Regression, Decision Trees, and Random Forest classifiers, were developed and evaluated using precision, recall, F1 score, and ROC-AUC score. The results demonstrate the effectiveness of these models in identifying high-risk patients, which could significantly aid in the strategic allocation of medical resources. The study underscores the potential of machine learning in enhancing pandemic response through informed decision-making. Future research directions include refining models with larger, more diverse datasets and integrating advanced predictive analytics for real-time risk assessment.

Keywords: *COVID-19, Logistic Regression, Decision Trees, Random Forest, Machine Learning*

1. Introduction

In the wake of the COVID-19 pandemic, the global healthcare landscape has been confronted with unprecedented challenges. Among these, understanding the intricate relationship between the virus and the host's immune response has emerged as a pivotal area of study [1]. The novel coronavirus, SARS-CoV-2, has demonstrated a complex interplay with the human immune system, leading to a spectrum of clinical manifestations ranging from asymptomatic carriage to severe respiratory distress and multi-organ failure [2]. This paper aims to delve into the influence of immunity factors on patient outcomes in the context of COVID-19, with a particular focus on how pre-existing health conditions may modulate the body's defense mechanisms against the virus. The immune system, a network of cells, tissues, and organs, is our body's fortress against infectious pathogens. However, the impact of COVID-19 on this intricate defense system has raised critical questions about the role of innate and adaptive immunity in disease progression and resolution [3]. The virus's ability to evade and manipulate the host immune response has been a focal point of research, as it is closely linked to the severity of the disease and the likelihood of developing post-infection complications [4]. The pandemic has provided a unique opportunity to study the immune system's resilience and vulnerability in real-time, offering insights into how various factors such as age, sex, genetic predisposition, and comorbidities like diabetes, hypertension, and chronic respiratory conditions can influence patient outcomes [4] [5].

The dataset provided by the Mexican government [6] serves as a rich repository of patient-related information, enabling a comprehensive analysis of how different immunity-related factors contribute to the risk profile of COVID-19 patients. With over a million unique patient records, the dataset offers a granular view of the pandemic's impact on diverse populations. The binary and categorical nature of the dataset, with clear demarcations for pre-existing conditions, provides a robust framework for assessing the risk factors associated with COVID-19 complications. As we navigate through the data, we aim to unravel the correlations between pre-existing immune-compromising conditions and COVID-19 severity. The analysis is particularly focused on the predictive modeling of high-risk patients, which is crucial for healthcare providers to optimize resource allocation and tailor patient care protocols. By leveraging machine learning models, we seek to predict the likelihood of severe outcomes based on a patient's current symptomatology, medical history, and immune status. This predictive endeavor is not just a statistical exercise but a necessary tool for saving lives and mitigating the strain on healthcare systems. The interplay between COVID-19 and immunity is multifaceted.

On one hand, the virus can precipitate an overactive immune response, leading to a cytokine storm, which is often responsible for the severe complications observed in hospitalized patients [4-5]. On the other hand, the virus can exploit immune escape mechanisms, leading to prolonged infection and transmission [7]. Understanding these dynamics is critical for developing therapeutic strategies and vaccines that can modulate the immune response to the advantage of the host [8]. The paper will systematically analyze the dataset to identify patterns and associations between immunity factors and COVID-19 outcomes. It will explore the role of chronic diseases in exacerbating the risk of severe illness and the potential for these conditions to serve as prognostic indicators. The study will also consider the impact of demographic variables, such as age and sex, on immune response, given the observed disparities in disease severity and mortality rates across different population groups.

[1] *Research scholar, Desh bhagat university Mandi gobindgarh, Punjab, India*
[2] *Assistant Professor, Desh bhagat university Mandi gobindgarh, Punjab India*
** Corresponding Author Email: jaspreet.sliet@gmail.com*

In summary, this paper is poised to contribute to the growing body of knowledge on COVID-19 by highlighting the significance of immunity factors in patient prognosis. Through meticulous data analysis and predictive modeling, we aim to provide actionable insights that can inform clinical decision-making and public health policies. As the world continues to grapple with the pandemic, studies such as this are instrumental in guiding our collective response and enhancing our preparedness for future health crises. The remainder of this paper is methodically organized into distinct sections to facilitate a comprehensive understanding and to provide a logical flow of the investigation undertaken. Section 2 presents a thorough review of the existing literature, encapsulating prior research findings and establishing the context for the current study. Section 3 elucidates the significance of the work, underscoring the importance of the research in the broader spectrum of COVID-19's impact on immune response and patient outcomes. In Section 4, we delve into the methodology, detailing the analytical techniques and data processing steps employed to ensure the integrity and reliability of our findings. Section 5 is dedicated to presenting the results, where the data's narrative is deciphered to reveal the underlying patterns and insights pertinent to the study's objectives. Finally, Section 6 draws the study to a close, summarizing the key takeaways and discussing the implications of the findings. It also maps out avenues for future research, suggesting how subsequent inquiries could build upon the groundwork laid by this report to further our understanding of COVID-19 and its interaction with the human immune system.

