Received: 17 January 2025 | Revised: 17 March 2025 | Accepted: 16 April 2025 | Published online: 30 April 2025

RESEARCH ARTICLE

Medinformatics
2025, Vol. 00(00) 1–10
DOI: 10.47852/bonviewMEDIN52025243

BON VIEW PUBLISHING

Leveraging Explainable AI for Drug Efficacy and Mortality Analysis: Insights from an Observational Dataset

Manu Kumar Shetty[1,*], Aaloke Mozumdar[2], Saurabh Gupta[2], Lalit Gupta[3], Kapil Chaudhary[3], Vandana Roy[1], Suresh Kumar[4], Bhupinder Singh Kalra[1], Sanjeev Khanth PE[1] and Anubha Gupta[2,*]

[1]*Department of Pharmacology, Maulana Azad Medical College and Lok Nayak Hospital, India*

[2]*Signal Processing and Biomedical Imaging Lab, Indraprastha Institute of Information Technology Delhi, India*

[3]*Department of Anesthesia, Maulana Azad Medical College and Lok Nayak Hospital, India*

[4]*Department of Medicine, Maulana Azad Medical College and Lok Nayak Hospital, India*

Abstract: Randomized control trials (RCTs) are the gold standard for establishing causality in drug efficacy, but they have limitations due to strict inclusion criteria and complexity. When RCTs are not feasible, researchers often turn to observational study analysis, where explainable AI (XAI) models offer a compliment to observational study approach for understanding cause-and-effect relationships. In this study, we employed an XAI model using a historical COVID-19 dataset consisting of 3,307 patients from a hospital in Delhi, India, to evaluate drug efficacy. By applying eight XAI models and traditional statistical methods, such as multivariate analysis, we identified key factors influencing COVID-19 survival. AI interpretability techniques were used to determine feature importance in the outcomes. The XGBoost classifier outperformed other models with a weighted *F*1 score of 91.7%, ROC-AUC of 92.2%, and sensitivity of 93.8%. However, both the XAI models and forest plot revealed that medications such as enoxaparin, remdesivir, and ivermectin did not show survival benefits. While XAI models provide valuable insights and individual-level interpretability, they should not replace RCTs in assessing the safety and efficacy of new treatments but can aid in clinical decision-making and suggest future research directions.

Keywords: randomized control trials, observational study, study designs, COVID-19, explainable AI, artificial intelligence, SARS-CoV-2

1. Introduction

Randomized controlled trial (RCT) is a gold standard study design to test the efficacy and safety of a medical treatment or a procedure [1]. RCT is designed to minimize bias and minimize confounding factors and provides evidence for efficacy of a drug to achieve its therapeutic objectives [2]. However, RCT can be complex and expensive and has strict inclusion criteria, making it difficult to conduct in all situations. When RCTs are not feasible, observational studies are used to understand the cause-and-effect relationship [3]. An observational study can be either retrospective or prospective [4]. Such a study provides valuable complementary information related to drug efficacy or treatment on real-world population and helps to inform clinical practice and public health policy [5].

During the early stages of the COVID-19 pandemic, observational studies became critical due to the immediate need for treatment insights, even though their inherent limitations, such as confounding factors, were well recognized [6, 7]. Early observational studies, often based on limited sample sizes, initially supported various treatments for COVID-19, but their findings were later contradicted by more rigorous RCTs. This underscores the potential discrepancies between observational and RCT-derived evidence. For example, observational data initially suggested benefits of simvastatin in treating moderate-to-severe chronic obstructive pulmonary disease exacerbations, yet the STATCOPE trial later showed no such efficacy [8].

An explainable AI (XAI) model can be built using historical data to predict a drug's efficacy and to find the relationship between the cause (drug) and effect (survival). This can provide valuable insights into the underlying mechanisms and processes that drive the observed associations [9]. An XAI model complements the observational study by using advanced machine learning algorithms to analyze large amounts of data and identify patterns and relationships that may not be apparent from observational data alone [10]. Moreover, a XAI model can provide a more comprehensive view of the relationship between the drug and the outcome because it considers multiple factors (comorbidity, symptoms) simultaneously. Additionally, XAI models can be used to make predictions about outcomes and to test the robustness and stability of the results [11]. This can be

*Corresponding authors: Manu Kumar Shetty, Department of Pharmacology, Maulana Azad Medical College and Lok Nayak Hospital, India. Email: dr.shetty09@delhi.gov.in and Anubha Gupta, Signal Processing and Biomedical Imaging Lab, Indraprastha Institute of Information Technology Delhi, India. Email: anubha@iiitd.ac.in

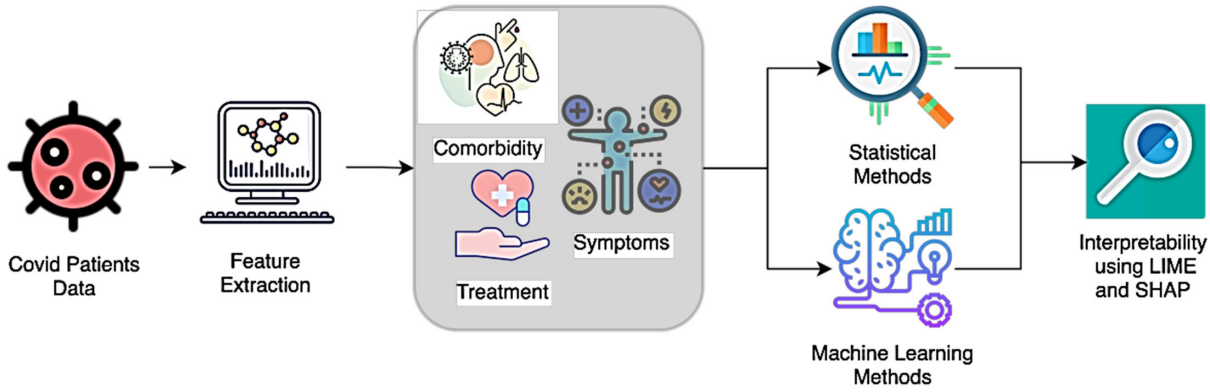

Figure 1. Flow diagram of an interpretable AI model for survival outcome prediction. The data of COVID-19 patients (death and discharged) were acquired. Four types of features were extracted, namely demographics, comorbidity, symptoms at the time of admission, and treatment given. Statistical and machine learning models were employed on the features to predict the survival outcome. Finally, LIME and SHAP are used to demonstrate the importance of each of the features in predicting the survival

especially useful in complex situations such as a new disease, for example the COVID-19 pandemic, where various unknown clinical factors may impact patient outcomes. In this study, we aim to explore drug efficacy hypotheses using an XAI model developed from historical observational COVID-19 data.

The contributions of our study are as follows (Figure 1):

1) We developed an XAI model to identify the factors affecting mortality in COVID-19 patients using historical COVID-19 data. We also hypothesize whether an XAI model can serve as a complement to Observational and RCTs when RCTs are not feasible.
2) We compared the results generated from observational study analytics with those from the ML-based XAI model. The ML-based XAI model demonstrated better accuracy compared to logistic or linear regression models.
3) Our proposed model provides individual-level risk interpretation using techniques such as SHAP (Shapley Additive Explanations) and LIME (Local Interpretable Model-agnostic Explanations) for enhanced interpretability.

