PrognosisAI

AI-Driven In-Hospital Mortality Prediction for Enhanced COVID-19 Patient Management

PROJECT SYNOPSIS
OF MAJOR PROJECT

BACHELOR OF TECHNOLOGY
Computer Science and Engineering (CSE)

SUBMITTED BY-

Name:	Mansi	Nandini Vashistha
University Roll No.:	2100290100091	2100290100106
Class Roll No.:	25	40

(Batch: 2021-2025)

KIET Group of Institutions, Delhi-NCR,
Ghaziabad (UP)
Department of Computer Science and Engineering

INTRODUCTION

The emergence of the coronavirus disease (COVID-19) has brought to the forefront a pressing concern: the vulnerability of hospitalized patients to the specter of mortality. In response to this challenge, the application of advanced Machine Learning (ML) algorithms emerges as a promising avenue for predicting mortality among COVID-19 hospitalized patients. In view of this, our study is driven by the overarching goal of conducting a comprehensive comparative analysis of diverse ML algorithms. This analysis is centered around their efficacy in predicting COVID-19 mortality using patient data collected at the critical juncture of initial admission. The ultimate objective is to discern the most adept algorithm among the array of contenders, thus furnishing a potent predictive instrument that can be judiciously wielded in the realm of healthcare decision-making.

In the wake of the global COVID-19 pandemic, the urgency of accurate predictive models has never been more pronounced. This project endeavors to contribute to the understanding of COVID-19 mortality prediction by employing advanced machine learning algorithms. Leveraging the power of technology and data-driven insights, this project seeks to assist healthcare systems in making informed decisions, managing resources effectively, and prioritizing patient care. The project operates at the intersection of data science and healthcare analytics, with a focus on developing predictive models that can provide valuable insights during times of crisis.

The project's technological foundation rests on the utilization of cutting-edge machine learning algorithms. As the pandemic continues to challenge healthcare systems worldwide, it becomes imperative to harness the capabilities of machine learning to augment decision-making processes. By extracting patterns from vast datasets and identifying features that significantly impact COVID-19 mortality, machine learning algorithms offer a toolset that has the potential to revolutionize healthcare practices.

Rationale:

The rationale underpinning this study is rooted in the urgent need for accurate and timely mortality prediction tools in the context of COVID-19. The dynamics of this pandemic have underscored the significance of rapid decision-making within the healthcare domain. Predicting the likelihood of mortality at the onset of hospitalization holds paramount importance for clinicians and healthcare administrators alike. Efficient allocation of resources, optimization of treatment strategies, and patient prioritization necessitate a precise understanding of potential outcomes. This pressing need aligns seamlessly with the capabilities of Machine Learning algorithms, which excel at discerning complex patterns within extensive datasets. By leveraging the power of ML, we can extract critical insights from patient data, contributing to enhanced prognostic accuracy and ultimately facilitating the deployment of tailored interventions.

Moreover, the versatility of Machine Learning algorithms allows for the identification of intricate relationships among a multitude of variables, some of which might elude human intuition. Traditional methods of mortality prediction might fall short in capturing these nuanced connections, thereby limiting their efficacy. In contrast, the data-driven nature of ML promises to unlock a deeper comprehension of the myriad factors contributing to COVID-19 mortality. By conducting a rigorous comparison of several ML algorithms, we aim to discern not only which approach yields the most accurate predictions, but also to unravel the underlying factors that drive these predictions. This study's rationale is anchored in the potential to harness sophisticated computational techniques to enhance our ability to foresee critical outcomes, offering tangible benefits in the race against COVID-19.

Objectives:

1. Algorithm Performance Evaluation:

The primary objective of this study is to rigorously assess the performance of multiple Machine Learning algorithms in predicting COVID-19 mortality among hospitalized patients. By subjecting various algorithms to a comprehensive evaluation, we aim to quantify their predictive accuracy and robustness. This entails measuring metrics such as sensitivity, specificity, accuracy, and the area under the receiver operating characteristic curve (AUC-ROC). Through this evaluation, we strive to identify the algorithms that exhibit consistently high predictive prowess across diverse patient profiles.

