STATISTICS, OPTIMIZATION AND INFORMATION COMPUTING
*Stat., Optim. Inf. Comput.*, Vol. **14**, August 2025, pp 677–703.
Published online in International Academic Press (www.IAPress.org)

# Machine Learning Models for Predicting COVID-19 Mortality Using Epidemiological Features

Sokaina El Khamlichi [1,2,*], Loubna Taidi [3]

[1] *Research Team in Science and Technology, Higher School of Technology of Laayoune, Ibn Zohr University, Quartier 25 Mars, P.O. Box 3007, Laayoune, Morocco*
[2] *LyRICA: Laboratory of Research in Informatics, Data Sciences and Artificial Intelligence, School of Information Sciences, B.P. 6204, Rabat-Instituts, Rabat, Morocco*
[3] *Laboratory of Innovative Technology, Faculty of Sciences and Technologies, Abdelmalek Essaadi University, Tangier, Morocco*

**Abstract** Identifying COVID-19 patients at high risk of fatality is critically important for healthcare professionals, as it supports informed decision-making and enhances the capacity to manage emerging crises within medical systems. Nevertheless, COVID-19 datasets are frequently highly imbalanced, with substantially fewer fatality cases presenting a challenge to the development of effective machine learning algorithms. This study aims to develop a high-performing machine learning approach to predict COVID-19 mortality using a Mexican epidemiological dataset. To tackle the class imbalance issue, numerous sampling techniques are applied, including SMOTE, SMOTE-ENN, ADASYN, SMOTE-Tomek, and Random Under-Sampling (RUS). Predictive models are created using several machine learning algorithms: Logistic Regression, Decision Tree, Gaussian Naïve Bayes, K-Nearest Neighbors, and Random Forest. Besides, we performed feature selection analysis using Shap technique to determine the main relevant attributes for predicting COVID-19 mortality. The results illustrate that Random Forest model, trained on balanced data with SMOTE-ENN technique yielded the best performance, with 89.44% accuracy, 87.88% Recall, and 88.74% ROC AUC score. Furthermore, feature selection analysis shows that Type of Patient, Age, Pneumonia, Intubation, having contact with COVID-19 infected patients are the key important attributes for predicting COVID-19 risk of fatality among hospitalized individuals.

**Keywords** Epidemiology, COVID-19, Imbalanced Dataset, Machine Learning, SMOTE, SMOTE-ENN, SMOTE-Tomek, ADASYN, Random Under-Sampling

**DOI:** 10.19139/soic-2310-5070-2159

## 1. Introduction

COVID-19 pandemic has recently evolved as global health crisis, and all nations throughout the entire globe are endeavoring to accurately contain it [1]. This disease generates different symptoms, such as headaches, coughs, fever, sore throats, as well as respiratory issues, which may sometimes cause death [2] [3]. Additionally, certain risk factors raise the ferocity of the virus. Older age is the most reported risk factor, alongside men's gender, which is a demographic characteristic affecting COVID-19 severity. Furthermore, prominent preexisting medical disorders such as hypertension, diabetes, and coronary heart disease are frequently linked to higher risk of mortality [4]. However, these risk factors may differ from a country to another. Indeed, prior study investigated the inequalities in case fatality rates within 93 nations, taking into consideration comorbidities, demographic risk factors such as older age, as well as social risk factors like overpopulation and poverty. Surprisingly, countries having greater social

[*] Correspondence to: Sokaina El Khamlichi (Email: s.elkhamlichi@uiz.ac.ma). Research Team in Science and Technology, Higher School of Technology of Laayoune, Ibn Zohr University, Quartier 25 Mars, P.O. Box 3007, Laayoune, Morocco. LyRICA: Laboratory of Research in Informatics, Data Sciences and Artificial Intelligence, School of Information Sciences, P.O. Box. 6204, Rabat-Instituts, Rabat, Morocco.

ISSN 2310-5070 (online) ISSN 2311-004X (print)

overpopulation and smaller socioeconomic progress had lower fatality rates [5]. Additionally, another research dealed with COVID-19 propagation across lower middle-income countries in the eastern Mediterranean region, finding that Tunisia, Egypt, Sudan, and Djibouti presented the highest case fatality rates when compared to their neighbors. This difference might be explained by Tunisia's comparatively older population when compared to other countries in the region, the simultaneous diseases in Egypt, as well as the failure of medical system in Sudan and Djibouti [6].

The prompt propagation of SARS-CoV-2 has left little time to investigate features influencing the spread of the virus, predictors of its intensity, and viable therapies. At the peak of the crisis, regions having a high number of COVID-19 infections faced the scarcity of resource and were obliged to prioritize life-saving cares like dialysis machines and ventilators [7] [8]. Moreover, with the scarcity of essential medical resources like hospital beds and ventilators, doctors are confronted with tough decisions on how to allocate these supplies among patients, frequently raising ethical concerns [9] [10] [11]. In response, machine learning techniques have been increasingly applied in the medical field, particularly for disease diagnosis and prognosis[12][13][14][15]. By leveraging advanced artificial intelligence approaches, medical systems may analyse vast datasets for predicting COVID-19 intensity. This information can assist doctors to effectively assign resources, carry out focused measures for individuals at high risk of mortality, as well as offering customized cares for patients in danger.

