Predictive Modelling of COVID-19 Severity Using Machine Learning Algorithms

Naralay Viji Paul[1], Amit Shah[2]

[1]PG Student, Computer Engineering, KITRC, Kalol, Gujarat, India
[2]Lecturer, Computer Engineering, LJ Polytechnic, LJ University, Gujarat, India

Abstract

The COVID-19 pandemic has placed unprecedented stress on global healthcare systems, amplifying the need for predictive tools to manage patient care more effectively. This study addresses the challenge of identifying patients at high risk of severe COVID-19 outcomes by developing machine learning models that leverage patient demographic data, clinical conditions, and pre-existing health issues. The dataset, obtained from the Mexican government's open-source COVID-19 patient records, includes comprehensive information on key comorbidities, hospitalization status, and patient demographics. Through a rigorous data preprocessing pipeline that includes cleaning missing values, encoding categorical variables, and scaling features, the dataset is prepared for advanced machine learning analysis. Various algorithms, including Logistic Regression, Random Forest, and Support Vector Machines, are applied to build predictive models. Model performance is measured using metrics such as accuracy, precision, recall, and the Area Under the Receiver Operating Characteristic Curve (AUC-ROC), with cross-validation techniques ensuring model generalizability. This research has the potential to provide actionable insights for healthcare professionals by enabling more efficient triaging of patients, improving resource allocation such as ICU beds and ventilators, and potentially reducing the burden on overwhelmed healthcare infrastructures during pandemics.

Keywords: **Include at least 5 keywords or phrases**

I. INTRODUCTION

The COVID-19 pandemic, caused by SARS-CoV-2, has presented one of the most severe global health crises in modern history, straining healthcare systems and disproportionately impacting individuals with pre-existing conditions such as cardiovascular disease, diabetes, and obesity. While many cases result in mild symptoms, a significant percentage of patients develop severe complications requiring hospitalization, intensive care, and mechanical ventilation. The unpredictable nature of disease progression has created an urgent need for predictive models that can help healthcare providers allocate resources effectively. Machine learning offers a promising solution by analyzing patient data to predict the severity of COVID-19 cases, enabling timely decision-making regarding hospital admissions, ICU prioritization, and ventilation support. This research explores the use of machine learning algorithms—Logistic Regression, Random Forest, and Support Vector Machines—to identify high-risk patients based on demographic, clinical, and comorbid factors. By leveraging data-driven insights, healthcare professionals can optimize patient management and improve outcomes, particularly in resource-limited settings where hospital capacity is a major concern.

Beyond physical health complications, the pandemic has also had a profound impact on mental health, triggering widespread psychological distress. Anxiety, depression, post-traumatic stress disorder (PTSD), and substance use disorders have become increasingly prevalent due to social isolation, financial instability, and uncertainty about the future. Vulnerable populations, including frontline healthcare workers and individuals with pre-existing mental health conditions, have been particularly affected. The stress of managing the crisis, coupled with the loss of loved ones and disrupted daily routines, has exacerbated mental health challenges globally. The integration of mental health considerations into predictive models is crucial, as it allows healthcare providers to assess not only the physical severity of COVID-19 cases but also the psychological

toll on patients. Public health campaigns emphasizing mental well-being, resilience-building strategies, and accessible support systems are essential in mitigating the long-term impact of the pandemic on mental health.

This study aims to develop a comprehensive predictive model that considers both physical and mental health factors to enhance risk assessment and optimize healthcare resource allocation. By analyzing key variables such as age, gender, comorbidities, and vital signs, the research seeks to determine the most significant predictors of severe COVID-19 outcomes. It also compares machine learning models to identify the most accurate and reliable approach for severity prediction. The expected outcome is a structured, interpretable, and efficient model that supports clinical decision-making, ensures optimal use of healthcare resources, and informs targeted interventions for high-risk patients. Furthermore, by incorporating mental health considerations, this research seeks to promote a more holistic understanding of patient needs during pandemics. Ultimately, this study aims to contribute to the development of a scalable and flexible predictive model that can be applied to other infectious diseases and public health emergencies, improving global healthcare preparedness and response.

II. OBJECTIVE

The first objective is to conduct a comprehensive analysis of patient data, including handling missing data and transforming categorical variables for machine learning models. Standardization and normalization of numerical features will be essential to ensure optimal model performance, as data quality directly impacts prediction reliability.

The next goal is to identify the most important patient attributes for predicting severe COVID-19 outcomes. Techniques such as correlation analysis, recursive feature elimination, and model-based importance ranking will be employed to focus on the most relevant features, improving predictive accuracy and efficiency.

The third objective involves developing machine learning models, including Logistic Regression, Random Forest, and Support Vector Machines, to predict COVID-19 severity. These models will be fine-tuned to address issues like data imbalance, which is common in healthcare datasets.

