Journal of Infrastructure, Policy and Development 2025, 9(1), 9877.
https://doi.org/10.24294/jipd9877

Article

An insight from machine learning perspective for COVID-19 survival prediction in Malaysia based on demographic factor

Nur Fatin Azwin A. Talib[1,†], Siti Meriam Zahari[2,3,*,†], Mahayaudin M. Mansor[2,4,†], Sumayyah Dzulkifly[5,†], Noryanti Nasir[2,†], S. Sarifah Radiah Shariff[2,3,6,†], Nurakmal Ahmad Mustaffa[7,†]

[1] Methodist College Kuala Lumpur, Kuala Lumpur 50470, Malaysia
[2] College of Computing, Informatics and Mathematics, Universiti Teknologi MARA, Shah Alam 40450, Malaysia
[3] RIG Logistics Modelling, Universiti Teknologi MARA, Shah Alam 40450, Malaysia
[4] Institute for Big Data Analytics and Artificial Intelligence, Universiti Teknologi MARA, Shah Alam 40450, Malaysia
[5] Faculty of Computing and Meta-Technology, Sultan Idris Education University, Tanjong Malim 35900, Malaysia
[6] Malaysia Institute of Transport, Universiti Teknologi MARA, Shah Alam 40450, Malaysia
[7] College of Arts and Sciences, Universiti Utara Malaysia, Sintok 06010, Malaysia
*** Corresponding author:** Siti Meriam Zahari, mariam@tmsk.uitm.edu.my
† These authors contributed equally to this work

CITATION

Talib NFAA, Zahari SM, Mansor MM, et al. (2025). An insight from machine learning perspective for COVID-19 survival prediction in Malaysia based on demographic factor. Journal of Infrastructure, Policy and Development. 9(1): 9877. https://doi.org/10.24294/jipd9877

ARTICLE INFO

Received: 25 October 2024
Accepted: 20 November 2024
Available online: 14 January 2025

COPYRIGHT

Copyright © 2025 by author(s). *Journal of Infrastructure, Policy and Development* is published by EnPress Publisher, LLC. This work is licensed under the Creative Commons Attribution (CC BY) license. https://creativecommons.org/licenses/by/4.0/

Abstract: Malaysia reported its first imported COVID-19 case on 23 January 2020, which marked the country's first confirmed positive case. The first case in Malaysia was from eight close contacts in Johor. The global health landscape has been significantly impacted by the COVID-19 pandemic, with mortality or survival being critical outcomes of interest. This study aims to predict COVID-19 survival occurrences in Malaysia by utilizing machine learning approaches based on demographic factors. The dataset used in this study comprises demographic information of 2,151,315 COVID-19 patients, including nationality, regions, age groups, gender, medical history, vaccine brands, and the number of vaccine doses received between 2020 and 2022. Four machine learning algorithms, namely Logistic Regression, Naïve Bayes, Support Vector Machine, and Artificial Neural Network were employed to assess the relationship between demographic factors and COVID-19 survival. To evaluate the model performance, the datasets are categorized into imbalanced and balanced (down-sampling). The results indicate that the balanced dataset (down-sampling) outperforms the imbalanced dataset in terms of overall accuracy, sensitivity, specificity, precision, and Area Under the Curve (AUC). Based on the analysis, the Artificial Neural Network (ANN) classifier exhibited the highest performance with a specificity 95.2% on a balanced dataset. The model excels in accurately identifying survivors, thereby minimizing false mortality predictions and is selected as the best model for predicting COVID-19 survival. Its capacity to process larger sample sizes, combined with numerous interconnected nodes, enables it to identify complex patterns and extract meaningful insights from diverse datasets, such as demographic factors. Additionally, the optimization of parameters, including the number of layers, learning rate, and activation functions, significantly contributed to its superior accuracy. The study identifies that those of chronic diseases, male, and aged 45 and above as the notable factors associated with lower survival rates among COVID-19 patients. The findings underscore the importance of completing the vaccination series by obtaining at least the second dose, as the first dose alone may not offer sufficient protection. In conclusion, this study successfully achieves its objectives by identifying the optimal dataset configuration and predictive model for forecasting COVID-19 survival based on demographic factors. This network could serve as a benchmark model classifier, offering a valuable tool to predict and promote vaccinations, as well as optimize the general healthcare system during the pandemic outbreak. The study not only contributes to the theoretical understanding of effective COVID-19 prediction but also emphasizes the practical implications of integrating advanced machine learning techniques into pandemic management strategies. Future research can build upon these findings by exploring additional machine learning techniques and considering geographical and environmental factors to further enhance the accuracy of long-term predictions.

Keywords: COVID-19; public health; machine learning; vaccination; prediction model; demographic factors

1. Introduction

A serious threat of Coronavirus Disease to public health surfaced in 2019. The unique SARS-CoV-2 epidemic outbreak first occurred in Wuhan City, Hubei Province, China, in December 2019, and since then, it quickly spread to the rest of the world. The World Health Organization defined this illness as Coronavirus Disease (COVID-19). A total of 4.57 million cases of COVID-19 had been reported globally until 3 July 2022, (World Health Organization, 2022). The major way that COVID-19 is spread is through inhalation of contaminated air that contains the virus in the form of droplets, aerosols, and tiny airborne particles (Jayaweera et al., 2020; Stadnytskyi et al., 2020). These particles are exhaled by infected individuals while they breathe, speak, cough, sneeze, or sing. The closer people are to one another, the higher the probability of transmission especially indoors, this illness also can spread across greater distances (Health Ontario, 2022). The World Health Organization defined a cluster of COVID-19 based on confirmed cases or called asymptomatic cases that are in close contact with positive cases or have visited a country that is having an epidemic. Mild symptoms including headaches, muscle soreness, a runny nose, a sore throat, or diarrhea may appear in some infected people. In addition, a person with positive reverse transcription–polymerase chain reaction (RT-PCR) results without any clinical symptoms is also considered an asymptomatic COVID-19 case (Lan et al., 2020).

The COVID-19 pandemic has brought attention to the capability and resilience of healthcare systems (Sagan et al., 2020). The healthcare system has faced immense burdens, including high ICU occupancy rates, shortages of medical staff, and disruptions to routine medical care. Vulnerable populations, such as the elderly and individuals with chronic illnesses, remain at greater risk, emphasizing the need for research to improve survival predictions and optimize healthcare resource allocation. These challenges have resulted in increased demands on both inpatient and outpatient healthcare services (Lal et al., 2022), as well as a rise in healthcare costs (An et al., 2022). However, there has been less emphasis on understanding how the strain on healthcare services during the pandemic has influenced overall performance (Bravata et al., 2021). Even if the healthcare systems are functioning within their capacity limits, the sheer volume of patients can strain their resources and lead to compromises in the quality of care. This is further exacerbated if there is insufficient funding to adequately support the healthcare services or if the healthcare system was not adequately prepared to handle the challenges posed by a pandemic like COVID-19 (Ahmad et al., 2021). Consequently, healthcare systems' capacities have been strained, impeding their ability to effectively provide routine services. It is worth noting that the burden of the COVID-19 pandemic extends beyond healthcare systems and affects general practitioners and all healthcare professionals involved in essential services (Jefferson et al., 2023; Papoutsi et al., 2020; Soares et al., 2021; Schrimpf et al., 2023).

In Malaysia, the threat of COVID-19 became increasingly apparent when neighboring Singapore reported its first imported COVID-19 case from Wuhan, China on 23 January 2020, which was also the first positive case in the republic. From this first case, eight close contacts were identified as being in Johor, Malaysia, (Shah et al., 2020). A study conducted by Zamzuri et al. (2020) on 214 COVID-19 cases in the Seremban district found that the highest mortality occurred during early 2020 among Malaysian individuals between the ages of 41 and 64, who also had a significant number of chronic illnesses as co-morbidities.

To control the spread of infectious diseases, widespread vaccination is a critical tool. Vaccination plays a crucial role in achieving herd immunity, which occurs when a significant portion of the population becomes immune to a particular infectious disease, either through vaccination or prior infection. This collective immunity significantly reduces the disease's spread, offering protection to those who are not immune or unable to receive the vaccine. To prevent the reintroduction of the disease, maintaining high vaccination coverage is essential. While no vaccine is 100\% effective, with some recipients not developing full protection and others experiencing diminishing immunity over time, the overall benefits of vaccination are clear. For example, studies such as those by Huang et al. (2020) and He et al. (2023) have investigated the effects of vaccines on specific populations, like diabetes patients, revealing nuanced impacts. Furthermore, research by Almufty et al. (2021) has explored the relationship between vaccines and blood clotting, contributing to our understanding of vaccine safety and efficacy. It is important to recognize that certain individuals, such as those with immune suppression, may not be eligible for vaccination, underscoring the need for widespread immunity within the community. In this context, the findings of Arifin et al. (2020), which suggest that vaccination is crucial in reducing mortality risk, become particularly relevant. Despite a small minority not achieving full protection, the collective effect of vaccination significantly contributes to community immunity, protecting vulnerable individuals and advancing the broader public health goal of controlling infectious diseases. In Malaysia, the epidemic response exemplifies a dynamic and adaptive strategy, evolving from strict containment measures to vaccination-led mitigation and eventual endemic management. This transition was facilitated by strong government coordination, widespread public compliance, and the effective use of digital tools such as MySejahtera, which collectively addressed challenges and helped minimize the pandemic's impact.

In response to the detrimental consequences of COVID-19, particularly the mortality outcomes, this study aims to investigate the risk factors associated with COVID-19 mortality in Malaysia. Zamzuri et al. (2020) utilized the Chi-square test to compare the sociodemographic characteristics of COVID-19 patients. However, the study only considered 214 COVID-19 cases in the Seremban district during the early months of 2020. In this study, we employ machine learning models such as Logistic Regression, Naïve Bayes, Support Vector Machine, and Artificial Neural Network to determine the mortality occurrence among COVID-19 patients in Malaysia. We evaluate and compare the performance of these models using an original imbalanced dataset and a balanced dataset achieved through downsampling. The aim is to identify key variables most strongly associated with COVID-19 survival in Malaysia.

This study is significant as it identifies key demographic factors influencing COVID-19 survival, including nationality, regional distribution, age groups, gender, medical history, vaccine brands, and the number of vaccine doses. By employing a machine learning model with the highest predictive accuracy, it provides actionable insights to enhance public health strategies. These insights include tailoring vaccination campaigns to prioritize vulnerable populations such as older adults and those with chronic conditions, designing triage protocols to allocate critical resources like ICU beds and ventilators to patients with the highest risk of mortality, and developing targeted policies to ensure equitable distribution of healthcare resources across regions with varying infection rates and healthcare capacities.