2. Materials

2.1. Data description

This study was approved by the Institutional Ethics Committee of Maulana Azad Medical College (see Appendix A for details). The study utilizes historical, digitally available datasets of 3,307 COVID-19 patients selected from approximately 30,000 total admissions at the Lok Nayak Hospital, New Delhi, India. All patients included in this dataset tested positive for COVID-19 and were admitted between March 2020 and July 2021. Mortality data from 2,955 deceased patients and data from 352 discharged patients were retrieved for analysis (Table 1).

Table 1 summarizes the results from a multivariate logistic regression analysis performed to identify factors significantly

Table 1. The results from a multivariate logistic regression analysis of survival outcomes (death vs. discharged) among COVID-19 patients

Feature group		Coefficient (Coef)	Standard error (SE)	z-value	P-value	Confidence interval (95%)
Demographics	Age	0.2097	0.069	3.028	0.002	0.109–0.159
	Gender_Female	0.0233	0.068	0.344	0.731	0.109–0.156
Comorbidity	Heart disorder	0.1111	0.081	1.374	0.17	0.109–0.164
	Kidney disorders	0.4844	0.086	5.641	<0.001	0.033–0.243
	Hypertension	−0.0395	0.078	−0.504	0.614	0.109–0.161
	Tuberculosis	0.1013	0.069	1.473	0.141	0.033–0.236
	Diabetes	0.0745	0.075	0.998	0.318	0.109–0.165
	Thyroid disorder	0.1186	0.082	1.443	0.149	0.033–0.244
	COPD	0.0141	0.07	0.201	0.841	0.109–0.157
Symptoms	cough	−0.1377	0.068	−2.023	0.043	0.033–0.238
	diarrhea	0.0393	0.066	0.6	0.549	0.109–0.158
	Breathlessness	0.3714	0.066	5.612	<0.001	0.033–0.245
	Headache	−0.1089	0.056	−1.944	0.052	0.109–0.162
	Body Weakness	Removed from multivariate analysis				
Treatment	Remdesivir	0.4923	0.097	5.096	<0.001	0.033–0.239
	Enoxaparin	1.062	0.073	14.619	<0.001	0.033–0.241
	Vit_d	−0.0871	0.059	−1.486	0.137	0.033–0.240
	Zinc	−0.6571	0.11	−5.979	<0.001	0.109–0.160
	Vit_c	−0.5432	0.116	−4.672	<0.001	0.109–0.163
	Ivermectin	0.1808	0.081	2.242	0.025	0.033–0.242
	on_steroids	−18.1649	0.621	−29.252	<0.001	0.033–0.237

associated with survival outcomes (death vs. discharged) among COVID-19 patients. Each feature is represented along with its corresponding regression coefficient, standard error, *z*-value, and *p*-value. The confidence intervals indicate the precision and reliability of each coefficient estimate. Positive coefficients imply increased odds of mortality, while negative coefficients indicate decreased odds. Features identified as statistically significant ($p < 0.05$) suggest a meaningful association with the patient survival outcome.

This dataset includes four types of features: two demographic features (age and gender), eleven symptoms noted at the time of admission (cough, body weakness, breathlessness, headache, diarrhea), seven comorbidity features detailing the existing ailments of the admitted patients (hypertension, diabetes, tuberculosis, thyroid, chronic obstructive pulmonary disease (COPD), kidney disease, and heart disease), and features comprising of the treatment given to the patients to cure COVID-19 (vitamin C, vitamin D zinc, Remdesivir, Ivermectin, Enoxaparin). Table 1 shows the distribution of data among the two classes: one class consisting of patients who died in the hospital after admission owing to COVID-19 (called mortality class) and, the other class of patients who survived COVID-19 and were discharged (called the survival class).

3. Methods

3.1. Statistical tests

Once the dataset is prepared and prepossessed, we implemented a multi-variable regression model to estimate the odds ratios for each predictor while adjusting for potential confounders. This model allows us to determine the independent effect of each variable on the outcome (mortality) by controlling for other variables in the model. The results, including odds ratios and confidence intervals, *p*-values for each predictor variable, indicating their statistical significance in the logistic regression model. We used Python and Jupyter Notebook to calculate the *P*-value. For calculating the odds ratio and confidence intervals, we utilized the "*statsmodels.api version 0.14.1*" module from the Python library. Additionally, we employed the "scikit-learn (sklearn)" library for machine learning development [12, 13].

The dataset exhibited a high degree of class imbalance, with a ratio of 2955 discharged patients to 352 deaths. This imbalance poses a challenge as models may over-fit in favor of the majority class. To address this, we applied random oversampling and undersampling techniques. Specifically, we utilized the synthetic minority oversampling technique (SMOTE) for random oversampling and random undersampling to balance the dataset [14]. These methods were implemented to address class imbalance and improve the performance of our models. We also employed K-fold cross-validation, a model validation strategy that evaluates how the findings of a statistical method will generalize to an independent dataset. Cross-validation is a resampling process that tests and trains a model using various chunks of the data in successive rounds. It is particularly useful in predictive modeling to assess model performance in practice.

3.2. Machine learning

We built eight machine learning classifiers, including logistic regression, random forest, extra tree classifier, support vector machine (SVM), Naive Bayes, and XGBoost. Logistic regression was chosen for its interpretability and simplicity, while random forest, extra tree classifier, and SVM were selected due to their robustness and effectiveness in handling complex, non-linear relationships. Naive Bayes was included for its computational efficiency and performance with categorical data. Additionally, boosting methods such as XGBoost were specifically employed due to their proven capability to achieve high predictive performance in classification tasks by iteratively improving weak learners. As a part of post-hoc analysis, SHAP [15] and LIME [16] algorithms were used for the interpretability of AI models.

SHAP uses different visuals to show the value of features and how they contribute to predictions. LIME stands for Local Interpretable model-agnostic explanations. LIME is model-independent, which means it may be used with any machine learning model. The technique tries to figure out what the model is doing by changing the input of data samples and seeing how the predictions change. Model-specific approaches examine the core components of the black box machine learning model and how they interact in order to gain a better understanding of it. Local model interpretability is provided by LIME. It evaluates every feature value in a single data sample to infer its influence on the output.

To further improve the model, we use sampling techniques to address the bias in the data using undersampling and oversampling techniques. We first employ random oversampling and random undersampling techniques to find that the oversampling is giving us better results. Subsequently, we applied more advanced oversampling techniques such as the adaptive synthetic (ADASYN) algorithm [17] and SMOTE, which effectively address class imbalance by generating synthetic examples of the minority class, thus enhancing the classifier's performance on imbalanced datasets. Among these methods, we achieved optimal classification performance using the AdaBoost classifier with XGBoost as the base estimator.

3.3. Evaluation metrics

The evaluation metrics of classifiers were computed using the numbers of true positive (TP), false positive (FP), true negative (TN), and false negative (FN) for a binary classifier, where TP indicates the number of samples of positive class (mortality class) identified correctly, FP indicates the number of samples of negative class (survival class) identified incorrectly as the positive class, TN indicates the number of samples of negative class identified correctly, and FN indicates the number of samples of positive class identified incorrectly as the negative class.