The models will be evaluated based on accuracy, precision, recall, and AUC-ROC, ensuring robust performance across unseen data using cross-validation and hyperparameter tuning to optimize predictions.

Finally, actionable insights will be derived from the models, helping healthcare professionals identify high-risk patients early. This will enable better resource management, such as ICU bed allocation, and provide valuable data for policy decisions during pandemics.

III. PROBLEM STATEMENT

The COVID-19 pandemic has significantly impacted global health systems, leading to a surge in cases requiring hospitalization and intensive care. Understanding the factors that contribute to the severity of COVID-19 outcomes is critical for effective patient management and resource allocation. However, healthcare providers often face challenges in identifying patients at high risk for severe complications, which can lead to overwhelmed healthcare facilities and poor patient outcomes.

The primary challenge lies in effectively analysing diverse patient datasets to predict the risk of severe COVID-19 outcomes. This process involves navigating the complexity of various factors, including demographic characteristics, pre-existing health conditions, and clinical presentations that influence disease severity. Traditional approaches may not adequately capture the intricate relationships between these factors and the likelihood of severe outcomes.

This research aims to address this gap by employing machine learning algorithms to analyze COVID-19 patient data, exploring patterns and correlations that could serve as predictive indicators for severe cases. By developing robust models that provide timely risk assessments, the study seeks to enhance decision-making processes in healthcare settings, ultimately contributing to improved patient care and efficient resource utilization during pandemics.

IV. LITERATURE REVIEW

Research on COVID-19 prediction and vaccine breakthrough using machine learning reveals gaps in generalizability, with models often relying on limited, region-specific datasets and short-term outcomes. The focus on supervised learning can miss the complexity of disease progression, and imbalanced datasets may skew mortality predictions. Many studies lack external validation and integration of granular data like demographics, genetics, and hospital capacity, reducing clinical applicability. Bridging these gaps would need diverse datasets and models capable of long-term, real-world predictions

Comparative study

Sr. No.	Title	Author(s)	Conclusion	Methodology
1	Forecast and Prediction of COVID-19 Using Machine Learning	Deepak Painuli, Divya Mishra, Suyash Bhardwaj	Extra Tree Classifier (ETC) achieved the highest accuracy (93.62%) for infection prediction. ARIMA was effective for forecasting case trends.	Machine Learning techniques (ETC, ARIMA) were used for prediction and time-series forecasting of COVID-19 cases in India.
2	COVID-19 Severity and Vaccine Breakthrough Infections in Idiopathic Inflammatory Myopathies and Other Autoimmune Diseases	Latika Gupta, Leonardo Santos Hoff, et al.	IIM patients had a higher risk of hospitalization but a lower risk of symptomatic pre-vaccination COVID-19 infection. Breakthrough infections were rare.	Self-reported electronic survey (COVAD) was used to analyze vaccine breakthrough infections and disease severity in autoimmune patients and healthy controls.
3	Analysis, Prediction, and Evaluation of COVID-19 Datasets Using Machine Learning	Kolla Bhanu Prakash, S. Sagar Imambi, Mohammed Ismail, T Pavan Kumar et al.	Random Forest Regressor and Random Forest Classifier were most effective in predicting COVID-19 infection rates and outcomes, particularly	Machine Learning models (Random Forest, SVM, XGBoost) were used to predict infection rates and identify demographic groups most affected by COVID-19

	Algorithms		for individuals aged 20-50 years.	using datasets from India.
4	Flares in Idiopathic Inflammatory Myopathies and the Timeline Following COVID-19 Vaccination	Naveen R., Parikshit Sen, Zoltán Griger, Jessica Day	9.6% of IIM patients reported flares post-vaccination. Risk factors like active disease and comorbidities influenced flare incidence.	Self-reported surveys (COVAD-1 and COVAD-2) were used to assess the incidence of post-vaccination flares in IIM patients and compare with other autoimmune disease patients.
5	Machine Learning Approaches in COVID-19 Diagnosis, Mortality, and Severity Risk Prediction	Norah Alballa, Isra Al-Turaiki	Models like Random Forest and XGBoost performed well in predicting COVID-19 mortality and severity based on clinical and laboratory data.	Review of supervised learning techniques (Random Forest, XGBoost, Logistic Regression) to predict COVID-19 diagnosis, severity, and mortality using clinical features.
6	Machine Learning-Based Prediction Models for the Prognosis of COVID-19 Patients with Diabeti	Zhongyuan Xiang, Shengfang Bu, Chen Xi	Logistic Regression achieved the highest accuracy (AUC 0.933) in predicting mortality and severe illness in COVID-19 patients with DKA.	Logistic Regression, Random Forest, XGBoost, SVM, and MLP were used to predict severity and mortality in COVID-19 patients with DKA based on clinical