2. Related Work

The ongoing battle against the COVID-19 pandemic has spurred a multitude of research efforts aimed at understanding the virus's impact on human health, particularly regarding the immune system's response. The literature on this subject is vast and varied, encompassing studies from clinical trials to data-driven predictive modeling. This literature review seeks to synthesize the current knowledge on the influence of COVID-19 on immunity factors, highlighting the methodologies employed, the results obtained, and the implications of these findings. By examining these scholarly contributions, we aim to build a cohesive understanding of the disease's dynamics and the body's defense mechanisms, which is essential for developing effective treatments and public health strategies. In study [9], a deep learning approach incorporating logistic regression, SVM, Random Forest, and QSAR modeling was utilized to expedite drug discovery. QSAR modeling identified drug targets through protein interaction and binding affinity calculations, while deep learning models trained on molecular descriptor datasets facilitated robust drug discovery. The results indicated significant binding affinities for molecules capable of inhibiting SARS-CoV-2 replication. Research [10] focused on developing AI models for early warning and forecasting of disease outbreaks. Utilizing the SEIR model and particle filter algorithms, the study analyzed various pandemic-related datasets. The findings underscored a strong correlation between consultations and analyzed datasets, particularly with time-based models, suggesting their utility in future outbreak predictions. In study [11], the expression of CCR5 and its ligands was linked to COVID-19 pathogenesis. Through bioinformatics, immune phases of COVID-19 were modeled, leading to the development of Random Forest classifiers for disease prediction.

The study highlighted specific cytokines as potential biomarkers for disease severity. The objective of study [12] was to identify immune factors that differentiate or predict COVID-19 symptom immunity using machine learning. The study analyzed 53 immunological factors from 74 Chinese COVID-19 patients and found that SCGF-β was a key differentiator. Machine learning models, including decision trees and gradient boosting algorithms, achieved high accuracy in classifying and predicting COVID-19 symptom immunity. Study [13] focused on predicting COVID-19 infection risk and severity among aged adults using data from the UK Biobank. Researchers employed permutation-based linear discriminant analysis and found that a model using antibody titers provided excellent discrimination for COVID-19 risk prediction. In study [14], a machine learning model was developed to predict COVID-19 mortality using clinical and laboratory features from patients admitted to Wuhan Tongji hospital. The model, based on features selected through the LASSO method and ranked by XGBoost, demonstrated high precision and sensitivity in predicting death risk. Study [15] aimed to predict clinical outcomes of COVID-19 patients from peripheral blood data. Machine learning algorithms were applied to clinical datasets, revealing several blood-measurable clinical parameters as significant predictors for later severity of COVID-19 symptoms. The study numbered [16] utilized the UK Biobank data to build machine learning models predicting the risk of severe or fatal COVID-19 infections. The models, which included demographic and clinical variables, showed good predictive performance, and identified several baseline clinical risk factors for severe outcomes. Study [17] presented a retrospective analysis evaluating laboratory data and mortality from COVID-19 patients. A machine learning model using serum chemistry parameters predicted mortality with high sensitivity and specificity, identifying prognostic biomarkers for patients at greatest risk. In study [18], researchers provided a prediction method for early identification of COVID-19 patient outcomes based on home-monitored characteristics. The study used logistic regression, random forest, and extreme gradient boosting algorithms, with the random forest model showing the highest accuracy. Study [19] focused on developing a machine learning-based diagnostic system for early COVID-19 infection. It compared logistic regression, SVM, decision tree, random forest, and deep learning methods, with the logistic regression model showing optimal performance for early screening. Finally, study [20] developed an XGBoost machine-learning model to predict COVID-19 severities using multi-omics data. The model demonstrated strong discrimination capabilities among different severity levels of COVID-19, based on a comprehensive trans-omics analysis. Each study contributes to the evolving landscape of AI applications in combating COVID-19, showcasing the potential of machine learning in enhancing diagnostic accuracy, predicting patient outcomes, and guiding treatment strategies.