This paper is structured as follows: Section II provides a review of related works on COVID-19 mortality and its prediction. Section III discusses the statistical models in data mining and machine learning employed in this study, along with a description of the dataset used for the variables and the evaluation criteria for the models. Section IV presents the results and discussion derived from the exploratory data analysis and modelling. Finally, Section V concludes the paper.

2. Related works

In 2020, the World Health Organization recognized COVID-19 as a pandemic because people could easily get infected through airborne transmission and high mortality rates. Studies indicate that over 85% of individuals with the disease are asymptomatic or have minor symptoms, while only 15% experience serious illness (10% with a case fatality rate of 15%) or critical conditions (5% with a case fatality rate of 50%). Globally, more than 3,750,000 confirmed cases and over 250,000 deaths reported across approximately 200 countries, territories, and areas (Eurosurveillance Editorial Team, 2020). The pandemic resulted in millions of deaths worldwide and has significantly increase the overall mortality rate. The mortality rate is influenced by various risk factors, and the impact varies across different regions. Zahid and Perna (2021) examine the highest number of cases, with Africa having the lowest number of COVID-19 cases. Mortality rates differ among the countries such as North America, Europe, South America, Africa, Oceania, and Asia, with the percentage of mortality rate generally below 4% based on daily cases ranging from 0 to 16,000 total cases. The mortality rate percentage has implications for both the health conditions of the citizen and the economy of the country. Furthermore, in Italy, a significant number of elderly COVID-19 patients die at home, which pose a challenge in accurately examining the actual risk factors of COVID-19 (Bhatraju et al., 2020).

Furthermore, a study analyzing a sample of 25,935 individuals who died after being infected with COVID-19 revealed that most deaths (69.9%) occurred among individuals who had not received at least one dose of the vaccine. Partially vaccinated individuals accounted for 22.5% of the deaths, while a smaller proportion (7.5%) of fully vaccinated individuals experienced mortality (Arifin et al., 2021). These findings highlight the association between vaccination status and mortality rate, with unvaccinated individuals contributing the highest number of deaths.

Machine learning techniques have played a significant role in disease prediction and analysis, including in the context of COVID-19. Despite being a ‘the first of its kind’ disease, considerable research has been conducted on COVID-19 prediction using machine learning. Machine learning algorithms analyze large datasets and provide multiple potential solutions. However, their outputs must be carefully interpreted to avoid errors and misapplications in decision-making, particularly when predicting COVID-19 survival. Prior studies highlight the critical need to balance interpretability and accuracy in forecasting COVID-19 outcomes. For example, Pourhomayoun and Shakibi (2021) emphasized the importance of interpretability in supporting medical decision-making, while Yan et al. (2020) employed interpretable models to identify biomarkers predictive of mortality. These findings underscore the necessity for transparency and reliability in healthcare models.

Ongoing contributions in this field have led to the development and utilization of various machine learning algorithms, such as Naive Bayes, Support Vector Machine (SVM), Logistic Regression, and Artificial Neural Network, for predicting and diagnosing diseases. These models have been employed to forecast the global impacts and trends of the COVID-19 outbreak.

The Support Vector Machine (SVM) is a powerful machine learning method used for prediction, classification, and regression tasks, particularly for time-series data. SVM exhibits excellent generalization capabilities and is well-suited for handling limited data (Khan et al., 2019). It employs kernel functions to transform data from a low-dimensional space to a high-dimensional space, enabling the creation of hyperplanes that separate classes in higher dimensions (Ivanciuc, 2007). While online algorithms are commonly employed to track time-varying changes and time-lagging characteristics in system modelling (Rehman, 2021), the kernel function plays a crucial role in determining the training period for data computation. The four most used kernel types are linear, radial basis, polynomial, and sigmoid. A linear kernel is suitable for datasets that can be separated linearly, while radial basis kernels are effective for circularly separable distributions.

When utilizing machine learning for outcome prediction, Naive Bayes is a straightforward and reliable technique. Many studies aim to select the best hypothesis (h) based on the available data (d). Naive Bayes is built upon Bayes’ Theorem, which provides a mechanism to determine the likelihood of a hypothesis given our prior information. The probability P(h|d) represents the likelihood of the hypothesis (h) being true, while P(d|h) represents the likelihood of the data (d) being true given the hypothesis. Additionally, P(h) represents the prior probability of hypothesis h, and P(d) represents the prior probability of data, d. By combining P(h) with P(d) and P(d|h), the posterior probability P(h|d) can be computed, which allows us to predict outcomes based on new data. Naive Bayes also provides a way to assess model uncertainty by considering the likelihoods of different outcomes, making it useful for both predictive and diagnostic tasks (Medhekar et al., 2013).

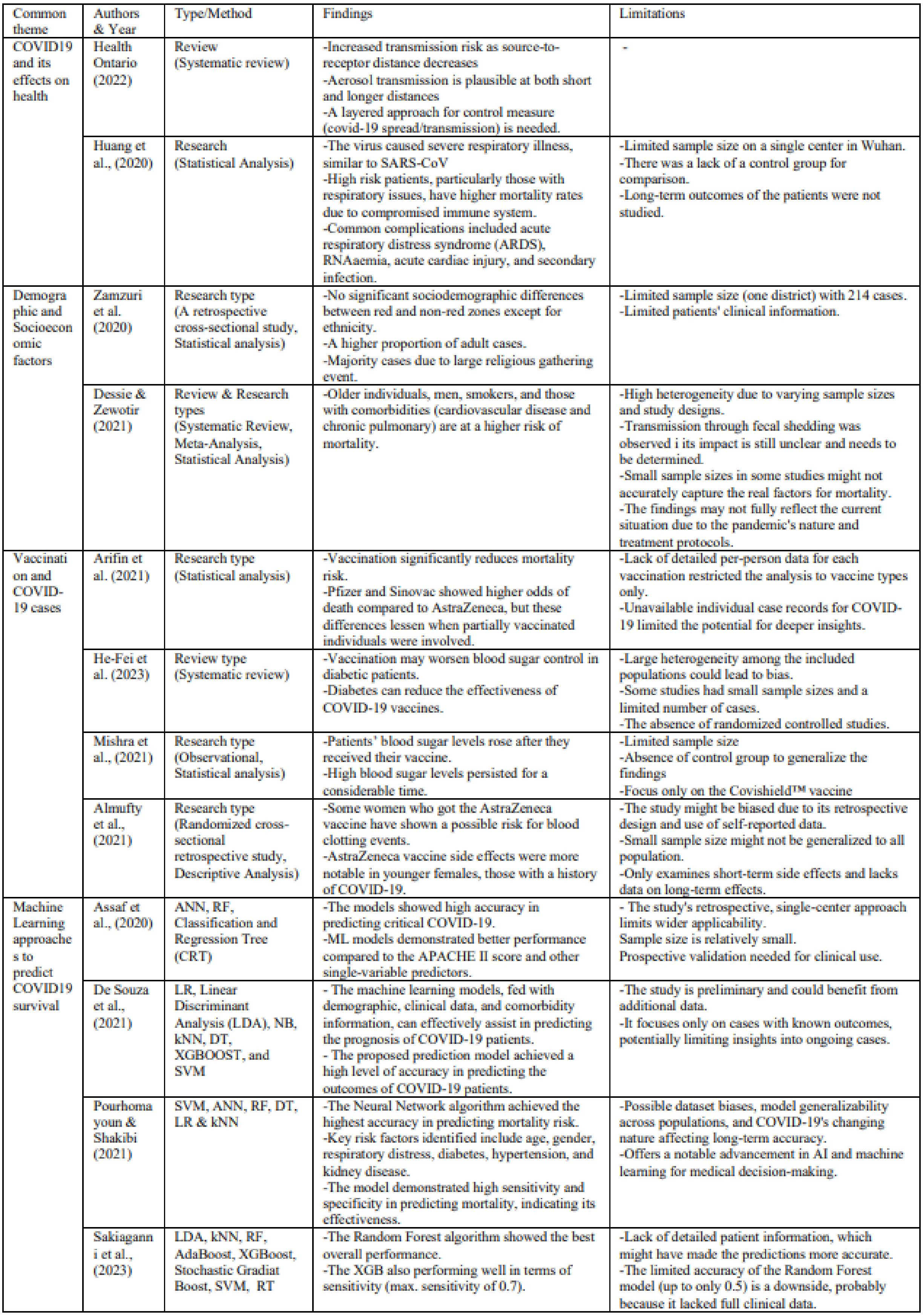
On the other hand, when the dependent variable (target) is categorical, a statistical approach and machine learning algorithm called Logistic Regression is frequently employed to solve classification difficulties. Pierre Francois Verhulst defined three parameters and the curve that passed through them in a paper that was published in the Proceedings of the Belgian Royal Academy, describing the logistic function and its characteristics (Delmas, 2004). Logistic Regression is simple and commonly used. The maximum likelihood estimation process is used to forecast binary classes using a statistical model. In Logistic Regression, the dependent variable has a Bernoulli distribution. The sigmoid or logistic function yields values between 0 and 1 and assumes a "S" form. The predicted value will be 1, however if the curve approaches negative infinity, the anticipated value will be 0.

In a study by Yan et al. (2020), blood samples from 404 patients in Wuhan, China, were analyzed to discover disease-predictive biomarkers. The authors proposed a COVID-19 mortality prediction model based on artificial intelligence techniques. Artificial Neural Networks, Random Forest, Decision Tree, Logistic Regression, and K-Nearest Neighbor (KNN) algorithms were employed, and the model successfully classified deep features extracted from chest X-ray images of COVID-19 and Pneumonia patients using ResNet152. The model achieved an accuracy of 0.973 on Random Forest and 0.977 on XGBoost predictive classifiers (Kumar et al., 2020). Another machine learning-based model has been developed to predict mortality due to COVID-19 by identifying individuals with a higher sensitivity based on their unique genetic and physiological traits (Alimadadi et al., 2020). Assaf et al. (2020) developed a COVID-19 risk prediction model utilizing machine learning techniques such as Artificial Neural Networks, random forests, and regression trees, and they explored various interpretable model properties.

In an attempt to forecast severe COVID-19 symptoms, Zhu et al. (2021) employed machine learning strategies, including a combined regression and classification algorithm. Their research achieved a prediction accuracy of 76.97% with a correlation coefficient of 0.524. Furthermore, Roland et al. (2020) utilized similar classification and regression techniques to analyze social media data and gather real-time information on COVID-19 symptoms and demographic data in order to improve prediction accuracy.