	c Ketoacidosis (DKA)			and laboratory features.
7	Supervised Machine Learning-Based Prediction of COVID-19 Using Cloud-Based SVM Model	Atta-ur-Rahman, Kiran Sultan, Iftikhar Naseer, Rizwan Majeed, et al.	CSDC-SVM achieved an accuracy of 98.4% in detecting COVID-19 severity, showing potential for cloud-based detection in healthcare systems.	Support Vector Machine (SVM) was used in a cloud-based model to detect and classify COVID-19 cases based on severity levels using data from sensors and electronic medical records (EMRs).
8	Prediction of COVID-19 Severity and Mortality in Hospitalized Children Using Machine Learning Tree-based Classifiers	Mehran Karimi, Zahra Nafei, Farimah Shamsi, Elahe Akbarian	Decision Tree and Random Forest models performed well for severity prediction, while Gradient Boosted Decision Trees were most accurate for mortality prediction.	Decision Tree, Random Forest, Gradient Boosted Decision Trees, AdaBoost, and SVM were used to predict COVID-19 severity and mortality in children based on clinical and laboratory features.
9	Prediction of COVID-19 Mortality Using Machine Learnin	Seyed Salman Zakariaee, Negar Naderi, Mahdi Ebrahimi, Hadi	Random Forest achieved the highest accuracy (97.2%) in predicting mortality, with CT	Machine Learning models (Random Forest, XGBoost, SVM, Logistic Regression)

	g Algorithms and Chest CT Severity Score	Kazemi-Arpanahi	severity score being a critical predictor.	were used to predict mortality in COVID-19 patients based on chest CT severity score and clinical/laboratory features.
10	Machine Learning Forecasting Model for COVID-19 Pandemic in India	R. Sujath, Jyotir Moy Chatterjee, Aboul Ella Hassanien	Multilayer Perceptron (MLP) outperformed Linear Regression and Vector Autoregression in predicting COVID-19 case trends.	Machine Learning models (MLP, Linear Regression, Vector Autoregression) were used to predict COVID-19 case trends in India based on Kaggle datasets of confirmed, death, and recovery cases.
11	Machine Learning Approaches in COVID-19 Diagnosis, Mortality, and Severity Risk Prediction: A Review	Norah Alballa, Isra Al-Turaiki	Logistic Regression was the most effective algorithm due to ease of modeling and interpretability. Random Forest and XGBoost also performed well for diagnostic and prognostic tasks.	Reviewed ML algorithms like Logistic Regression, Random Forest, and XGBoost for COVID-19 applications using clinical and laboratory data.
12	COVID-19 Severity	Rambola, et al.	The mResNet-50 model outperforme	Combined PSEC classifier for detection and

	Prediction Using Combined Machine Learning and Transfer Learning Approaches		d other architectures with 97.79% accuracy for severity prediction, and PSEC achieved 95.5% accuracy for COVID-19 detection.	mResNet-50 for severity prediction using clinical data and CT-scan images.
13	Prediction of COVID-19 Using Machine Learning Techniques	Raja, et al.	Gradient Boosting achieved 90% accuracy for confirmed and cured cases, and 92% for death predictions, outperforming other models.	Used Random Forest, Decision Tree, SVM, Gradient Boost, and XGBoost with Kaggle datasets.
14	COVID-19 Severity Risk Prediction Using Machine Learning	Laatifi, et al.	UMAP-enabled Random Forest, XGBoost, and AdaBoost achieved perfect classification performance for severity estimation.	Feature extraction using UMAP and classification using ensemble models with clinical and laboratory data.
15	Severity Prediction of COVID-19 Patients Using Machin	Gull, et al.	SVM achieved the highest accuracy (60%) for severity prediction, helping prioritize	Applied seven classifiers, including SVM, Random Forest, and Naive Bayes, on Kaggle datasets with

	e Learning Classification Algorithms		healthcare resources.	clinical attributes.

