AI-Driven Mortality Prediction in COVID-19 Patients Using Advanced Feature Selection

Indika Rajakaruna
Department of Computer Science and Digital technologies
University of East London
London, United Kingdom
u2083478@uel.ac.uk

Mohammad Hossein Amirhosseini
Department of Computer Science and Digital Technologies
University of East London
London, United Kingdom
m.h.amirhosseini@uel.ac.uk

Yang Li
Department of Computer Science and Digital Technologies
University of East London
London, United Kingdom
y.li@uel.ac.uk

Amin Karami
Department of Computer Science and Digital technologies
University of East London
London, United Kingdom
a.karami@uel.ac.uk

Deepa Jayakody Arachchillage
department of Immunology and Imflammation
Imperial College London
London, United Kingdom
d.arachchillage@imperial.ac.uk

***Abstract*—COVID-19 has caused significant global mortality, with early risk stratification being critical for effective clinical management. Using a dataset of 8,032 COVID-19 hospitalized patients from a multicenter UK study, we developed and evaluated seven AI models, including deep and machine learning techniques, to predict in-hospital mortality. Key predictors were identified through a rigorous feature selection process combining statistical analysis, clinical expertise, and literature review. The Support Vector Classifier (SVC) achieved the best performance with 84% accuracy, 86% precision, and an AUC of 0.858, outperforming other methods in robustness and predictive accuracy. This study presents a novel application of AI on a large and diverse dataset, offering valuable insights for managing future pandemics/other clinical setting and improving clinical decision-making to reduce mortality.**

Keywords—Deep Learning, Machine Learning, Combined Feature Selection, Predictive Models, Mortality, Covid-19

I. Introduction

In early 2020, COVID-19 was officially recognised as a worldwide health emergency [1,2]. By December 2024, the World Health Organization reported 7.1 million deaths globally and 232,000 deaths in the UK alone due to COVID-19. The disease was associated with severe complications, including thrombosis, multi-organ failure (MOF), and major bleeding, all of which significantly increased the risk of mortality. COVID-19 patients experienced a three- to six-fold higher risk of thrombosis compared to individuals hospitalized for other reasons [3]. This unprecedented pandemic placed immense strain on healthcare systems, particularly in resource-limited settings, emphasizing the need for tools to streamline diagnoses, predict clinical outcomes, and optimize treatment strategies [4-7].

During the pandemic, artificial intelligence (AI) proved to be an essential asset, supporting areas such as disease diagnosis, public health planning, clinical guidance, and treatment development [8].

Several studies utilized machine learning (ML) and deep learning (DL) techniques to predict mortality and complications in COVID-19 patients [9-20]. However, these models often use small datasets or lack robust validation, leaving a gap in reliable mortality prediction methods for COVID-19 patients, which this study aims to address.

For instance, Li et al. [10] demonstrated the utility of autoencoders and traditional ML models, while Shahid et al. [11] highlighted the predictive power of Bi-LSTM models for recovery and mortality. Fang et al. [12] evaluated deep neural networks alongside traditional ML models to predict ICU admissions and mortality across datasets from Iran and the USA. Similarly, Zhang et al. [13] developed a novel FKNN-based model for deep venous thrombosis prediction, achieving 91.02% accuracy but limited by a small dataset.

In other studies, Liang et al. [14] used a Cox proportional hazards model to predict critical illness, achieving a concordance index (C-index) of 0.894 and an AUC of 0.911. Wu et al. [15] leveraged logistic regression models incorporating clinical, laboratory, and CT imaging features to predict COVID-19 severity, with the best model attaining an AUC of 0.90 on validation data. Furthermore, Mushtaq et al. [16] used convolutional neural networks (CNNs) to analyse chest X-rays (CXRs) for clinical outcomes, achieving AUC scores between 0.89 and 0.98. Jin et al. [17] employed deep learning to analyse CT data across diverse respiratory conditions, achieving an AUC of 0.9745 for COVID-19 diagnoses. Zandehshahvar et al. [18] and Sayed et al. [19] applied transfer learning and feature selection techniques like PCA and RFE, while Ucar et al. [20] optimized SqueezeNet CNNs, achieving 98.3% accuracy in COVID-19 diagnosis.

Although COVID-19 has diminished as a significant public health concern thanks to mass vaccinations and effective antiviral therapies, the insights gained during the pandemic continue to be crucial for managing future health crises and forecasting disease outcomes.

This study aims to compare the various AI models for mortality prediction in COVID-19 hospitalized patients using a large dataset from the multicentre "Coagulopathy in COVID-19" study conducted across 26 UK NHS Trusts (NCT04405232). This dataset, which includes demographic, clinical, and laboratory features, has previously been utilized to evaluate outcomes like thrombosis, MOF, and mortality using standard statistical methods [21-26]. By applying advanced ML and DL techniques, this research seeks to enhance predictive accuracy and provide valuable insights for developing models to assess the disease outcomes in future pandemics or other large scale diseases.