2. Feature Identification and Analysis:

Beyond assessing algorithmic performance, another pivotal objective is the identification and analysis of key features that significantly contribute to accurate COVID-19 mortality predictions. By scrutinizing the importance and relevance of various input variables, we seek to unravel the intricate relationships that underlie mortality outcomes. This process entails employing techniques such as feature importance ranking and correlation analysis. The insights garnered from this objective can potentially shed light on critical clinical indicators that might have otherwise been overlooked, enriching our understanding of mortality determinants in COVID-19 cases.

3. Algorithm Suitability Analysis:

In the heterogeneous landscape of COVID-19 patient data, the performance of Machine Learning algorithms can vary. Hence, a central goal of this study is to offer a nuanced analysis that recommends the most suitable algorithms for different categories of patient data. We aim to provide a pragmatic guide that healthcare practitioners can consult when selecting an appropriate algorithm for specific patient demographics or data characteristics. This objective aligns with the imperative of ensuring that predictive models are tailored to the unique profiles of patients,

thereby enhancing the practicality and effectiveness of these tools in real-world clinical settings.

4. Decision-Making Tool Development:
Building on the outcomes of the preceding objectives, the final goal of this study is to develop a robust and reliable predictive model for COVID-19 mortality. By selecting the algorithm that consistently demonstrates superior performance and integrating it with the identified key features, we aspire to craft a decision-making tool that empowers healthcare professionals with timely and accurate mortality predictions. This tool holds the potential to influence critical aspects of patient care, resource allocation, and treatment strategies, fostering more informed decision-making in the challenging landscape of COVID-19 healthcare management.

Literature Review

Paper 1

Coronavirus disease (COVID-19) cases analysis using machine-learning applications

- Ameer Sardar Kwekha-Rashid,
- Heamn N. Abduljabbar &
- Bilal Alhayani

Applied Nanoscience **volume 13**, pages2013–2025 (2023)
doi: https://doi.org/10.1007/s13204-021-01868-7

1.Coronavirus Disease analysis using machine learning applications

This paper shows the role of machine learning and its algorithms in detecting and investigating coronavirus and various purposes related to it. The general deep learning based feature extraction is used to complete the process. Algorithms such as MobileNet, DenseNet, Xception, ResNet, InceptionV3, InceptionResNetV2, VGGNet, NASNet were selected among a group of DCNN for the recognition of novel Coronavirus from machine learning supporting radiologists in analysing and recognising COVID-19. It was seen that supervised learning methods were more efficient in the process than unsupervised and reinforcement learning methods. The study of 14 original articles in year 2020 that used supervised learning models gives us the review of different models and their response in analysing different cases and percentage of accuracy.

S.no	Country and Dataset	TASKS AND ALGORITHM	RESULTS
1.	India 212 reports	Classification, LR, Naïve bayes	It gives 96% accuracy obtained from other findings.
2.	USA US health systems 197 patients	Classification and Logical Regression	This algo gives higher diagnostic odds ratio and 0.90 sensitivity than MEWS(0.78 sensitivity) ,minimizes false positive results.
3.	USA 184,319 reported cases	Classification, CNN	This helps frontline physicians in the emergency dept to better risk-assess patients and predict who would require intubation and mechanical ventilation.
4.	Italy 85 chest X-rays	Classification, KNN	A method to automatically detect by analysing medical images and differentiating covid with other pulmonary disease
5.	Germany 150 covid patients 500 chest CTs	Classification, CNN	Segmentation of lung and opacity quantification.
6.	Egypt 5000 covid	Regression, Linear Regression	The fourth degree polynomial regression model helps govt to

