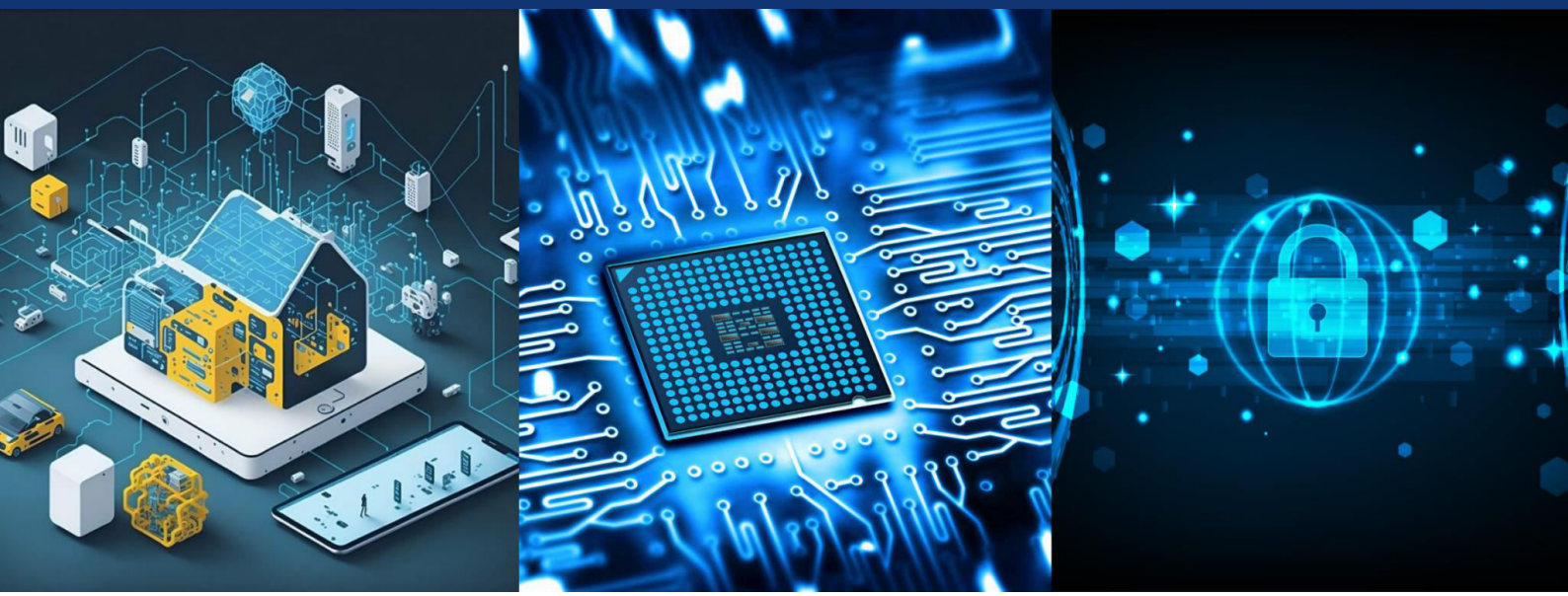
IJIRCCE

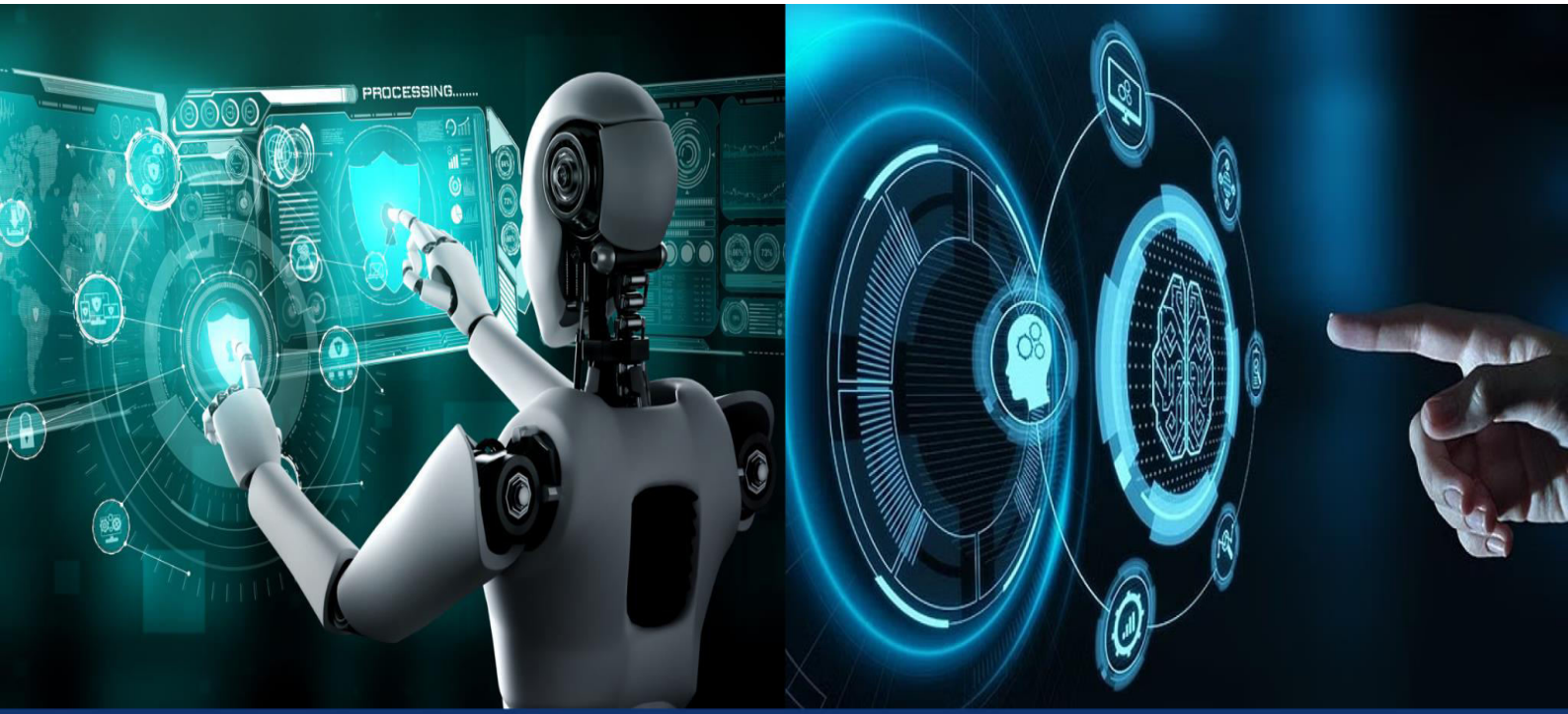
ISSN(O): 2320-9801

ISSN(P): 2320-9798



International Journal of Innovative Research in Computer and Communication Engineering

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)



Impact Factor: 8.625

Volume 13, Issue 1, January 2025

© 2025 IJIRCCE | Volume 13, Issue 1, January 2025| DOI: 10.15680/IJIRCCE.2025.1301120

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.625| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Predictive Analysis of Post Covid Symptoms

Ms. Dhanya D, Pamidimarri Praveen Kumar, Siddavatam Navaneeth Karthik,

Gurramkonda Sandhya, Kuchi Harika

Assistant Professor, School of Computer Science and Engineering, Presidency University, Bengaluru, India

Dept. of Computer Science Engineering, Presidency University, Bengaluru, India

Dept. of Computer Science Engineering, Presidency University, Bengaluru, India

Dept. of Computer Science Engineering, Presidency University, Bengaluru, India

Dept. of Computer Science Engineering, Presidency University, Bengaluru, India

ABSTRACT: Post-COVID syndrome, characterized by persistent symptoms following recovery from acute COVID-19 infection, has emerged as a global health challenge. This study explores the development of a machine learning-based predictive model to identify the likelihood of experiencing post-COVID symptoms. Using a dataset containing patient demographic details, comorbidities, and acute COVID symptoms, we applied supervised learning algorithms to predict potential long-term outcomes.