II. METHODOLOGY

A. Data Source

Ethical clearance for this research was granted by multiple regulatory bodies, including the Health and Care Research Wales (HCRW), Health Research Authority (HRA), and Scotland's Caldicott Guardian (approval number: 20/HRA/1785). The data utilised in this investigation consisted of information from 8,027 COVID-19 hospitalized patients, all aged 18 years or older, admitted over the period from April to July 2020.

Outliers and invalid data points were detected during the preprocessing phase and through scatter plot analysis, data sorting techniques, and interquartile range (IQR) assessments. Specific thresholds were set for variables like body weight, height, and laboratory test outcomes to eliminate implausible entries. To handle missing values—particularly for laboratory markers such as Ferritin, D-dimer, Troponin I, and lactate levels (with missingness under 10%)—we employed the k-Nearest Neighbours (k-NN) imputation method. This approach was not applied to clinical outcomes or comorbidities. After completing the imputation, each affected variable was carefully examined to confirm that the substituted values were realistic and within expected clinical ranges.

Age and Body mass index (BMI) were categorized into clinically relevant groups: BMI ('<18.5', '18.6-24.9', '25-29.9', '30-39.9', '>40') and age ('18-29 years', '30-49 years', '50-69 years', '70-89 years', '>90 years'). Table 1 presents a comprehensive list of features.

TABLE I. LABORATORY FEATURES, CLINICAL CHARACTERISTICS AND DEMOGRAPHICS OF THE PATIENTS

Feature	**Subcategory**	**Number (Total n = 8027)**	**Percentage**
Gender	Male	4403	55%
	Female	3624	45%
Age (Years)	18-29	207	3%
	30-49	991	12%
	50-69	2237	28%
	70-89	3864	48%
	>90	728	9%
Ethnicity	White	5811	72%
	Black	313	4%
	Asian	428	5%
	Other	1475	19%
Body Mass Index (kg/m2)	< 18.5 (Underweight)	215	3%
	18.6 – 24.9 (Healthy weight)	979	12%
	25.0 – 29.9 (Overweight)	5596	69%
	30.0 – 39.9 (Obese)	1007	13%
	> 40.0 (Severe obesity)	230	3%
History of Liver Disease	Present	295	4%
	Absent	7732	96%
History of Lung Disease	Present	1964	24%
	Absent	6063	76%
History of Diabetes	Present	2256	28%
	Absent	5771	72%
History of Heart Disease	Present	1837	23%
	Absent	6190	77%
History of Hypercholesterolemia	Present	1265	16%
	Absent	6762	84%
History of Hypertension	Present	3740	47%
	Absent	4287	53%
History of Malignancy	Present	873	11%
	Absent	7154	89%
History of Autoimmune disease	Present	604	8%
	Absent	7423	92%
History of Bleeding Disorders	Present	59	1%
	Absent	7968	99%
Laboratory features			
Laboratory Results	**Median**	**Inter Quartile**	**Reference Range**
Hemoglobin (g/L)	130 110*	114 -143 98 - 134*	130 – 160 (*115 –150)*
Platelets (10^9/L)	220	168 - 289	150 - 400
D-dimer (ng/mL)	1077	585-2851.5	0 - 500
White Cell Count (10^9 /L)	7.68	5.5 - 7.8	4.1 – 11.1
Neutrophils (10^9/L)	5.89	3.9 - 8.8	2.1 – 6.7
Lymphocytes (10^9/L)	0.9	0.6 - 1.3	1.3 – 3.7
Fibrinogen (g/L)	5.6	4.3 - 6.8	1.5 – 4.5
Alanine transferase (IU/L)	26	17 - 43	8 - 40
Bilirubin (µmol/L)	10	7 - 14	0 - 20

*Female hemoglobin

To prepare the clinical and demographic categorical variables for modelling, one-hot encoding was applied, transforming each category into separate binary indicators. Concurrently, numerical attributes—such as laboratory biomarkers—were standardised using conventional scaling techniques to ensure all feature values conformed to a uniform range. This standardisation was crucial in maintaining the integrity of the modelling process, as unbalanced feature scales could otherwise skew predictions, leading to elevated misclassification rates and diminished accuracy. A comprehensive summary of the selected variables and their relevance to mortality prediction is presented in Table 2.

TABLE II. FEATURES USED TO IDENTIFY THEIR SIGNIFICANCE FOR MORTALITY

Demographics Features	**Comorbidities and Clinical Conditions**	**Laboratory Markers**
Gender (Male/Female)	Multi-Organ Failure	Hemoglobin
Ethnicity (White/Asian/Black)	Thrombosis	Platelets
Age (Years)	Major Bleeding	D-dimer
Age groups (18–29, 30–49, 50–69, 70–89, 90+)	History of Smoking	White Cell Count
BMI categories (kg/m²): <18.5, 18.6–24.9, 25–29.9, 30–39.9, >40	History of liver disease	Neutrophils
	History of lung disease	Lymphocytes
	History of diabetes	Fibrinogen
	History of heart disease	Alanine transferase (ALT)
	History of hypercholesterolemia	Bilirubin
	History of hypertension	Creatinine
	History of malignancy	C-reactive protein
	History of autoimmune disease	Lactate dehydrogenase
	History of bleeding disorder	Troponin I
		Ferritin
		Prothrombin time

		Activated partial thromboplastin time
		Lactate

B. Feature Selection Approach

Following an initial variable assessment guided by clinical expertise, a range of advanced analytical techniques was employed to identify the predictors most closely linked to COVID-19 mortality. These included: (i) statistical significance testing methods such as Chi-square test, Mann–Whitney U test, and T-test; (ii) Pearson's correlation to examine linear associations between variables and outcomes; (iii) feature prioritisation through recursive feature elimination (RFE) leveraging logistic regression as the base model; and (iv) importance ranking using random forest algorithms. The features analysed encompassed patient demographics, pre-existing medical conditions, and laboratory parameters, which are detailed comprehensively in Table 1.