4. Results

4.1. Baseline characteristics

The COVID-19 patients' data consisted of two classes, the mortality class (Positive class or the class 1) and the survival class (Negative class or the class-0). The data are summarized in Table 1 and Figure 2.

The forest plot illustrates that several factors are significantly associated with increased mortality risk in COVID-19 patients. Notably, breathlessness, kidney disorders, and advanced age all show odds ratios (ORs) greater than 1, with confidence intervals (CIs) that do not cross 1, indicating these are strong predictors of mortality. Enoxaparin, remdesivir, and ivermectin demonstrate a markedly increased risk with an OR greater than 1, highlighting their potential adverse association with mortality. In other words, these medications did not prevent the mortality. Conversely, zinc and Vit C show an OR below 1, suggesting a potential protective effect, and did not cross the CI line indicating that this finding is statistically significant. These results underscore the importance of

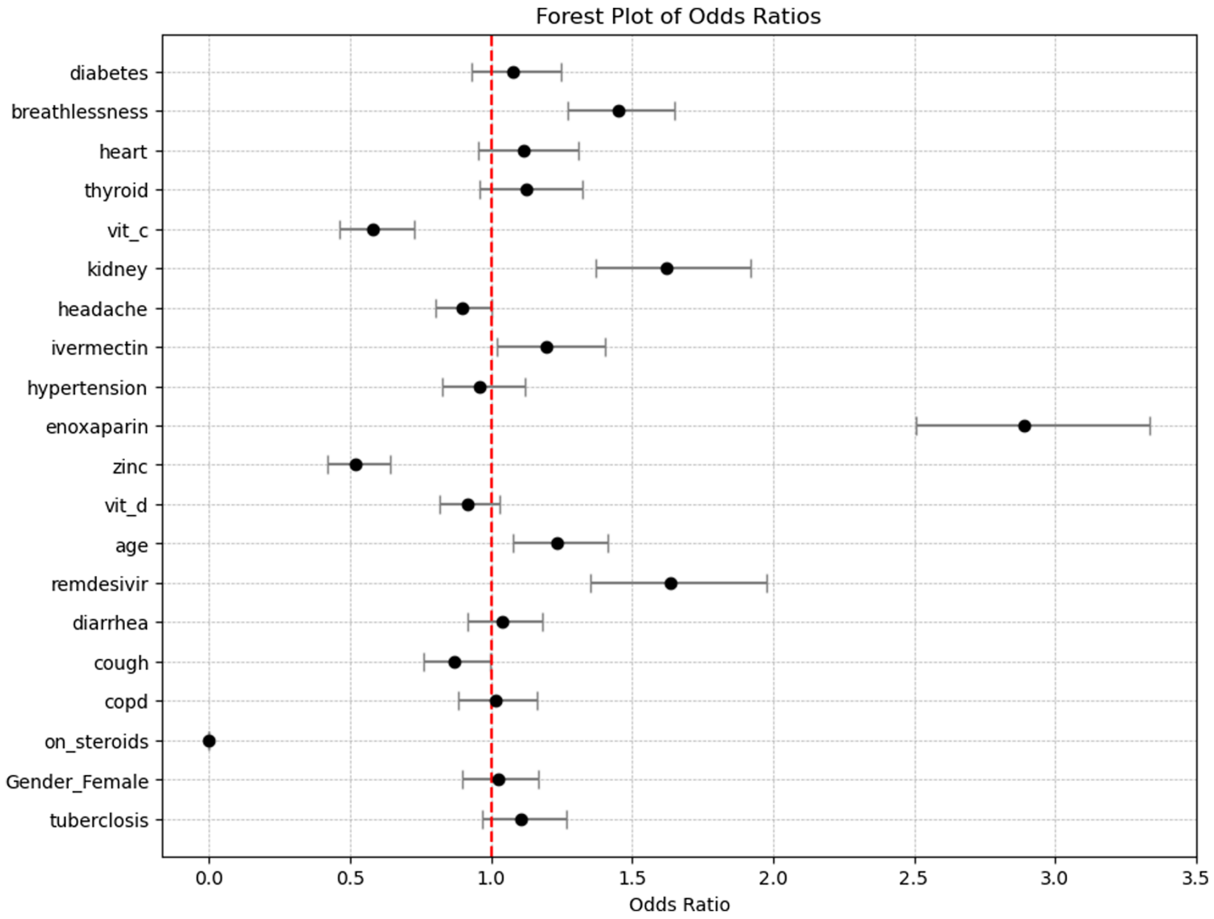

Figure 2. This forest plot displays the odds ratios and 95% confidence intervals for various features assessed in the study. Each line represents a different feature, with the odds ratio depicted as dot and the confidence interval shown as a horizontal line extending from the lower to the upper bound. The vertical line indicates the null value (odds ratio of 1), where features with confidence intervals crossing this line are not statistically significant. This plot allows for a visual comparison of the effectiveness of each feature in relation to the reference

managing comorbidities and patient characteristics to improve COVID-19 outcomes.

The mean age of COVID-19 patients who survived is 49 years, and the mean age of mortality class is 57 years. The majority of COVID-19 patients admitted to hospitals were males (64%). A similar gender ratio was observed in survival and mortality class patients. The most common symptoms were breathlessness and cough lesser presentation with other symptoms including diarrhea and headache. Among all symptoms, breathlessness and cough were observed to have statistically significant differences between the survival and the mortality classes. Comorbidity of kidney disorder was found to be statistically significant between the two classes.

4.2. ML model performance

We first trained the ML models (Logistic Regression, Kernel SVM, Complement NB, Random Forest, Extra Trees Classifier, XGBoost, AdaBoost) using cost-sensitive loss functions. Based on the performance of these ML models on AUC, sensitivity, and specificity, random forest (AUC 90%, sensitivity 97%, and specificity 51%) and XGBoost models (AUC 93%, sensitivity 96%, and specificity 56%) were the best performing model among all models (Table 2). To further improve the model performance in terms of specificity, we trained the random forest and XGboost models using undersampling and oversampling techniques. We first employed random oversampling and random undersampling techniques and observed that oversampling yielded better results on our dataset. Next, we tried advance oversampling techniques such as ADASYN algorithm and SMOTE [14, 17]. We have used 5-fold cross-validation to generate five different sets (or folds) of data. Five classifiers were trained, each choosing a different fold as the test set (unseen data) for validating the classifier, while the remaining four-folds were used with sampling techniques to train the respective classifiers. Cross-validated results (one-fold) were generated for evaluating the classifiers (Figure 3).

Table 3 shows the results generated using different sampling techniques and averaged over all the 5-folds. We have shown results of top two classifiers (Random forest and XGBoost) after applying various sampling techniques. Among these, we got the best overall classification performance with XGBoost classifier and SMOTE as the sampling technique, which provided weighted *F*1 score, MCC, accuracy, ROC-AUC, sensitivity, and specificity score of 91.7%, 58.8%, 91.3%, 92.2% 93.8%, and 70.2%, respectively.