	cases		prepare their procedures and predicts epidemic peak and its ending time.
7.	Israel 6995 patients	Classification, ANN	This helps in accurate risk prediction, optimize patients triage, better prioritization of medical resources and overall management of covid 19 pandemic.
8.	Netherlands 319 patients	Classification, Logistic Regression	Improves accuracy of diagnostic chest CT for covid.
9.	Germany 368 independent variables	Classification, Naïve bayes	This focuses on variables and factors that increase contamination.
10.	USA 290 patients	Classification, Logistic Regression	There is no correlation between mortality and treatment. The algo detects 31% of improving among the covid 19 population.
11.	International Food for each of 170 countries.	Clustering, KNN (Unsupervised learning)	The countries with highest death ratio were those with high fat consumption and with lower death rate have a higher level of cereal consumption.
12.	USA 790 Korean Immigrants	Classification, ANN	It predicted the person's flexibility, familiarities of everyday discernments and the racism actions towards the Asians in US
13.	Southern Ethiopia 244 samples	Classification, Logistic Regression	Shows the affect of coronavirus on the health care staff
14.	USA (centre of Disease and Control and John Hopkins University) Database of 57 candidates	Classification , ANN	It describes the presence/absence of hotspot of covid in geographic system.

Table-1

The above table shows that the ML models shows promising results in treating the coronavirus and also giving important results for all other aspects related to it other than its treatment. There is only one unsupervised learning method used in the articles which shows that supervised learning models give more accurate results i.e.92% accuracy compared to 7.1% accuracy of unsupervised ML model. This shows that ML models are an effective way to study this huge disease and also improving the treatment techniques.

Paper-2

Comparing machine learning algorithms for predicting COVID-19 mortality

- Khadijeh Moulaei,
- Mostafa Shanbehzadeh,
- Zahra Mohammadi-Taghiabad &
- Hadi Kazemi-Arpanahi

BMC Medical Informatics and Decision Making **volume 22**, Article number: 2 (2022)

https://doi.org/10.1186/s12911-021-01742-0

2.Machine Learning-Based Mortality Prediction for Hospitalized COVID-19 Patients

This study addresses the critical need for accurate mortality prediction among hospitalized COVID-19 patients using machine learning (ML) techniques. The research involved the creation and evaluation of predictive models based on 38 essential clinical features. The primary objective was to enhance patient care and optimize the allocation of constrained hospital resources by identifying high-risk cases as early as admission or during hospitalization.