The choice between the T-test and Mann–Whitney U test was determined by the underlying distribution of the data, with each method used to assess group-level differences accordingly. To examine linear associations between variables, Pearson correlation analysis was conducted, with the strength of relationships measured by Pearson's correlation coefficient in relation to the target outcome. By performing pairwise correlation assessments, clusters of highly correlated features were identified, allowing for the refinement of the feature set to enhance predictive accuracy while reducing redundancy. This strategy improved the efficiency of the model by focusing on the most impactful variables.

Recursive Feature Elimination (RFE), functioning as a wrapper-based feature selection approach, was applied alongside logistic regression and random forest algorithms. This technique operates by repeatedly training the model, evaluating the importance of each feature, and discarding the least influential variables in successive rounds until the target number of features is retained. In contrast to filter-based approaches that assess features individually, RFE leverages the predictive power of a machine learning algorithm to progressively refine the feature set. The process started with the complete set of training variables and systematically eliminated less critical features, continuously re-ranking and retraining the model until an optimised subset was identified. This strategy ensures that the final selection focuses on the most relevant predictors, enhancing model effectiveness and efficiency. Table 3 presents the features selected through each of the applied feature selection methods.

TABLE III. KEY FEATURES IDENTIFIED BY DIFFERENT FEATURE SELECTION METHODS

Feature Selection Method	Identified significant features
Statistical Tests (T-test, Mann–Whitney U test, Chi-squared test)	Multi-organ failure, Age, Thrombosis, Elevated levels of Fibrinogen, ALT (Alanine Aminotransferase), and Bilirubin.
Recursive Feature Elimination (RFE) with Logistic Regression	Multi-organ failure, Major bleeding, Smoking status, Asian ethnicity, Age, and history of autoimmune disease.
Recursive Feature Elimination (RFE) with Random Forest Regressor	Raised levels of D-Dimer, Ferritin, Lactate Dehydrogenase (LDH), and Troponin I.

Drawing on the overlap of features identified by the selection techniques (as outlined in Table 3), combined with clinical expertise and established findings from COVID-19 research, a final set of critical predictors was chosen for model development. This refined feature set comprised factors such as a history of autoimmune disorders, smoking status, age, Asian ethnicity, and elevated biomarkers including D-Dimer, Ferritin, Troponin I, Fibrinogen, LDH, ALT, and Bilirubin. Additionally, the occurrence of significant bleeding events, thrombosis and multi-organ failure were also incorporated as important indicators.

C. Model Development

In this study, seven distinct machine learning and deep learning techniques were utilised for prediction of mortality outcomes in COVID-19 patients. The selected models included: (i) Multi-Layer Perceptron (MLP) classifier, (ii) Artificial Neural Network (ANN) classifier leveraging backpropagation for learning, (iii) Extreme Gradient Boosting (XGBoost) classifier, (iv) Support Vector Classifier (SVC), (v) Stochastic Gradient Descent classifier (SGD), (vi) Random Forest (RF), and (vii) Logistic Regression (LR).

(i) Multi-layer Perceptron (MLP) Classifier:

The Artificial Neural Network (ANN) implemented in this research adopts a feedforward architecture, where information progresses sequentially from the input nodes through intermediate layers to the final output layer. The use of nonlinear activation functions within the hidden layers empowers the network to detect and model complex patterns within the data. The inclusion of multiple hidden layers between the input and output stages enhances the network's ability to represent sophisticated relationships, thereby improving its predictive capability and accuracy. Specifically, the architecture consists of an input layer connected to a single hidden layer with 100 neurons. To mitigate the risk of overfitting, dropout regularization is applied immediately after the hidden layer. The final output layer uses a logistic (sigmoid) activation function to produce probabilities across the target classes.

For optimisation, the Adam algorithm was selected due to its ability to adaptively adjust learning rates while maintaining computational efficiency. The ANN's effectiveness was assessed on both training and testing datasets to validate its learning capability and generalisation to unseen data.

(ii) ANN with backpropagation:

Backpropagation serves as a pivotal algorithm in neural network training, aiming to optimise the network's internal parameters — namely weights and biases — by evaluating the difference between predicted results and true values. This error is propagated in reverse through the layers of the network, beginning at the output layer and progressing back to the input. Throughout this backward traversal, the algorithm systematically updates the parameters of each neuron to minimise the cumulative error, ultimately improving the model's predictive accuracy and overall performance.