Nevertheless, imbalanced datasets is a challenging issue for predicting COVID-19 mortality. As the number of severe infections is disproportionately greater than mild or moderate cases. This imbalance may hinder the model's capacity to accurately predict patients at high-risk of mortality, since it may tend to be biased towards the majority class. Addressing this matter is essential for improving the accuracy of predictions and ensuring that life-saving resources and interventions reach the patients who need them most. In fact, class imbalance is a frequent challenge in several real-world scenarios, impacting the accuracy and the quality of machine learning methods [16][52][18]. Imbalanced data corresponds to classification challenges where the sample sizes of different classes are unequal. A common example of class imbalance arises in medical diagnosis, where most individuals are healthy, making the accurate prediction of rare diseases paramount [19]. Furthermore, biological datasets are often imbalanced due to various challenges in creating, handling, and obtaining new observations, particularly clinical data, which relies heavily on individuals' willingness to disclose their information or participate in clinical experiments. Analyzing these imbalanced datasets can be challenging, often requiring the use of advanced machine learning techniques to attain satisfactory outcomes, particularly in cases involving low-prevalence illnesses or medical disorders [20]. In particular, the issue of COVID-19 imbalanced datasets presents significant challenges. COVID-19 datasets are frequently highly imbalanced, which can reduce predictive performance. To improve prediction outcomes with these types of datasets, statisticians commonly apply sampling techniques, feature selection, as well as cost-sensitive learning. Nevertheless, there is no recommendations for determining the adequate method and prediction algorithm based on specific dataset characteristics [21]. The majority of COVID-19 cases result in recovery or mild symptoms, while only a small fraction lead to severe outcomes, including death. This imbalance hinders the ability of machine learning algorithms to thoroughly predict mortality, as they tend to be biased toward the more frequent non-fatal cases. As a result, models may not adequately figure out individuals at high risk of mortality, leading to underprediction of critical cases. The scarcity of data on fatal outcomes further complicates model training. To tackle this matter, sampling techniques are often required to improve the predictive performance for mortality and ensure better identification of at-risk patients.

In this study, we introduce a data-driven machine learning algorithm to predict mortality risk in hospitalized COVID-19 patients. This technique may assist healthcare systems in managing resource distribution, favor care for patients at high risk of fatality, and reduce delays in offering adequate medical attention, particularly when the infrastructure is overburdened with people. Consequently, it supports AI-driven crisis management, by enhancing resilience within medical systems. However, the development of an effective machine learning algorithm is particularly difficult due to the significant imbalance in COVID-19 datasets, which is characterized by a significantly lower fatalities than non-fatal cases. This issue is compounded by the limited number of studies that specifically address this data imbalance. Therefore, our investigation focuses on building a high-performing machine learning model, which predicts COVID-19 mortality employing various balancing data techniques to boost model performance and improve the accuracy of mortality predictions for hospitalized cases.

The remainder of this study is arranged as follows: section 2 displays the related work; section 3 shows the background of the present work; section 4 exhibits the data and methods; section 5 illustrates the results; section 6 deals with the discussion; section 7 concludes this research.

## 2. Related work

The shortage of health care equipment with respect to the rising number of COVID-19 infections has made the prediction of COVID-19 fatality critical. Accurate predictions help strengthen healthcare resilience, manage the distribution of healthcare supplies, and deliver a customized care to hospitalized cases. However, COVID-19 datasets are often highly imbalanced, leading to biased model performance. In fact, Some studies predicted the risk of COVID-19 severity without focusing directly on data balancing strategies. For instance, Iwendi et al. [22] built multiple machine learning models to predict COVID-19 fatality using datasets from Brazil and Mexico. In the Mexican dataset, logistic regression achieved the best outcome, with an accuracy of 92.272 %, a precision of 62.169 %, an F1-score of 53.215 %, and a recall of 46.516%. Nevertheless, in the Brazilian dataset, The best outcome was achieved by decision trees with an accuracy of 69.158 %, a precision of 74.276 %, a recall of 38.502 %, and an F1-score of 50.715%. Wollenstein-Betech et al. [23] employed Support Vector Machine, Logistic Regression, Random Forest, and XGBoost to predict COVID-19 severity in two different situations: prior to and after visiting a hospital. The first situation examines cases where only the fundamental characteristics of an infected person have been identified. The prediction obtained 73% accuracy and 69% AUC. The second situation covers additional information regarding hospitalization like ICU, pneumonia, and ventilator use. This improves the model's accuracy to 76%, and AUC to 74%. Furthermore, Bolourani et al. [24] developed machine learning models predicting respiratory issues among COVID-19 patients during 48 hours of hospitalization. They appraised three classifiers: XGBoost, Logistic Regression, and XGBoost+SMOTEENN. Thereby, XGBoost model surpassed the other techniques with a mean accuracy of 91.9%, and an AUC of 77%.