Algorithms and their efficiency in Covid mortality rate

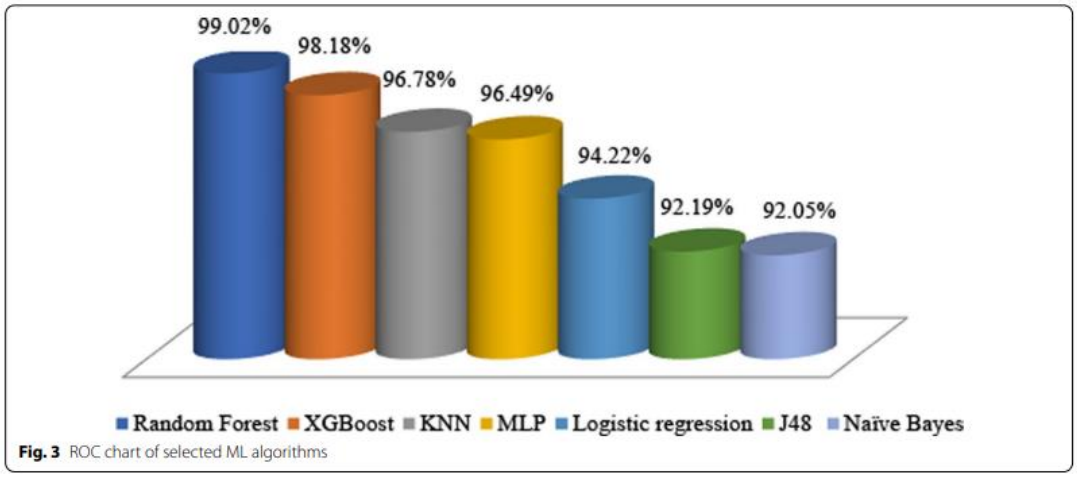

Fig. 3 ROC chart of selected ML algorithms

Fig 1

The study assessed the performance of five ML algorithms in predicting in-hospital mortality. Among these, the Random Forest (RF) algorithm demonstrated the most promising results, achieving superior classification accuracy compared to the other methods. The RF model's ability to effectively process complex interactions among multiple predictors contributed to its success in mortality prediction.

The proposed ML model holds immense potential for real-world implementation. By automating the identification of high-risk patients, healthcare providers can proactively allocate

resources, prioritize treatments, and tailor interventions to individuals who are most likely to face adverse outcomes. This timely intervention can significantly impact patient survival rates and alleviate the strain on hospital resources during the ongoing COVID-19 pandemic.

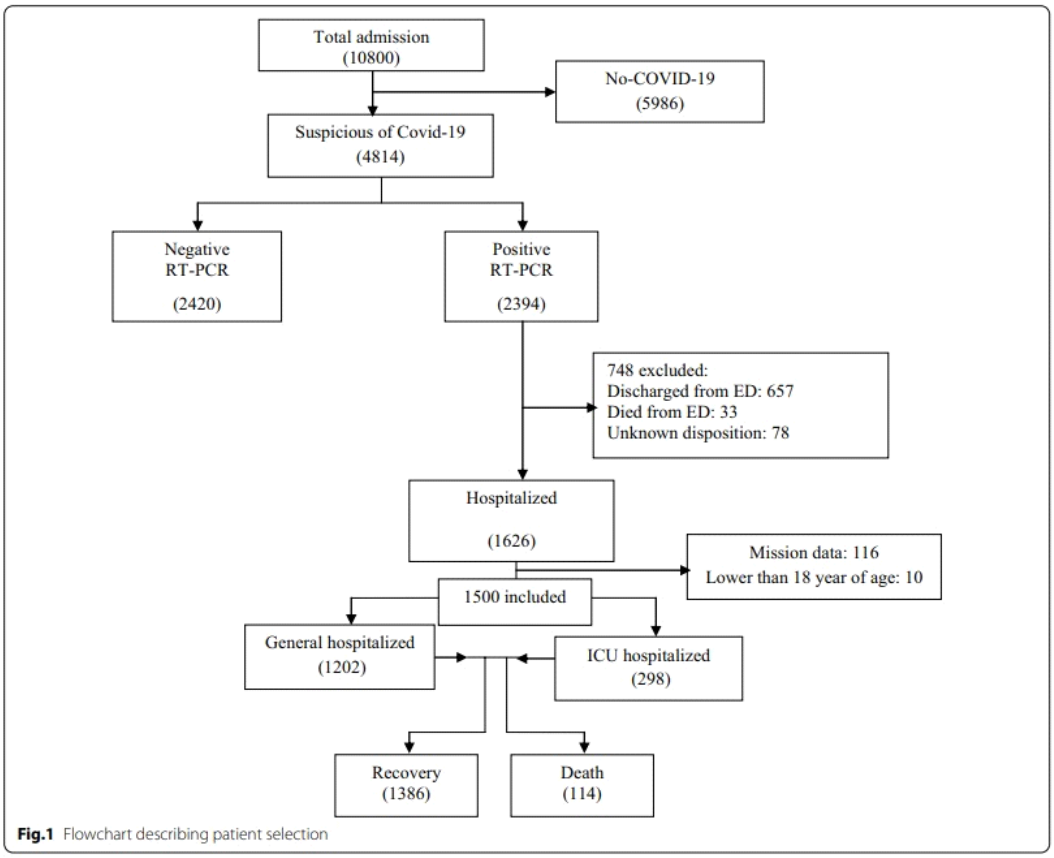

Fig.1 Flowchart describing patient selection

Fig 2

A notable aspect of the study is its emphasis on the value of comprehensive patient databases, such as patient registries, in combination with ML algorithms. The integration of qualitative and quantitative data allows for robust predictions and informed decision-making. The study also highlights the significance of ongoing data collection and refinement to enhance the model's accuracy and generalizability.