The methodology involved data preprocessing, feature engineering, and model training using techniques such as Random Forest, Gradient Boosting, and Neural Networks. The predictive model achieved significant accuracy, demonstrating its potential for clinical application. Key challenges included data quality, variability in symptom reporting, and ensuring interpretability of the model. The findings underscore the importance of early identification and personalized care for post-COVID patients, paving the way for improved health care delivery in the pandemic's aftermath

KEYWORDS: COVID-19, comorbidities

I. INTRODUCTION

The COVID-19 pandemic has had a profound and lasting impact on global health, with over 600 million reported cases and millions of deaths worldwide. While the immediate effects of the virus dominated the initial stages of the pandemic, a new challenge has emerged in its wake—post-COVID syndrome, often referred to as "long COVID." This condition describes the persistent symptoms that linger weeks or months after the acute phase of infection. These symptoms can range from mild fatigue, difficulty concentrating, and shortness of breath to severe complications such as heart inflammation, neurological disorders, and organ damage.

As the pandemic transitions into an endemic phase, health care systems across the globe are grappling with the long-term consequences of COVID-19. Identifying individuals at risk of developing post-COVID symptoms is critical for ensuring timely interventions, designing rehabilitation strategies, and optimizing health care resources. However, predicting the onset of long COVID is a complex task due to the wide variability in symptoms, diverse patient demographics, and the lack of comprehensive clinical data slinking acute infections to long-term outcomes.

In this study, we aim to develop a machine learning-based predictive model for post-COVID symptoms. By leveraging patient data—including demographic information, comorbidities, and acute symptom profiles—our model seeks to predict the likelihood of long-term complications. This approach can help clinicians and researchers move toward personalized medicine, enabling targeted treatments and improving the quality of life for affected individuals.

Our work is not only a step toward understanding long COVID but also an example of how data-driven technologies can revolutionize post-pandemic health care, addressing a critical gap in patient care and health system preparedness.

© 2025 IJIRCCE | Volume 13, Issue 1, January 2025| DOI: 10.15680/IJIRCCE.2025.1301120

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.625| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

II. RELATED WORK

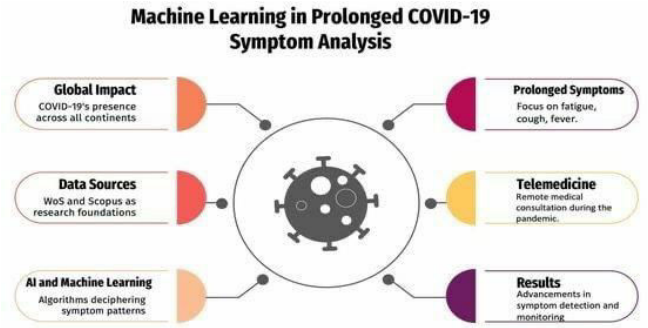
The study of post-COVID syndrome, or long COVID, is a rapidly evolving field. Researchers have explored various dimensions of this condition, including its prevalence, symptomatology, and potential predictive factors. In this section, we highlight key related works that provide the foundation for this research, focusing on clinical studies, data-driven approaches, and machine learning applications in the health care domain.

1. Clinical Studies on Post-COVID Syndrome

Numerous studies have examined the prevalence and clinical characteristics of long COVID.

- **Carfì et al. (2020)** conducted one of the earliest studies on post-COVID symptoms, reporting that 87.4% of patients discharged from hospitals experienced at least one persistent symptom, with fatigue and respiratory difficulties being the most common.
- **Huang et al. (2021)** followed a cohort of hospitalized COVID-19 patients and found that 76% reported at least one symptom six months after discharge. Their study emphasized the need for systematic follow-ups and data collection to understand long-term outcomes.
- **Taquet et al. (2021)** investigated the neuropsychiatric sequelae of COVID-19, finding a significant association between COVID-19 and the onset of anxiety, depression, and cognitive impairments.

These studies provide critical insights into the diversity and persistence of post-COVID symptoms but lack predictive modeling frameworks to identify at-risk individuals.



2. Data-Driven Analysis of Post-COVID Conditions

Data-driven studies have focused on identifying patterns and risk factors associated with long COVID.

