Engineering Systems and Intelligent Technologies
ESIT

Journal Homepage
https://pub.scientificirg.com/index.php/ESIT

A Machine Learning Framework for COVID-19 Mortality Risk Prediction Using Clinical and Demographic Data

Rehab Ahmed[a,1] **, Khalid S. Alqarni** [b]**, Yasser AbdelSatar**[c]**, Amr A. Hassanain**[c]

[a] *Faculty of Computers and Artificial Intelligence, Sohag University, Sohag 82524, Egypt. E-mail: rehabahmed23234@gmail.com.*
[b] *College of Computers and Information Technology, Taif University, Saudi Arabia. E-mail:Ksalqarni99@gmail.com.*
[c] *Department of Artificial Intelligence, Faculty of Computers and Artificial Intelligence,Sphinx University, Assiut, 71511, Egypt. E-mail: yasser.selim@sphinx.edu.eg, Amr.Hassanain@sphinx.edu.eg.*

ABSTRACT

The COVID-19 pandemic revealed an immediate need for trustworthy mortality prediction tools for effective allocation of health care resources. This paper presents the creation and validation of a machine learning model for predicting mortality risk using clinical and demographic information from the Mexican national registry (1,048,575 records, February 2020-October 2021). In this study we utilized six different algorithms: Logistic Regression, Support Vector Machine, Random Forest, AdaBoost, Decision Tree, and Deep Neural Networks. The models were taught 21 clinically relevant variables (including comorbidities, treatment, and demographic factors). For model assessment, we implemented stratified 80-20 training/testing splits with 5-fold cross-validation, and measured accuracy, F1-score, precision, and recall. The best value of accuracy (0.9181) was obtained by the Deep Neural Networks, and by Logistic Regression with value 0.9107 and a better value of interpretability. The model performance difference was only 0.7 percentage points which implies that the interpretable models strike a good deal for clinical usage. The more interpretable a model is; the more a clinician is likely to trust it. Importance of certain features measures (pneumonia, hospitalization type, age) were found to be predictors of interest, clinically plausible. The framework shows the value of clinically useful interpretable ML models and offers support for their use in sensitive medical decision making.

PAPER INFORMATION

HISTORY
Received: 30 October 2025
Revised: 15 December 2025
Accepted: 17 February 2026
Online: 27 February 2026

MSC
62K05; 62K15

KEYWORDS
Deep Learning;
Medical Prognosis;
COVID-19 Risk;
Machine Learning;
Predictive Healthcare.

1 Introduction

The COVID-19 pandemic, caused by the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), has been one of the most profound public health crises in history. The virus was first reported in Wuhan, China, in December of 2019, and soon crossed all borders. It wasn 't until March 11, 2020, that the World

[1]Faculty of Computers and Artificial Intelligence, Sohag University, Sohag 82524, Egypt.
E-mail: rehabahmed23234@gmail.com.

Health Organization classified the outbreak as a pandemic [1]. In 2021, confirmed cases of COVID-19 were reported at approximately 270 million and over 5.3 million deaths. However, these numbers are most certainly underreported [2]. The pandemic also resulted in devastating stress on the health care system worldwide and revealed a lack of hospital, acute care, and critical care resources and medical supplies [3]. The demand on the health care system was unprecedented, and that created a need for tools that supported clinical decision-making by identifying patients at risk for decline and death [4].

Prediction modeling in medicine has undergone a considerable amount of change in the past two decades, moving away from solely combing traditional statistics and toward the use and integration of machine learning algorithms [5]. For numerous years, clinical prediction relied on regression modeling, specifically Logit regression [6, 7]. While models such as these hold promise in the prediction space from a transparency and interpretability point of view, they do poorly when a predictions success relies on the models ability to navigate and capture complex, interwoven, nonlinear relationships [8]. The use of and reliance on machine learning as a means of capturing and defining patterns in datasets of greater dimensionality from a clinical perspective has been a game changer. Clinical prediction modeling has also benefited from the use of ensemble learning methods including Random Forest and Gradient Boosting [9]. Deep learning (DL) models are, however, the more current customary predictors of choice when the predictive task at hand is the analysis of medical images (and to a lesser but increasing extent, in the prediction of structured clinical data) [10]. The COVID-19 pandemic has been a real world opportunity for the implementation of machine learning in the clinical space (and more specifically, predictive modeling) to assist in resolving public health crises. To provide a high-level descriptive context beyond the patient-level registry, COVID-19 cases, deaths, and recoveries per million in the Mexico database in 2020 were summarized by continent as shown in **Figure 1**.