V. METHODOLOGY

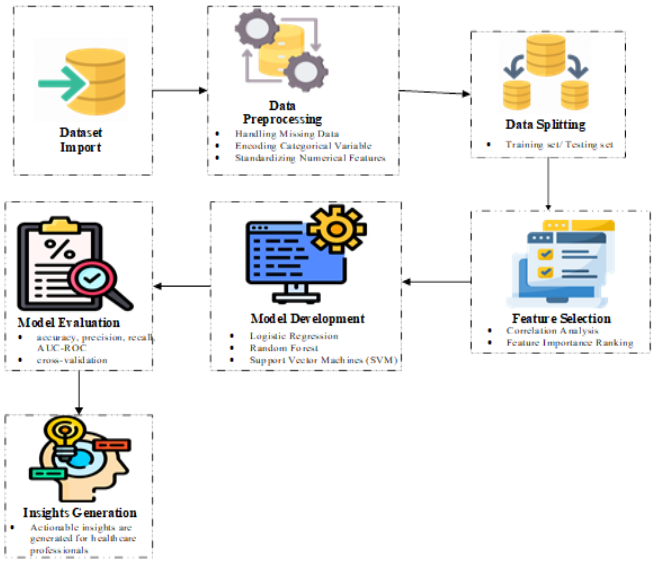

Fig 1: Proposed Flow of the Model

Data Overview

The provided dataset contains 566,602 entries with 23 columns that include patient information and medical conditions and medical test results for COVID-19. The system seeks to establish COVID-19 test results by analyzing patients' medical histories along with their symptom indications. The target variable covid_res shows three possible outcomes which include positive (1) and negative (2) test results as well as pending results tagged as (3).

The gathered data revealed several important points. The patient demographic reveals an imbalance because males outnumber females so gender-based biases may exist in the testing process. The data demonstrates that male patients show increased positivity rates when compared to female patients possibly due to the effects of biological or lifestyle components. The COVID-19 testing results of pneumonia-diagnosed patients revealed increased positivity rates that support existing understanding about COVID-19 complications. The dataset contains no null or missing data although three particular category entries (Feig's sign and Hutt's sign) serve to indicate unknown data points in the analysis stage. Supervisor intervention is required before further data analysis.

Data Pre-processing and Exploratory Data Analysis (EDA)

After data collection the information requires preparative processing for analytic purposes. The pre-processing step handles the resolution of technical problems like missing data values as well as handles inconsistent patterns and applies normalization procedures for numerical features. Machine learning algorithms need categorical variables which must undergo encoding before usage.

A. Data Summary

	count	mean	std	min	25%	50%	75%	max
sex	566602.0	1.506726	0.499955	1.0	1.0	2.0	2.0	2.0
patient_type	566602.0	1.215165	0.410937	1.0	1.0	1.0	1.0	2.0
intubed	566602.0	76.562952	39.058676	1.0	97.0	97.0	97.0	99.0
pneumonia	566602.0	1.846262	0.560939	1.0	2.0	2.0	2.0	99.0
age	566602.0	42.622483	16.659973	0.0	31.0	41.0	53.0	120.0
pregnancy	566602.0	50.400692	47.501579	1.0	2.0	97.0	97.0	98.0
diabetes	566602.0	2.210633	5.683523	1.0	2.0	2.0	2.0	98.0
copd	566602.0	2.280221	5.327832	1.0	2.0	2.0	2.0	98.0
asthma	566602.0	2.265029	5.334658	1.0	2.0	2.0	2.0	98.0
inmsupr	566602.0	2.319628	5.667381	1.0	2.0	2.0	2.0	98.0
hypertension	566602.0	2.145774	5.459866	1.0	2.0	2.0	2.0	98.0
other_disease	566602.0	2.410030	6.489959	1.0	2.0	2.0	2.0	98.0
cardiovascular	566602.0	2.286157	5.438405	1.0	2.0	2.0	2.0	98.0
obesity	566602.0	2.138905	5.395578	1.0	2.0	2.0	2.0	98.0
renal_chronic	566602.0	2.283765	5.393232	1.0	2.0	2.0	2.0	98.0
tobacco	566602.0	2.238360	5.571901	1.0	2.0	2.0	2.0	98.0
contact_other_covid	566602.0	31.573034	45.082123	1.0	1.0	2.0	99.0	99.0
covid_res	566602.0	1.728651	0.658710	1.0	1.0	2.0	2.0	3.0
icu	566602.0	76.562864	39.059060	1.0	97.0	97.0	97.0	99.0

Fig 2: Data statistics

The statistical breakdown of numerical columns in the dataset appears through the output generated by .describe().T function. Every item in the summary shows how the data is distributed and what characteristics these data points possess. A count shows the quantity of non-empty values across each column and the mean shows computed average values. SD (std) indicates the extent of data points dispersing around average values. The data shows its smallest and largest observed values in the min and max sections. The 25th, 50th (median), and 75th percentiles show the values which represent the lowest points where 25% and 50% and 75% of the data reside.

Majority of categorical data entries use numerical values for presentation. Several variables including sex, patient_type, intubed, pneumonia apart from missing or special-case records (represented by 97, 98, 99) have numerical values of 1, 2 in their data fields. The middle value of numerous encoded features rests at 2.0 which indicates binary numeric representation using 1 for "No" and 2 for "Yes" or another equivalent scheme of "Male/Female."

The average patient age stands at 42.62 years old but the dataset contains possible new borns identified by a zero value along with potential errors in an extreme 120-year-old entry. The dataset presents a bias toward middle-aged individuals based on the median age of 41 and the 53-years-old 75th percentile value.