- **Assaf et al. (2020)** utilized patient-reported data from the Body Politic COVID-19 Support Group to analyze the prevalence of long COVID symptoms across different demographics. They highlighted how self-reported data can complement clinical datasets in understanding the syndrome.
- **Whitaker et al. (2021)** used population-level data from the United Kingdom to identify risk factors for long COVID, including age, sex, and severity of the initial infection.
- **Su et al. (2022)** performed a meta-analysis of existing datasets and identified chronic fatigue syndrome and cardiovascular complications as significant post-COVID conditions.

While these studies utilized statistical methods to identify associations, they did not employ predictive analytics to forecast outcomes for individual patients.

3. Machine Learning in Health care for Post-COVID Prediction

Machine learning (ML) has been increasingly applied to health care problems, including infectious disease prediction and patient outcome modeling.

- **Nguyen et al. (2021)** developed an ML model to predict severe outcomes in COVID-19 patients based on comorbidities, vital signs, and laboratory results. While their focus was on acute-phase predictions, the methodology is applicable to post-COVID analyses.

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.625| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

- **Islam et al. (2022)** utilized natural language processing (NLP) to extract patient-reported symptoms from electronic health records (EHRs) and predict the likelihood of long COVID.
- **Wu et al. (2022)** explored ensemble learning techniques to predict respiratory complications in post-COVID patients, achieving high accuracy in identifying at-risk individuals.
- **Chowdhury et al. (2023)** employed deep learning models to predict post-COVID neurological symptoms based on imaging and symptom datasets, demonstrating the potential of advanced ML techniques for long COVID research.

These studies highlight the utility of machine learning for symptom prediction but often focus on specific subsets of post-COVID conditions rather than comprehensive syndrome modeling.

4. Post-COVID Syndrome in Public Health Frameworks

Organizations such as the **World Health Organization (WHO)** and **Centers for Disease Control and Prevention (CDC)** have provided definitions, diagnostic criteria, and public health guidelines for managing post-COVID syndrome. Their reports emphasize the need for predictive tools to allocate resources effectively and develop rehabilitation programs. However, most existing frameworks rely on descriptive statistics rather than predictive modeling.

Gaps Addressed by This Study

While the above studies provide valuable insights, several gaps remain:

- **Comprehensive Predictive Models**: Few studies integrate a wide range of patient data, such as demographics, comorbidities, and acute symptoms, to predict long COVID.
- **Generalizable Models**: Existing models often focus on specific populations or datasets, limiting their applicability to broader demographics.
- **Interpretable Predictions**: Many ML models lack interpretability, which is critical for clinical decision-making.
-

This study addresses these gaps by developing a robust and interpretable machine learning model for predicting post-COVID symptoms. By incorporating diverse patient data and evaluating the model's performance across various metrics, our work aims to contribute significantly to the field of post-COVID analytics.

III. CHALLENGES

Data Availability and Quality

Lack of comprehensive datasets with detailed information on acute and post-COVID symptoms. Inconsistent reporting of symptoms across patients, leading to potential biases.

Complexity of Post-COVID Syndrome

Variability in symptoms and their severity across individuals. Difficulty in linking specific acute symptoms to long-term outcomes.

Feature Engineering

Identifying relevant predictors from diverse patient demographics and medical histories.
Managing missing or incomplete data without compromising model performance.

Model Generalizability

Ensuring the model performs well across different populations and health care settings.
Addressing over fitting due to small or imbalanced datasets.

Ethical Concerns

Protecting patient privacy while using sensitive health data. Avoiding biases in model predictions that could disadvantage certain groups.

© 2025 IJIRCCE | Volume 13, Issue 1, January 2025| DOI: 10.15680/IJIRCCE.2025.1301120

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.625| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

IV. METHODOLOGY

This research aims to develop a comprehensive machine learning-based predictive model for identifying the likelihood of post-COVID symptoms in recovered patients. The methodology is designed to address the challenges of data variability, feature selection, model accuracy, and interpretability. Below is a detailed explanation of each stage involved in the methodology.