Prakash (2020) provided a brief overview of different machine learning techniques applied to COVID-19 datasets, focusing on determining the vulnerability of different age groups. The study compared the performance of eight alternative algorithms and found that Random Forest achieved the highest accuracy rate of 96% for COVID-19 prediction. De Souza et al. (2021) conducted a study using supervised machine learning methods such as Logistic Regression, linear discriminant analysis, Naive Bayes, k-nearest neighbors, decision trees, XGBOOST, and Support Vector Machine to identify patients at risk of experiencing severe COVID-19 symptoms early. The techniques were trained using a database from various social media platforms, including basic details of individuals such as gender, age range, symptoms, comorbidities, and recent travel history. The study reported an ROC area under curve (AUC) of 0.92, sensitivity of 0.88, and specificity of 0.82 for predicting disease outcomes.

Pourhomayoun and Shakibi (2021) proposed an Artificial Intelligent (AI) model to assist hospitals and healthcare facilities in determining priority patients managing patient overload, and reducing wait times for necessary care. The study utilized machine learning techniques such as Support Vector Machine, Artificial Neural Network, Random Forest, Decision Tree, Logistic Regression, and K-Nearest Neighbor to predict the mortality rate of COVID-19 patients. They analyzed a dataset of over 2,670,000 COVID-19 patients from 146 countries, including 307,382 labeled samples, and achieved an overall mortality rate prediction accuracy of 89.98%. Logistic Regression analysis revealed a significant association between mortality and a history of chronic diseases such as Hypertension (HTN) and Diabetes Mellitus (DM). Mollalo et al. (2020) highlighted the use of Artificial Neural Networks (ANNs) in simulating complex non-linear connections in structural epidemiology. The methods have been used in several disciplines, including epidemiology and public health (Kiang et al., 2006; Kawka et al., 2021; Mollalo et al., 2018), agriculture (Abdipour et al., 2019), finance (Bae, 2012), and environmental science (Marohasy and Abbot, 2015). Recent studies focusing on machine learning algorithms for predicting survival in the context of the COVID-19 pandemic include Sharma et al. (2022), Andonov et al. (2023), and Sakagianni et al. (2023). A brief summary of the literature review from the main studies is presented in **Figure 1**.



Common theme	Authors & Year	Type/Method	Findings	Limitations
COVID19 and its effects on health	Health Ontario (2022)	Review (Systematic review)	-Increased transmission risk as source-to-receptor distance decreases -Aerosol transmission is plausible at both short and longer distances -A layered approach for control measure (covid-19 spread/transmission) is needed.	-
	Huang et al., (2020)	Research (Statistical Analysis)	-The virus caused severe respiratory illness, similar to SARS-CoV -High risk patients, particularly those with respiratory issues, have higher mortality rates due to compromised immune system. -Common complications included acute respiratory distress syndrome (ARDS), RNAaemia, acute cardiac injury, and secondary infection.	-Limited sample size on a single center in Wuhan. -There was a lack of a control group for comparison. -Long-term outcomes of the patients were not studied.
Demographic and Socioeconomic factors	Zamzuri et al. (2020)	Research type (A retrospective cross-sectional study, Statistical analysis)	-No significant sociodemographic differences between red and non-red zones except for ethnicity. -A higher proportion of adult cases. -Majority cases due to large religious gathering event.	-Limited sample size (one district) with 214 cases. -Limited patients' clinical information.
	Dessie & Zewotir (2021)	Review & Research types (Systematic Review, Meta-Analysis, Statistical Analysis)	-Older individuals, men, smokers, and those with comorbidities (cardiovascular disease and chronic pulmonary) are at a higher risk of mortality.	-High heterogeneity due to varying sample sizes and study designs. -Transmission through fecal shedding was observed i its impact is still unclear and needs to be determined. -Small sample sizes in some studies might not accurately capture the real factors for mortality. -The findings may not fully reflect the current situation due to the pandemic's nature and treatment protocols.
Vaccination and COVID-19 cases	Arifin et al. (2021)	Research type (Statistical analysis)	-Vaccination significantly reduces mortality risk. -Pfizer and Sinovac showed higher odds of death compared to AstraZeneca, but these differences lessen when partially vaccinated individuals were involved.	-Lack of detailed per-person data for each vaccination restricted the analysis to vaccine types only. -Unavailable individual case records for COVID-19 limited the potential for deeper insights.
	He-Fei et al. (2023)	Review type (Systematic review)	-Vaccination may worsen blood sugar control in diabetic patients. -Diabetes can reduce the effectiveness of COVID-19 vaccines.	-Large heterogeneity among the included populations could lead to bias. -Some studies had small sample sizes and a limited number of cases. -The absence of randomized controlled studies.
	Mishra et al., (2021)	Research type (Observational, Statistical analysis)	-Patients' blood sugar levels rose after they received their vaccine. -High blood sugar levels persisted for a considerable time.	-Limited sample size -Absence of control group to generalize the findings -Focus only on the Covishield™ vaccine
	Almufty et al., (2021)	Research type (Randomized cross-sectional retrospective study, Descriptive Analysis)	-Some women who got the AstraZeneca vaccine have shown a possible risk for blood clotting events. -AstraZeneca vaccine side effects were more notable in younger females, those with a history of COVID-19.	-The study might be biased due to its retrospective design and use of self-reported data. -Small sample size might not be generalized to all population. -Only examines short-term side effects and lacks data on long-term effects.
Machine Learning approaches to predict COVID19 survival	Assaf et al., (2020)	ANN, RF, Classification and Regression Tree (CRT)	-The models showed high accuracy in predicting critical COVID-19. -ML models demonstrated better performance compared to the APACHE II score and other single-variable predictors.	- The study's retrospective, single-center approach limits wider applicability. Sample size is relatively small. Prospective validation needed for clinical use.
	De Souza et al., (2021)	LR, Linear Discriminant Analysis (LDA), NB, kNN, DT, XGBOOST, and SVM	- The machine learning models, fed with demographic, clinical data, and comorbidity information, can effectively assist in predicting the prognosis of COVID-19 patients. - The proposed prediction model achieved a high level of accuracy in predicting the outcomes of COVID-19 patients.	-The study is preliminary and could benefit from additional data. -It focuses only on cases with known outcomes, potentially limiting insights into ongoing cases.
	Pourhomayoun & Shakibi (2021)	SVM, ANN, RF, DT, LR & kNN	-The Neural Network algorithm achieved the highest accuracy in predicting mortality risk. -Key risk factors identified include age, gender, respiratory distress, diabetes, hypertension, and kidney disease. -The model demonstrated high sensitivity and specificity in predicting mortality, indicating its effectiveness.	-Possible dataset biases, model generalizability across populations, and COVID-19's changing nature affecting long-term accuracy. -Offers a notable advancement in AI and machine learning for medical decision-making.
	Sakiagianni et al., (2023)	LDA, kNN, RF, AdaBoost, XGBoost, Stochastic Gradiat Boost, SVM, RT	-The Random Forest algorithm showed the best overall performance. -The XGB also performing well in terms of sensitivity (max. sensitivity of 0.7).	-Lack of detailed patient information, which might have made the predictions more accurate. -The limited accuracy of the Random Forest model (up to only 0.5) is a downside, probably because it lacked full clinical data.

Figure 1. Brief summary of main findings and limitations based on common themes in COVID-19 cases.

3. Methodology

In this study, the secondary dataset is obtained from the Ministry of Health (MOH) through the GitHub website. The dataset is publicly shared and can be accessed at the following link: https://github.com/MoHMalaysia/covid19public/tree/main/epidemic/linelist.

The dataset contains information on COVID-19 patients in Malaysia, including their demographics and medical histories. The data used in the analysis represents the daily confirmed cases of COVID-19 patients in Malaysia from 2020 to 2022. There are a total of 2,151,315 observations in the dataset, which represent all the recorded COVID-19 cases during that period. The dataset used in the study consists of 12 variables that capture the demographic profile of COVID-19 patients. These variables include age, gender, region, comorbidities, nationality, and several other variables denoted as vaccine brands. Pfizer, AstraZeneca, Sinovac, no vaccine, and others are the four categories used to classify vaccine brands. All vaccine brands are analyzed for each of their respective doses, including dose 1, dose 2, and dose 3 (booster). The `Others' category in this study includes unverified vaccine brands as well as vaccines from Cansino, Moderna, Sinopharm, and Janssen. Patients who have not received any vaccinations are categorized under the `no-vaccine' category. The research flowchart is presented in **Figure 2**. In this study, all states in Malaysia have been classified into three regions: East Coast, West Coast, and East Malaysia. The West Coast region covers the states of Johor, Melaka, Negeri Sembilan, Selangor, Wilayah Persekutuan Kuala Lumpur, Wilayah Persekutuan Putrajaya, Perak, Perlis, Kedah, and Pulau Pinang. The East Coast region includes Kelantan, Terengganu, and Pahang. East Malaysia comprises Sabah, Sarawak, and Labuan. Patients with chronic diseases are categorized into Group 1 for the comorbidity's variable, while those without any recorded chronic diseases are placed in Group 0. The target variable, survival, indicates whether the patient has survived due to COVID-19, where 1 represents survival and 0 represents death. **Table 1** provides a summarized overview of the dataset, highlighting the variables, their descriptions, and the level of measurement for each attribute.