4.3. Interpretability

We applied interpretability techniques to the best performing classifier, namely XGBoost with SMOTE as the sampling

Table 2. Performance of chosen learning algorithms on the dataset using 5-fold cross-validation and class weights

Classifier	wF1-score	MCC	Accuracy	ROC-AUC	Sensitivity	Specificity
Logistic Regression	0.848 ± 0.011	0.475 ± 0.029	0.819 ± 0.014	0.929 ± 0.017	0.814 ± 0.017	**0.858 ± 0.047**
Kernel SVM	0.623 ± 0.042	0.223 ± 0.017	0.543 ± 0.044	0.800 ± 0.035	0.506 ± 0.053	0.852 ± 0.035
Complement NB	0.855 ± 0.006	0.469 ± 0.039	0.829 ± 0.008	0.903 ± 0.032	0.830 ± 0.009	0.818 ± 0.066
RandomForest	0.919 ± 0.009	0.557 ± 0.051	**0.924 ± 0.007**	0.903 ± 0.022	**0.973 ± 0.003**	0.517 ± 0.069
ExtraTreesClassifier	0.909 ± 0.008	0.508 ± 0.049	0.912 ± 0.007	0.867 ± 0.025	0.958 ± 0.004	0.520 ± 0.063
XGBoost	0.920 ± 0.010	0.568 ± 0.057	0.923 ± 0.009	**0.935 ± 0.015**	0.966 ± 0.004	0.562 ± 0.060
Adaboost*	0.916 ± 0.007	0.543 ± 0.046	0.920 ± 0.005	0.922 ± 0.012	0.968 ± 0.004	0.523 ± 0.068
Adaboost**	**0.921 ± 0.007**	**0.568 ± 0.040**	0.924 ± 0.007	0.925 ± 0.015	0.969 ± 0.005	0.548 ± 0.031

Note: *=XGBoost as base estimator; **=Decision Tree as base estimator; PPV = Positive Predicted Value; NPV = Negative Predicted Value

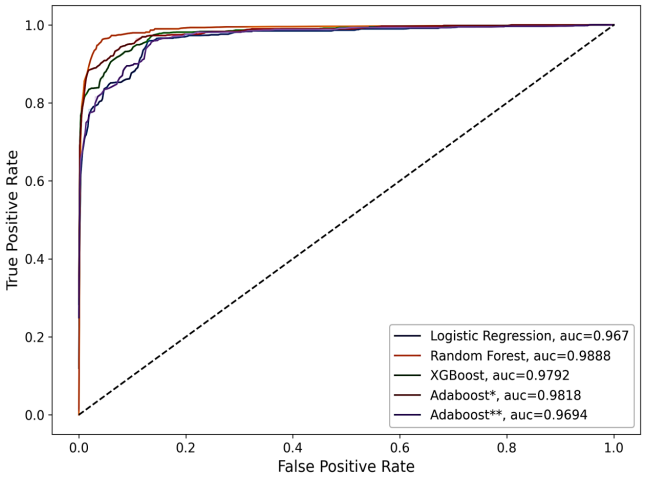

Figure 3. ROC curve plot of Top 5 ML model results. Results are reported on a single 4:1 train test stratified split instead of 5-fold cross-validation

technique, including all the features. We looked at the ML interpretability using two approaches – global and local interpretability. In the case of the former, we explained the predictions of our ML models at a population level, whereas in the case of the latter, we explained the model prediction for a specific patient.

For the purposes of global interpretability, we used SHAP to analyze our model. SHAP uses Shapley values [16] to determine the importance and the role that each feature in the model predictions. The SHAP summary plot, Figure 4 indicates how value of each feature affects the classifier predictions and also ranks the features on the basis of the mean absolute value of their Shapley values, consequently showing their mean impact on the classification.

The SHAP summary plot (Figure 4) obtained from the trained XGBoost model revealed the below observations on our dataset: (1) *Demographic features*: males were observed to have a higher likelihood of mortality compared to females. Also, a positive correlation was observed between age and mortality (i.e., higher the age, more is the risk to mortality). (2) *Comorbidity features*: The occurrence of kidney disease, diabetes, or heart disorders in a patient was noted to significantly increase the mortality risk of a patient. Other comorbidity such as thyroid disorders, hypertension, and tuberculosis also played a moderate role in increasing the mortality risk of a patient as observed from the SHAP plot. (3) *Symptoms*: Breathlessness, diarrhea, and body weakness stood out among the symptoms as the ones most strongly linked to an increased risk of mortality. Although cough was observed to have a little correlation with mortality, it was not thought to be as important a characteristic as the others. (4) *Treatment*: Among drugs used for COVID-19 patients, enoxaparin, remdesivir, and ivermectin did not prevent mortality. These drugs were observed to be in the favor of mortality class, while the other drugs including zinc, vitamin C, and vitamin D were the most important features for the survival class.

We used LIME for the local interpretability (patient-level interpretability) of the model. We have presented LIME results on two samples – one of mortality and the other of a discharged patient. The former patient (Figure 5) had comorbidity of heart disease and tuberculosis and reported the symptom of body weakness. Patient was administered several drugs, namely, enoxaparin and remdesivir. Our model also assigned a very high mortality score (100%) to this patient. In the case of latter patient (Figure 6), we clearly observed that the patient did not have any major symptoms and comorbidities and was not administered any drug such as enoxaparin or remdesivir. ML model assigned a very high discharge score (98%) to this patient.

Table 3. Classification report after applying sampling techniques like random under/oversampling, ADASYN, and SMOTE

Sampling	Classifier	wF1-score	MCC	Accuracy	ROC-AUC	Sensitivity	Specificity
Over-sampling	**RandomForest**	0.914 ± 0.006	0.546 ± 0.040	0.914 ± 0.005	0.903 ± 0.017	**0.953 ± 0.005**	0.591 ± 0.065
	XGBoost	0.894 ± 0.009	0.554 ± 0.025	0.881 ± 0.013	**0.931 ± 0.015**	0.890 ± 0.019	0.804 ± 0.057
Undersampling	**RandomForest**	0.853 ± 0.013	0.478 ± 0.025	0.826 ± 0.018	0.920 ± 0.019	0.824 ± 0.024	**0.844 ± 0.051**
	XGBoost	0.855 ± 0.007	0.479 ± 0.030	0.828 ± 0.010	0.921 ± 0.019	0.826 ± 0.012	0.841 ± 0.052
ADASYN	**RandomForest**	0.913 ± 0.003	0.552 ± 0.026	0.911 ± 0.002	0.904 ± 0.019	0.945 ± 0.004	0.628 ± 0.047
	XGBoost	0.905 ± 0.010	0.556 ± 0.057	0.897 ± 0.011	0.919 ± 0.018	0.916 ± 0.006	0.733 ± 0.076
SMOTE	**RandomForest**	0.916 ± 0.009	0.565 ± 0.053	**0.915 ± 0.008**	0.902 ± 0.018	0.949 ± 0.001	0.633 ± 0.071
	XGBoost	**0.917 ± 0.011**	**0.588 ± 0.055**	0.913 ± 0.012	0.922 ± 0.019	0.938 ± 0.008	0.702 ± 0.058

Note: *=XGBoost as base estimator; **=Decision Tree as base estimator; MCC = Matthew's Correlation Coefficient