(iii) XGBoost:

This gradient boosting algorithm trains multiple decision trees in sequence, with each tree addressing errors from the previous one. The results are aggregated to improve predictive accuracy and reduce overfitting. XGBoost improves upon the traditional gradient boosting framework by introducing advanced features such as L1 and L2 regularization, efficient handling of sparse data, and parallel processing. These enhancements make XGBoost highly suitable for predictive modelling tasks, especially for high-dimensional datasets like those in text-based personality prediction. The model minimizes the following objective function:

$$\mathcal{L} = \sum_{i=1}^{n} \ell(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k)$$

Within this formulation, $\ell(y_i, \hat{y}_i)$ represents the error function that quantifies the difference between the actual outcome y_i and the predicted value $\hat{y}_i$. Commonly applied functions include mean squared error (MSE) for regression analysis and logarithmic loss for classification tasks. Furthermore, $\Omega(f_k)$ embodies the regularization term, which discourages excessive model complexity to prevent overfitting. The regularization expression is defined as:

$$\Omega(f_k) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2$$

In this formula, T refers to the total count of terminal nodes (leaves) in the decision tree, w_j represents the assigned weight for the leaf j, γ acts as the threshold for the minimum loss decrease necessary to justify splitting a leaf, and λ determines the strength of the L2 regularization applied to the leaf weights.

XGBoost begins with a constant prediction, often the mean of the target variable. At each iteration, it adds a tree to reduce residual errors. The gradient (first derivative) and Hessian (second derivative) of the loss function guide tree construction, ensuring efficient optimization. Predictions are refined by combining outputs from previous iterations with the current tree. XGBoost's strength lies in its ability to handle non-linear relationships and complex interactions within the data. Its use of regularization techniques minimizes overfitting, enabling robust generalization across diverse samples. Additionally, XGBoost offers insights into feature importance, aiding in the identification of key patterns linked to the labels. These capabilities, combined with its efficiency and scalability, establish XGBoost as a top-performing model for this task.

(iv) Support Vector Classifier (SVC):

This model constructs an optimal hyperplane within a multidimensional feature space to distinguish between different classes. By leveraging the kernel method, it effectively enables the separation of classes even when the relationship between variables is non-linear. The primary goal of SVC is to maximise the margin between the closest data points of each class, ensuring robust generalisation of the decision boundary. The separating function of this boundary is mathematically defined as:

$$f(x) = w^T x + b$$

where w represents the weight vector, x is the input vector, and b is the bias term. SVC minimizes hinge loss using the following objective function:

$$L(w, b) = \frac{1}{2}||w||^2 + C\sum_{i=1}^{n} \max(0, 1 - y_i(w^T x_i + b))$$

Here, C is a regularization parameter that balances margin maximization and classification error. The algorithm projects data points into a higher-dimensional feature space by applying kernel functions, enabling it to manage complex class boundaries. It then addresses an optimisation task to determine the hyperplane that offers the greatest separation between classes. Data points are assigned to categories based on which side of the hyperplane they fall, with their proximity to the boundary influencing the classification decision.

(v) Stochastic Gradient Descent (SGD) Classifier:

This method of linear classification improves its decision boundary by continuously adjusting parameters to lower the value of the cost function. By leveraging stochastic gradient descent, the algorithm ensures rapid and efficient convergence, which is especially advantageous when working with large-scale and complex data.

(vi) Random Forest (RF):

Random Forest builds multiple decision trees, with each tree trained on a different randomly sampled subset of both the data and the features. Every tree in the ensemble produces its own classification output, and the overall model prediction is determined by combining these individual results, most commonly through majority voting. The algorithm enhances its predictive accuracy by aiming to reduce the average error rate across all trees within the forest.

$$E = \frac{1}{T}\sum_{t=1}^{T} E_t$$

where T is the total number of trees, and E_t represents the error of the t-th tree. This algorithm is particularly effective for complex datasets, as it can model non-linear relationships and identify intricate patterns within the data. Moreover, Random Forest offers interpretability by assessing feature importance, allowing researchers to identify the attributes most strongly associated with labels. This combination of interpretability, accuracy, and resilience makes Random Forest a strong baseline model for comparison with more sophisticated algorithms.

(vii) Logistic Regression (LR):

Logistic regression predicts the probability of a class label using the logistic function, making it particularly suitable for linearly separable data. The logistic function is expressed as:

$$P(y = 1|x) = \frac{1}{1 + e^{-(w^T x + b)}}$$

The model employs binary cross-entropy as its loss function, mathematically expressed as follows:

$$L(w,b) = -\frac{1}{n}\sum_{i=1}^{n}[y_i \log(p_i) + (1-y_i)\log(1-p_i)]$$

The model begins by initializing weights and biases. It then calculates the probability for each class using the logistic function. To minimize the cross-entropy loss, the weights and biases are optimized iteratively through gradient descent. Logistic regression's ability to handle non-linear relationships through the logistic function is particularly beneficial for capturing the complexities of data. Its simplicity and interpretability make it an ideal baseline model for comparing with more advanced techniques, while also providing valuable insights into the relationship between selected features and labels.