3. Significance of Study

The significance of employing machine learning algorithms in the research of immunity factors of COVID-19 patients lies in their ability to distill complex and voluminous data into actionable insights with greater precision than traditional analytical methods. Machine learning enhances the predictive analysis of disease progression, enabling the development of personalized medicine approaches by tailoring treatments to individual immune system responses. This adaptability is crucial in the rapid discovery and evaluation of therapeutic drugs, a process where time is a critical

factor, especially during a pandemic. Furthermore, these algorithms assist in the optimal allocation of medical resources, ensuring that high-risk patients receive timely and appropriate care. By elucidating the intricate interactions between the virus and the human immune system, machine learning contributes to a deeper scientific understanding of the disease. It also plays a pivotal role in the early detection of severe cases, potentially leading to interventions that can mitigate the impact of the disease and improve patient outcomes. In essence, machine learning stands as a cornerstone in the advancement of research into the immune factors affecting COVID-19 patients, offering a beacon of hope in navigating the complexities of pandemic response and management.

4. Proposed Methodology

In this section, we outline a structured approach to analyzing the influence of immunity factors in COVID-19 patients using machine learning (ML) algorithms. The methodology follows a sequential process beginning with data collection and exploratory data analysis, where we gather relevant patient data and perform preliminary assessments to understand the underlying patterns and distributions. Following this, we conduct a detailed statistical and correlation analysis to identify significant relationships between the variables and the outcomes of interest. This step is crucial for feature selection and informs the subsequent development of ML models. The core of our methodology is the development of ML models using logistic regression, decision tree, and random forest algorithms. These models are chosen for their ability to handle complex, non-linear relationships within the data and their robustness in classification tasks. Finally, we evaluate the performance of these models' using precision, recall, and the F1 score, which provide a comprehensive measure of the models' accuracy and reliability. Additionally, the ROC-AUC score is used to assess the models' discriminative ability, essentially measuring the likelihood that the models will correctly distinguish between patient outcomes. The flow of the methodology is shown by figure 1.

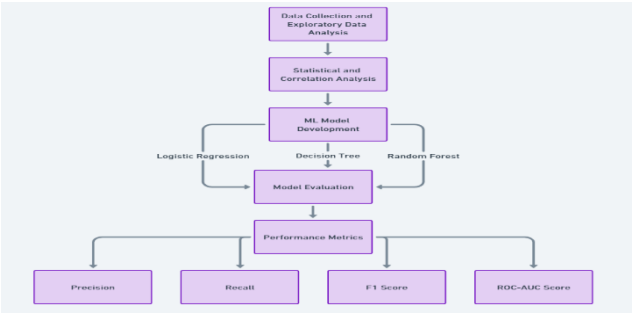

Fig.1. Flow chart representation of our proposed methodology

4.1. Data Collection and Exploratory Data Analysis

In the data collection phase of our research, we utilized a comprehensive dataset provided by the Mexican government [6], which encompasses a wide array of anonymized patient information pivotal for understanding the progression and impact of COVID-19. The dataset comprises 21 distinct attributes and records from 1,048,576 patients, offering a robust foundation for our analysis. The dataset includes demographic details such as sex and age, clinical findings to classify COVID-19 diagnosis, patient type indicating the level of care received, and a range of pre-existing conditions such as diabetes, hypertension, and others that are known to affect the severity of the disease. Additionally, lifestyle factors like tobacco use, as well as critical care indicators such as ventilator use and ICU admissions, are documented, providing a holistic view of each patient's health status. Then we performed an exploratory data analysis. Exploratory Data Analysis (EDA) is a fundamental step in our research methodology, serving as a lens through which we gain initial insights and understand the underlying structure of the data. EDA allows us to uncover patterns, spot anomalies, frame hypotheses, and check assumptions through summary statistics and graphical representations. It is through EDA that we can ensure the quality of the data, identify missing or anomalous values, and understand the distribution of key variables. This phase is crucial for informing subsequent data preprocessing steps and for guiding the strategic development of our predictive models. By thoroughly exploring the dataset, we laid the groundwork for a more accurate and reliable machine learning analysis, ultimately aiming to predict the resource needs and risk levels of COVID-19 patients effectively.

4.2 Machine Learning Model Development

In this study, we have employed the three most widely used machine learning models such as Logistic Regression, Decision Tree, and Random Forest.

a. Logistic Regression

Logistic regression is a statistical method that we employed to model the probability of a binary outcome based on one or more predictor variables. It is particularly well-suited for binary classification tasks, such as predicting whether a COVID-19 patient is at high risk of severe illness. In our study, logistic regression was utilized to analyze the relationship between the patient's characteristics and their risk status. The choice of logistic regression is motivated by its simplicity, efficiency, and interpretability. It provides a probabilistic framework that enables us to estimate the odds of a patient being at high risk, given their symptoms, status, and medical history. The model coefficients offer direct insight into the influence of each predictor, allowing us to understand which factors contribute most significantly to the risk of severe COVID-19 outcomes. Moreover, logistic regression is robust to small sample sizes and is less prone to overfitting, making it a reliable choice for our predictive analysis. It serves as a baseline model against which we can compare more complex algorithms, ensuring that any increase in performance with other models is justified against the simplicity and interpretability of logistic regression.