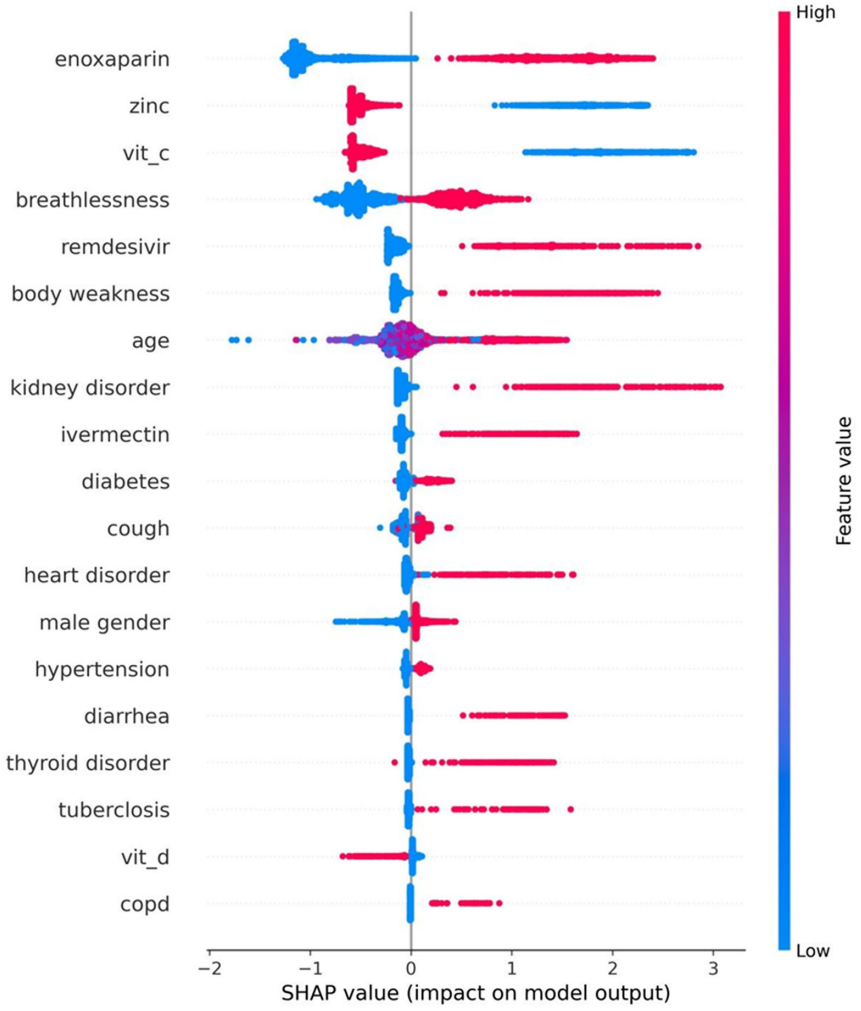

Figure 4. Global interpretability SHAP summary plot using the XGBoost machine learning model. Each point on the plot represents the SHAP value for an individual patient, indicating the contribution of each feature to the prediction of patient mortality. The y-axis lists features ranked by their overall importance in predicting mortality, from highest (top) to lowest (bottom). The x-axis indicates the SHAP value; positive values (right side) suggest an increased risk of mortality, whereas negative values (left side) indicate a decreased risk. The color gradient represents feature values from low (blue) to high (red). For instance, features like enoxaparin, breathlessness, remdesivir, age, kidney disorder, and ivermectin have high feature values (red dots) on the positive SHAP value side, suggesting these factors increase mortality risk. Conversely, features such as zinc, vitamin C, and vitamin D have high values (red dots) on the negative side, indicating their protective effect against mortality

5. Discussion

Our study showed the potential of XAI models in providing the valuable insights into the clinical decision-making process, thereby identifying factors that contribute to outcomes. Consequently, XAI models can serve as a valuable complement to RCT, providing more information and enhancing our understanding of drug efficacy causal relationships, particularly in situations where conducting an RCT may be impractical or unfeasible. Notably, XAI models can be used in assessing the efficacy of drugs when the traditional RCT approach encounters limitations.

In our study, we found that among all the ML algorithms, XGBoost classifier outperformed with demographic, comorbidity, symptomatic, and treatment features. From XAI, it is evident that the *demographic feature* of age was significant, with individuals of higher age are shown to have an increased likelihood of mortality. This observation aligns with well-established findings from prior studies that indicate the older population is more susceptible to severe COVID-19 outcomes and mortality. Furthermore, it is observed that males face double the risk of COVID-19 compared to their female counterparts. Hence, the mortality is higher in male gender [18].

From the SHAP summary plot of the XAI model, it is shown that *Comorbidity factors* like kidney disease, diabetes, heart disease, hypertension, thyroid disorder, and tuberculosis contributed to mortality in COVID-19 patients. These results were true based on the latest studies [19–21]. In the meta-analysis study of [20], it was found that the presence of cardiovascular, cerebrovascular, and kidney-related comorbidities in COVID-19 patients is strongly associated with an elevated risk of mortality [21].

Next, we checked the literature for the feature of *Symptoms*. A study with 100 COVID-19 hospital patients was done in Northern Ethiopia. The researchers discovered that individuals who experienced shortness of breath and bodily weakness had