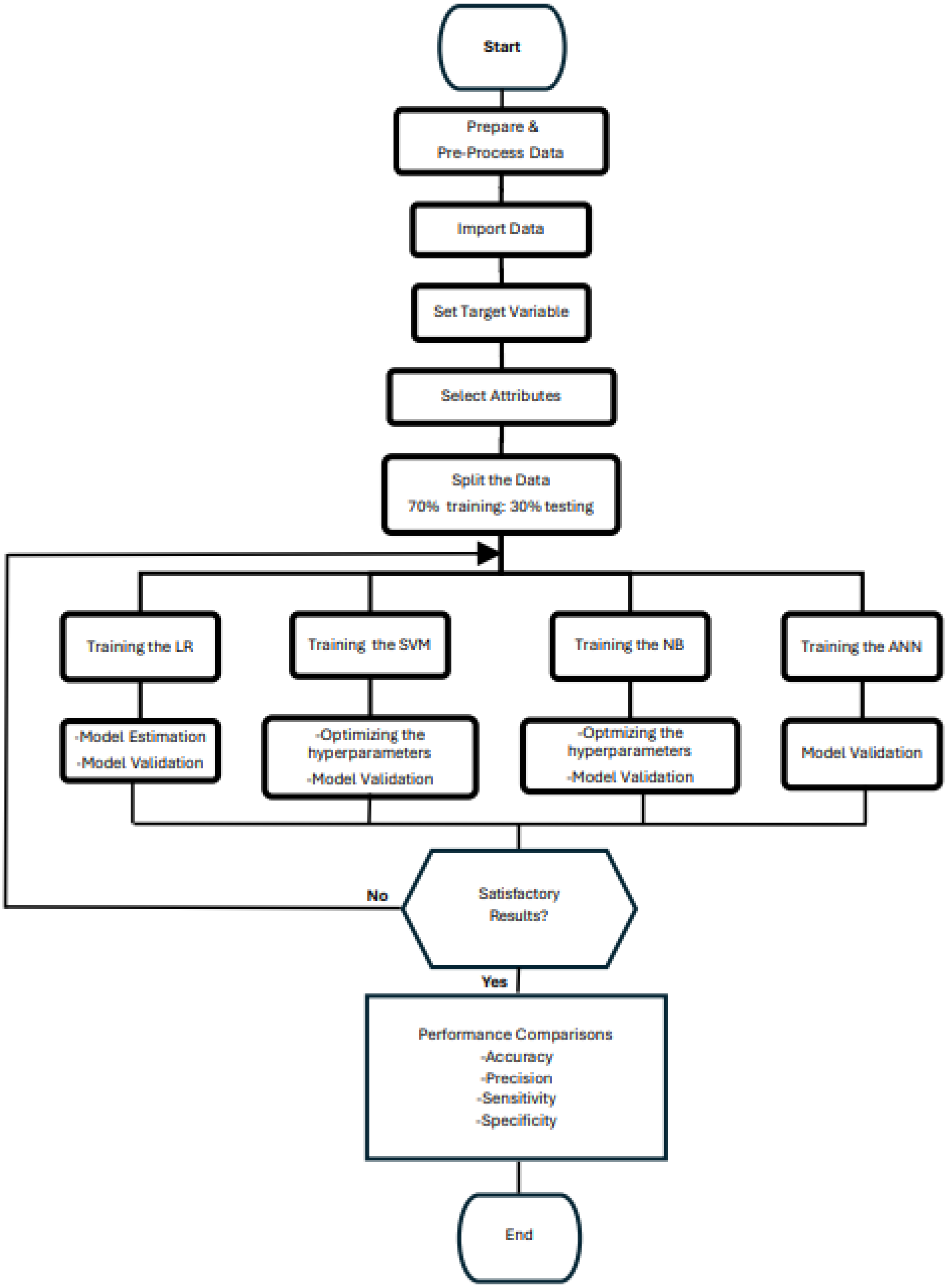


Figure 2. Research flowchart.

In this study, the analysis was done using both balanced (down-sampling) and imbalanced (original) to select the best of dataset and make a decision of best model classifiers: Logistic Regression, Naïve Bayes, Support Vector Machine and, Artificial Neural Network. The dataset was split into two parts, with 70% allocated to the training set and 30% to the validation set. Data analysis was conducted using RapidMiner Studio version 10.3.1, which is available at https://docs.rapidminer.com/latest/studio/.

Table 1. Description of variables.

Variable	Description	Scale of measurement
Age	The age of COVID-19 patient is divided into 5 age groups	Nominal
Gender	Of COVID-19 patient, 1 denotes Male and 0 for female	Nominal
Nationality	Nationality of COVID-19 patient, 1 denoted Malaysian, and 0 denotes non-Malaysian	Nominal
Commodities	COVID-19 patient with chronic disease is denotes as 1, and 0 for without chronic disease	Nominal
Brand vaccine 1	First dose of vaccine received by the COVID-19 patients	Nominal
Brand vaccine 2	Second dose of vaccine received by the COVID-19 patients	Nominal
Brand vaccine 3	Booster dose of vaccine received by the COVID-19 patients	Nominal
Region	Location of COVID-19 patients infected by the virus based on three regions	Nominal
Survival	The survival of COVID-19 patients: 1 denotes survival, and 0 denotes dead	Binary

3.1. Modelling

3.1.1. Logistic Regression

Logistic Regression (LR) is a type of classification algorithm that is used when the dependent variable is in binary format. Logistic Regression evaluates the relationship between the dependent variable (binary target) and the independent variables by estimating probabilities. Next, the attribute weights for each independent variable (attributes) are identified. The odds ratio produces the likelihood of the outcome when there is a constant effect on a predictor. A Logistic Regression can be written as:

$$P_i = \frac{1}{1 + e^{-z_i}} \tag{1}$$

where P_i represents the probability of survival, and z is the linear combinations of the predictor variables, are the model parameters.

$$z_i = \beta_0 + \beta_1 X_{1_i} + \beta_2 X_{2_i} + \ \dots \ + \beta_k X_{k_i}$$

where:

$$X_1 = \begin{cases} 1, Dose1Az \\ 0, Otherwise \end{cases} \quad X_2 = \begin{cases} 1, Dose1Sz \\ 0, Otherwise \end{cases} \quad X_3 = \begin{cases} 1, Dose1Pf \\ 0, Otherwise \end{cases}$$

$$X_4 = \begin{cases} 1, Dose1Ot \\ 0, Otherwise \end{cases} \quad X_5 = \begin{cases} 1, Dose2Az \\ 0, Otherwise \end{cases} \quad X_6 = \begin{cases} 1, Dose2Az \\ 0, Otherwise \end{cases}$$

$$X_7 = \begin{cases} 1, Dose2Pf \\ 0, Otherwise \end{cases} \quad X_8 = \begin{cases} 1, Dose2Ot \\ 0, Otherwise \end{cases} \quad X_9 = \begin{cases} 1, BoostAz \\ 0, Otherwise \end{cases}$$

$$X_{10} = \begin{cases} 1, BoostSz \\ 0, Otherwise \end{cases} \quad X_{11} = \begin{cases} 1, BoostPf \\ 0, Otherwise \end{cases} \quad X_{12} = \begin{cases} 1, BoostOt \\ 0, Otherwise \end{cases}$$

$$X_{13} = \begin{cases} 1, Children \\ 0, Otherwise \end{cases} \quad X_{14} = \begin{cases} 1, YoungA \\ 0, Otherwise \end{cases} \quad X_{15} = \begin{cases} 1, Middle \\ 0, Otherwise \end{cases}$$

$$X_{16} = \begin{cases} 1, OldA \\ 0, Otherwise \end{cases} \quad X_{17} = \begin{cases} 1, Male \\ 0, Otherwise \end{cases} \quad X_{18} = \begin{cases} 1, Comorbid \\ 0, Otherwise \end{cases}$$

$$X_{19} = \begin{cases} 1, Malaysian \\ 0, Otherwise \end{cases} \quad X_{20} = \begin{cases} 1, WestCoast \\ 0, Otherwise \end{cases} \quad X_{21} = \begin{cases} 1, EastCoast \\ 0, Otherwise \end{cases}$$

3.1.2. Support Vector Machine

Next, this study used a Support Vector Machine (SVM) algorithm to identify a hyperplane that clearly divides the data points into different classes in an *N*-dimensional space (*N* is the number of features). The hyperplane can be represented as:

$$w \times x + b = 0 \tag{3}$$

where *w* is the weight vector, *x* is the feature vector, and *b* is the bias. Multiple potential hyperplanes might be selected to divide the two classes of data points. The goal is to identify a plane with the largest margin, or the greatest separation between data points from both classes. To improve the classification accuracy of upcoming data points, the margin distance should be maximized. Support vectors are the data points nearest to the dividing line and they determine where and how this line is positioned. These points satisfy the conditions:

$$\mathrm{w} \times x_p + b \geq 1$$

$$\mathrm{w} \times x_n + b \leq -1$$

where x_p are support vectors from the positive class (Category A) and x_n from the negative class (Category B). SVM was originally designed for numerical variables, but it can also automatically convert nominal data to numerical and normalize the input data before use. *A* and *B* here represent two categories in a target variable (dependent variable). SVM is utilized to segregate data points and determine the hyperplane for the target variable, coded as 0 for 'survived' and 1 for 'not survived', based on independent variables. The *C* parameter, denoting the margin, signals to the SVM optimizer the level of priority given to avoiding misclassification. A high *C* value results in choosing a hyperplane with a smaller margin, focusing on precise classification of training examples. Conversely, a lower *C* value prompts the selection of a hyperplane with a larger margin, accepting more misclassifications to ensure broader class separation. This balance, achieved by adjusting the *C* parameter, is instrumental in reducing misclassification while improving overall model efficacy. This method is effective for both linear and non-linear datasets, as it emphasizes clear differentiation between classes. The ideal hyperplane, marked by the maximum margin, is the one farthest from the nearest data points on either side. In RapidMiner, setting the *C*-value to zero in the SVM operator's parameters is a tactic to minimize misclassification.

3.1.3. Naïve Bayes

This study also employed The Naïve Bayes (NB) algorithm as it is the most popular and simple machine learning which is based on Bayes theorem:

$$\mathrm{P(Z|X_i, \dots, X_{21})} = \frac{\mathrm{P(X_i, \dots, X_{21}|Z)Pz}}{\mathrm{P(X_i, \dots, X_{21})}} \tag{3}$$

where $P(Z|X_i, \dots, X_{21})$ represents the posterior probability of the target variable, where *z* represents the class of interest (survival), and are the predictor variables. $P(z)$ corresponds to the prior probability of survival which reflects the initial probability without considering any predictor information. $P(X_i, \dots, X_{21}|Z)$ represents the likelihood, indicating the probability of observing the predictor variables given the survival outcome. $P(X_i, \dots, X_{21})$ represents the prior probability of the predictor variables, indicating the overall probability of observing the specific combination of predictor values. In this study, the theorem will be calculated automatically using the RapidMiner software. The Naïve Bayes operator in RapidMiner makes it easier to conduct calculations with a large number of observations in the dataset.

3.1.4. Artificial Neural Network

Artificial Neural Network (ANN) model is also considered in this study. ANN is a field of research that seeks to mimic and learn from the functioning of the human brain. The network consists of interconnected nodes, similar to neurons. However, the effectiveness of an ANN depends on the type and number of neurons it contains. A smaller number of neurons typically leads to higher performance of the system. In essence, a Neural Network algorithm aims to create a function that maps input data to the desired output. The concept of Artificial Neural Networks is inspired by how biological neural systems process information and learn to generate knowledge. The key aspect of this concept is the development of new structures for information processing. **Figure 3** illustrates the architecture of the Artificial Neural Network.