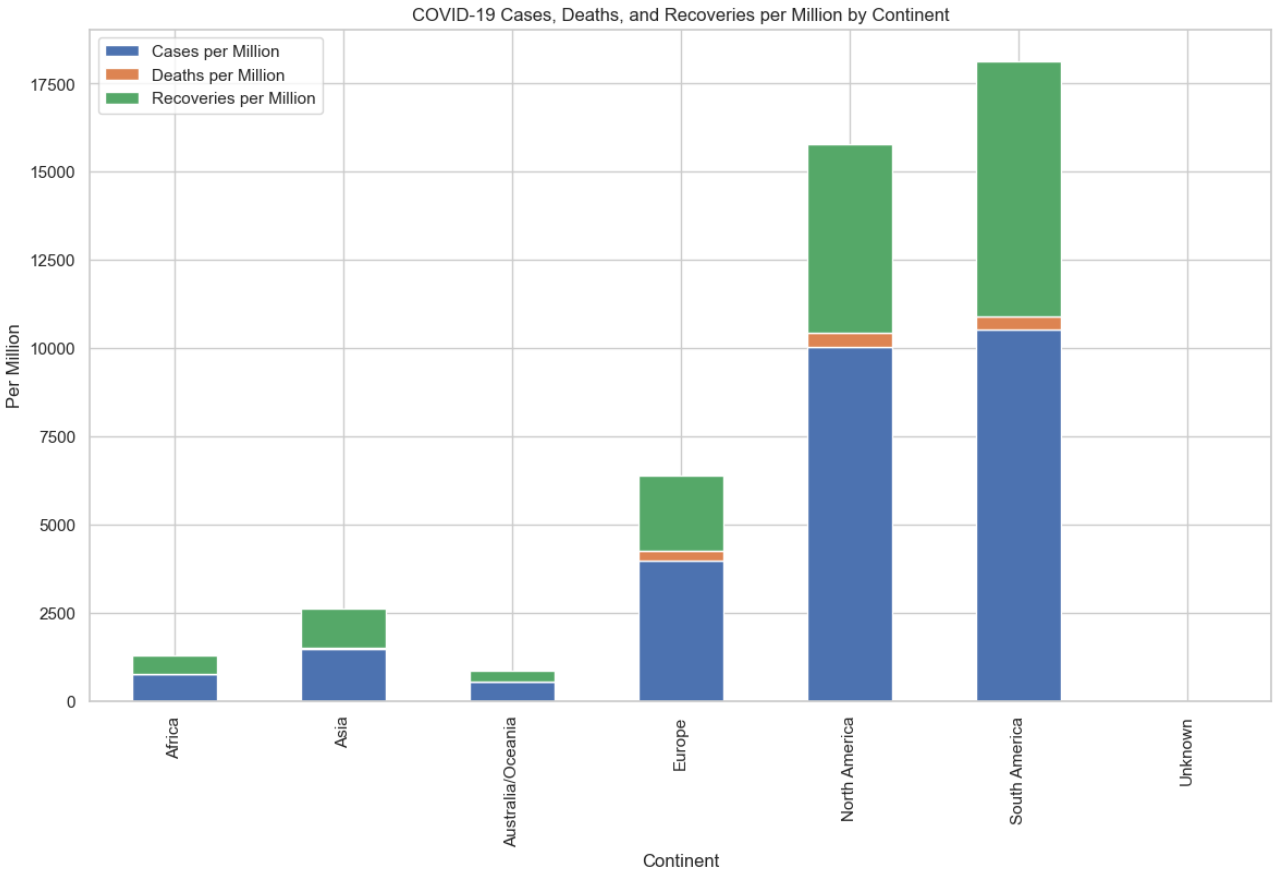


Figure 1: Continent-level distribution from COVID-19 Mexico database.

Machine learning has been used for prognostication of COVID-19 in many research studies. The initial studies used only radiographic data where COVID-19 related pneumonia were identified in chest X-ray and CT scan images using convolutional neural networks (CNNs) [11]. Although these studies assisted and improved the diagnostic process and workflow, they were not designed for and hence did not predict mortality, which would involve more complex resource planning than most health systems could afford. Later studies proposed models using routine clinical and demographic data that are collected in all health care systems in the > 95% of the patients they examine. One of those studies was done by Liang et al. where they designed a deep learning

model using 1,590 patients data collected from 575 hospitals in China which achieved an AUC of 0.88 for predicting critical illness.[12] Along the same lines is the study done by Yan et al. [13] where a random forest model was designed using laboratory markers of 485 patients and achieved a predictive accuracy of 90% for mortality. Although these studies proved a model of mortality predicting COVID-19, the small number of patients (region/city specific) was a great limitation.

The literature demonstrates some remaining challenges concerning the predictive modeling of COVID-19. Most of the newer articles lack strong methodology concerning missing data,These methodological concerns are further illustrated in **Figure 2**, which summarizes the distribution of risk-of-bias assessments across evaluated domains. some articles don't even justify the use of temporal validation, and external validation is also usually underutilized [14]. Next, there is wide variability in the predictors used in the modeling of the COVID-19 outcome predictors as some of the clinical variables are known and some are not. Additionally, there are multiple modeling frameworks that have been proposed for the clinical utility and the most critical is the poor implementation requirement as poorly implemented frameworks are known to lack in clinical utility [15]. Most of the COVID-19 predictive analytics frameworks are 'black box' models that are poorly implemented without an explanation and the opaque nature of the models has been an impediment to the use models in clinical practice. Clinicians lack the confidence to apply the modeling prediction analytics to clinical decision making and this is the most critical aspect that the models should have [16]. Clinical machine learning faces a

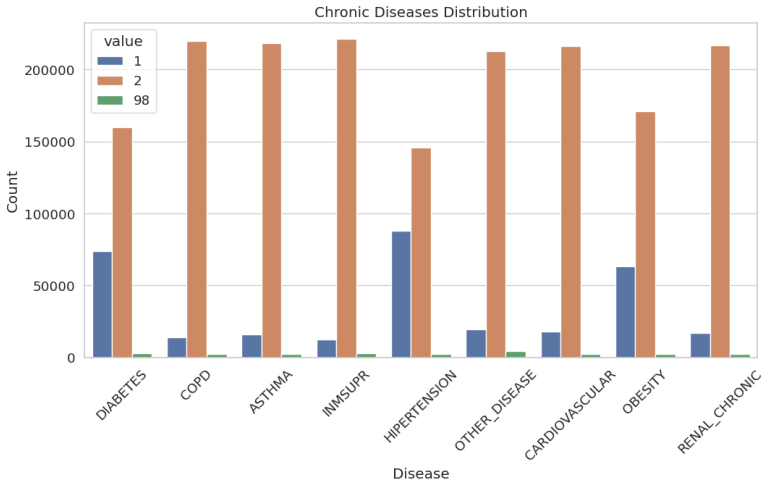


Figure 2: Chronic disease of data distribution of covid-19 Mexico database.

challenge when attempting to balance interpretability and complexity of the model. Caruana et al. [17] state that even though ensemble methods and deep neural networks often have excellent predictive accuracy, the decision-making process becomes a black-box and is hard to interpret, which in turn causes a lack of trust and accountability from the clinician. Simplistic models such as logistic regression may provide explanatory power, but their accuracy may leave much to be desired when complex predictive tasks are considered. The problem is more pronouncd in the case of medicine, especially when wrong predictions may have a disastrous impact on the state of the patient. Explainable Artificial Intelligence, or XAI, is a relatively newer domain that aims to provide techniques to improve interpretability of models. One such example is the use of SHAP (SHapley Additive exPlanations) values and LIME (Local Interpretable Model-agnostic Explanations) (Lundberg et al. 2017). Despite the fact that such models may provide (post-hoc) explanations to the users, these explanations are more or less approximations to a model's behavior and are not a substitute for true interpretability and may very well, in fact, increase the uncertainty present in the clinical interpretation.

The COVID-19 pandemic has brought to the fore the need for modelers to have consideration for how representative the population is for the purpose of the modeling as well as the population's data diversity [18]. Modeling studies published during the pandemic often relied on data from one or two geographic regions or healthcare systems. Such models would have limited utility for populations that differ in demographics, comorbidities, and healthcare resources. This is especially the case due to the large disparities observed in the COVID-19 pandemic's impact and outcomes across different racial, ethnic, and socioeconomic populations [19]. There is a risk that models, developed from datasets that are too homogenous, will result in sustaining and even worsening health inequities if such models are utilized without prior validation and calibration for the target population and/or other diverse populations [20]. For prediction models to be robust and generalizable, their intended use populations should be diverse, large, and representative of the intended model population [21, 22].