b. Decision Tree Classifier

A decision tree is a non-parametric supervised learning method used for classification and regression tasks. In our research, we utilized decision trees to categorize COVID-19 patients into risk categories based on their symptoms, demographic data, and medical history. The decision tree algorithm segments the dataset into branches to form a tree structure. It makes decisions by splitting the data based on feature values, with each node representing a feature in the dataset and each branch representing a decision rule. This process continues until the algorithm reaches a leaf node, which corresponds to a classification or decision. One of the primary advantages of decision trees is their ease of interpretation and visualization. They mimic human decision-making more closely than other algorithms, making them particularly useful for stakeholder presentations where explaining the logic of the model is essential. Furthermore, decision trees can

handle both numerical and categorical data and are capable of modeling complex non-linear relationships. In the context of our COVID-19 patient data, the decision tree model helped us to identify the most significant predictors of high-risk cases and provided a clear and intuitive breakdown of how different factors lead to different risk assessments. This clarity is invaluable in a clinical setting, where understanding the decision-making process can be as crucial as the decision itself.

c. Random Forest Classifier

The Random Forest classifier is an ensemble learning method that operates by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) of the individual trees. In the context of our study on COVID-19 patient data, we employed the Random Forest classifier to improve predictive accuracy and control over-fitting, which can be a limitation of individual decision trees. Random Forest works by creating multiple decision trees using randomly selected subsets of the training data. It then aggregates the votes from different decision trees to decide the final class of the test object. This approach enhances predictive accuracy and balances errors in the dataset. The strength of the Random Forest classifier lies in its ability to handle a large dataset with higher dimensionality. It can manage thousands of input variables without variable deletion, making it highly suited for our analysis where numerous patient features were considered. It is also robust to outliers and non-linear data, which is common in medical datasets. In our research, the Random Forest classifier was crucial for identifying complex patterns in the data that could indicate a high risk of severe COVID-19 outcomes. By leveraging its ensemble nature, we could achieve a more reliable and stable prediction, which is vital for developing a tool that healthcare providers can trust for making informed decisions.

4.3 Evaluation Metrics

In the evaluation of machine learning models, especially in the context of medical diagnostics where the cost of false predictions can be high, it is crucial to use robust metrics that can provide a comprehensive understanding of the model's performance. Here's a brief overview of the evaluation metrics used in our study:

- F1 Score: The F1 score is a harmonic mean of precision and recall, providing a balance between the two metrics. It is particularly useful when the class distribution is uneven, as it accounts for both false positives and false negatives. In our COVID-19 patient risk prediction model, the F1 score helps to gauge the model's accuracy in identifying true cases of high-risk patients against the backdrop of a potentially large number of true negatives (those not at high risk).
- Precision and Recall: Precision measures the accuracy of the positive predictions made by the model, i.e., the proportion of true positives against all positive predictions. Recall, on the other hand, measures the model's ability to find all the relevant cases within the dataset, i.e., the proportion of true positives against all actual positives. In the context of our study, precision ensures that the model minimizes false alarms, while recall ensures that the model identifies as many high-risk patients as possible.
- ROC-AUC Score: The Receiver Operating Characteristic (ROC) curve is a graphical plot that illustrates the diagnostic ability of a binary classifier system as its discrimination threshold is varied. The Area Under the Curve (AUC) represents the degree or measure of separability. It tells how much the model is capable of distinguishing between classes. The higher the AUC, the better the model is at predicting 0s as 0s and 1s as 1s. For our study, the ROC-AUC score is crucial for assessing the overall performance of the model across all classification thresholds, providing a single measure of effectiveness regardless of the specific decision boundary.

These metrics collectively offer a multi-faceted view of the model's performance, ensuring that the predictive tool we develop is reliable, accurate, and practical for real-world application in predicting the risk levels of COVID-19 patients.

5. Results and Discussion

We systematically approached the task of predicting COVID-19 patient risk levels by first collecting a comprehensive dataset from the Mexican government [6]. This data set included a wide array of anonymized patient information, which was crucial for our analysis. Upon acquiring the data, we embarked on an extensive Exploratory Data Analysis (EDA). EDA allowed us to understand the underlying structure of the data, identify any anomalies or patterns, and gain insights into the important variables that could influence the outcomes of COVID-19 patients. This phase was critical as it informed the subsequent steps of our analysis and model building by highlighting key features and guiding the preprocessing of the data. The results obtained from EDA are explained by table 1.