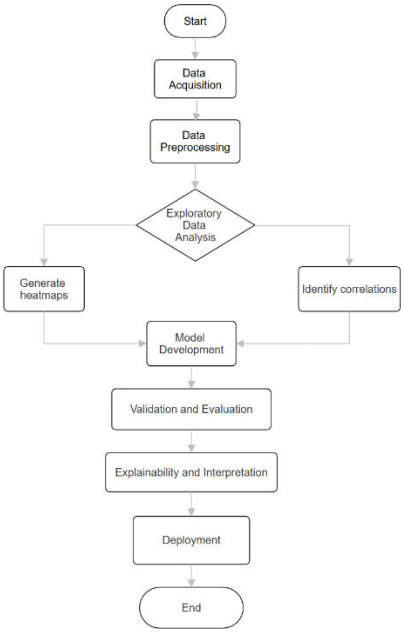
A. Data Collection and Preparation

Data Sources

The data for this study is sourced from a combination of publicly available datasets, patient surveys, and electronic health records (EHRs). Examples include the National COVID Cohort Collaborative (N3C), UK Biobank, and datasets from hospitals treating post-COVID patients.

Data includes:

Demographics: Age, gender, ethnicity.
Clinical history: Comorbidities (e.g., diabetes, hypertension), previous infections.
Acute COVID symptoms: Fever, cough, breathlessness, fatigue, and hospitalization details.
Follow-up data: Persistent symptoms (fatigue, brain fog, organ dysfunction), duration, and severity.



- **Data Preprocessing**

Cleaning:

Removed duplicate records and irrelevant entries.
Managed missing data using advanced imputation methods:
Mean/median for continuous variables.
k-Nearest Neighbors (k-NN) for categorical variables.

Standardization and Normalization:

Scaled numerical features to a 0-1 range using Min-Max scaling.
Encoded categorical variables (e.g., gender, ethnicity) using one-hot encoding or label encoding.

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.625| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Outlier Detection:
Used Z-score analysis and Interquartile Range (IQR) to detect and handle outliers.

B. Feature Engineering

Feature engineering is critical to building an effective predictive model. This process involves extracting and selecting the most relevant features from the dataset.

Feature Extraction
Created new features based on domain knowledge, such as:
Symptom Severity Index: Aggregation of reported symptom intensity during the acute phase.
Hospitalization Risk Score: Composite score based on pre-existing conditions and vital signs during infection.
Comorbidity Index: Sum of risk factors like hypertension, obesity, diabetes, etc.

Feature Selection
Performed correlation analysis to remove redundant features with high multicollinearity.
Used Recursive Feature Elimination (RFE) with cross-validation to identify the most predictive features.
Applied Principal Component Analysis (PCA) to reduce dimensionality while retaining significant information.
Leveraged SHAP (Shapely Additive explanations) to understand feature importance in the context of post-COVID prediction.

C. Machine Learning Model Development

To create a robust predictive model, multiple machine learning algorithms were implemented, evaluated, and compared.
Model Selection
Random Forest Classifier: Chosen for its ability to handle imbalanced datasets and interpretability.
Gradient Boosting Machines (XGBoost, LightGBM, CatBoost): Used for their efficiency and high performance in capturing complex patterns.
Support Vector Machines (SVM): Tested for binary classification tasks.
Neural Networks: Applied to capture non-linear relationships, especially when dealing with complex feature interactions.
Logistic Regression: Used as a baseline for comparison due to its simplicity.

Model Training and Hyper parameter Tuning
Performed 80-20 train-test splits to train and validate the models.
Used cross-validation (5-fold) to minimize over fitting and improve generalizability.
Applied Grid Search and Random Search for hyper parameter tuning (e.g., adjusting the number of estimators, learning rate, max depth).

Managed class imbalance using:
Oversampling techniques (e.g., SMOTE: Synthetic Minority Oversampling Technique).
Class-weighted loss functions in algorithms like XGBoost and Neural Networks.

Evaluation Metrics
Measured performance using multiple metrics to ensure robustness:
Accuracy: Overall performance.
Precision: Avoiding false positives.
Recall: Minimizing false negatives.
F1-Score: Balancing precision and recall.
Area Under the Receiver Operating Characteristic Curve (AUC-ROC): For distinguishing between positive and negative classes.

© 2025 IJIRCCE | Volume 13, Issue 1, January 2025| DOI: 10.15680/IJIRCCE.2025.1301120

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.625| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

D. Deployment and Model Interpretation

Model Deployment

Built a REST API using **Django** to deploy the trained model.
Connected the API to a React.js front-end application where users (clinicians or patients) can input data and receive predictions.