The target variable covid_res shows a mean value of 1.72 which indicates that most results cluster around 1 and 2 based on binary encoding formats such as positive results coded as 1 and negative results as 2. The dataset contains 1 as its minimum value whereas the highest observation at 97 indicates possibly unknown data points.

The ICU variable shows a mean of 76.56 and standard deviation of 39.05 as its values spread from 99 down to 99. The high value points further validate that recorded information contains encoded blanks or unknown entries. The mean values for patients with diabetes hypertension and obesity fall between 2.1 and 2.3 while their median result is 2.0. The recorded data supports the assumption that 1 and 2 indicate binary categories because many patients were documented as having these health conditions.

B. Covid 19 Test Results by gender

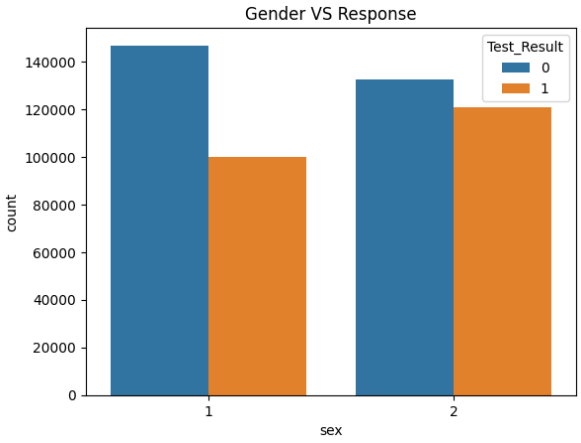

Fig 3: Covid 19 Test Results by gender

The "Gender vs Response" bar chart shows COVID-19 test results between female and male participants. Gender values within the dataset consist of numerical codes that classify females by '1' and males by '2' and test results demonstrate '0' for negative COVID results and '1' for positive COVID results.

The provided chart and data points reveal various important findings. There were more COVID-19 positive results among male test subjects who numbered 120,799 compared to female test subjects who totalled 99,858. A higher number of COVID-19 case confirmations existed amongst male participants in the study. According to the visual representation both negative test outcomes exceeded those of females. Females exhibit a substantial height difference between their blue negative case bar and orange positive case bar whereas males display smaller relative discrepancies between the bars.

The data shows that females received more negative test results although men experienced more positive results. The disparity shows that male individuals had greater probabilities of catching COVID-19 in this dataset when compared to females as infection rates were higher for men. The research data matches certain epidemiological studies that document gender variations in COVID-19 rates and exposure risks together with reporting norms.

C. Intubation vs. COVID-19 Test Result

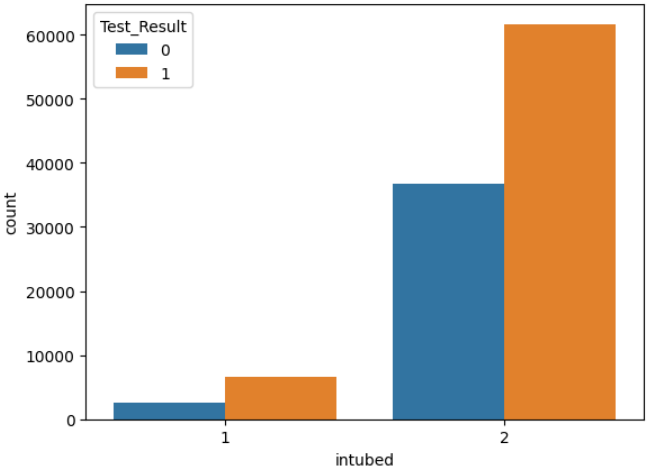

Fig.4: Intubation vs. COVID-19 Test Result

The bar chart demonstrates how patients who were or were not intubated performed in COVID-19 testing. The chart uses intubed as its x-axis variable with a value of 1 indicating non-intubated patients and value of 2 representing intubated patients requiring ventilation support. The number of individuals is displayed on the y-axis with the bars using blue for negative (0) and orange for positive (1) COVID-19 test results.

Reviewing this chart provides multiple important discoveries. The highest bar representing intubated patients with COVID-19 stands out as orange on the chart which demonstrates a high number of intubated individuals who tested positive for the virus. The nature of respiratory complications demonstrates the seriousness of

COVID-19 infections among patients. The non-intubated group displayed few cases of positive tests since the orange bar remained comparatively short in that section. Most of the testing positive individuals who did not receive intubation treatment exhibited either mild symptoms or no symptoms at all.

The data indicates that intubative procedures occurred infrequently among both COVID-positive and COVID-negative patients because their blue bars remain shorter. The general pattern reveals COVID-19 infected patients made up most of the patients who required intubation thus strengthening the link between the virus and severe respiratory failure in patients. The data presented in this chart demonstrates that intubation procedures strongly match the appearance of severe COVID-19 infections among patients.