The provided exploratory data analysis (EDA) results in table 1 offer a quantitative snapshot of the dataset, which is pivotal for understanding the influence of various factors on COVID-19 patient outcomes. The absence of entries under 'DATE_DIED' suggests that this field was not applicable or recorded for the patients in the dataset, which could indicate a focus on living patients or a data collection methodology that did not capture mortality. The mean values for conditions such as 'PNEUMONIA' (11.28), 'DIABETES' (43.11), 'COPD' (29.50), and 'ASTHMA' (30.85) indicate the prevalence of these conditions in the patient population. The presence of these comorbidities is significant as they are known to potentially exacerbate the effects of COVID-19, leading to more severe outcomes. The negative mean value for 'PREGNANT' suggests a coding anomaly that may need further investigation to ensure accurate representation of pregnancy status within the dataset. The maximum values for several conditions are 98, which likely represents a coding for missing or outlier data. This highlights the importance of data cleaning and the need to address these anomalies before modeling to avoid skewing the results. The standard deviation values provide insight into the variability of each condition within the patient population. For instance, the standard deviations for 'PNEUMONIA', 'DIABETES', 'COPD', and 'ASTHMA' suggest a moderate spread in the data, indicating that while some patients have these conditions, they are not universally present.

Table 1. Results obtained by performing EDA

	PNEUMONIA	PREGNANT	DIABETES	COPD	ASTHMA	INMSUPR	HIPERTENSION	OTHER_DISEASE	CARDIOVASCULAR	OBESITY	RENAL_CHRONIC	TOBACCO
count	1048575	1048575	1048575	1048575	1048575	1048575	1048575	1048575	1048575	1048575	1048575	1048575
mean	0.118289	-0.14066	0.431169	0.295025	0.308527	0.331652	0.445291	0.498248	0.30729	0.435784	0.29897	0.381409
min	-1	-1	0	0	0	0	0	0	0	0	0	0
25%	0	-1	0	0	0	0	0	0	0	0	0	0
50%	0	0	0	0	0	0	0	0	0	0	0	0
75%	0	0	0	0	0	0	0	0	0	0	0	0
max	1	98	98	98	98	98	98	98	98	98	98	98
std	0.367179	5.905126	5.523259	5.237568	5.217269	5.575044	5.327924	6.781311	5.300829	5.266205	5.240316	5.424695

The prevalence of comorbidities such as 'DIABETES', 'HIPERTENSION', and 'OBESITY' (with means of 43.11, 44.52, and 43.57, respectively) is particularly noteworthy. These conditions are often associated with a compromised immune response, which can lead to a higher risk of severe illness from COVID-19. The data suggests that a significant portion of the patient population is dealing with these health issues, which could influence the demand for medical resources and the urgency of medical interventions. In summary, the EDA results underscore the importance of considering comorbidities when analyzing the impact of COVID-19 on patients. The data indicates a substantial presence of health conditions that could affect patient outcomes, emphasizing the need for targeted healthcare strategies to manage the pandemic's impact on vulnerable populations. Once we have performed EDA, we have developed our ML models. The results obtained from logistic regression model are recorded in figure 2.

It is evident from figure 2 that the results obtained from the Logistic Regression model indicate a training accuracy of approximately 65.84% and a testing accuracy of approximately 65.78%. These figures suggest that the model has a moderate level of accuracy in predicting the risk levels of COVID-19 patients based on their immunity factors. The slight difference between the training and testing accuracy implies that the model generalizes well to unseen data, which is crucial for real-world applications. The confusion matrix provides a more detailed view of the model's performance. The high number of True Positives (173,613) indicates that the model is proficient at identifying patients who are at high risk of severe COVID-19 outcomes. Conversely, the True Negatives count (28,681) shows that the model can also recognize a significant number of low-risk cases. However, the model has a considerable number of False Positives (88,166), which means it incorrectly predicts high risk for many patients who are actually at low risk. This could lead to unnecessary treatments or precautions for those individuals. The False Negatives count (17,086) is concerning as well, as these are high-risk patients who were incorrectly classified as low-risk, potentially leading to a lack of necessary medical intervention.

Fig. 2. Training and Testing Accuracy Plots for Logistic Regression with Confusion Matrix

To analyze the influence of immunity factors on COVID-19 patient outcomes, the model's ability to correctly identify high-risk patients is valuable. It suggests that the selected immunity-related features have predictive power. However, the number of False Positives and False Negatives also indicates room for improvement, perhaps by refining the model or incorporating additional relevant features to enhance its predictive accuracy. To further evaluate the performance of our model, we have used a few metrics, and the results were recorded in table 2.