The neurons in a ANN are connected closely and organized into layers, including the input layer, hidden layer(s), and output layer. The input layer receives the data, and the output layer generates the results. Each connection between neurons has a connection weight, and each neuron has a threshold value and an activation function. The connection weights determine the significance or influence of each input on the neuron's output. The weights can be positive or negative, affecting the strength of the signal transmitted through the connection. The output of a neuron is determined by the activation function applied to the weighted sum of its inputs, also known as the summing unit. The equation for a neuron's output can be represented as:

$$O_j = f\left(\sum_i \left(W_{ij} \times X_i\right) + b_j\right) \tag{4}$$

where (O_j) is the output of the j-th neuron, (W_{ij}) are the weights, (X_i) are the input values, and (b_j) is the bias. The activation function, $f()$, used is the Sigmoid function, as in (Equation (1)). For prediction and classification tasks, feed-forward neural networks trained using the Back Propagation algorithm, known as multilayer perceptron, are used. In this algorithm, each input feature, (X_i), is multiplied by its associated weight, (W_{ij}), and the weighted features are summed up along with a bias term, (b_j). During the training process, the output of the weighted sum is passed through the activation function, to obtain the final output of the neural network. The difference between the predicted output and the desired output is then used to calculate the error. The error is then propagated back through the network, and the weights and biases are adjusted using gradient descent to minimize the error. By iteratively adjusting the weights and biases using the backpropagation of errors, the neural network can learn to better recognize patterns and make accurate predictions.

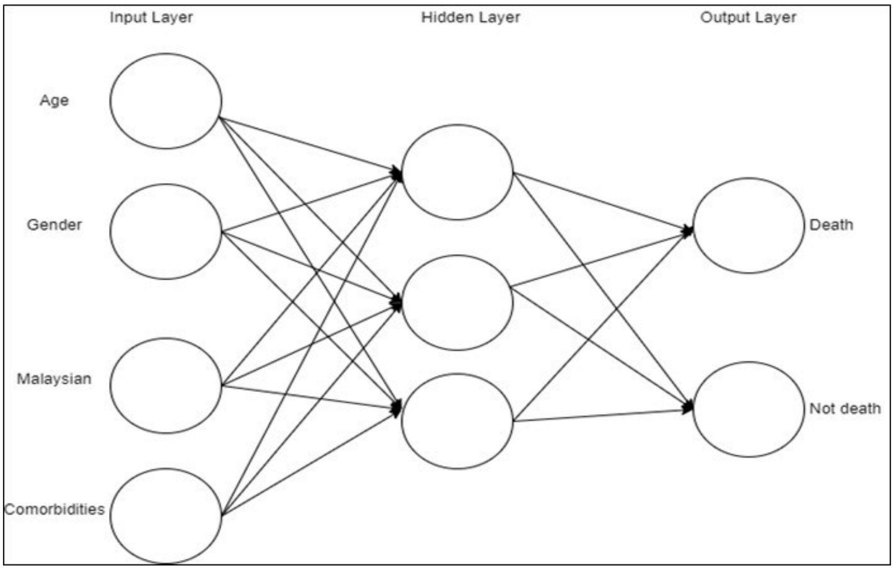


Figure 3. Artificial Neural Network architecture.

3.2. Model evaluation

This stage of model evaluation aims to assess the best model for predicting patient survival based on overall performance metrics.

Table 2 represents the confusion matrix, which is used to evaluate the performance of a machine learning model based on categorization. The confusion matrix provides a summarized table of accurate and inaccurate predictions made by the classifier. The accuracy value represents the percentage of correctly classified instances. In the confusion matrix, TP refers to True Positive, representing the correct classification of survival (Survival = 1, Dead = 0); TN stands for True Negative, representing the correct classification of non-survival; FP represents False Positive, indicating incorrect classifications; and FN denotes False Negative, indicating incorrect classifications. In addition to accuracy, this study will assess other performance metrics such as specificity, sensitivity, and precision. The formulas for these performance metrics are provided in **Table 3**. The performance of all predictive models, including LR, NB, SVM and ANN, will be discussed in the next section.

Table 2. Confusion matrix.

		Actual	
		Survival = 1	**Dead = 0**
Predicted	Survival = 1	True Positive (TP)	False Positive (FP)

	Dead = 0	False Negative (FN)	True Negative (TN)
	Recall	Sensitivity	Specificity

Table 3. Performance measurement.

Performance	**Formulae**
Overall accuracy	$\frac{\sum_{i=1}^{n} TP_i + TN_i}{\sum_{i=1}^{n} TP_i + TN_i + FP_i + FN_i}$
Precision	$\frac{TP}{\sum TP_i + FP_i}$
Sensitivity	$\frac{TP_i}{\sum TP_i + FN_i}$
Specificity	$\frac{TN}{\sum TN_i + FP_i}$

4. Results and discussion

4.1. Exploratory data analysis

Through exploratory data analysis, this study discovered that out of 2,151,315 COVID-19 patients, 2,114,483 have successfully recovered, whereas 36,831 individuals have passed away. This significant disparity between the number of deceased and surviving patients highlights the existing class imbalance. To address this issue, two different types of experiments were conducted in this study. The first experiment examined the impact of imbalanced data, while the second experiment involved applying a down-sampling method to achieve a balanced dataset. The results obtained from both experiments were utilized to select a model with high-quality performance metrics, including accuracy, sensitivity, specificity, and other relevant measures. **Table 4** provides a summary of patients based on the doses of the vaccine they have received, including the first dose, second dose, and third dose (booster). The data indicates that the majority of patients received Pfizer as their first dose of the vaccine, accounting for 43% of the total. Similarly, for the second dose, Pfizer was also the most common brand, with 41.3% of patients receiving it. Sinovac was the second-highest administered brand for both the first and second doses, with 30.6% and 30.3% of patients receiving it, respectively. AstraZeneca and other vaccines were administered to a lower percentage of patients. As for the booster dose, approximately 41% of patients received Pfizer, while the majority, 47.8%, did not receive a booster dose.

Table 4. Descriptive statistics of vaccination brands based on frequency.

Type of vaccine	**First dose**	**Second dose**	**Booster dose**
Pfizer	924281 (43%)	886380 (41.2%)	882580 (41%)
AstraZeneca	232822 (10.8%)	231583 (10.8%)	155684 (7.2%)
Sinovac	658062 (30.6%)	652881 (30.3%)	80963 (3%)
Others	60330 (0.3%)	3769 (0.2%)	2537 (1%)

4.2. Modelling results

In this section, the performance of different supervised machine learning models implemented in the study, namely LR, NB, SVM and ANN, will be discussed. These classifier algorithms were utilized to examine the relationship between demographic factors and the incidence of mortality related to COVID-19. By utilizing these classifier algorithms, the study aimed to gain insights into the demographic factors that contribute to COVID-19 mortality and provide a better understanding of the disease's impact on different population groups. The evaluation of model performance will provide valuable information on the predictive capabilities of these algorithms in identifying individuals at higher risk of mortality, aiding in the development of targeted interventions and strategies for disease management.

4.2.1. Logistic Regression

In addition to addressing the objective, further discussions will focus on exploring the association between demographic factors and the survival due to COVID-19 in Malaysia. **Table 5** shows the estimated model of Logistic Regression.

- The odds of survival due to COVID-19 are 0.64 times lower in the West Coast compared to East Malaysia, while the East Coast shows no significant difference in the odds of survival compared to East Malaysia.
- For patients that have comorbidities, the odds of survival due to COVID-19 is 0.03 times lower than the odds of patients with no comorbidities.
- The odds of survival due to COVID-19 for male patients are 0.66 times lower than for female patients with COVID-19.
- For Malaysians, the odds of survival are 2.39 times higher than for non-Malaysians
- MiddleA refers to COVID-19 patients between the ages of 31 and 45, while OldA includes patients older than 45 years old. Both MiddleA and OldA exhibit odds of survival that are 0.001 and 0.01 lower, respectively, compared to infants with COVID-19.
- Young adults have odds of survival 0.06 lower than infants with COVID-19, whereas children have odds of survival 2.01 higher than infants with COVID-19.
- Among the different doses of vaccines received, only Dose1others, which represents patients who receive other brand doses of vaccine, shows a significant increase in the odds of survival. For Dose1Az (first dose of AstraZeneca), the odds of survival due to COVID-19 are 0.27 times lower than the odds of no vaccine. For Dose1Pf (first dose of Pfizer), the odds of survival are 0.94 times lower than the odds of no vaccine. Similarly, for Dose1Sv (first dose of Sinovac), the odds of survival are 0.5 times lower than the odds of no vaccine
- However, the results differ for the second dose of the vaccine. For patients who have received the AstraZeneca brand vaccine, the odds of survival are 172 times higher than the odds of no vaccine. Sinovac shows the second highest odds of survival, with a 124 times higher rate, followed by Pfizer with 59 times higher, and other brands with a 32 times higher rate compared to no vaccine.
- For patients who received a third dose (booster), the odds of survival also increase. Pfizer, Sinovac, and AstraZeneca show similar increases, with odds of survival approximately 13–14 times higher. Other brands demonstrate a 4 times higher rate of survival compared to no vaccine.

Table 5. Estimated model of Logistic Regression.