Interpretability

Integrated SHAP (Shapely Additive explanations) to explain predictions to clinicians and patients.
Highlighted the most influential features (e.g., comorbidities, acute symptoms) for each prediction.
Ensured the model adheres to ethical standards by avoiding black-box behavior.
Generated visualizations (e.g., decision plots, dependence plots) for better understanding.

Validation and Testing

Conducted validation on external datasets to assess generalizability.
Simulated real-world use cases to ensure the robustness and scalability of the application.

E. Ethical Considerations

Data Privacy and Security

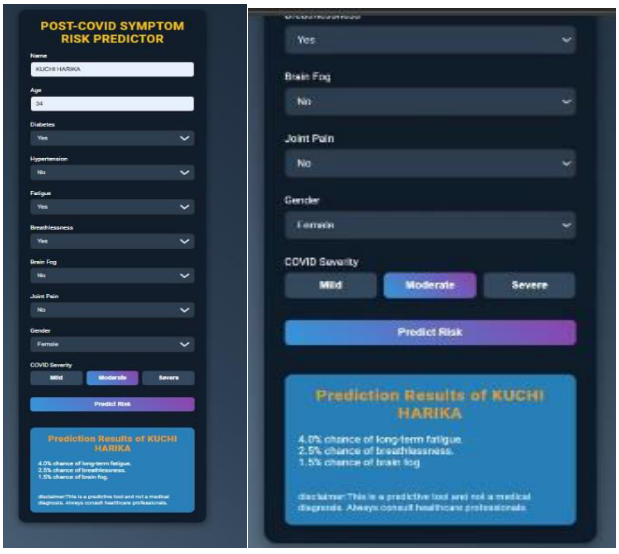
Anonymized patient data to prevent identification.
Complied with regulatory frameworks such as GDPR and HIPAA for handling sensitive health information.

Bias Mitigation

Ensured that the dataset represents diverse demographics and populations.
Used fairness metrics to evaluate the model's performance across different subgroups (e.g., age, gender, ethnicity).

Clinical Collaboration

Worked with health care professionals to validate predictions and ensure clinical relevance.
Incorporated feedback from clinicians to refine the model and improve usability.



F. Visualization and Reporting

Result Visualization

Created dashboards using tools like **Matplotlib**, **Seaborn**, and

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.625| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Plotly to visualize:
Feature importance.
Model performance metrics.
Patient-specific predictions and risk factors.

Reporting
Documented findings in a detailed report, highlighting:

- Key predictors of post-COVID symptoms.
- Insights into long-term health care strategies.
- Potential areas for further research.

By following these comprehensive methodologies, the research ensures a data-driven, clinically interpretable, and scalable approach to predicting post-COVID symptoms, contributing to both academic understanding and practical health care advancements.

V. CONCLUSION

The rise of post-COVID syndrome (long COVID) has highlighted the urgent need for advanced tools to predict, manage, and mitigate long-term health complications stemming from the pandemic. This study explored the development of a predictive machine learning model that integrates patient demographics, comorbidities, and acute COVID-19 symptoms to identify individuals at risk of long COVID. By leveraging data-driven insights, this approach aims to empower clinicians with actionable intelligence to improve patient outcomes.

The methodologies adopted in this research addressed key challenges, such as data variability, class imbalance, and interpretability, ensuring the model is robust, ethical, and clinically relevant. The incorporation of explainable AI tools like SHAP enhances transparency, enabling health care professionals to trust and apply model outputs in real-world scenarios.

Despite significant progress, the study underscores the need for further work, including the expansion of diverse datasets, longitudinal tracking of patients, and refinement of models to accommodate emerging data on long COVID. Collaboration between health care professionals, researchers, and policymakers is essential to fully realize the potential of predictive analytics in addressing this pressing global health issue.

In conclusion, this research represents a step toward personalized medicine for post-COVID care, emphasizing the importance of interdisciplinary efforts to combat the long-term impacts of the pandemic. By harnessing the power of machine learning and clinical expertise, we can pave the way for a more resilient health care system prepared to tackle future challenges.