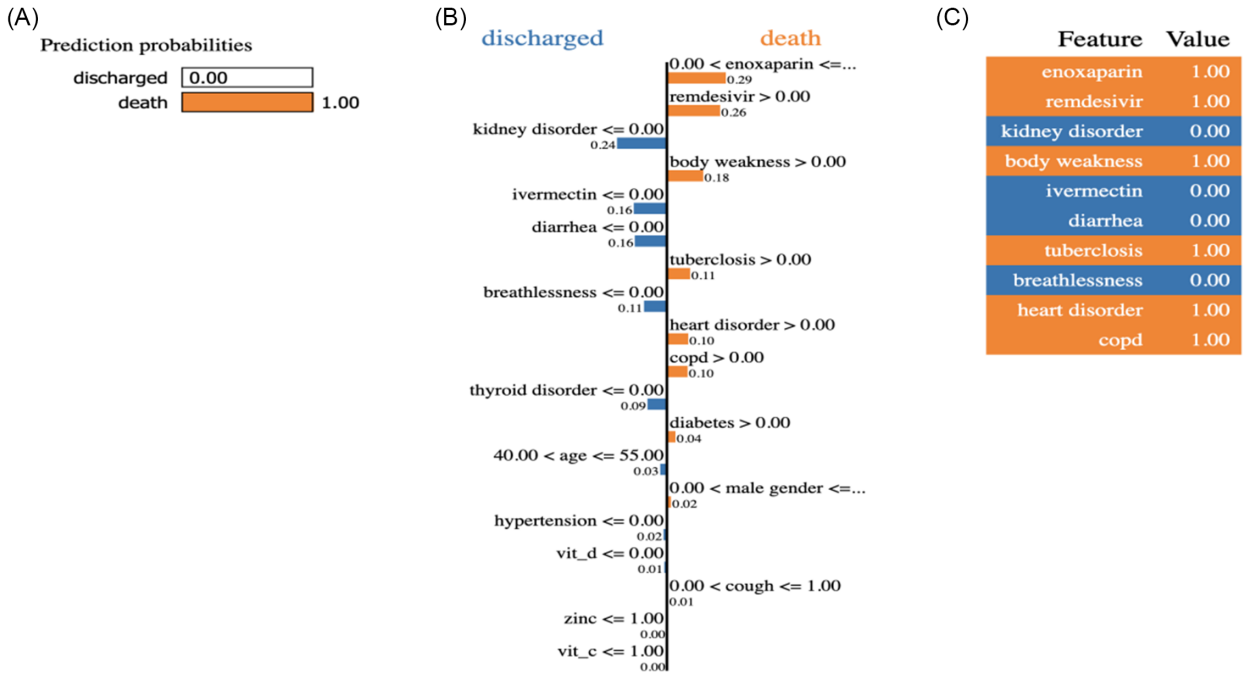

Figure 5. Local interpretability using LIME on a dead patient. (A) It shows the prediction probabilities of the class for a given patient. Here, it is predicted to be a 100% death class. (B) Features on the right side of the central line (shown with orange color) have a higher impact on mortality. Treatment using enoxaparin and remdesivir, presence of body weakness, heart disease, and tuberculosis indicate mortality for this patient). Features on the left side of the central line (shown in the blue color) have an impact on the survival. Absence of the disorders of kidney, thyroid disorder, diarrhea, and breathlessness indicate the survival of this patient. Value of each feature indicates its impact on the outcome. (C) It presents feature-importance rank in the form of a table for this particular patient, where value 1 indicates that the feature is present, while the value 0 indicates that the feature is absent

considerably increased mortality odds [19]. This is confirmed by our model as well because breathlessness is observed as the fourth ranked in features correlating with the mortality class in our study.

The *treatment* enoxaparin, ivermectin, and remdesivir, which were intended to prevent mortality, exhibited a positive association with mortality in this study. Our study utilizing the XAI model revealed that these drugs do not prevent mortality effectively, while other features such as vitamin C, vitamin D, and zinc are associated with improved survival rates. During the initial stages of the pandemic, case studies and observational studies in the literature indicated that anticoagulation therapy could potentially decrease the severity and reduce mortality rates among hospitalized COVID-19 patients [22]. Therefore, despite the lack of RCTs, enoxaparin was routinely used in severe COVID-19 patients. Consequently, prophylactic-dose enoxaparin was widely adopted as the standard of care and was included in the WHO treatment guidelines as a recommended therapy [23]. Based on these limited evidence, clinicians administered enoxaparin to severe COVID-19 patients, which led to an increased risk of bleeding and hence mortality rather than reducing the mortality [24]. In the later stages of COVID-19, the RCT was done [25] and a meta-analysis publication [26] was also published. These studies did not support the use of anticoagulant to lower mortality in all COVID-19 patients including those who were critically ill. These studies reported that the greater anticoagulant doses increased the risk of bleeding, while decreasing thrombotic events [25, 26]. Indeed, our XAI model revealed that the presence of drug enoxaparin favored mortality class, indicating that enoxaparin could not prevent mortality in severe COVID-19 patients. Thus, the results of our XAI model about the use of enoxaparin are consistent and proven with these latest RCTs and meta-analysis [25, 26]. In fact, XAI can provide important insights about the factors that may impact outcomes in COVID-19 patients in the early stages of a pandemic.

Similarly, our study has shown that remdesivir and ivermectin drugs used on COVID-19 patients were actually not beneficial in preventing the mortality. These results of remdesivir and ivermectin, of not being beneficial in preventing mortality, were confirmed from a systematic review [27]. Similarly, based on the SHAP values obtained from our XAI model, vitamin C and zinc were beneficial in reducing the severity and mortality in COVID-19 patients and thus had an overall beneficial effect. These results were also validated in the later studies [28].

These findings suggest that XAI cannot be a replacement for RCTs, but rather is a compliment to observational studies in cases where RCTs are not feasible or practical. Analysis generated by XAI provides better initial insights similar to those of observational studies. Furthermore, XAI provides the outcome prediction score with a significance level of each feature via SHAP score. SHAP summary plot has an important advantage over observational studies. SHAP values provide a way to interpret the results of the AI model, giving a more nuanced understanding of how different factors contribute to the overall predictions made by the model. This information can help healthcare professionals and researchers to better understand the factors that are most strongly associated with poor outcomes in COVID-19 patients. Thus, our novel approach to use XAI model as supplementary to observational study is clearly beneficial. If the XAI model is properly validated with domain knowledge, then this XAI model can be used in clinical scenario. Our study

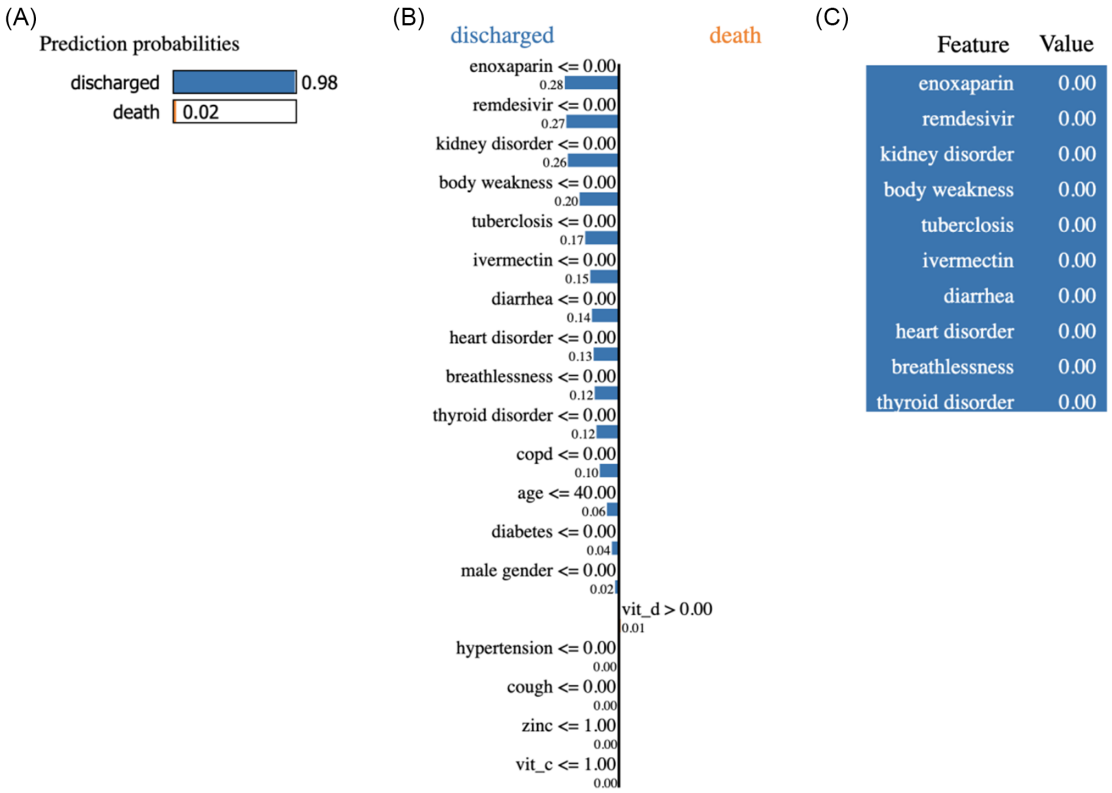

Figure 6. Local Interpretability using LIME on a discharged patient. (A) It shows the prediction probabilities of the class for a patient. Here, the patient is predicted to have 98% chances of discharge. (B) Right side of the central line indicates the impact of features on mortality. Left side of the central line indicates the impact of features on survival. (C) It presents feature-importance rank in this particular patient, where value 1 indicates that the feature is present, while value 0 indicates that the feature is absent

demonstrates that a XAI model without external validation can prove to be equivalent or better than an observational study.