D. Hyperparameter Tuning and Cross Validation

To assess the models' effectiveness, we employed a 5-fold cross-validation strategy (K=5), which ensures that each training and testing partition accurately reflects the overall dataset. During this procedure, the data was segmented into five equally sized parts. In every iteration, the model was trained on four of these subsets, while the remaining one was reserved for testing. After completing all five rounds, the performance metrics were averaged to yield an overall evaluation. This method helps to mitigate overfitting and delivers a more dependable estimation of the model's generalisation to unseen data.

To optimise model performance, we implemented Grid Search as our hyperparameter tuning strategy. This technique systematically evaluates various combinations of hyperparameter values to determine the most effective setup for each algorithm, with the goal of maximising predictive accuracy. Hyperparameter selection plays a crucial role in shaping model outcomes, as these parameters have a substantial impact on performance. During this process, we used 5,378 samples from the total dataset of 8,027 records, applying stratified K-fold cross-validation to maintain balanced class distributions within each fold. This approach successfully identified the optimal hyperparameter sets for each model, as summarised in Table 4.

TABLE IV. OPTIMAL HYPERPARAMETERS FOR EACH MODEL

Model	**Optimum value for each parameter**
Multilayer Perceptron (MLP) Classifier	Activation Function: logistic, Hidden layer structure: (100, 1), Learning rate schedule: invscaling.
Artificial Neural Network (ANN) with Backpropagation	Implemented using Keras Sequential API, Optimizer: RMSprop, Batch size: 25, Epochs: 10, Loss function: binary cross-entropy, Activation functions: *ReLU* and *sigmoid*.
Extreme Gradient Boosting (XGBoost)	Learning rate: 0.05, Maximum tree depth: 3, Minimum child weight: 1,
Support Vector Classifier (SVC)	Regularization strength (C): 3.406, Kernel coefficient (gamma): 0.332; Probability estimates enabled: *True*.
Stochastic Gradient decent (SGD) Classifier	Elastic Net mixing parameter (*l1_ratio*): 0.14, Loss function: *log_loss,* Penalty term: *elasticnet*.
Random Forest Classifier	Minimum samples per leaf: 5, Maximum tree depth: 6, Split criterion: *entropy*.
Logistic Regression Classifier	Inverse regularization strength (C): 100.

E. Performance Measurments

To evaluate the effectiveness and validity of the developed models, a range of performance metrics and diagnostic tools were employed. These included the accuracy, F1 score, recall, precision, log loss, ROC AUC, and confusion matrix.

The confusion matrix provides a clear snapshot of the model's classification results by aligning the predicted categories with the actual ones, highlighting areas where the model made correct decisions as well as where it misclassified instances, including errors like false positives and false negatives. It forms the foundation for calculating crucial evaluation metrics, including accuracy, precision, recall, and the F1 score. Additionally, the ROC curve and its corresponding AUC provide insight into the model's capacity to differentiate between the positive and negative classes, with higher AUC values reflecting better discrimination. Precision reflects the percentage of predicted positives that were correct, while recall (also known as sensitivity) indicates the fraction of actual positive cases accurately detected. The F1 score acts as a harmonised metric, balancing precision and recall into a single value. Finally, log loss measures the reliability of probability estimates, penalising the model more heavily for confident but incorrect predictions.

Together, the selected evaluation metrics can provide a well-rounded evaluation of each model's predictive performance and generalisation capability. In addition, the ROC curve analysis was utilised to fine-tune the classification threshold. To ensure practical applicability, this threshold selection was guided by expert input, aligning the model's output with the real-world demands of the binary classification task (mortality versus survival).

III. RESULTS

Of the total of 8,027 COVID-19 patients included in this study,1,748 patients (21.8%) succumbed to the disease. 5,378 patient records (out of the 8,027 total),were used for the model training whilst the remaining 2,649 records used for testing. A stratified shuffle split method was used to split the data for training and testing and to ensure balanced class distributions across the datasets. As outlined earlier, the development of the models involved the application of 5-fold cross-validation alongside hyperparameter optimisation to enhance predictive performance. The outcomes of these models are summarised in Table 5, with visual representations displayed in Figures 1 and 2.

Among the various models assessed, the Support Vector Classifier (SVC) proved to be the most proficient in predicting mortality among patients, achieving a commendable accuracy rate of 84%. It particularly excelled in accurately identifying cases of survival (true negatives) and consistently surpassed its counterparts in correctly detecting mortality cases (true positives). In addition to its

overall accuracy, the SVC demonstrated leading performance across key evaluation metrics, with a precision of 86%, a recall of 84%, and an F1 score of 80%. Impressively, it recorded the lowest log-loss value of 0.476, indicating excellent model calibration and effective learning from the data. Furthermore, the SVC attained the highest ROC AUC value of 0.858, confirming its strong capability to distinguish effectively between survivors and non-survivors.

Following the SVC, models like Logistic Regression, XGBoost, Random Forest, and SGDClassifier also demonstrated strong overall performance across crucial evaluation indicators, such as accuracy, F1 score, recall, precision, and the ROC AUC. XGBoost, in particular, showed promising outcomes, recording a log-loss of 0.496, which suggests good model calibration and low training error. The Random Forest algorithm achieved an ROC AUC of 0.69, reflecting a fair capacity for class separation, although it did not reach the effectiveness displayed by the SVC. Similarly, XGBoost maintained credible results with an ROC AUC score of 0.66, reinforcing its utility for this classification challenge.