Table 3 Identifying the initial list of features affecting mortality in patients with COVID-19

Classes	Number of suggested features	Delphi round		Final features	Included features	Excluded features
		<75%	75%<			
Demographic	9	6	3	3	Gender, age, length of hospitalization	Body mass index, blood group, marital status, ethnicity, place of birth, level of education
Risk factors	10	2	7	7	Smoking, ICU admission, hypertension, pneumonia, diabetes, cardiac disease, another underline disease	Recent travel, exposure type
Clinical manifestations	23	9	14	14	Dyspnea, sore throat, runny nose, loss of taste, loss of smell, contusion, muscular pain, chill, fever, cough, nausea/ vomiting, chest pain and pressure, headache, gastrointestinal symptoms	Weakness, sneezing, exudative pharyngitis, mucus or phlegm, conjunctivitis, hemoptysis, anorexia, dry mouth, decrease consciousness
Laboratory results	24	11	13	13	Blood urea nitrogen, white cell count, C-reactive protein, hypersensitive troponin, glucose, erythrocyte sedimentation rate, creatinine, alkaline phosphatase, aspartate aminotransferase, alanine aminotransferase, absolute lymphocyte count, absolute neutrophil count,	Hematocrit, red cell count, hemoglobin, total bilirubin, thromboplastin time, prothrombin time, albumin calcium, phosphorus, magnesium, sodium, potassium
Therapeutic plan	1	0	1	1	Oxygen therapy	

Looking ahead, the researchers suggest that further advancements can be achieved by expanding the study's scope. This includes testing the model on larger datasets from multiple medical centers to ensure its effectiveness across diverse patient populations. Additionally, exploring alternative classification techniques may unveil additional insights and improve predictive accuracy.

Table 5 Descriptive statistics of the Features

Features (quantitative)	Range	Mean (SD)
Age (year)	18–100	57.25 (17.8)
Leng of hospitalization	1–32	61.89 (13.25)
Creatinine (mg/dL)	0.1–17.9	51.39 (14.4)
White-cell count	1300–63,000	82.34 (4897.4)
Platelet count	108,000–691,000	66.2 (38.1)
Absolute lymphocyte count	2–95	23.74 (11.8)
Absolute neutrophil count	8–98	74.52 (12.3)
Blood urea nitrogen	0.5–251	42.52 (31.7)
Aspartate aminotransferase	3.8–924	44.45 (53.5)
Alanine aminotransferase	2–672	38.29 (41.6)
Glucose	18–994	36.09 (74.2)
Lactate dehydrogenase	4.6–6973	55.68 (339.0)
Prothrombin time	0.9–46.8	42.82 (23.9)
Alkaline phosphatase	9.6–2846	21.12 (39.2)
Erythrocyte sedimentation rate	2–258	40.65 (28.8)
Features (qualitative)	**Values**	**Frequencies**
Gender	Male, Female	836, 664
Cough	Yes, No	736, 764
Contusion	Yes, No	363, 1137
Nausea/vomiting	Yes, No	802, 848
Headache	Yes, No	899, 601
Gastrointestinal symptoms	Yes, No	976, 524
Muscular pain	Yes, No	1021, 479
Chill	Yes, No	878, 622
Fever	Yes, No	728, 772
Pneumonia	Yes, No	1044,456
Oxygen therapy	Yes, No	1053, 447
Dyspnea	Yes, No	1078, 422
Loss of taste	Yes, No	272, 1228
Loss of smell	Yes, No	1195, 305
Runny Noise	Yes, No	637, 863
Sore throat	Yes, No	444, 1056
Other underline disease	Yes, No	360, 1140
Cardiac disease	Yes, No	118, 1382
Hypertension	Yes, No	395, 1105
Diabetes	Yes, No	268, 1232
Smoking	Yes, No	141,1359
Alcohol addiction	Yes, No	26, 1474
C-reactive protein	Positive, Negative	1163, 337
Oxygen therapy	Yes, No	363, 1137
ICU admission	Yes, No	298, 1202
Deceased	Yes, No	114, 1386