Variable	Coef.	*P*-value	Conclusion	Odd ratio
Dose1Pf	−0.06	0.695	Not significant	0.94
Dose1Sv	−1.90	0.000	Significant	0.50
Dose1others	2.58	0.000	Significant	13.2
Dose1Az	−1.32	0.003	Significant	0.27
Dose2Pf	4.07	0.000	Not significant	58.56
Dose2Sv	4.82	0.000	Significant	123.97
Dose2Az	5.15	0.000	Significant	172.43
Dose2others	3.48	0.000	Significant	32.46
Dose3Pf	2.64	0.000	Significant	14.01
Dose3Sv	2.56	0.000	Significant	12.94
Dose3	2.53	0.000	Significant	12.55
Dose3others	1.47	0.063	Not significant	4.35
Comorb	−3.68	0.000	Significant	0.03
OldA	−6.63	0.000	Significant	0.001
MiddleA	−4.61	0.000	Significant	0.01
YoungA	−2.86	0.000	Significant	0.06
Child	0.70	0.001	Significant	2.01
Gender	−0.42	0.000	Significant	0.66
Nationality	0.87	0.000	Significant	2.39
WestCoast	−0.44	0.000	Significant	0.64
EastCoast	0.02	0.801	Not significant	1.02
Intercept	3.80	0.000	Significant	44.70

4.2.2. Performances of Logistic Regression, Naïve Bayes, Support Vector Machine and Artificial Neural Network

Table 6 displays the performance of each model for both balanced and imbalanced (in brackets) datasets. Using the imbalanced dataset, the Artificial Neural Network (ANN) achieved the highest accuracy of 98.95%, followed by Logistic Regression (LR) and Support Vector Machine (SVM) with equal accuracies of 96.9%, and Naïve Bayes (NB) with an accuracy of 86.62%. While accuracy is a widely used measure, it may not be the most suitable metric for imbalanced datasets. Therefore, this study addressed the imbalanced problem by employing down-sampling techniques. For the balanced dataset, the performance based on accuracy shows that ANN attained the highest accuracy of 94.9%, followed by LR with 94.57%, SVM with 92.25%, and NB with 88.77%. In addition to accuracy, sensitivity is a crucial measure for predicting COVID-19-related survival (the positive class). In the case of the imbalanced dataset, ANN demonstrated the highest ability to predict survival with 99.7%, followed by LR and SVM with equal sensitivities of 96.56%, and NB with a sensitivity of 86.53%. For the balanced dataset, SVM exhibited the highest sensitivity of 96.56%, followed by LR with 94.91%, ANN with 94.56%, and NB with 83.17%. On the other hand, specificity measures the ability of a model to accurately predict the negative class, which is COVID-19 survival. In the imbalanced dataset, most models achieved a specificity of around 99%, with NB having the highest specificity of 99.84%, followed by LR and SVM with 99.52%, and ANN with 99.24%. In the case of the balanced dataset, ANN demonstrated the highest specificity of 95.21%, followed by LR with 94.26%, NB with 93.66%, and SVM with 89.33%.

Figures 4–11 show the AUC curve for all models. It measures the overall performance of a model across all possible classification thresholds. The red line in the ROC curve represents the false positive rate (1-specificity) on the X-axis and the true positive rate (sensitivity) on the Y-axis. It demonstrates the trade-off between these two rates as the classification threshold varies. On the other hand, the blue line represents the false positive rate (1-specificity) on the X-axis and the threshold values on the Y-axis. It illustrates the relationship between the false positive rate and the threshold values. The transparent regions around the mean values, plotted with a solid line, represent the standard deviation regions. These regions help visualize the variability or uncertainty associated with the performance metrics at different threshold values. The closer is the red line to the top-left corner of the plot indicates a model with higher sensitivity and lower false positive rate, indicating better performance. The AUC results indicate that the balanced dataset outperformed the imbalanced dataset in terms of distinguishing between mortality and survival instances for LR and SVM, while showing equal performance for ANN and NB. Both ANN and LR achieved the best performance with an AUC of 98.6%.

After executing all predictive models and evaluating their performance on the imbalanced dataset and the balanced dataset (down-sampling), the results have helped the researchers in determining the best configuration for each algorithm. This optimal configuration aims to yield the highest possible results in addressing the objectives of this study, which involve identifying the best dataset and selecting the best predictive model for predicting the survival due to COVID-19.

Table 6. Performance measures of LR, NB, SVM, and ANN on imbalanced (in brackets) and balanced datasets.

Model/performance	LR (%)	NB (%)	SVM (%)	ANN (%)
Accuracy	94.57 [96.94]	88.77 [86.62]	92.25 [96.94]	94.90 [98.95]
Rank	2 [2]	4 [3]	3 [2]	1 [1]
Sensitivity	94.91 [97.36]	83.17 [86.53]	96.56 [97.36]	94.56 [99.70]
Rank	2 [2]	4 [3]	1 [2]	3 [1]
Specificity	94.22 [72.92]	94.37 [92.27]	88.47 [72.92]	95.24 [56.00]
Rank	3 [2]	2 [1]	4 [2]	1 [3]
Precision	94.26 [99.52]	93.66 [99.84]	89.33 [99.52]	95.21 [99.24]
Rank	2 [2]	3 [1]	4 [2]	1 [3]
AUC	98.6 [97.30]	95.80 [95.80]	98.40 [97.30]	98.60 [98.60]
Rank	1	3	2	1

	[2]	[3]	[2]	[1]

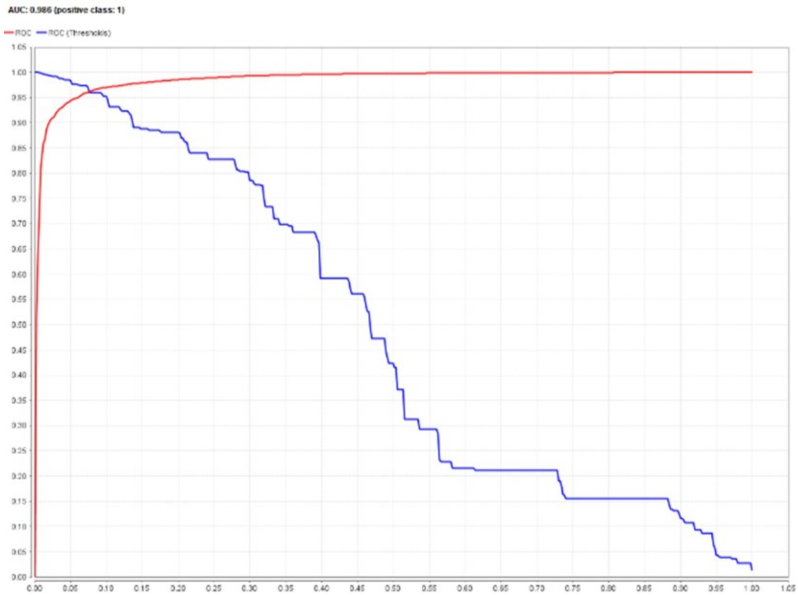


Figure 4. AUC 97.3% Logistic Regression (imbalanced).

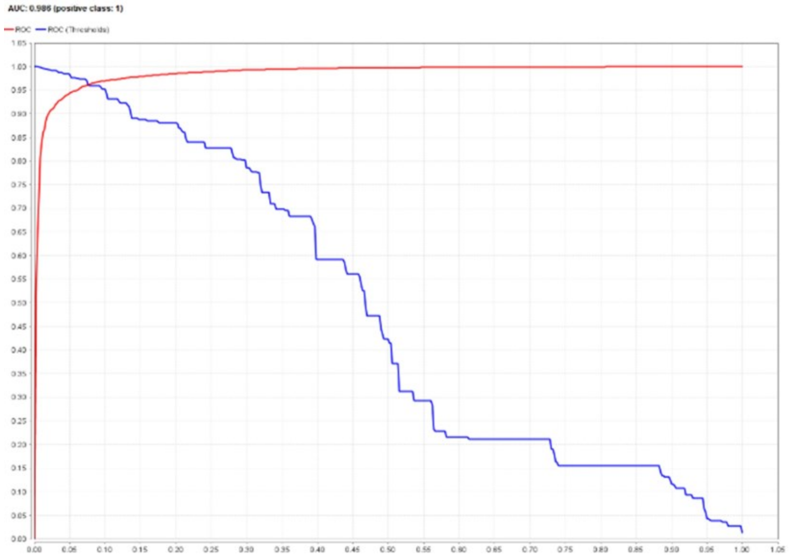


Figure 5. AUC 98.6% Logistic Regression (balanced).

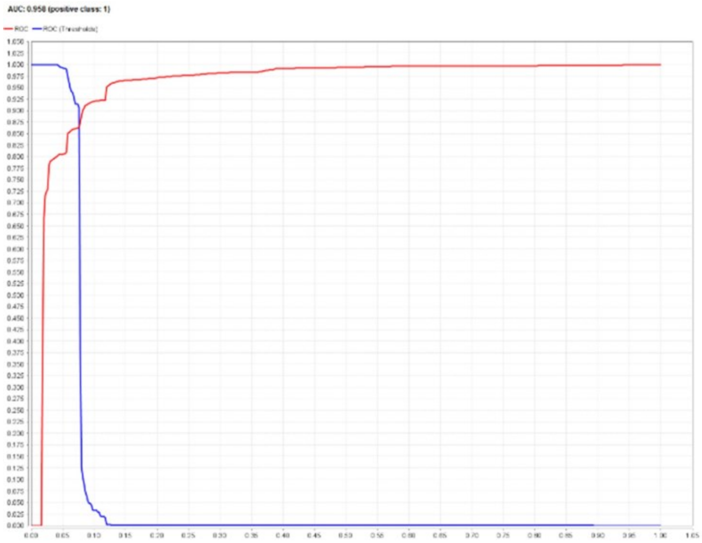


Figure 6. AUC 95.8% Naïve Bayes (imbalanced).

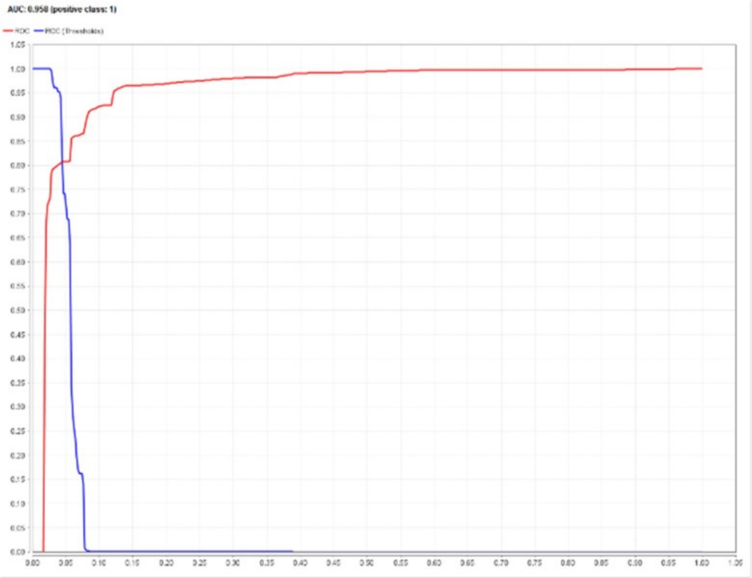


Figure 7. AUC 95.8% Naïve Bayes (balanced).