D. Pneumonia vs. COVID-19 Test Result

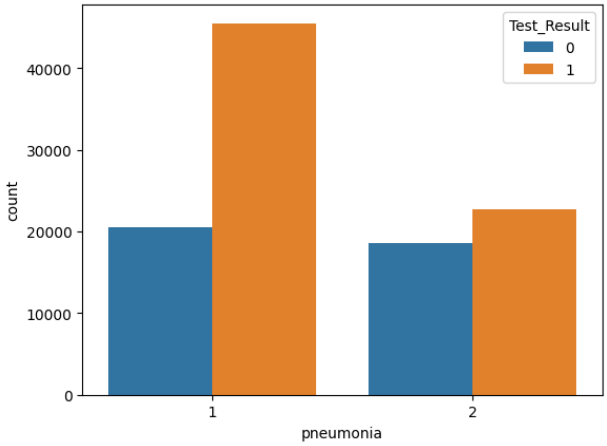

Fig 5.2.5: Pneumonia vs. COVID-19 Test Result

Pneumonia status reveals its correlation with the COVID-19 test results in the presented bar chart. In the chart pneumonia data points are positioned along the horizontal axis through numeric labels where 1 indicates no pneumonia disease and 2 denotes pneumonia condition. The horizontal axis displays person counts as numbers and diagnostic results use blue for negative COVID results (0) and orange for positive COVID results (1).

The presented chart reveals significant observations. Most coronavirus-positive patients without pneumonia fall under the tallest orange bar category according to the data. The data shows that a large number of COVID-19 patients were not presenting pneumonia symptoms during the time of testing positive for the virus. Pneumonia occurred with greater frequency in COVID-positive cases as compared to COVID-negative individuals. The orange bar that represents pneumonia = 2 surpasses the blue bar representing COVID-negative patients because significant numbers of COVID-positive patients experienced this complication.

The total number of COVID-negative people with pneumonia was minimal since the short blue bar appears under pneumonia = 2. The presented data distribution reinforces the notion that pneumonia diagnoses primarily involved COVID-19 infection instead of alternative causes of disease. The condition of pneumonia occurred more often among COVID-positive patients who did not start with pneumonia during the course of their illness in spite of its wide prevalence.

E. Age Distribution

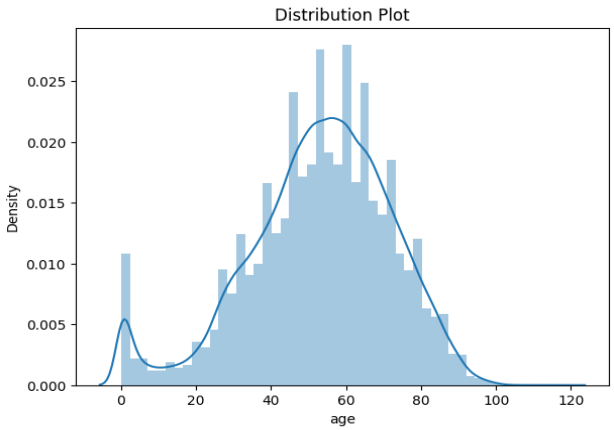

Fig 6: Density of different Age Groups

The distribution plot displays the age distribution of patients through a histogram and Kernel Density Estimation curve combination. The distribution plot demonstrates patient age groups using the x-axis and shows density concentration through the y-axis which illustrates data frequency patterns. Patient distribution by age shows as histogram bars while the blue KDE line delivers a smooth curve that represents the underlining age distribution of patients.

The plot reveals that most patients have ages between thirty to seventy showing older people account for the highest case numbers. Ethnicity presents itself as a key risk factor that makes elderly populations more susceptible to contracting COVID-19 with severe outcomes. Predictive modeling of COVID-19 incorporates age as a primary feature because it delivers crucial value to process-based forecasting solutions.

The dataset reveals that adult and elderly patients represent the most numerous entries since children and young people form much smaller parts of the overall distribution. This contrast points to the importance of age-based risk stratification in both medical decision-making and public health policy. A substantial cluster of extremely elderly patients extending beyond age 100 indicates a distribution that extends significantly on its right side. The right-skewed pattern of data distribution necessitates proper calibration models designed to handle extreme age values while preventing model prediction biases.

F. Covid 19 results by Age

Fig 7: Covid 19 by Age

This density plot displays how COVID-19 positive and negative persons were distributed according to their ages. The age values appear on the x-axis with the probability or density measurements presented on the y-axis across the separate age ranges. The blue and red shaded sections represent negative and positive COVID-19 outcomes which are characterized by the density distribution in each specified age segment.