The COVID-19 prediction research needs good data that comes from serious impact Mexico experienced with the pandemic and also their unusual demographics. By October 2021, in Mexico, the confirmed cases were 3.7 million and the deaths were about 280,000. However, the true deaths from COVID-19 may be much higher due to the estimates from excess mortality [23]. Mexico has a population with the high levels of some comorbidities that may lead to severe outcomes from COVID-19. These comorbidities include diabetes (16%), hypertension (25%), and obesity (36%) [24]. Mexico's healthcare system is also a challenge to predicting COVID-19 impact accurately, as the system is a mixture of public and private healthcare options that makes it difficult to standardize care and accurate impact prediction[25, 26] . This context makes predicting the impact of COVID-19 critically important for Mexico's healthcare system to make effective decisions about its limited resources.

Though numerous COVID-19 forecasting models have been published, there are still critical gaps in the literature. First, there is a lack of evidence from studies with large, nationally representative, clinical, and demographic datasets. Second, there are few examples of studies that take an in-depth, head-to-head, and rigorous approach with multiple machine learning algorithms applied to the same datasets with the same evaluation criteria. Third, the COVID-19 mortality predictive literature is limited in regarding the balance between predictive accuracy and clinical interpretability. Fourth, literature is scant regarding the cross-modeling approach to understanding clinically predictive robustness beyond the algorithm with respect to the various modeling approaches. Finally, the literature is lacking regarding the resource-poor practical healthcare settings evidence of the model selection for clinical practice.

Each mentioned area of research works to develop and evaluate a machine learning framework to evaluate the predictions of mortality for COVID-19 using a national patient registry for Mexico's COVID-19 patients. This gives us enough concrete reasons to address the following objectives. First, there is a need to evaluate the performance of six machine learning algorithms on a national representative dataset. These range from models based on the description of reasoning, to models based on the description of processing neural networks. Second, there is a need to evaluate the extent to which the predictive and clinical imbalanced trade of the mortality predictions of COVID-19 is justifiable. Third, a need exists to evaluate the clinical predictors of mortality to identify the predictors that consistently performing across different models. We propose that although complex models can achieve a minimal improvement in predictive accuracy, simple models that can be interpreted will, on balance, have the same predictive performance and far greater clinical utility for everyday practice.

This research offers numerous contributions. Methodologically, we present a methodologically sound comparison of machine learning algorithms on one of the largest COVID-19 datasets used to date which allows the largest degree of statistical power and generalizability. Clinically, we provide patterns of feature importance that offer insight that both corroborate and refine understanding of the risk factors of COVID-19. From the implementation perspective, we offer evidence that supports the choice of interpretable models in clinical use where model transparency and trust are critical. From the ethics perspective, using a dataset that is nationally representative improves the predictive framework's fairness and equity compared to models that are trained on more homogeneous datasets.

2 Related Work

The last three years have evidencedd significant adaptation and growth in the field of applying machine learning (ML) to predicting the possible disease trajectories in patients infected with the novel coronavirus. There has been the development of various forms that tackle a plethora of different challenges. This section compiles the important predictive value and methodology of different works and models, and tries to tackle the problematic area of a model's mere existence, and the degree to which a model will be interpretable. This will serve to provide the relevant context to our study and the multitude of works that are of similar nature.

2.1 Machine Learning for COVID-19 Mortality Prediction

ML research has considered mortality prediction as one of its primary focuses. The early models demonstrated that algorithms could accurately identify risks in patients. Take for example a study that utilized electronic health records (EHR) of more than 86,000 patients in Spain. The study employed a specific ensemble logistic regression model for predicting patient status (deceased/discharged) with an accuracy of 90 - 93 % and a ROC-AUC of 0.94. The study pointed out that age and sex and specific comorbidities (e.g. obesity, renal failure) provided highly relevant explanatory variabilities. The study highlighted the significance of predictive analytics that are derived from sufficiently available and administratively provided combined clinical data.

Follow-up studies focused on the predictive capacity of the models and added data from biomarkers. In the study of the Deep Neural Networks (DNN), Random Forest (RF), Support Vector Machine (SVM) models, and from more than five thousand patients immunological and metabolic biomarkers, one of the studies of 2023, focused on the comparison of the three, models measurement, analysis and comparison of the three. Their Random Forest achived the highest ROC-AUC of 0.98 and the selected biomarkers of renal function (glomerular filtration rate, urea), inflammatory (C-reactive protein, procalcitonin), and oxygenation) [28]. This study emphasized the overperformace that could be gained from being data rich, yet it also showed the increased complexity and expense of obtaining an array of biomarkers for routine triage.

In addition to acute COVID-19, ML methods have been used to anticipate death in patients with multi-morbidity, another group heavily affected by the pandemic. One recent example is a 2023 retrospective study that applied Extreme Gradient Boosting (XGBoost) to a set of easily obtainable variables and health resource usage to forecast 4-year mortality with 87% accuracy and an AUC of 0.88[30]. This is a classic example of a recent trend in health systems to develop novel methods and tools using simple real world data.

2.2 The Imperative for Explainability and Interpretability

As predictive modelling gets better, they become more complicated, leading to their classification as "Black Boxes" which result in something called clinical distrust and in turn aversion to use them[29]. Explainable Artificial Intelligence (XAI) ascertains this situation. Studies have shown that the integration of workflow in clinical practice relies auf the interpretability of predictive models; the absence of this crucial element creates resistance.

Some studies have started to carefully incorporate different methods of XAI in their research. One such study used Local Interpretable Model-agnostic Explanations (LIME) to explain And Random Forest models to create a COVID-19 diagnostic tool based on blood testing models, coupled with the GBDT (Gradient Boosted Decision Tree) model that, in isolation, yielded an AUC of 86.4%, and the LIME models were able to provide clinicians Case- specific explanations (predictive) by showing a highlight of an elevated lactate dehydrogenase or a high number of white blood cells)[29]. Other studies have attempted to close the gap between high performance and clinical explainability by incorporating deep learning models with SHAP (SHapley Additive exPlanations). These frameworks used global feature importance and local explaintations at the patient/individual level as a way of offering both.

2.3 Synthesis of Methodological Trends and Identified Gaps

A systematic review of ML prediction models for 2020-2023 suggests that while hybrid and ensemble modeling approaches tend to give the best predictive accuracy, the field continues to struggle with the quality of the data, generalizibility of the models, and the need to strike a balance between model interpretability and predictive performance [31]. This review, along with some bibliography studies, indicates that the field is still evolving, and the central focus of the shifting field is moving away from modeling performance to implement models that are dependable and safe.

Table 1 synthesizes the core characteristics of several pivotal studies discussed, illustrating the evolution in data types, methodologies, and the growing emphasis on explainability.