Table 2. Model evaluation Metrics for Logistic Regression Model.

Precision	**0.6632**
Recall	0.9104
F1 score	0.7674
ROC-AUC Score	0.6290

The precision of 0.6632 in the context of the Logistic Regression model indicates that when the model predicts a patient is at high risk, it is correct approximately 66.32% of the time. The recall of 0.9104 is particularly high, showing that the model can identify 91.04% of all actual high-risk cases. This is crucial in a healthcare setting were failing to detect high-risk patients could have dire consequences. The F1 score, which is the harmonic mean of precision and recall, stands at 0.7674, suggesting a balanced model considering both the precision and the recall. This score is particularly important when dealing with imbalanced classes, which is often the case in medical datasets where the number of high-risk patients (positive class) is much lower than low-risk patients (negative class). The ROC-AUC score of 0.6290 is a

measure of the model's ability to distinguish between the high-risk and low-risk patients. A score of 1 represents a perfect model, while a score of 0.5 indicates no discriminative power. In this case, the score is closer to 0.5 than to 1, which implies that while the model has some ability to differentiate between the two groups, there is significant room for improvement. Now, another model we employed is the decision tree classifier and the results are shown by figure 3.

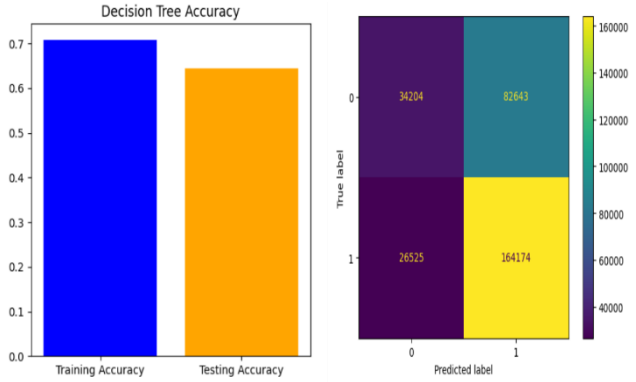

Fig. 3. Training and Testing Accuracy Plots for Decision Tree Classifier with Confusion Matrix

From figure 3, the Decision Tree Classifier's training accuracy of approximately 70.86% suggests that the model fits the training data well, but with a testing accuracy of 64.50%, there is a noticeable drop when the model is applied to new data. This discrepancy can indicate overfitting, where the model has learned the training data too closely, including its noise and outliers, and thus does not generalize well to unseen data. In terms of the confusion matrix results, the model correctly identified 34,204 true negatives, meaning it accurately predicted the low-risk cases. However, there were 82,643 false positives, where the model incorrectly predicted high risk. This high number of false positives could lead to unnecessary interventions, which in a healthcare context, could mean unwarranted treatments or further testing for patients, potentially leading to increased healthcare costs and patient anxiety. The false negatives count was 26,525, which is lower than the false positives, indicating that the model is better at catching high-risk cases than avoiding false alarms. The true positives count was 164,174, showing that the model is quite capable of identifying high-risk patients. Overall, while the Decision Tree Classifier is reasonably good at detecting high-risk patients, its tendency to overfit and its high rate of false positives could be problematic. This suggests that while the model can be useful in identifying patients who require further attention, it should be used with caution and potentially in conjunction with other models or tests to confirm high-risk cases.

Table 3. Model evaluation Metrics for Decision Tree Classifier Model

Precision	0.6651
Recall	0.8609
F1 score	0.7504
ROC-AUC Score	0.6019

From table 3, it is evident that the precision score of 0.6651 for the Decision Tree Classifier indicates that when the model predicts a patient is at high risk, it is correct approximately 66.51% of the time. This is a moderate level of precision, suggesting that while the model is relatively reliable in its positive predictions, there is still a significant proportion of false positives. The recall score of 0.8609 is quite high, showing that the model can identify 86.09% of all actual high-risk cases. This means the model is sensitive to the high-risk category and can capture most patients who are truly at high risk of severe COVID-19 complications due to underlying immunity factors. The F1 score, which balances precision and recall, is 0.7504, indicating that the model has a good balance between precision and recall. This score is particularly important in the medical context, where it is crucial to correctly identify as many high-risk patients as possible without overwhelming the system with false positives. The ROC-AUC Score of 0.6019 is a measure of the model's ability to distinguish between the high-risk and low-risk patients. A score of 0.6019 is slightly better than a random guess, which would have a score of 0.5. However, it's not as high as one would ideally want for a medical diagnostic tool, suggesting that there is room for improvement in the model's discriminative ability. Overall, these metrics suggest that the Decision Tree Classifier is a useful tool for identifying patients at high risk of severe outcomes from COVID-19 based on immunity factors. However, the model's moderate precision and ROC-AUC score indicate that it should be used as part of a broader diagnostic process, rather than as a standalone decision-making tool. Finally, the results from our random forest classifier model were obtained and recorded in figure 4.