When RCT evidence of drug efficacy is not available, clinical judgment can be made with the help of XAI models. Even though XAI is not a replacement of RCT, XAI developed using data without validation is valuable and can consequently have direct impact on the lives of patients [29, 30]. It is relatively easier to comprehend and explain the predictions of a XAI model using large dataset with existing domain knowledge. This enables healthcare experts to make reasonable and data-driven decisions to provide personalized decisions that can ultimately lead to higher quality of service in healthcare. This study also has a limitation. Since we included only the digitally available data, the ML model was built on an imbalanced dataset, wherein majority of the patients belonged to the mortality class. Also, the dataset had an uneven number of males and females, with males being twice as many as females. This imbalance of the data could be the cause that our model showed that males were more susceptible to mortality. Due to this gender difference, it is important to be careful when interpreting these results. Nevertheless, care was taken to build a robust model that took care of this class imbalance.

6. Conclusion

Our study demonstrates the potential of XAI models as complementary tools to observational studies, particularly in scenarios where RCTs are impractical or infeasible. The XAI approach identified critical factors influencing mortality among COVID-19 patients, highlighting demographic variables like age and gender, as well as key comorbidities and symptomatic presentations. Interestingly, our analysis revealed that commonly administered treatments such as enoxaparin, ivermectin, and remdesivir were associated with increased mortality, contradicting their intended therapeutic roles, while vitamin C, vitamin D, and zinc showed protective effects.

Limitations of our study include combining data from two distinct COVID-19 waves, potentially introducing variability due to evolving treatment protocols across waves. Additionally, the lack of severity-level stratification prevented more granular analysis of treatment effectiveness based on disease severity. Variables such as obesity, education level, or socioeconomic status were excluded due to incomplete data, representing a limitation in comprehensively understanding patient outcomes.

Future research should validate our XAI findings using external datasets and explore the integration of additional socioeconomic and lifestyle variables to refine model accuracy and applicability in clinical scenarios.

Acknowledgement

We thank Medical Director, Lok Nayak Hospital, New Delhi, India, for providing the COVID-19 data, and Dr. Sapna, Chief Medical Officer, for her assistance in data collection. All authors would like to thank the Centre for Excellence in Healthcare at IIIT-Delhi for providing the technical support.

Ethical Statement

Ethics approval for this study was granted by the Institutional Ethics Committee, Maulana Azad Medical College, New Delhi, India, with reference number F.I/IEC/MAMC/84/02/2021/no 382, dated May 18, 2021.

Conflicts of Interest

The authors declare that they have no conflicts of interest to this work.

Data Availability Statement

The data that support the findings of this study are openly available in GitHub at https://github.com/Survival-SBILab/Leveraging-XAI-for-Drug-Efficacy-and-Mortality-Analysis-Insights-from-an-Observational-Dataset. The data that support this work are available upon reasonable request to the corresponding author.

Author Contribution Statement

Manu Kumar Shetty: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Resources, Data curation, Writing – original draft, Writing – review & editing, Visualization, Project administration. **Aaloke Mozumdar:** Methodology, Software, Validation, Formal analysis, Writing – original draft, Writing – review & editing, Visualization. **Saurabh Gupta:** Methodology, Software, Validation, Formal analysis, Writing – original draft, Writing – review & editing, Visualization. **Lalit Gupta:** Investigation, Resources, Data curation, Writing – original draft, Writing – review & editing, Supervision, Project administration. **Kapil Chaudhary:** Investigation, Resources, Data curation, Writing – review & editing, Writing – review & editing, Supervision, Project administration. **Vandana Roy:** Conceptualization, Investigation, Resources, Data curation, Writing – review & editing, Supervision. **Suresh Kumar:** Investigation, Resources, Data curation, Writing – review & editing, Supervision, Project administration. **Bhupinder Singh Kalra:** Investigation, Resources, Data curation, Writing – review & editing, Supervision, Project administration. **Sanjeev Khanth PE:** Formal analysis, Resources. Data curation, Writing – review & editing. **Anubha Gupta:** Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Writing – original draft, Writing – review & editing, Visualization, Supervision, Project administration.

References

[1] Hariton, E., & Locascio, J. J. (2018). Randomised controlled trials—The gold standard for effectiveness research. *BJOG: An International Journal of Obstetrics and Gynaecology*, *125*(13), 1716. https://doi.org/10.1111/1471-0528.15199

[2] Franklin, J. M., Platt, R., Dreyer, N. A., London, A. J., Simon, G. E., Watanabe, J. H., . . . , & Califf, R. M. (2022). When can nonrandomized studies support valid inference regarding effectiveness or safety of new medical treatments? *Clinical Pharmacology & Therapeutics*, *111*(1), 108–115. https://doi.org/10.1002/cpt.2255

[3] Zhou, Q., Chen, Z. H., Cao, Y. H., & Peng, S. (2021). Clinical impact and quality of randomized controlled trials involving interventions evaluating artificial intelligence prediction tools: A systematic review. *NPJ Digital Medicine*, *4*(1), 154. https://doi.org/10.1038/s41746-021-00524-2

[4] Talari, K., & Goyal, M. (2020). Retrospective studies–utility and caveats. *Journal of the Royal College of Physicians of Edinburgh*, *50*(4), 398–402. https://doi.org/10.4997/jrcpe.2020.409

[5] Blonde, L., Khunti, K., Harris, S. B., Meizinger, C., & Skolnik, N. S. (2018). Interpretation and impact of real-world clinical data for the practicing clinician. *Advances in Therapy*, *35*, 1763–1774. https://doi.org/10.1007/s12325-018-0805-y

[6] Singhal, S., Bansal, S., Negi, A., Kalra, B. S., Gupta, L., Garg, S., . . . , & Chawla, S. (2022). Drug utilization pattern in the treatment of severe acute respiratory syndrome coronavirus-2 (SARS CoV 2) patients at dedicated COVID tertiary care teaching hospital: An observational study. *MAMC Journal of Medical Sciences*, *8*(3), 218–223.