The bar graph provides multiple significant observations regarding the data. Positive COVID-19 test results (in red) reach their maximum occurrence rate among people between 50 to 70 years of age. The data

demonstrates that patients aged between 50 and 70 years collectively obtain more COVID-19 diagnosis results therefore making older people more prone to virus infection. The data reveals a little peak of tests performed in children between 0-10 years of age but few results came back positive.

The positive cases contain a denser cluster of people in the 60-80 year age segment relative to their negative counterparts in this range. COVID-19 presents a greater risk to the elderly population since they experience the virus more frequently than younger adults.

G. COVID-19 Test Results Based on Tobacco Usage

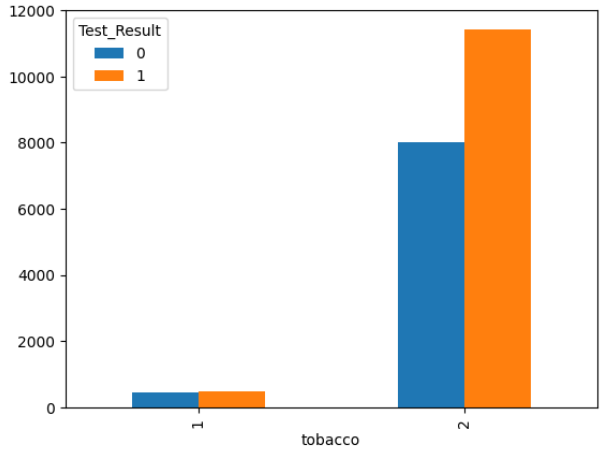

Fig 8: COVID-19 Test Results Based on Tobacco Usage

This visual displays the connection between tobacco intake and COVID-19 tests performed. The tobacco variable appears on the x-axis with non-smokers marked as 1 but smokers designated as 2 and the y-axis shows the total counts of individuals under consideration. Each bar contains different color directions indicating COVID-19 test results where negative cases (0) are shown in blue and positive results (1) appear in orange.

Multiple essential findings emerge from studying the presented chart. The highest number of confirmed COVID-positive patients belongs to smokers (category 2) because the orange bar indicating smokers reaches an elevated height in this category. The data indicates smoking serves as a possible risk factor for developing COVID-19 possibly because cigarette smoking has negative effects on respiratory health and impairs immune responses. The COVID-positive case numbers are considerably lower in non-smokers thus suggesting non-smokers face reduced chances of positive tests or severe COVID symptoms.

The visual graph confirms established health dangers which exist for smokers. The respiratory system of smokers suffers from weakening while their immune system defense decreases thereby making these patients more likely to fall sick with COVID-19. The medical research shows smoking as a risk factor that raises severe respiratory disease rates and physicians' findings support that smokers experience higher COVID-19 infection rates.

Data Splitting and Feature Selection

The COVID-19 datasets present an extreme unbalance of class occurrences because COVID-negative patients outnumber COVID-positive patients substantially. The datasets contain sparse distribution of critical outcomes such as ICU admission and intubation. Such class distribution imbalance creates predictions that benefit the majority class while producing poor results for the minority class which makes model detection of critical cases unreliable. The model becomes ineffective in proper pattern learning which leads to poor overall performance when dealing with the minority class.

The minority class population can be improved using SMOTE (Synthetic Minority Over-sampling Technique) which produces artificial examples. The method selects one random point from the minority class after which

it identifies k-nearest neighbors from the same class for interpolation to produce synthetic data points. The combination of SMOTE produces enhanced model performance because it allows the model to discover valuable patterns contained within the minority class.

For developing COVID-19 outcome prediction models via machine learning techniques it is mandatory to divide the dataset into parts that will be used for training and testing. Applying this approach leads to better data generalization because it prevents the model from using pattern memorization instead of genuine data analysis. The train_test_split() function splits the dataset into parts where training forms 80% while testing constitutes 20% of the available data. The usage of random_state=0 ensures that the data divisions persist identically from test to test to guarantee repeatable outcomes. A distinct testing set helps model evaluation through unseen data analysis to detect overfitting potentials and verify generalization abilities.

Feature scaling takes place on the split dataset through the StandardScaler() method. The dataset requires numerical feature scaling due to its different feature value intervals. The age feature spans from zero to one hundred but binary sex and diabetes features retain only no/yes values. Machine learning algorithms tend to favor more important features within their predictions when feature values have not been scaled. This results in prediction biases. StandardScaler transforms the data to have 0 as the mean and 1 as the standard deviation which results in identical feature significance during model development. The initialization of the transformation involves fit_transform(x_train) to calculate mean and standard deviation from training data and then standardize it. The testing data receives the previously applied transformation through transform(x_test) for maintaining data consistency and avoiding information leakage.

Standardization yields exceptional performance benefits for Logistic Regression alongside Support Vector Machines (SVM) and Neural Networks because these models need normalized numerical values. K-Nearest Neighbors (KNN) and SVM require feature scaling for proper feature comparison because they use Euclidean distance computations in their operations.