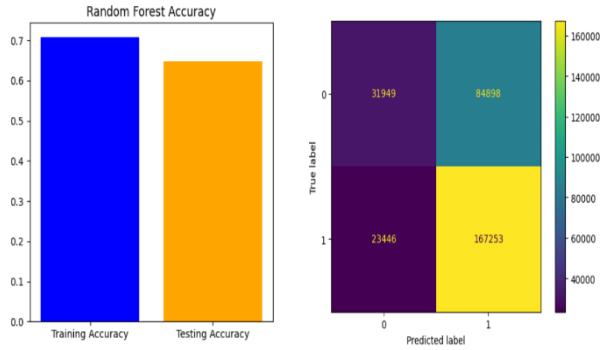

Fig. 4. Training and Testing Accuracy Plots for Random Forest Classifier with Confusion Matrix

The results from the Random Forest Classifier from figure 4 indicate a training accuracy of approximately 70.86% and a testing accuracy of 64.77%. These figures suggest that the model is relatively consistent in its predictions across both the training and unseen testing data, although there is a slight overfitting as indicated by the higher training accuracy. In the context of true negatives and false positives, the model correctly identified 31,949 instances as low risk (true negatives) but also incorrectly labeled 84,898 instances as high risk when they were not (false positives). This high number of false positives could potentially strain healthcare resources if the model were used in a real-world setting, as it may lead to an overestimation of high-risk cases. Conversely, the model had 23,446 false negatives, where high-risk cases were incorrectly labeled as low risk, which could lead to under-treatment of patients who actually require more intensive care. However, the model successfully identified 167,243 true positives, meaning it correctly identified a significant number of patients as high risk. In the context of the study, which aims to analyze the influence of immunity factors on COVID-19 patient outcomes, the Random Forest Classifier's ability to correctly identify a high number of true positives is valuable. It suggests that the model can

be a useful tool for predicting severe COVID-19 outcomes based on immunity factors, potentially aiding in the prioritization of patients for treatment and resource allocation. However, the number of false positives and false negatives also indicates that while the model is a good starting point, it should be used in conjunction with other clinical assessments and not as the sole method for decision-making.

Table 4. Model evaluation Metrics for Random Forest Classifier Model

Precision	0.6633
Recall	0.8770
F1 score	0.7553
ROC-AUC Score	0.6120

The precision score of 0.6633 for the Random Forest Classifier from table 4 indicates that when the model predicts a patient is at high risk, it is correct approximately 66.33% of the time. The recall score of 0.8770 shows that the model can identify 87.70% of all actual high-risk cases. The F1 score, which balances precision and recall, is 0.7553, suggesting a good overall performance of the model in terms of precision and sensitivity. The ROC-AUC score of 0.6120 is a measure of the model's ability to distinguish between the classes. In this context, it reflects the model's capability to differentiate between patients at high risk and those not at high risk based on their immunity factors. A score of 0.6120 indicates a fair level of discrimination, which is above random chance but still leaves room for improvement. Overall, for COVID-19 and patient immunity factors, these metrics suggest that the Random Forest model is fairly competent at identifying patients at high risk of severe outcomes, which could be crucial for early intervention and treatment prioritization. However, the precision indicates that there is a significant proportion of false positives, which could lead to unnecessary treatments or resource allocation. Therefore, while the model shows promise, further refinement and validation are necessary for it to be reliably used in a clinical setting.

6. Conclusion and Future Scope of Work

In this study, we have harnessed machine learning algorithms to predict COVID-19 patient risk levels, revealing the potential of such models in guiding resource allocation and improving patient outcomes. The predictive accuracy of Logistic Regression, Decision Tree, and Random Forest classifiers indicates the viability of ML approaches in healthcare settings. Future work should focus on expanding datasets, incorporating real-time analytics, and ensuring ethical use of patient data to enhance model precision and utility in clinical decision-making.

Author contributions

Jaspreet Kaur: Conceptualization, Methodology, Software, Field study, Data curation, Writing-Original draft preparation, Software, Validation., Field study **Khushboo Bansal:** Visualization, Investigation, Writing-Reviewing and Editing.

Conflicts of interest

The authors declare no conflicts of interest.