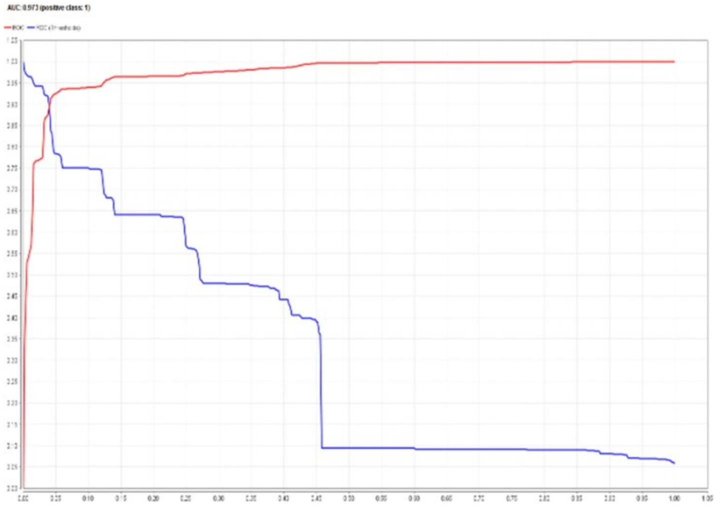


Figure 8. AUC 97.3% Support Vector Machine (imbalanced).

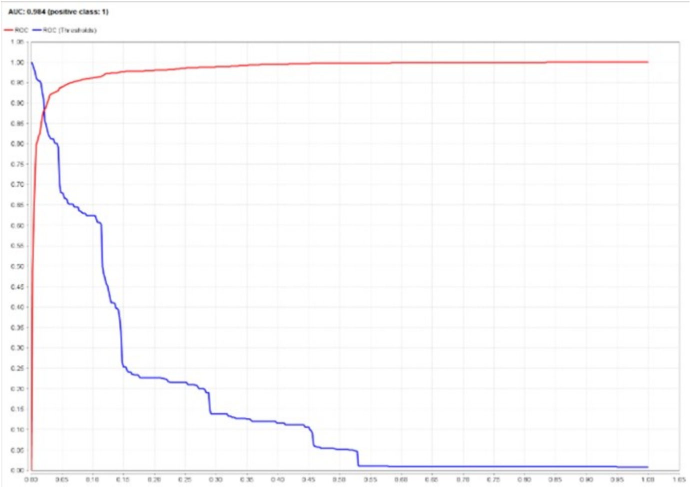


Figure 9. AUC 98.4% Support Vector Machine (balanced).

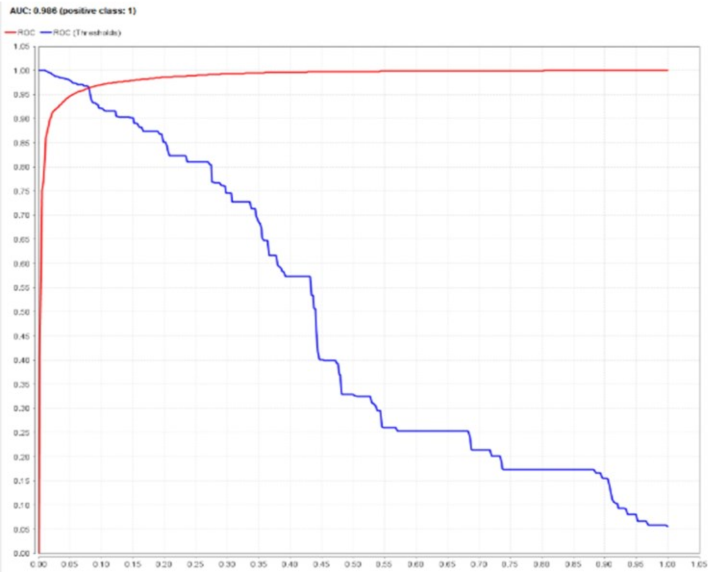


Figure 10. AUC 98.6% Artificial Neural Network (imbalanced).

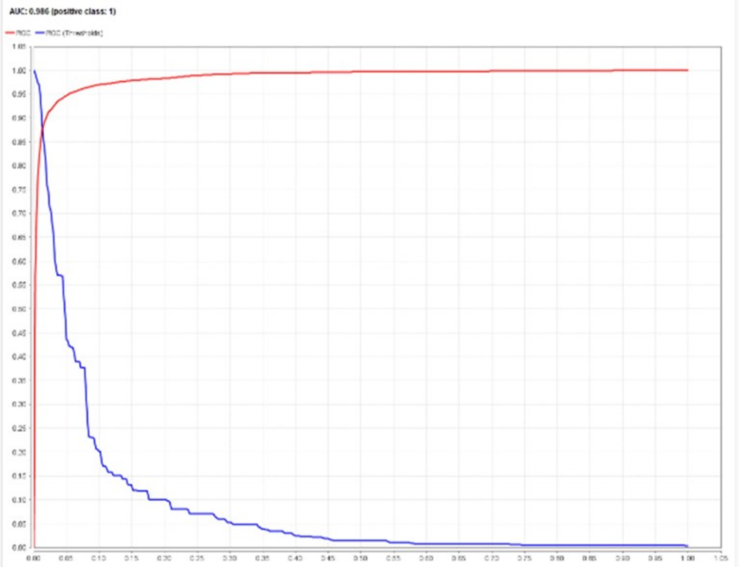


Figure 11. AUC 98.6% Artificial Neural Network (balanced).

5. Conclusion

The COVID-19 pandemic has reached an endemic stage as of 24 January 2023. This means that the virus is still spreading and infecting people, but the severity of the illness is significantly reduced. This shift is attributed to the high vaccination coverage, with approximately 84.2% of the global population having received at least two doses of the COVID-19 vaccine as of 6 October 2022. In Malaysia, the daily confirmed cases have plateaued since 1 January 2023, with an average of 460 cases over a seven-day period as of 11 January 2023. Hospital admissions for ventilated COVID-19 patients and COVID-19 patients in intensive care have also decreased since the beginning of the previous month. In conclusion, this study has successfully addressed all of its objectives. The first objective determined that the balanced dataset (down-sampling) yielded the best performance for the machine learning models compared to the imbalanced dataset (original dataset). The second objective identified several factors associated with the likelihood of survival due to COVID-19. We found that the odds of survival from COVID-19 are significantly lower in the West Coast of Malaysia, particularly in Selangor and Kuala Lumpur, compared to East Malaysia. On the other hand, the East Coast shows no significant difference in survival odds compared to East Malaysia. This discrepancy is believed to be attributed to the high population density in the West Coast, leading to a faster spread of COVID-19 and a higher number of reported cases. Malaysians exhibit significantly higher odds of survival compared to non-Malaysians, with the most affected non-Malaysians being those who migrated to work in various sectors within the community. In line with Arifin et al. (2021), our findings support the notion that the different doses of vaccines have varying impacts on the odds of survival. Some doses demonstrate higher odds of survival compared to no vaccine, while others show lower odds.

Our results suggest that receiving the second dose of the vaccine is associated with improved survival rates. The analysis underscores the importance of completing the vaccination series by obtaining at least the second dose, as the first dose alone may not provide sufficient protection. This underscores the critical role of appropriate vaccine doses in enhancing the chances of survival during the COVID-19 pandemic and is instrumental in achieving herd immunity. This immunity reduces the disease’s spread, safeguarding individuals who cannot be vaccinated or are not yet immune, which leads to more people survive from COVID-19.

In this study, we employed four different models; Artificial Neural Network (ANN), Logistic Regression (LR), Naïve Bayes (NB), and Support Vector Machine (SVM) for predicting COVID-19 survival, each with its own advantages and limitations. Logistic Regression was chosen for its simplicity and interpretability, especially in providing probabilities and odds ratios for survival outcomes. It effectively assessed relationships between binary survival outcomes and key demographic predictors such as age, gender, vaccination doses, and comorbidities. However, its linear assumptions can limit its ability to model non-linear relationships, making it less suitable for capturing the complexities in the data when compared to more advanced models like ANN. Naïve Bayes was utilized for its efficiency in handling categorical data, such as vaccine types and dose counts. This model was adept at calculating survival probabilities based on independent demographic variables. However, it faces challenges with imbalanced datasets, and its assumption of feature independence may not always hold in real-world data, potentially affecting its performance. SVMs were included for their strong capability in modeling complex relationships between predictors, such as interactions between comorbidities and vaccination history. They have shown effective performance in many classification tasks. However, the computational intensity of SVMs, particularly when applied to large datasets, is a notable limitation, which can make them less efficient compared to ANN. The third objective achieved the selection of ANN as the best model classifier for predicting COVID-19 survival, with high-quality performance metrics such as accuracy (94.90%), sensitivity (94.56%) specificity (95.24%), precision (95.21%), and AUC (98.65%). The ability of ANN to capture complex, non-linear relationships was crucial, particularly in understanding the influence of factors like comorbidities and age on survival outcomes. Based on the ANN model, the most influential variables on survival patients of COVID-19 measured by using features weight revealed that comorbid and age are the most influential factors for COVID-19 survival. Our findings show that patients with comorbidities have significantly lower odds of survival compared to patients with no comorbidities. In contrast to newborns, people with COVID-19 who are 31 years of age or older have lower odds of survival. In comparison to female patients with COVID-19, male patients have significantly lower odds of survival. These findings align with those of Ahmad et al. (2021), Dessie and Zewotir (2021) and Zamzuri et al. (2020).

The findings from this study offer valuable insights for health, economic, and social policy development, with actionable implications for healthcare systems. Policymakers can leverage these insights to design targeted vaccination campaigns, protect vulnerable populations, and strengthen healthcare infrastructure. Integrating these findings with machine learning models further enhances decision-making, particularly in resource-limited settings. Models such as LR, ANN and SVM highlight key demographic factors, enabling healthcare providers to prioritize high-risk patients like the elderly or those with comorbidities. By predicting mortality risks, these tools guide early interventions, optimize ICU bed allocation, and inform vaccination strategies in underserved areas. These combined efforts not only improve patient outcomes but also align with Malaysia’s broader development goals of building a resilient society and economy prepared for future public health challenges.

This study has made significant contributions to understanding the global regional distribution and spread of COVID-19, underscoring the need for focused attention and care in managing the pandemic in densely populated areas. Our findings suggest that future research should explore the impact of geographic factors like humidity and temperature on COVID-19 transmission, both within urban settings and across different countries. Additionally, this research highlights the necessity for more detailed studies on various machine learning models tailored to specific national contexts, aiming to improve the accuracy of long-term predictive models.