2.4 Positioning of the Current Study

The current research addresses all three themes outlined above. First, there is a clear need for large-scale validation, as evaluation with single-center or regional cohort datasets [27] is far less generalizable. Most national datasets are limited to regional coherence while the dataset from Mexico is more representative of the entire country. The second response involves what is most likely the first direct, complete, and rigorous comparison of multiple algorithmic families (from simple interpretable linear models to more sophisticated ensembles: Random Forest, AdaBoost, and deep learning) on the same dataset ([32]). Most importantly, this research examines the one case in which the trade-off between predictive performance and model interpretability is a priority. If complex DNNs or XGBoost models ([30], [28]) are generally agreed to report high accuracy, XAI advocates ([29], [18]) seem to overlook the fact that very few research attempts to measure or quantify what is lost when a more interpretable model is used. We are interested in identifying whether the marginal gains from a model classified as a 'black box' are outweighed by a model which is more transparent in its operations while optimising the trade-off between performance and clinical utility in high-stake decision-making scenario's during the allocation of limited or restricted resources.

Table 1: Summary of Key Related Work in ML for COVID-19 Prognosis

Study (Year)	Primary Focus / Contribution	Key Algorithm(s)	Key Performance	Data Type & Size
Cisterna et al. (2022)	Mortality prediction using a novel method (LR-IPIP) to handle extreme class imbalance.	Ensemble Logistic Regression (LR-IPIP)	Acc: 90–93%, AUC: 0.94	Demographics & 19 comorbidities (N=86,867) [27]
Hong Kong Study (2023)	Identification of immunological/metabolic biomarkers for high-accuracy mortality prediction.	Random Forest, DNN, SVM	RF AUC: 0.98 (0.96–0.98)	63 biomarkers from EHR (N=5,059) [28]
Explainable AI Study (2022)	COVID-19 diagnosis using routine blood tests with model-agnostic explanations (LIME).	GBDT, RF, AdaBoost, XGBoost	GBDT AUC: 86.4%	32 blood test indices (N=1,374) [29]
López Seguí et al. (2023)	Mortality prediction in complex chronic patients, highlighting healthcare utilization variables.	XGBoost, Gradient Boosting	XGBoost Acc: 87%, AUC: 0.88	Demographics, diagnoses, resource use (N=264) [30]
Systematic Review (2025)	Analysis of ML trends (2020-2023), highlighting hybrid models and current challenges.	Review of RF, SVM, DNN, LSTM, Hybrids	N/A (Review)	Analysis of 136 studies [31]

3 Methodology

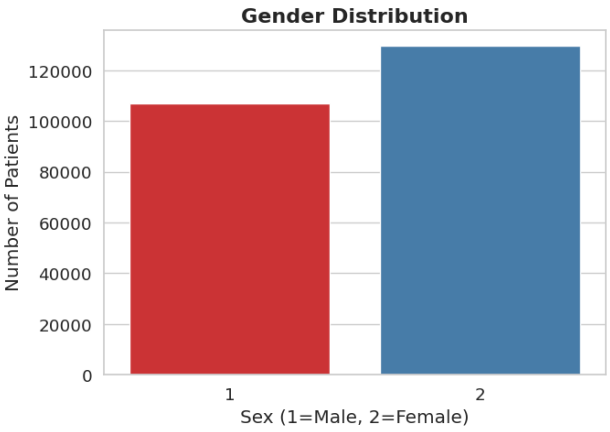
The proposed model followed a retrospective, observational design, analyzing an anonymized national COVID-19 patient registry. The study goal is to provide a supervised machine learning pipeline that predicts whether a patient is high risk from routinely recorded demographics, clinical status, and prior medical conditions. An observational approach is appropriate here because the research question is predictive rather than interventional and relies on existing surveillance data collected at scale. All analytical decisions were fixed before model training to improve transparency and to reduce opportunities for inadvertent information leakage.

3.1 Data Source and Context

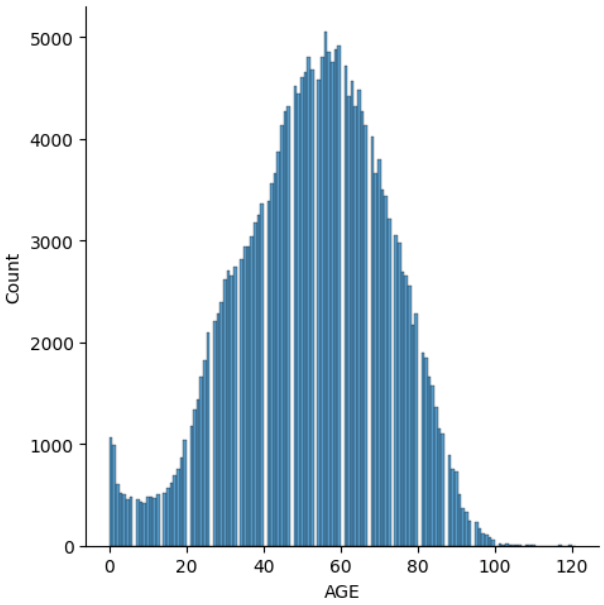
This study utilizes a national COVID-19 patient registry provided by the Mexican Government through the General Directorate of Epidemiology. The dataset was downloaded from Kaggle [32]. The original data were provided by the Mexican government [33]. The raw data contains 1,048,576 anonymized patient records and 21 recorded features, capturing demographics, comorbidities, care pathway indicators, and clinical status variables. Because the study relied exclusively on publicly available, anonymized secondary data, no direct patient contact or prospective enrollment was involved. The dataset encompasses the complete epidemiological surveillance records from February 2020 to October 2021. It was specifically compiled to support research on disease progression and resource allocation during the pandemic. The primary objective of the original data collection was to enable the development of predictive models for identifying high-risk patients requiring intensive medical intervention.

All personally identifiable information was removed prior to public release to ensure patient confidentiality. Boolean variables in the dataset were originally encoded as 1 for "Yes" and 2 for "No," with values 97 and 99 indicating missing data. These were transformed into standard binary encoding (1/0) with appropriate handling of missing values. The target variable, patient mortality, was derived from the date of death field: '9999-99-99' indicated survival, and any other date indicated death. The demographic characteristics of the dataset are summarized in **Figure 3** . Specifically, sex distribution is shown in subfigure (a), age distribution in (b), and patient type distribution in (c).