Taking into account the full range of evaluation criteria, including (i) accuracy, (ii) precision, (iii) recall, (iv) F1 score, (v) log-loss, and (vi) ROC AUC, the SVC model distinctly emerged as the top performer for mortality prediction in COVID-19 patients. Its excellent convergence throughout training, combined with its strong ability to distinguish between survival and mortality outcomes, solidifies its position as the leading model in this research.

TABLE V. EVALUATION METRICS FOR EACH MODEL

Model	Accuracy	Precision	Recall	F1 Score	Log Loss (training/test)	ROC AUC (training/test)
MLP Classifier	0.78	0.39/0.61	0.50/0.78	0.44/0.69	0.52/0.52	0.566/0.55
ANN	0.77	0.61/0.71	0.54/0.77	0.53/0.72	0.50/0.52	0.63/0.57
XGBoost	0.78	0.65/0.73	0.53/0.78	0.51/0.72	0.49/0.51	0.66/0.58
SVC	0.84	0.91/0.86	0.63/0.84	0.66/0.80	0.47/0.47	0.86/0.85
SGD Classifier	0.78	0.64/0.73	0.53/0.78	0.52/0.72	0.51/0.51	0.60/0.56
Random Forest	0.78	0.64/0.73	0.53/0.78	0.52/0.72	0.49/0.51	0.69/0.58
Logistic Regression	0.79	0.62/0.72	0.54/0.78	0.53/0.72	0.50/0.51	0.68/0.57

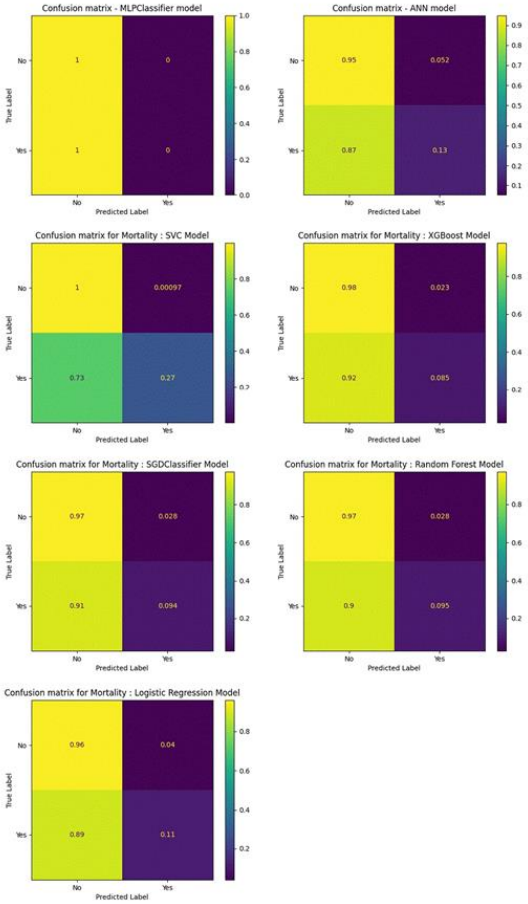

Fig. 1. Confusion matrix for each model

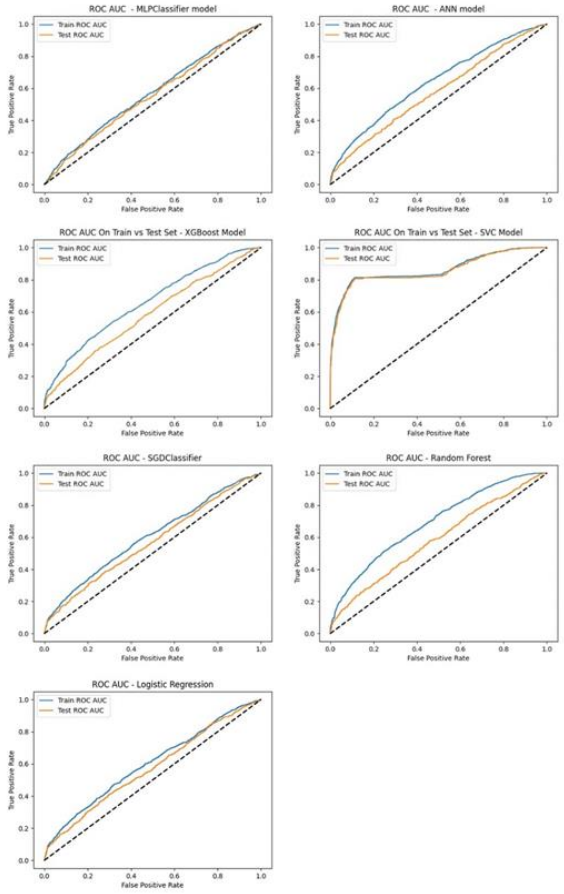

Fig 2. AUC curves for implemented models

IV. Discussion

In this research, we combined artificial intelligence techniques, conventional statistical methods, and clinical expert insights to pinpoint critical predictors of mortality in patients hospitalised with COVID-19. The influential factors identified encompassed demographic characteristics (such as Asian ethnicity and advancing age), underlying health conditions (including smoking history and autoimmune diseases), key laboratory indicators (notably elevated D-Dimer, Ferritin, LDH, Troponin I, Fibrinogen, ALT, and Bilirubin levels), as well as severe clinical complications like multi-organ failure, bleeding episodes, and thrombosis. These selected variables formed the foundation for training and evaluating seven distinct AI-based models aimed at forecasting mortality risk among COVID-19 patients.