However, it is important to recognize the challenges encountered during this research, particularly in data processing due to the sheer volume of the dataset, which exceeds one million entries. This complexity was further compounded by the initial incompleteness of patient records, although the minimal proportion of missing values compared to the overall sample size allowed us to proceed with discarding incomplete datasets without significantly impacting the study's robustness. Despite these hurdles, the study's findings pave the way for enhanced preparedness against potential future pandemics, taking into account the unpredictable nature of such events.

Author contributions: N.F.A.T. and S.M.Z., performed all the analyses; and S.M.Z. wrote the initial draft of the manuscript with input; M.M.M. and S.S.R.S verified the interpretation of the analyses; N.N and N.A.M assisted with the manuscript writing, formatting and coordinating comments from co-authors and S.D. provided comments on the manuscript from its initial drafts to the final version. All authors reviewed the manuscript.

Acknowledgment: We thank the Ministry of Health, Malaysia for granting access to the data for the study in the public domain, and we acknowledge the College of Computing, Informatics and Mathematics, Universiti Teknologi MARA for their support.

Data availability: The original datasets used in the current study can be obtained by following this link: https://github.com/MoH-Malaysia/covid19-public. Additionally, the datasets and analysis results are also available upon reasonable request from the corresponding author.

Conflict of interest: The authors declare no conflict of interest

References

Abdipour, M., Younessi-Hmazekhanlu, M., Ramazani, S. H. R., et al. (2019). Artificial Neural Networks and multiple linear regression as potential methods for modeling seed yield of safflower (Carthamus tinctorius L.). Industrial Crops and Products, 127, 185–194.

Ahmad, W. M. A. W., et al. (2021). COVID-19: A scenario of Malaysian mortality. International Medical Journal, 28.

Alimadadi, A., et al. (2020). Artificial intelligence and machine learning to fight COVID-19.

Almufty, H. B., Mohammed, S. A., Abdullah, A. M., & Merza, M. A. (2021). Potential adverse effects of COVID-19 vaccines among Iraqi population: A comparison between the three available vaccines in Iraq; A retrospective cross-sectional study. Diabetes & Metabolic Syndrome: Clinical Research & Reviews, 15, 102207.

An, X., et al. (2022). Economic burden of public health care and hospitalization associated with COVID-19 in China. Public Health, 203, 65–74.

Andonov, D., et al. (2023). Impact of the COVID-19 pandemic on the performance of machine learning algorithms for predicting perioperative mortality. BMC Medical Informatics and Decision Making, 23, 67.

Arifin, W. N., et al. (2021). A brief analysis of the COVID-19 death data in Malaysia. medRxiv.

Assaf, D., et al. (2020). Utilization of machine-learning models to accurately predict the risk for critical COVID-19. Internal and Emergency Medicine, 15, 1435–1443.

Bae, J. K. (2012). Predicting financial distress of the South Korean manufacturing industries. Expert Systems with Applications, 39, 9159–9165.

Bhatraju, P. K., et al. (2020). COVID-19 in critically ill patients in the Seattle region—Case series. New England Journal of Medicine, 382, 2012–2022.

Bravata, D. M., et al. (2021). Association of intensive care unit patient load and demand with mortality rates in US Department of Veterans Affairs hospitals during the COVID-19 pandemic. JAMA Network Open, 4, e2034266.

De Souza, F. S. H., Hojo-Souza, N. S., Dos Santos, E. B., Da Silva, C. M., & Guidoni, D. L. (2021). Predicting the disease outcome in COVID-19 positive patients through machine learning: A retrospective cohort study with Brazilian data. Frontiers in Artificial Intelligence, 4, 579931.

Delmas, B. (2004). Pierre-François Verhulst et la loi logistique de la population. Mathématiques et Sciences Humaines. Mathématiques et Sciences Sociales.

Dessie, Z. G., & Zewotir, T. (2021). Mortality-related risk factors of COVID-19: A systematic review and meta-analysis of 42 studies and 423,117 patients. BMC Infectious Diseases, 21, 855.

European Centre for Disease Prevention and Control, et al. (2020). Latest updates on COVID-19 from the European Centre for Disease Prevention and Control. Eurosurveillance, 25, 2002131.

Figueredo, A. J., & Wolf, P. S. A. (2009). Assortative pairing and life history strategy – A cross-cultural study. Human Nature, 20, 317–330. https://doi.org/10.1007/s12110-009-9068-2.

He, Y.-F., et al. (2023). Correlation between COVID-19 vaccination and diabetes mellitus: A systematic review. World Journal of Diabetes, 14, 892–918.

Health Ontario. (2022). COVID-19 transmission through short and long-range respiratory particles.

Huang, C., et al. (2020). Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. The Lancet, 395, 497–506.

Ivanciuc, O., et al. (2007). Applications of support vector machines in chemistry. Reviews in Computational Chemistry, 23, 291.

Jayaweera, M., Perera, H., Gunawardana, B., & Manatunge, J. (2020). Transmission of COVID-19 virus by droplets and aerosols: A critical review on the unresolved dichotomy. Environmental Research, 188, 109819.

Jefferson, L., Heathcote, C., & Bloor, K. (2023). General practitioner well-being during the COVID-19 pandemic: A qualitative interview study. BMJ Open, 13, e061531.

Kawka, M., Dawidziuk, A., Jiao, L. R., & Gall, T. M. (2022). Artificial intelligence in the detection, characterisation, and prediction of hepatocellular carcinoma: A narrative review. Translational Gastroenterology and Hepatology, 7.

Khan, S. A., et al. (2019). Lungs nodule detection framework from computed tomography images using support vector machine. Microscopy Research and Technique, 82, 1256–1266.

Kiang, R., et al. (2006). Meteorological, environmental remote sensing and neural network analysis of the epidemiology of malaria transmission in Thailand. Geospatial Health, 1, 71–84.

Kumar, R., et al. (2020). Accurate prediction of COVID-19 using chest X-ray images through deep feature learning model with SMOTE and machine learning classifiers. medRxiv.

Lal, A., Lim, C., Almeida, G., & Fitzgerald, J. (2022). Minimizing COVID-19 disruption: Ensuring the supply of essential health products for health emergencies and routine health services. The Lancet Regional Health, 6.

Lan, L., et al. (2020). Positive RT-PCR test results in patients recovered from COVID-19. JAMA, 323, 1502–1503.

Marohasy, J., & Abbot, J. (2015). Assessing the quality of eight different maximum temperature time series as inputs when using Artificial Neural Networks to forecast monthly rainfall at Cape Otway, Australia. Atmospheric Research, 166, 141–149.

Medhekar, D. S., Bote, M. P., & Deshmukh, S. D. (2013). Heart disease prediction system using naive Bayes. International Journal of Enhanced Research in Science Technology & Engineering, 2.

Mollalo, A., Rivera, K. M., & Vahedi, B. (2020). Artificial Neural Network modeling of novel coronavirus (COVID-19) incidence rates across the continental United States. International Journal of Environmental Research and Public Health, 17, 4204.

Papoutsi, E., Giannakoulis, V. G., Ntella, V., Pappa, S., & Katsaounou, P. (2020). Global burden of COVID-19 pandemic on healthcare workers.

Pourhomayoun, M., & Shakibi, M. (2021). Predicting mortality risk in patients with COVID-19 using machine learning to help medical decision-making. Smart Health, 20, 100178.

Prakash, K. B., Imambi, S. S., Ismail, M., Kumar, T. P., & Pawan, Y. (2020). Analysis, prediction, and evaluation of COVID-19 datasets using machine learning algorithms. International Journal, 8, 2199–2204.

Rehman, A. (2021). Light microscopic iris classification using ensemble multi-class support vector machine. Microscopy Research and Technique, 84, 982–991.

Roland, L. T., Gurrola, J. G., Loftus, P. A., Cheung, S. W., & Chang, J. L. (2020). Smell and taste symptom-based predictive model for COVID-19 diagnosis. International Forum of Allergy & Rhinology, 10, 832–838.

Sagan, A., et al. (2020). COVID-19 and health systems resilience: Lessons going forwards. Eurohealth, 26, 20–24.

Sakagianni, A., et al. (2023). Prediction of COVID-19 mortality in the intensive care unit using machine learning. Caring Is Sharing–Exploiting the Value in Data for Health and Innovation, 536.

Schrimpf, A., Bleckwenn, M., & Braesigk, A. (2023). COVID-19 continues to burden general practitioners: Impact on workload, provision of care, and intention to leave. Healthcare, 11, 320.

Schröer, C., Kruse, F., & Gómez, J. M. (2021). A systematic literature review on applying CRISP-DM process model. Procedia Computer Science, 181, 526–534.

Shah, A. U. M., et al. (2020). COVID-19 outbreak in Malaysia: Actions taken by the Malaysian government. International Journal of Infectious Diseases, 97, 108–116.

Sharma, S., Alsmadi, I., Alkhawaldeh, R. S., & Al-Ahmad, B. (2022). Data-driven analysis and predictive modeling on COVID-19. Concurrency and Computation: Practice and Experience, 34, e7390.

Shipe, M. E., Deppen, S. A., Farjah, F., & Grogan, E. L. (2019). Developing prediction models for clinical use using Logistic Regression: An overview. Journal of Thoracic Disease, 11(Suppl 4), S574.

Soares, A., Thakker, P., Deych, E., Jain, S., & Bhayani, R. K. (2021). The impact of COVID-19 on dual-physician couples: A disproportionate burden on women physicians. Journal of Women's Health, 30, 665–671.

Stadnytskyi, V., Bax, C. E., Bax, A., & Anfinrud, P. (2020). The airborne lifetime of small speech droplets and their potential importance in SARS-CoV-2 transmission. Proceedings of the National Academy of Sciences, 117, 11875–11877.

World Health Organization, et al. (2022). COVID-19 weekly epidemiological update, edition 101, 20 July 2022.

Yan, L., et al. (2020). A machine learning-based model for survival prediction in patients with severe COVID-19 infection.

Zahid, M. N., & Perna, S. (2021). Continent-wide analysis of COVID-19: Total cases, deaths, tests, socio-economic, and morbidity factors associated with the mortality rate, and forecasting analysis in 2020–2021. International Journal of Environmental Research and Public Health, 18, 5350.

Zamzuri, M. I. A., et al. (2020). Epidemiological characteristics of COVID-19 in Seremban, Negeri Sembilan, Malaysia. Open Access Macedonian Journal of Medical Sciences, 8, 471–475.

Zhu, X., et al. (2021). Joint prediction and time estimation of COVID-19 developing severe symptoms using chest CT scan. Medical Image Analysis, 67, 101824.