In conclusion, this study underscores the potential of machine learning to transform healthcare practices during the COVID-19 pandemic. The developed RF-based mortality prediction model stands as a valuable tool for hospitals and healthcare systems, enabling them to identify high-risk patients promptly and optimize resource allocation, thereby enhancing patient outcomes and contributing to more effective pandemic management.

Feasibility Study:

The feasibility study conducted for this project underscores its viability and relevance in the context of COVID-19 mortality prediction. The technical, economic, and operational aspects have been thoroughly assessed to ascertain the project's feasibility.

Technical Feasibility: With the advancements in machine learning and the availability of diverse open-source libraries, the technical feasibility of the project is high. Various algorithms and tools required for data preprocessing, feature selection, model training, and evaluation are readily accessible. The project's technical feasibility is further reinforced by the substantial volume of COVID-19 patient data available for analysis.

Economic Feasibility: The project's economic feasibility is notable due to its reliance on cost-effective open-source software and readily available hardware. The use of tools such as Python and libraries incurs minimal financial burden. Additionally, the project's potential to enhance healthcare decision-making and patient outcomes adds intrinsic value, strengthening its economic feasibility.

Operational Feasibility: The project's operational feasibility is well-established, given its alignment with the urgent need for predictive tools in the current healthcare landscape. The integration of machine learning models into healthcare decision-making processes is a progressive step toward informed resource allocation and optimized patient care. The project's outcomes have the potential to resonate with healthcare practitioners and administrators seeking data-driven solutions to address the challenges posed by COVID-19.

Methodology/Planning of work:

Research Type: The project will follow an analytical research approach, involving the analysis of existing datasets and the development of predictive models.

Research Unit: The unit of analysis will be individual COVID-19 cases within the dataset.

Data Collection/Analysis Tools: Python programming language will be used for data collection, preprocessing, and analysis. Libraries such as Pandas, Scikit-learn, and TensorFlow will facilitate data manipulation and model development.

Methods: The project will involve data preprocessing, exploratory data analysis, feature selection, model training and evaluation, and cross-validation. Machine learning algorithms including decision trees, random forests, neural networks, and support vector machines will be implemented and compared.

Facilities Required:

Software: A computer with necessary software tools such as Python (with libraries like Pandas, Scikit-learn, and TensorFlow), an integrated development environment (IDE) like Jupyter Notebook or Visual Studio Code, and software for documentation (e.g., Microsoft Word or LaTeX).

Hardware: A computer system with sufficient processing power and memory to accommodate data preprocessing, model training, and analysis tasks.

Expected Outcomes:

The anticipated outcomes of this study are multifaceted and hold the potential to significantly impact both the academic and practical realms of COVID-19 mortality prediction. Firstly, a comprehensive comparative analysis of diverse Machine Learning algorithms will yield insights into their varying predictive performances. This analysis will not only identify the most accurate algorithms but also shed light on the nuances of their strengths and weaknesses. The identification of key features contributing to accurate predictions will contribute to a deeper understanding of mortality determinants. Additionally, the study's emphasis on algorithm suitability will provide tailored recommendations, empowering healthcare practitioners to choose the most effective algorithms for specific patient profiles. Ultimately, the culmination of these outcomes will pave the way for the development of a robust predictive model that can serve as a valuable decision-making tool for healthcare professionals, thereby enhancing the quality of care, resource allocation, and overall management of COVID-19 patients.