The COVID-19 prediction model establishes optimal conditions for training through data splitting while applying feature scaling from a prepared dataset. Data preprocessing methods stabilize models by maintaining them while removing feature dominance problems and let algorithms recognize proper patterns instead of fluctuating because of varying feature magnitudes.

VI. CONCLUSIONS

The research utilized machine learning models—Logistic Regression, Random Forest, and XGBoost—to predict COVID-19 severity, aiding physicians in identifying high-risk cases. Random Forest (88.50% accuracy) and XGBoost (87.53%) outperformed Logistic Regression (68.91%), with superior recall and AUC-ROC values. Key risk factors included age (50-70 years), male gender, smoking, and comorbidities like diabetes and hypertension. SMOTE addressed dataset imbalance, improving severe case detection. Feature scaling ensured unbiased predictions, and key variables enhanced model efficiency. These predictive systems optimize patient triage, resource allocation, and early intervention, proving machine learning's vital role in improving COVID-19 healthcare management.

REFERENCES

[1] Alballa, N. & Al-Turaiki, I., 2021. Machine learning approaches in COVID-19 diagnosis, mortality, and severity risk prediction: A review. *Informatics in Medicine Unlocked,* Volume 24, pp. 1-17.

[2] Atta-ur-Rahman, et al., 2020. Supervised Machine Learning-based Prediction of COVID-19. *Computers, Materials & Continua,* pp. 1-12.

[3] Deepak, P., Divya, M., Suyash, B. & Mayank, A., 2021. Forecast and prediction of COVID-19 using machine learning. *Elsevier,* pp. 381-393.

[4] Gull, H., Krishna, G., Aldossary, M. I. & Iqbal, S. Z., 2020. *Severity Prediction of COVID-19 Patients Using Machine Learning Classification Algorithms: A Case Study of Small City in Pakistan with Minimal Health Facility.* s.l., IEEE 6th International Conference on Computer and Communications.

[5] Gupta, L., Hoff, L. S., Sen, P. & Shinjo, S. K., 2022. POS0201 COVID-19 SEVERITY AND VACCINE BREAKTHROUGH INFECTIONS IN IDIOPATHIC INFLAMMATORY MYOPATHIES, OTHER SYSTEMIC AUTOIMMUNE AND INFLAMMATORY DISEASES, AND HEALTHY. *Annals of the Rheumatic Diseases*, pp. 334-335.

[6] Karimi, M., Nafei, Z., Shamsi, F. & Akbarian, E., 2024. Prediction of COVID-19 Severity and Mortality in HospitalizedChildren Using Machine Learning Tree-based Classifiers. *Research Square*, pp. 1-23.

[7] Laatifi, M. et al., 2022. Machine learning approaches in Covid-19 severity risk prediction in Morocco. *Journal of Big Data*, pp. 1-21.

[8] Naveen, Sen, P., Griger, Z. & Day, J., 2023. Flares in IIMs and the timeline following COVID-19 vaccination: a combined analysis of the COVAD-1 and 2 surveys. *British Journal of Rheumatology*, pp. 1-43.

[9] Norah, A. & Al-Turaiki, I., 2021. Machine learning approaches in COVID-19 diagnosis, mortality, and severity risk prediction: A review. *Informatics in Medicine Unlocked*, pp. 1-17.

[10] Prakash, K. B. et al., 2020. Analysis, Prediction and Evaluation of COVID-19 Datasets using Machine Learning Algorithms. *International Journal of Emerging Trends in Engineering Research*, pp. 1-6.

[11] Raja, M. A., Ullah, I., Babar, M. & Aziz, T., 2023. Prediction of COVID-19 using machine learning techniques. *ASIAN BULLETIN OF BIG DATA MANAGMENT*, pp. 1-15.

[12] Rambola, A. R., Andavar, S. & Raj, R. S. P., 2024. COVID-19 Severity Prediction Using Combined Machine Learning and Transfer Learning Approaches. *Engineering, Technology and Techniques*, pp. 1-18.

[13] Sujath, Chatterjee, J. M. & Hassanien, A. E., 2020. A machine learning forecasting model for COVID-19 pandemic in India. *Stochastic Environmental Research and Risk Assessment*, pp. 959-972.

[14] Xiang, Z., Hu, J., Bu, S. & Ding, J., 2024. Machine Learning-Based Prediction Models for thePrognosis of COVID-19 Patients with DKA. *Research Square*, pp. 1-14.

[15] Zakariaee, S. S., Naderi, N., Ebrahimi, M. & Kazemi-Arpanahi, H., 2023. Comparing machine learning algorithms to predict COVID-19 mortality using a dataset including chest computed tomography severity score data. *Scientific Report*, pp. 1-12.