[7] Gueyffier, F., & Cucherat, M. (2019). The limitations of observation studies for decision making regarding drugs efficacy and safety. *Therapies*, *74*(2), 181–185. https://doi.org/10.1016/j.therap.2018.11.001

[8] Criner, G. J., Connett, J. E., Aaron, S. D., Albert, R. K., Bailey, W. C., Casaburi, R., . . . , & Lazarus, S. C. (2014). Simvastatin for the prevention of exacerbations in moderate-to-severe COPD. *New England Journal of Medicine*, *370*(23), 2201–2210. https://doi.org/10.1056/NEJMoa1403086

[9] Payrovnaziri, S. N., Chen, Z., Rengifo-Moreno, P., Miller, T., Bian, J., Chen, J. H., . . . , & He, Z. (2020). Explainable artificial intelligence models using real-world electronic health record data: A systematic scoping review. *Journal of the American Medical Informatics Association*, *27*(7), 1173–1185. https://doi.org/10.1093/jamia/ocaa053

[10] Saeed, W., & Omlin, C. (2023). Explainable AI (XAI): A systematic meta-survey of current challenges and future opportunities. *Knowledge-based Systems*, *263*, 110273. https://doi.org/10.1016/j.knosys.2023.110273

[11] Qiu, W., Chen, H., Dincer, A. B., Lundberg, S., Kaeberlein, M., & Lee, S. I. (2022). Interpretable machine learning prediction of all-cause mortality. *Communications Medicine*, *2*(1), 125. https://doi.org/10.1038/s43856-022-00180-x

[12] Seabold, S., & Perktold, J. (2010). Statsmodels: Econometric and statistical modeling with python. In *The Proceedings of the 9th Python in Science Conference*, 92–96.

[13] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., . . . , & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

[14] Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, *16*, 321–357. https://doi.org/10.1613/jair.953

[15] Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. In *NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems*, 4768–4777. https://dl.acm.org/doi/abs/10.5555/3295222.3295230

[16] Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?" Explaining the predictions of any classifier. In *KDD '16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135–1144. https://doi.org/10.1145/2939672.2939778

[17] He, H., Bai, Y., Garcia, E. A., & Li, S. (2008). ADASYN: Adaptive synthetic sampling approach for imbalanced learning. In *2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*, 1322–1328. https://doi.org/10.1109/IJCNN.2008.4633969

[18] Kebede, F., Kebede, T., & Gizaw, T. (2022). Predictors for adult COVID-19 hospitalized inpatient mortality rate in North West Ethiopia. *SAGE Open Medicine*, *10*, 20503121221081756. https://doi.org/10.1177/20503121221081756

[19] Sifaat, M., Patel, P., Sheikh, R., Ghaffar, D., Vaishnav, H., Nahar, L., . . . , & Quadri, S. (2022). Cardiorenal disease in COVID-19 patients. *Journal of the Renin-Angiotensin-Aldosterone System*, *2022*, 4640788. https://doi.org/10.1155/2022/4640788

[20] Lee, A. C., Li, W. T., Apostol, L., Ma, J., Taub, P. R., Chang, E. Y., . . . , & Ongkeko, W. M. (2021). Cardiovascular, cerebrovascular, and renal co-morbidities in COVID-19 patients: A systematic-review and meta-analysis. *Computational and Structural Biotechnology Journal*, *19*, 3755–3764. https://doi.org/10.1016/j.csbj.2021.06.038

[21] Péterfi, A., Mészáros, Á., Szarvas, Z., Pénzes, M., Fekete, M., Fehér, Á., . . . , & Fazekas-Pongor, V. (2022). Comorbidities and increased mortality of COVID-19 among the elderly: A systematic review. *Physiology International*, *109*(2), 163–176. https://doi.org/10.1556/2060.2022.00206

[22] Albani, F., Sepe, L., Fusina, F., Prezioso, C., Baronio, M., Caminiti, F., . . . , & Natalini, G. (2020). Thromboprophylaxis with enoxaparin is associated with a lower death rate in patients hospitalized with SARS-CoV-2 infection. A cohort study. *eClinical Medicine*, *27*, 100562. https://doi.org/10.1016/j.eclinm.2020.100562

[23] Cuker, A., Tseng, E. K., Nieuwlaat, R., Angchaisuksiri, P., Blair, C., Dane, K., . . . , & Schünemann, H. J. (2021). American Society of Hematology living guidelines on the use of anticoagulation for thromboprophylaxis in patients with COVID-19: May 2021 update on the use of intermediate-intensity anticoagulation in critically ill patients. *Blood Advances*, *5*(20), 3951–3959. https://doi.org/10.1182/bloodadvances.2021005493

[24] Al-Samkari, H., Gupta, S., Karp Leaf, R., Wang, W., Rosovsky, R., Bauer, K., & Leaf, D. (2020). Thrombosis, bleeding, and the effect of anticoagulation on survival in critically ill patients with COVID-19. *Annals of Internal Medicine*, *174*(5), 622–632. https://doi.org/10.7326/M20-6739

[25] Mazloomzadeh, S., Khaleghparast, S., Ghadrdoost, B., Mousavizadeh, M., Baay, M. R., Noohi, F., . . . , & INSPIRATION Investigators. (2021). Effect of intermediate-dose vs standard-dose prophylactic anticoagulation on thrombotic events, extracorporeal membrane oxygenation treatment, or mortality among patients with COVID-19 admitted to the intensive care unit: The INSPIRATION randomized clinical trial. *Jama*, *325*(16), 1620–1630. https://doi.org/10.1001/jama.2021.4152

[26] Bonfim, L. C., Guerini, I. S., Zambon, M. G., Pires, G. B., Silva, A. C., Gobatto, A. L., . . . , & Brosnahan, S. B. (2023). Optimal dosing of heparin for prophylactic anticoagulation in critically ill COVID-19 patients a systematic review and meta-analysis of randomized controlled trials. *Journal of Critical Care*, *77*, 154344. https://doi.org/10.1016/j.jcrc.2023.154344

[27] Batista, D. R., Floriano, I., Silvinato, A., Bacha, H. A., Barbosa, A. N., Tanni, S. E., & Bernardo, W. M. (2022). Use of anticoagulants in patients with COVID-19: A living systematic review and meta-analysis. *Jornal Brasileiro de Pneumologia*, *48*(04), e20220041. https://doi.org/10.36416/1806-3756/e20220041

[28] Pedrosa, L. F., Barros, A. N., & Leite-Lais, L. (2022). Nutritional risk of vitamin D, vitamin C, zinc, and selenium deficiency on risk and clinical outcomes of COVID-19: A narrative review. *Clinical Nutrition ESPEN*, *47*, 9–27. https://doi.org/10.1016/j.clnesp.2021.11.003

[29] Chadaga, K., Prabhu, S., Umakanth, S., Sampathila, N., & Chadaga, R. (2021). COVID-19 mortality prediction among patients using epidemiological parameters: An ensemble machine learning approach. *Engineered Science*, *16*(10), 221–233. http://dx.doi.org/10.30919/es8d579

[30] Stiglic, G., Kocbek, P., Fijacko, N., Zitnik, M., Verbert, K., & Cilar, L. (2020). Interpretability of machine learning-based prediction models in healthcare. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *10*(5), e1379. https://doi.org/10.1002/widm.1379

How to Cite: Shetty, M. K., Mozumdar, A., Gupta, S., Gupta, L., Chaudhary, K., Roy, V., . . . , & Gupta, A. (2025). Leveraging Explainable AI for Drug Efficacy and Mortality Analysis: Insights from an Observational Dataset. *Medinformatics*. https://doi.org/10.47852/bonviewMEDIN52025243