Key contributors to mortality among hospitalised COVID-19 patients were identified as demographic variables (such as Asian ethnicity and advancing age), underlying health conditions (including smoking and autoimmune disorders), critical laboratory markers (elevated levels of D-Dimer, Ferritin, LDH, Troponin I, Fibrinogen, ALT, and Bilirubin), alongside severe clinical complications like multi-organ failure, bleeding events, and thrombosis. Taking these variables into account, we developed and assessed seven separate AI-based models to predict mortality risk in patients diagnosed with COVID-19. Of all the models evaluated, the Support Vector Classifier (SVC) emerged as the top performer, reaching an accuracy rate of 84%. Notably, the SVC demonstrated flawless accuracy in correctly predicting survival outcomes (true negatives) and surpassed its counterparts in effectively identifying fatal cases (true positives).

Furthermore, the Support Vector Classifier (SVC) excelled in multiple key performance measures, securing a precision rate of 86%, recall of 84%, an F1 score of 80%, and the lowest log-loss value of 0.476 on the test set—indicating both accurate predictions and excellent calibration. The model also achieved the highest AUC score at 0.858, highlighting its robust capacity to distinguish between different outcome categories. By capitalising on the foundational concepts of Support Vector Machines, the SVC effectively identified the most suitable hyperplane to separate the classes. While other algorithms delivered satisfactory results, the SVC consistently outperformed them across nearly all evaluation metrics.

Numerous AI-focused investigations have been conducted to forecast severe outcomes such as critical illness, ICU admission, or mortality among COVID-19 patients [16, 17, 18, 19]. One such example is the study [10] titled *"Individual-Level Fatality Prediction of COVID-19 Patients Using AI Methods,"* which reported impressive performance, achieving over 90% accuracy and specificity with its leading autoencoder model. Nonetheless, that research primarily depended on publicly accessible datasets, which lacked comprehensive, case-specific details. This limitation notably hindered the model's predictive strength. The authors themselves acknowledged that the scarcity of rich, high-quality data represented a significant constraint on the effectiveness of their predictive approach.

A major strength of this study lies in the utilisation of a large, diverse dataset sourced from 26 NHS Trusts across England, Wales, and Scotland, collected during the height of the COVID-19 pandemic. The data's reliability was reinforced by its collection by qualified clinical professionals, ensuring both accuracy and clinical relevance. Moreover, the dataset reflected a broad and representative sample of the UK patient population, enhancing the generalisability of the findings. In addition to leveraging this robust dataset, we conducted a comprehensive evaluation of seven distinct AI models to identify the most effective approach for mortality risk prediction.

One of the main constraints of this research relates to the temporal context and its applicability to present-day clinical environments. The dataset underpinning model development was gathered in the early phases of the pandemic in 2020, a time characterised by severe disease presentations and elevated mortality rates. Given that COVID-19 has since evolved into less severe forms, with significantly lower fatality rates, the direct applicability of these models to contemporary clinical practice may be limited. However, the comprehensive methodological framework established in this study—including rigorous data cleansing, AI-based imputation techniques, feature selection processes, model construction, cross-validation procedures, hyperparameter optimisation, and performance evaluation—remains highly transferable. With appropriate adaptations, this approach could be effectively applied to build binary classification models for forecasting outcomes in other clinical conditions.

V. Conclusion

In summary, this study successfully designed and evaluated seven artificial intelligence models aimed at forecasting mortality among hospitalised COVID-19 patients, utilising patient demographic details, underlying health conditions, and laboratory findings collected at admission. The top-performing model demonstrated an accuracy rate of 84%, highlighting the potential of AI to support clinical decision-making. These findings emphasise the significant contribution of AI technologies in healthcare, particularly in situations with limited resources. Furthermore, this research establishes a foundation for the future development of adaptable AI-driven solutions capable of predicting clinical outcomes across both emerging infectious diseases and existing healthcare challenges.

References

[1] Zhou P, Yang XL, Wang XG, et al. A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature* 2020; 579(7798): 270-3.

[2] Reeves JJ, Hollandsworth HM, Torriani FJ, et al. Rapid response to COVID-19: health informatics support for outbreak management in an academic health system. *J Am Med Inform Assoc* 2020; 27(6): 853-9.

[3] Thomas MR, Scully M. Clinical features of thrombosis and bleeding in COVID-19. *Blood* 2022; 140(3): 184-95.

[4] Emanuel EJ, Persad G, Upshur R, et al. Fair Allocation of Scarce Medical Resources in the Time of Covid-19. *N Engl J Med* 2020; 382(21): 2049-55.