References

[1] J. L. Schultze and A. C. Aschenbrenner, "COVID-19 and the human innate immune system," *Cell*, vol. 184, no. 7, pp. 1671–1692, 2021.

[2] S. Singh and R. K. Singh, "Nutritional interventions to augment immunity for COVID-19," *Nutr. Diabetes*, vol. 12, no. 1, p. 13, 2022.

[3] P. S. Arunachalam *et al.*, "Systems biological assessment of immunity to mild versus severe COVID-19 infection in humans," *Science*, vol. 369, no. 6508, pp. 1210–1220, 2020.

[4] M. A. Chowdhury, N. Hossain, M. A. Kashem, M. A. Shahid, and A. Alam, "Immune response in COVID-19: A review," *J. Infect. Public Health*, vol. 13, no. 11, pp. 1619–1629, 2020.

[5] M. Z. Tay, C. M. Poh, L. Rénia, P. A. MacAry, and L. F. P. Ng, "The trinity of COVID-19: immunity, inflammation and intervention," *Nat. Rev. Immunol.*, vol. 20, no. 6, pp. 363–374, 2020.

[6] M. Government Of, *COVID-19 Case Information in Mexico [Data set]. Deputy Director of Notification and Epidemiological Records*. 2020.

[7] R. Jayawardena, P. Sooriyaarachchi, M. Chourdakis, C. Jeewandara, and P. Ranasinghe, "Enhancing immunity in viral infections, with special emphasis on COVID-19: A review," *Diabetes Metab. Syndr.*, vol. 14, no. 4, pp. 367–382, 2020.

[8] J. Paces, Z. Strizova, D. Smrz, and J. Cerny, "COVID-19 and the immune system," *Physiol. Res.*, vol. 69, no. 3, pp. 379–388, 2020.

[9] N. Jha *et al.*, "Deep learning approach for discovery of in silico drugs for combating COVID-19," *J. Healthc. Eng.*, vol. 2021, p. 6668985, 2021.

[10] N. Jha and D. Prashar, "Rapid Forecasting of Pandemic Outbreak Using Machine Learning: The Case of COVID-19. Enabling Healthcare 4.0 for Pandemics: A Roadmap Using AI," in *Machine Learning, IoT and Cognitive Technologies*, 2021, pp. 75–90.

[11] B. K. Patterson *et al.*, "Immune-based prediction of COVID-19 severity and chronicity decoded using machine learning," *Front. Immunol.*, vol. 12, p. 700782, 2021.

[12] E. Luellen, "A machine learning explanation of the pathogen-immune relationship of SARS-CoV-2 (COVID-19), and a model to predict immunity and therapeutic opportunity: A comparative effectiveness research study," *JMIRx Med*, vol. 1, no. 1, p. e23582, 2020.

[13] A. A. Willette *et al.*, "Using machine learning to predict COVID-19 infection and severity risk among 4510 aged adults: a UK Biobank cohort study," *Sci. Rep.*, vol. 12, no. 1, p. 7736, 2022.

[14] X. Guan *et al.*, "Clinical and inflammatory features based machine learning model for fatal risk prediction of hospitalized COVID-19 patients: results from a retrospective cohort study," *Ann. Med.*, vol. 53, no. 1, pp. 257–266, 2021.

[15] S. Aktar *et al.*, "Machine learning approach to predicting COVID-19 disease severity based on clinical blood test data: Statistical analysis and model development," *JMIR Med. Inform.*, vol. 9, no. 4, p. e25884, 2021.

[16] K. C.-Y. Wong, Y. Xiang, L. Yin, and H.-C. So, "Uncovering clinical risk factors and predicting severe COVID-19 cases using UK Biobank data: Machine learning approach," *JMIR Public Health Surveill.*, vol. 7, no. 9, p. e29544, 2021.

[17] A. L. Booth, E. Abels, and P. McCaffrey, "Development of a prognostic model for mortality in COVID-19 infection using machine learning," *Mod. Pathol.*, vol. 34, no. 3, pp. 522–531, 2021.

[18] A. K. Dubey, S. Narang, A. Kumar, S. M. Sasubilli, and V. García Díaz, "Performance estimation of machine learning algorithms in the factor analysis of COVID-19 dataset," *Computers, Materials and Continua*, 2020.

[19] N.-N. Sun *et al.*, "A prediction model based on machine learning for diagnosing the early COVID-19 patients," *bioRxiv*, 2020.

[20] Y. M. Mueller *et al.*, "Stratification of hospitalized COVID-19 patients into clinical severity progression groups by immuno-phenotyping and machine learning," *Nat. Commun.*, vol. 13, no. 1, p. 915, 2022.