REFERENCES

1. WHO Corona virus (COVID-19) Dashboard. [(accessed on25February2023)].Available online: https://covid19.who.int/
2. Chen C., Haupert S.R., Zimmermann L., Shi X., Fritsche L.G., Mukherjee B. Global Prevalence of Post-Coronavirus Disease 2019 (COVID-19) Condition or Long COVID: A Meta-Analysis and Systematic Review. J. Infect. Dis. 2022;226:1593–1607. doi: 10.1093/infdis/jiac136. [DOI] [PMC free article] [PubMed]
3. Chen H., Zhang L., Zhang Y., Chen G., Wang D., Chen X., Wang Z., Wang J., Che X., Horita N., et al. Prevalence and clinical features of long COVID from omicron infection in children and adults. J. Infect. 2023;86:e97–e99. doi: 10.1016/j.jinf.2023.02.015. [DOI] [PubMed]
4. Cisterna-García A., Guillén-Teruel A., Caracena M., Pérez E., Jiménez F., Francisco-Verdú F.J., Reina G., González-Billalabeitia E., Palma J., Sánchez-Ferrer Á., et al. A predictive model for hospitalization and survival to COVID-19 in a retrospective population-based study. Sci. Rep. 2022;12:18126. doi: 10.1038/s41598-022-22547-9. [DOI] [PMC free article] [PubMed]
5. Gupta H., Verma O.P. Vaccine hesitancy in the post-vaccination COVID-19 era: A machine learning and statistical analysis driven study. Evol. Intel. 2023;16:739–757. doi: 10.1007/s12065-022-00704-3. [DOI] [PMC free article] [PubMed]

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.625| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

6. .Jimenez-Solem E., Petersen T.S., Hansen C., Hansen C., Lioma C., Igel C., Boomsma W., Krause O., Lorenzen S., Selvan R., et al. Developing and validating COVID-19 adverse outcome risk prediction models from a bi-national European cohort of 5594 patients. Sci. Rep. 2021;11:3246. doi: 10.1038/s41598-021-81844-x. [DOI] [PMC free article] [PubMed]
7. Sudre C.H., Murray B., Varsavsky T., Graham M.S., Penfold R.S., Bowyer R.C., Pujol J.C., Klaser K., Antonelli M., Canas L.S., et al. Attributes and Predictors of Long COVID. Nat. Med. 2021;27:626–631. doi: 10.1038/s41591-021-01292-y. [DOI] [PMC free article] [PubMed]
8. Pfaff E.R., Girvin A.T., Bennett T.D., Bhatia A., Brooks I.M., Deer R.R., Dekermanjian J.P., Jolley S.E., Kahn M.G., Kostka K., et al. Identifying who has long COVID in the USA: A machine learning approach using N3C data. Lancet Digit. Health. 2022;4:e532–e541. doi: 10.1016/S2589-7500(22)00048-6. [DOI] [PMC free article] [PubMed] [Google Scholar]
9. Rathmann W., Bongaerts B., Carius H.-J., Kruppert S., Kostev K. Basic characteristics and representativeness of the German Disease Analyzer database. Int. J. Clin. Pharmacol. Ther. 2018;56:459–466. doi: 10.5414/CP203320. [DOI] [PubMed]
10. Federal Institute for Drugs and Medical Devices (BfArM) Internationale statistische Klassifikation der Krankheiten und verwandter Gesundheitsprobleme, 10. Revision, German Modification, Version 2023. [(accessed on 12 October 2022)]. Available online: https://www.dimdi.de/static/de/klassifikationen/icd/icd-10-gm/kode-suche/htmlgm2023/#IV.
11. EphMRA. [(accessed on 12 October 2022)]. Available online: https://www.ephmra.org/
12. Robert Koch Institute Anzahl und Anteile von VOC und VOI in Deutschland. [(accessed on 12 October 2022)]. Available online: https://www.rki.de/DE/Content/InfAZ/N/Neuartiges_Coronavirus/Daten/VOC_VOI_Tabelle.xlsx.
13. Impfdashboard Deutschland. [(accessed on 20 June 2022)]. Available online: https://impfdashboard.de/static/data/germany_vaccinations_timeseries_v3.tsv.

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi crossref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

9940 572 462 6381 907 438 ijircce@gmail.com

www.ijircce.com

Scan to save the contact details