[5] Mashamba-Thompson TP, Crayton ED. Blockchain and Artificial Intelligence Technology for Novel Coronavirus Disease-19 Self-Testing. *Diagnostics (Basel)* 2020; 10(4).

[6] Siow WT, Liew MF, Shrestha BR, Muchtar F, See KC. Managing COVID-19 in resource-limited settings: critical care considerations. *Crit Care* 2020; 24(1): 167.

[7] Yassine HM, Shah Z. How could artificial intelligence aid in the fight against coronavirus? *Expert Rev Anti Infect Ther* 2020; 18(6): 493-7.

[8] Chen J, See KC. Artificial Intelligence for COVID-19: Rapid Review. *J Med Internet Res* 2020; 22(10): e21476.

[9] Ryan L, Mataraso S, Siefkas A, et al. A Machine Learning Approach to Predict Deep Venous Thrombosis Among Hospitalized Patients. *Clin Appl Thromb Hemost* 2021; 27: 1076029621991185.

[10] Li Y, Horowitz MA, Liu J, et al. Individual-Level Fatality Prediction of COVID-19 Patients Using AI Methods. *Front Public Health* 2020; 8: 587937.

[11] Shahid F, Zameer A, Muneeb M. Predictions for COVID-19 with deep learning models of LSTM, GRU and Bi-LSTM. *Chaos Solitons Fractals* 2020; 140: 110212.

[12] Fang X, Kruger U, Homayounieh F, et al. Association of AI quantified COVID-19 chest CT and patient outcome. *Int J Comput Assist Radiol Surg* 2021; 16(3): 435-45.

[13] Zhang L, Yu R, Chen K, Zhang Y, Li Q, Chen Y. Enhancing deep vein thrombosis prediction in patients with coronavirus disease 2019 using improved machine learning model. *Comput Biol Med* 2024; 173: 108294.

[14] Liang W, Yao J, Chen A, et al. Early triage of critically ill COVID-19 patients using deep learning. *Nat Commun* 2020; 11(1): 3543.

[15] Wu G, Yang P, Xie Y, et al. Development of a clinical decision support system for severity risk prediction and triage of COVID-19 patients at hospital admission: an international multicentre study. *Eur Respir J* 2020; 56(2).

[16] Mushtaq J, Pennella R, Lavalle S, et al. Initial chest radiographs and artificial intelligence (AI) predict clinical outcomes in COVID-19 patients: analysis of 697 Italian patients. *Eur Radiol* 2021; 31(3): 1770-9.

[17] Jin C, Chen W, Cao Y, et al. Development and evaluation of an artificial intelligence system for COVID-19 diagnosis. *Nat Commun* 2020; 11(1): 5088.

[18] Zandehshahvar M, van Assen M, Maleki H, Kiarashi Y, De Cecco CN, Adibi A. Toward understanding COVID-19 pneumonia: a deep-learning-based approach for severity analysis and monitoring the disease. *Sci Rep* 2021; 11(1): 11112.

[19] Sayed SA, Elkorany AM, Sayed Mohammad S. Applying Different Machine Learning Techniques for Prediction of COVID-19 Severity. *IEEE Access* 2021; 9: 135697-707.

[20] Ucar F, Korkmaz D. COVIDiagnosis-Net: Deep Bayes-SqueezeNet based diagnosis of the coronavirus disease 2019 (COVID-19) from X-ray images. *Med Hypotheses* 2020; 140: 109761.

[21] Arachchillage DJ, Rajakaruna I, Odho Z, et al. Clinical outcomes and the impact of prior oral anticoagulant use in patients with coronavirus disease 2019 admitted to hospitals in the UK - a multicentre observational study. *Br J Haematol* 2022; 196(1): 79-94.

[22] Arachchillage DJ, Rajakaruna I, Scott I, et al. Impact of major bleeding and thrombosis on 180-day survival in patients with severe COVID-19 supported with veno-venous extracorporeal membrane oxygenation in the United Kingdom: a multicentre observational study. *Br J Haematol* 2022; 196(3): 566-76.

[23] Arachchillage DJ, Weatherill A, Rajakaruna I, et al. Thombosis, major bleeding, and survival in COVID-19 supported by veno-venous extracorporeal membrane oxygenation in the first vs second wave: a multicenter observational study in the United Kingdom. *J Thromb Haemost* 2023; 21(10): 2735-46.

[24] Arachchillage DJ, Rajakaruna I, Pericleous C, Nicolson PLR, Makris M, Laffan M. Autoimmune disease and COVID-19: a multicentre observational study in the United Kingdom. *Rheumatology (Oxford)* 2022; 61(12): 4643-55.

[25] DJ, Rajakaruna I, Odho Z, Makris M, Laffan M. Impact of thromboprophylaxis on hospital acquired thrombosis following discharge in patients admitted with COVID-19: Multicentre observational study in the UK. *Br J Haematol* 2023; 202(3): 485-97.

[26] Crossette-Thambiah C, Nicolson P, Rajakaruna I, et al. The clinical course of COVID-19 in pregnant versus non-pregnant women requiring hospitalisation: results from the multicentre UK CA-COVID-19 study. *Br J Haematol* 2021; 195(1): 85-9.