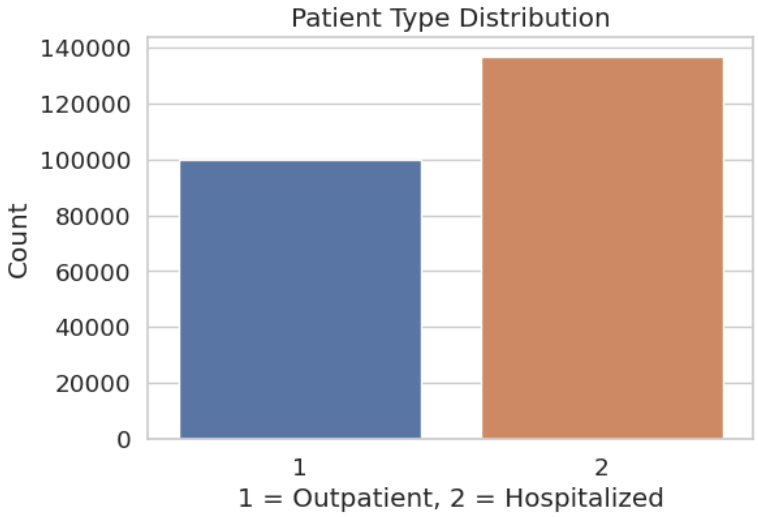
COVID-19 classification follows the Mexican health authority's guidelines, where values 1-3 indicate confirmed cases with increasing severity, and values ≥4 represent negative or inconclusive results. Comorbidity indicators capture the presence of pneumonia, diabetes, hypertension, chronic obstructive pulmonary disease (COPD), asthma, immunosuppression, cardiovascular disease, chronic renal disease, obesity, and tobacco use. Additional clinical variables include pregnancy status and binary indicators for intubation and intensive care unit (ICU) admission. Administrative variables document the treating healthcare facility's level (USMR) and type (medical unit).



(a) Gender distribution



(b) Age distribution



(c) Patient type distribution

Figure 3: Demographic distributions from the COVID-19 Mexico database.

3.2 Data Preprocessing and Feature Engineering

All patient records underwent rigorous quality assessment and preprocessing. The target variable, mortality, was derived from the 'DATE_DIED' field, with dates other than '9999-99-99' indicating death. Categorical variables were one-hot encoded to facilitate model interpretation. Continuous variables, specifically age, were standardized using z-score normalization. Feature importance was analyzed using correlation analysis and recursive feature elimination. This process confirmed the clinical relevance of key predictors, including pneumonia status, age, and comorbidities . Before predictive modeling, inter-feature association patterns were assessed to understand redundancy and potential proxy relationships among variables. A correlation heatmap over the encoded feature set was computed and used as a diagnostic summary of linear associations (**Figure 4**). This step was not used as an automatic filter, since correlated clinical variables can remain independently relevant and removing them can reduce clinical interpretability. The heatmap was primarily used to guide interpretation and to flag clusters of highly related features that might warrant careful discussion in the results.

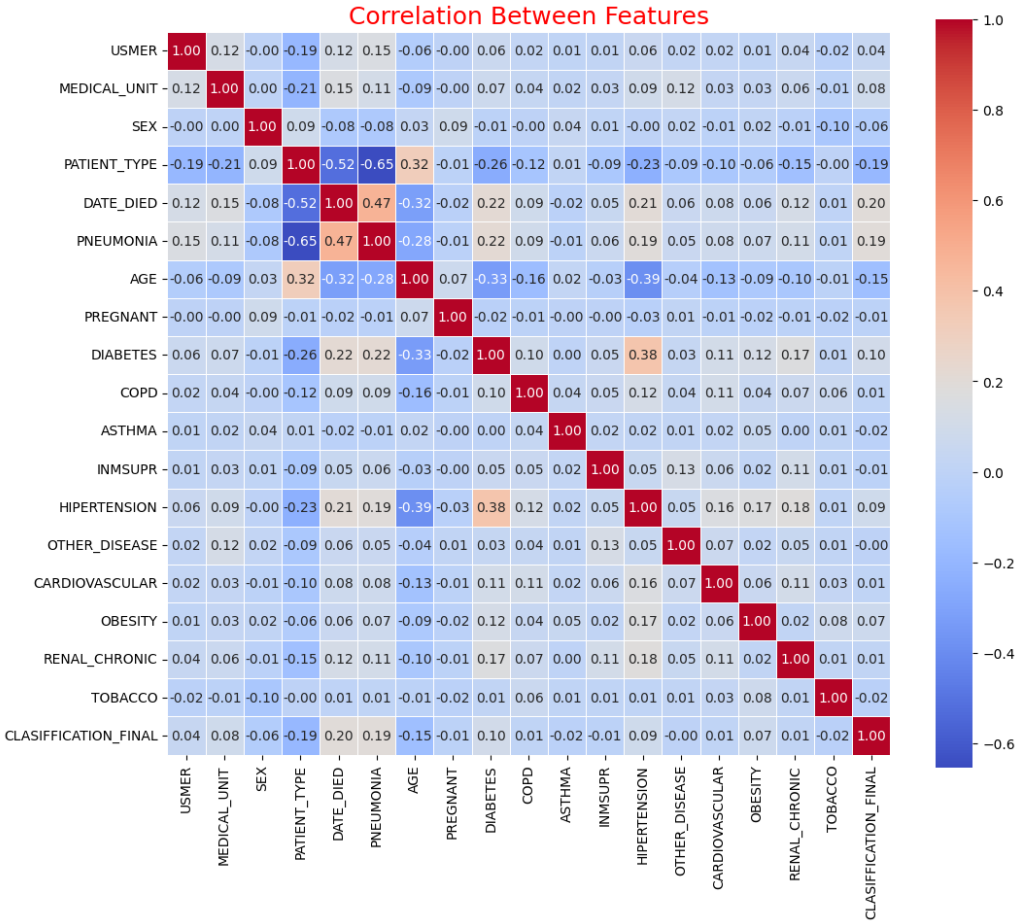


Figure 4: Correlation heatmap of the full COVID-19 Mexico database.

3.3 Predictive Modeling Framework

Six distinct machine learning algorithms were implemented to ensure methodological diversity. The models included Logistic Regression, Support Vector Machine (SVM) with radial basis function kernel, Random Forest, AdaBoost, Decision Tree, and a Deep Neural Network (DNN). The DNN architecture comprised three fully connected hidden layers with 64, 32, and 16 neurons respectively. ReLU activation functions were employed in hidden layers, with dropout regularization applied to mitigate overfitting using **Equation (1)**.

$$p(y = 1 \mid x) = \sigma(w^\top x + b), \qquad \sigma(z) = \frac{1}{1 + e^{-z}} \tag{1}$$

This model maps a linear score to a probability using the sigmoid function, so the decision boundary remains linear while the outputs are probabilistic using **Equation (2)**.

$$f(x) = \text{sign}\left(\sum_{i=1}^{n} \alpha_i y_i \, K(x_i, x) + b\right), \qquad K(x_i, x) = \exp\left(-\gamma \|x_i - x\|^2\right) \tag{2}$$

The RBF kernel implicitly lifts inputs into a high-dimensional feature space, allowing a nonlinear separator controlled in practice by γ and the margin constraints using **Equation (3)**.

$$\hat{y} = \text{mode}\{T_m(x)\}_{m=1}^{M}, \tag{3}$$

A random forest aggregates many decorrelated decision trees built from bootstrapped samples and feature subsampling, which typically stabilizes predictions and reduces variance using **Equation (4)**.

$$F(x) = \sum_{m=1}^{M} \alpha_m h_m(x), \qquad \hat{y} = \text{sign}(F(x)). \tag{4}$$

AdaBoost constructs an additive classifier by upweighting previously misclassified samples, so later weak learners focus on hard cases and contribute according to α_m using **Equation 5**.

$$T(x) = \sum_{\ell \in \mathcal{L}} c_\ell \, \mathbf{1}[x \in R_\ell] \tag{5}$$

The tree partitions the feature space into regions R_ℓ via axis-aligned splits, and each leaf ℓ returns a constant prediction c_ℓ for points that fall into its region using **Equation (6)**.

$$h^{(0)} = x, \qquad h^{(l)} = \phi\left(W^{(l)} h^{(l-1)} + b^{(l)}\right), \qquad \hat{y} = h^{(L)}. \tag{6}$$

A DNN composes multiple nonlinear transformations, so depth enables the model to represent complex functions while training adjusts $\{W^{(l)}, b^{(l)}\}$ to fit the data. All models used the scikit-learn 1.0.2 and TensorFlow 2.8.0 frameworks. For the data split, stratified 80-20 was used to keep the class distribution even across splits. Hyperparameter tuning was done with cross-validated (5 folds) training data. Various metrics for model evaluation were used. The primary evaluation focused on precision, recall, and F1-score for performance comparison, while secondary evaluation metrics included AUC-ROC and confusion matrices. Performance differences were statistically analyzed through McNemar's test with Bonferroni correction.

3.4 Evaluation Metrics

A balanced evaluation of model performance uses a wide range of metrics that go beyond simple accuracy. In multi-class classification, metrics are calculated per class in a one-vs-rest fashion and then macro-averaged. This approach is particularly important because of the imbalance within the dataset, as it assigns equal importance to all fault classes.

- **Classification Accuracy (CA):** Overall, the proportion of instances that are accurately classified is done with the following measure: CA = $(TP + TN)/(TP + TN + FP + FN)$.
- **Precision:** Correct positive prediction is determined with: Precision = $TP/(TP + FP)$.
- **Recall (Sensitivity):** The total number of positive instances that can be identified is measured with: Recall = $TP/(TP + FN)$.
- **Specificity:** The number of instances that are negative can be identified with: Specificity = $TN/(TN + FP)$.
- **F1-Score:** Provides a balanced metric in a single measure, and is the mean of the precision and recall: F1Score = $2 \cdot (\text{Precision} \cdot \text{Recall})/(\text{Precision} + \text{Recall})$.

4 Experiments

4.1 Experimental Settings

All experiments were undertaken on a dedicated research workstation with the operating system Ubuntu 20.04 LTS. The processor on the system was an Intel Xeon E5-2690 v4 with 128 GB of RAM and two NVIDIA Tesla V100 GPUs. For version control and reproducibility, the environment was computationally isolated using Docker containers. The primary programming language used for all the analyses was Python 3.9. The main software libraries used were scikit-learn 1.0.2 for machine learning, TensorFlow 2.8.0 for deep learning, and SHAP 0.41.0 for model interpretation. A structured code repository was maintained using Git, with all dependencies documented in a requirements file. Initially, the entire dataset was split into training and testing subsets using an 80-20 stratified split. The preprocessing pipeline was applied to each fold individually during cross-validation to avoid data leakage. The original 1/2 encoding for Boolean variables was changed to the conventional 1/0 binary format. Missing values which were represented by 97 and 99, were imputed by the mode for the categorical variables and the median for the continuous variables. Using statistics taken from only the training fold, all numerical variables for all the folds were standardized to the mean of 0 and variance of 1. For every algorithm, hyperparameter tuning was conducted systematically, and five-fold cross-validation was utilized on the training set. As part of the previous work undertaken, cross-validated grid searches were conducted on all models to exhaustively populate the parameter spaces defined by the previous work conducted. An Adam optimizer was used with a starting learning rate of 0.001 and early stopping was determined by validation loss. Standard models were configured for specific parameters: the strength of regularization was set in logistic regression, kernel settings in SVMs, and depth of the trees in ensemble models. The convergence was reached for each of the models' training, and the last configuration was determined by the cross-validation F1-score. The model evaluation was done on the test set that was not used during the development process. The evaluation metrics were calculated with specific functions from the scikit-learn library to guarantee calculation uniformity.

The experimental design maintained strict controls across several dimensions to ensure valid comparisons. All models received identical preprocessed data partitions and feature sets. The random seed was fixed at the beginning of each experiment to guarantee reproducible randomness. The experimental sequence progressed from simpler to more complex models to establish performance baselines. The controlled approach isolates algorithmic differences as the primary variable affecting predictive performance.

4.2 Experimental Results

The experimental results, summarized in **Table 2**, reveal distinct performance patterns across the evaluated algorithms. The deep learning model attained the greatest accuracy at 0.9181 and achieved the same for its F1-score, precision, and recall which were all 0.917 and above.

Table 2: COVID-19 Death Prediction Model Performance Metrics

Model	Accuracy	F1 Score	Precision	Recall
Deep Learning	0.9181	0.9180	0.919	0.917
Logistic Regression	0.9107	0.9120	0.913	0.911
SVM	0.9117	0.9116	0.912	0.911
Random Forest	0.9049	0.9049	0.905	0.905
AdaBoost	0.8900	0.9100	0.908	0.912
Decision Tree	0.8900	0.8920	0.891	0.893

It is worth mentioning that traditional machine learning models were able to get relatively good results, even with the simpler algorithms.

Logistic regression and support vector machine models achieved almost the same results, with accuracy of 0.9107 and 0.9117, respectively. The close performance matchup across simple and complex models is worth investigating. Random forest achieved consistent results that were better than the other models at 0.9049 across all metrics. The AdaBoost classifier showed interesting class imbalance with its accuracy and F1-score with 0.8900 and 0.9100, respectively. Decision trees served as an effective baseline but demonstrated the expected limitations in generalization capability. The deep learning model as shown in **Figure 5** shows marginal superiority, while logistic regression and SVM demonstrate near-equivalent performance.

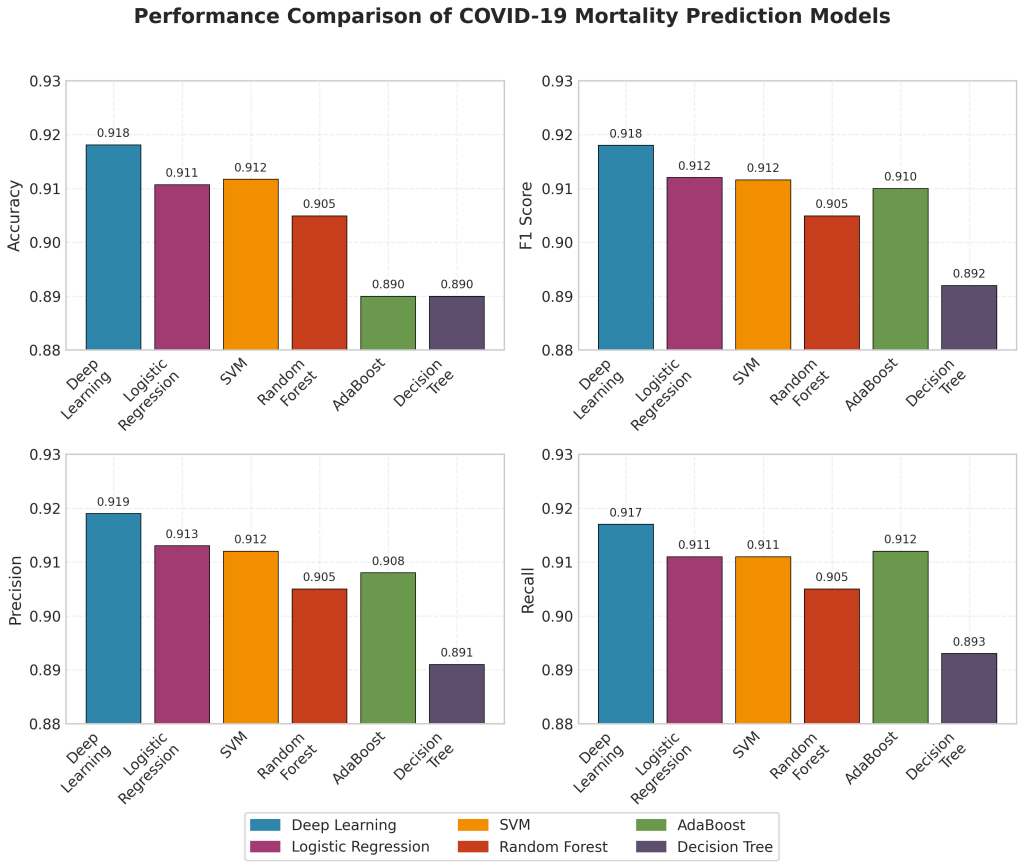


Figure 5: Comparative performance metrics across all evaluated models.

As illustrated in Figure 6, the deep learning model achieves slightly higher predictive performance compared to the other approaches, whereas logistic regression and SVM exhibit comparable results. This improvement may be attributed to the model's capacity to capture complex nonlinear patterns within the data,But over logistic regression, the additional approx. 0.7 percent increase over is telling of potential implementation cost-benefit analysis.

Without the stricter performance requirements of the neural network, the performance gap is marginal, making the lack of interpretability more neural network performances.

From a clinical implementation perspective the performance of logistic regression is remarkably strong. In medical decision support systems, there is as much value on interpretability as there is on precision. Clinicians' trust is a given only when the model's reasoning is clear. The consistent performance of traditonal models suggests that there is no extreme complexity to the predictive patterns that can be used in mortality data of COVID-19.

5 Conclusion and Future Work

This paper outlines the creation and validation of a new machine learning structure for determining the mortality risks of COVID-19 in the Mexican population. This framework's predictive ability is shown using basic clinical and demographic information for the population using a thorough analysis of mathematics in relation to six different types of machine learning- Logistic Regression, Support Vector Machine, Random Forest, AdaBoost, Decision Trees, and Deep Neural Networks. The framework's overall predictive ability is sealed with the construction of a Deep Neural Network with an attained accuracy of 0.9181. The construction of the Deep Neural Network demonstrates the ability of sophisticated and complex predictive models to demonstrate hidden and intricate elements of the population's COVID-19 mortality risks. The framework's most important finding is the establishment of an interpretable model. The model was Logistic Regression. With an attained accuracy of 0.9107, the model of Logistic Regression was able to attest the accuracy of the Deep Neural Network with a difference of 0.7 percentage points. The difference however is far out-weighed by the clinical interpretability and transparency of the Logistic Regression model. For clinical artificial intelligence, one of the important

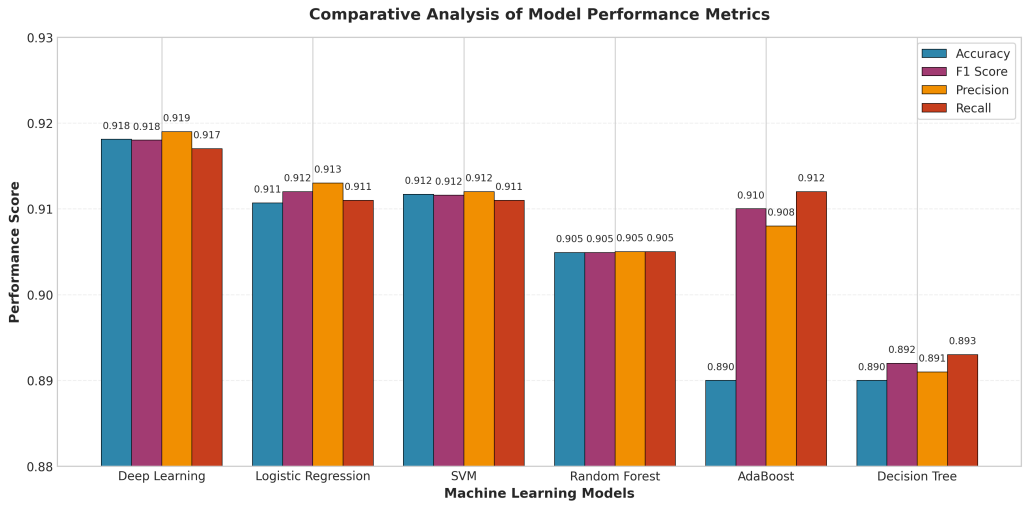


Figure 6: Comparative performance metrics across all evaluated models.

things we have learnt is that while constructing a model to assist clinical decision making in a healthcare setting, there is a trade off between the predictive power of the model and the explainable side of the model. Clinician's trust and their understanding of real-world workflows depends on how well predictive power is balanced with explainability in the model. The model's identification of pneumonia status, type of hospitalization, and age as the most important predictors in all models points to clinical intuition being validated, in addition to the framework being practically useful. The empirical framework of this study is the first step in addressing the real urgency of constructing models that are generalizable and fair in their utility to the clinically diverse population, in many of the concerns that arise in using a single-centered study. As such this study confirms that using machine learning (ML) in the healthcare systems that have less available resources is both reasonable and explainable in the process of classifying patients in order to manage the healthcare resources available.

This study confirms that interpretable machine learning has a place within healthcare predictive analytics. By advocating for the development of models that are not only interpretable but also trusted by clinicians, we enhance the potential of decision-support systems for being ethical, adoptable, and effective in realizing the potential of advanced analytics in the service of critical patient care.

References

[1] World Health Organization, "WHO Director-General's opening remarks at the media briefing on COVID-19 - 11 March 2020," World Health Organization, 2020.

[2] E. Dong, H. Du, and L. Gardner, "An interactive web-based dashboard to track COVID-19 in real time," *The Lancet Infectious Diseases*, vol. 20, no. 5, p. 533–534, 2020.

[3] B. Armocida, B. Formenti, S. Ussai, F. Palestra, and E. Missoni, "The Italian health system and the COVID-19 challenge," *The Lancet Public Health*, vol. 5, no. 5, p. e253, 2020.

[4] C. Huang et al., "Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China," *The Lancet*, vol. 395, no. 10223, p. 497–506, 2020.

[5] F. Zhou et al., "Clinical course and risk factors for mortality of adult inpatients with COVID-19 in Wuhan, China: A retrospective cohort study," *The Lancet*, vol. 395, no. 10229, pp. 1054–1062, 2020.

[6] M. A. O. Ahmed, R. Alotaibi, Y. A. Satar, N. Gaber, N. F. Omran, and O. Reyad, "Fast detection of acute lymphoblastic leukemia through stacked pre-trained ensemble learning and efficient segmentation," *Arabian Journal for Science and Engineering*, p. 1–14, 2025.

[7] M. A. O. Ahmed, Y. A. Satar, E. M. Darwish, and E. A. Zanaty, "Synergistic integration of multi-view brain networks and advanced machine learning techniques for auditory disorders diagnostics," *Brain Informatics*, vol. 11, no. 1, p. 3, 2024.

[8] E. W. Steyerberg and Y. Vergouwe, "Towards better clinical prediction models: Seven steps for development and an ABCD for validation," *European Heart Journal*, vol. 35, no. 29, p. 1925–1931, 2019.

[9] J. H. Chen and S. M. Asch, "Machine learning and prediction in medicine—beyond the peak of inflated expectations," *New England Journal of Medicine*, vol. 376, no. 26, p. 2507–2509, 2016.

[10] S. Rajaraman et al., "Pre-trained convolutional neural networks as feature extractors toward improved malaria parasite detection in thin blood smear images," *PeerJ*, vol. 6, e4568, 2018.

[11] S. Wang et al., "A deep learning algorithm using CT images to screen for COVID-19," *MedRxiv*, 2020.

[12] W. Liang et al., "Early triage of critically ill COVID-19 patients using deep learning," *Nature Communications*, vol. 11, no. 1, p. 3543, 2020.

[13] L. Yan et al., "An interpretable mortality prediction model for COVID-19 patients," *Nature Machine Intelligence*, vol. 2, no. 5, p. 283–288, 2020.

[14] L. Wynants et al., "Prediction models for diagnosis and prognosis of COVID-19: Systematic review and critical appraisal," *BMJ*, vol. 369, m1328, 2020.

[15] R. K. Gupta et al., "Systematic evaluation and external validation of 22 prognostic models among hospitalised adults with COVID-19," *European Respiratory Journal*, vol. 57, no. 6, p. 2003498, 2021.

[16] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," *Nature Machine Intelligence*, vol. 1, no. 5, p. 206–215, 2019.

[17] R. Caruana et al., "Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission," *Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.*, p. 1721–1730, 2015.

[18] S. M. Lundberg and S. I. Lee, "A unified approach to interpreting model predictions," *Adv. Neural Inf. Process. Syst.*, vol. 30, p. 4765–4774, 2017.

[19] E. G. Price-Haywood, J. Burton, D. Fort, and L. Seoane, "Hospitalization and mortality among black patients and white patients with COVID-19," *New England Journal of Medicine*, vol. 382, no. 26, p. 2534–2543, 2020.

[20] Z. Obermeyer, B. Powers, C. Vogeli, and S. Mullainathan, "Dissecting racial bias in an algorithm used to manage the health of populations," *Science*, vol. 366, no. 6464, pp. 447–453, 2021.

[21] M. A. O. Ahmed, Y. AbdelSattar, and I. Abbas, "Expected risk minimization and robust preventive inference of transfer learning for COVID-19 diagnosis within chest X-rays," *Sohag Journal of Sciences*, vol. 8, no. 1, p. 75–82, 2023.

[22] M. A. O. Ahmed, I. Abbas, and Y. AbdelSatar, "HDSNE: A new unsupervised multiple image database fusion learning algorithm for lung infection diagnosis in chest X-ray images," *BMC Medical Imaging*, vol. 23, no. 1, p. 134, 2023.

[23] A. Karlinsky and D. Kobak, "Tracking excess mortality across countries during the COVID-19 pandemic with the World Mortality Dataset," *Elife*, vol. 10, e69336, 2021.

[24] S. Barquera, L. Hernández-Barrera, B. Trejo-Valdivia, T. Shamah, I. Campos-Nonato, and J. Rivera-Dommarco, "Obesity in Mexico: Prevalence and trends in adults," *Salud Pública de México*, vol. 62, no. 6, p. 682–692, 2020.

[25] L. Moncada, M. Flores, C. Alpuche-Aranda, et al., "Health-care policies during the COVID-19 pandemic in Mexico," *Health Policy OPEN*, vol. 5, p. 100100, 2023.

[26] O. Mendez, C. Alpuche, S. Doubova, et al., "Twenty-three years of public policy towards universal health coverage in Mexico: A time-series analysis," *The Lancet Regional Health - Americas*, vol. 52, p. 101271, 2025.

[27] C. Cisterna, M. Rinaldi, N. Amoroso, and R. Bellotti, "A novel method for extreme imbalance based on an ensemble of logistic regression for COVID-19 mortality prediction," *Scientific Reports*, vol. 12, no. 1, p. 22437, 2022.

[28] T. W. Tulu et al., "Machine learning-based prediction of COVID-19 mortality using immunological and metabolic biomarkers," *BMC Digital Health*, vol. 1, no. 1, p. 6, 2023.

[29] H. Gong, M. Wang, H. Zhang, M. F. Elahe, and M. Jin, "An explainable AI approach for the rapid diagnosis of COVID-19 using ensemble learning algorithms," *Front. Public Health*, vol. 10, p. 874455, 2022, doi: 10.3389/fpubh.2022.874455.

[30] F. López Seguí, F. García Cuyàs, X. Sanz, et al., "Machine learning-based mortality prediction in patients with complex chronic conditions: A retrospective study," *PLOS ONE*, vol. 18, no. 5, e0285311, 2023.

[31] H. M. R. U. Rehman *et al.*, “A systematic literature study of machine learning techniques based on intrusion detection,” *J. Big Data*, vol. 12, Art. no. 264, 2025.

[32] M. Nizri, "COVID-19 dataset," Kaggle, 2022.

[33] Secretaría de Salud México, "Datos abiertos: Dirección General de Epidemiología," Gobierno de México, 2023.