Additionally, few investigations have tackled the problem of imbalanced datasets for predicting COVID-19 mortality. For example, Moulaei et al. [25] compared seven machine learning methods, namely, Naïve Bayes, k-nearest neighbors, logistic regression, random forest, multi-layer perceptron, J48 decision tree, and eXtreme gradient boosting (XGBoost), for predicting COVID-19 mortality employing patients information at admission gathered from the medical record of Ayatollah Taleghani Hospital in Abadan, Iran. The synthetic minority oversampling technique (SMOTE) was applied to handle the class imbalance. The outcome of this study revealed that RF outperformed the other models with 95.03% accuracy, 90.70% sensitivity, 94.23% precision, 95.10% specificity, 99.02% Receiver Operating Characteristic (ROC). Moreover, subudhi et al. [26] developed 18 machine learning models to predict COVID-19-related ICU admission and mortality outcomes. They used an imbalanced healthcare dataset from the integrated health care system, Mass General Brigham, in New England, USA. To handle the class imbalance, they apply Random Under Sampling Technique. The results showed that ensemble-based approaches surpass the other kinds of classifiers in predicting ICU admission and mortality. Specifically, all ensemble models achieved F1-scores above 0.8 for ICU prediction and exceeded 0.83 for mortality prediction. Furthermore, Chadaga et al. [27] have built ensemble learning techniques, namely adaboost, catboost, extreme gradient boosting, light GBM, gradient boost, and Random Forest to predict COVID-19 fatality in hospitalized patients from Mexico. They performed SMOTE Technique for balancing data. The outcome revealed that Extreme Gradient Boosting attained the best performance with 96% accuracy, 95% precision, 95% recall, 95% F1-score, 96% ROC_AUC, 99% AP.

On the other hand, more researches have addressed the problem of imbalanced datasets for identifying COVID-19 infections. For example, Wu et al. [28] presented a novel hybrid dynamic ensemble selection (DES) algorithm to detect COVID-19 in imbalanced dataset, using whole blood counts. They employed synthetic minority oversampling technique with edited nearest neighbor (SMOTE-ENN) for balancing data and removing the noise. Afterward, DES outcome was improved using a novel hybrid multiple clustering and bagging classifier generation (HMCBCG). Thereby, the HMCBCG + k-nearest oracles eliminate algorithm yielded the best results with 99.81% accuracy, 99.78% G-mean, 99.86% F1-score, and 99.81% AUC. Additionally, AlJame et al. [29] suggested an ensemble learning approach called ERLX, for early COVID-19 screening using regular blood testing. This approach comprises two levels of classification techniques. The first level encompasses extra trees, random forest, and logistic regression. The outputs of these models are fed into extreme gradient boosting algorithm in the second level. This proposed model employed isolation forest for removing aberrant data, KNNImputer model for imputing missing data, and synthetic minority oversampling technique (SMOTE) for balancing the dataset. Moreover, SHapley Additive exPlanations (SHAP) model was implemented for feature selection. The ensemble model achieved the best results with an overall accuracy of 99.88%, a specificity of 99.99%, a sensitivity of 98.72%, and an AUC of 99.38%. Furthermore, Mohammedqasem et al. [30] created a deep learning model able to deal with imbalanced dataset to improve the detection of COVID-19 cases. They applied the Synthetic Minority Oversampling Technique (SMOTE) to balance the data and used the Recursive Feature Elimination technique to figure out the most prominent features. Thus, the classifier attained the best outcomes with a maximum accuracy of 98% and a precision of 97%. Besides, dorn et al. [20] investigated several widely used machine learning techniques, including Logistic Regression (LR), Decision Trees (DT), Random Forest (RF), Support Vector Machines (SVM), K-Nearest Neighbors (KNN), Multi-Layer Perceptron (MLP), eXtreme Gradient Boosting (XGBoost), and Naïve Bayes (NB) applied to complete blood count (CBC) data. They also examined the most famous sampling techniques to handle the class imbalance, namely Random Over Sampling (ROS), Random Under Sampling (RUS), Synthetic Minority Over Sampling TEchnique (SMOTE), Adaptive Synthetic Sampling (ADASYN), as well as Synthetic Minority Over Sampling TEchnique Tomek links (SMOTETomek). The findings demonstrate that SVM, LR, and RF attained the best outcome, even if the performance of each model will rely on the studied datasets and metrics. When it comes to the sampling methods, they may reduce the bias about the majority class and enhance the classification results. However, no specific approach can be considered as the best choice. Additionally, soares et al. [31] created an approach named ER-CoV through combining three machines learning algorithms: Support Vector Machine, SMOTEBoost, and ensembling methods, to detect COVID-19 negative cases from suspected patients in the Emergency Room. The study used an imbalanced dataset of 599 hospitalized patients in Brazil.encompassing basic blood test results as key predictors. The findings attained 85.98% specificity, 70.25% sensitivity, 94.92% Negative Predictive Value, 44.96% Positive Predictive Value, and 86.78% ROC_AUC (Table 1).

## 3. Background

### 3.1. Sampling techniques

*3.1.1. Oversampling techniques* :

Oversampling techniques are applied to generate additional samples in the minority group to balance data [32]. These methods can be divided into synthetic and random oversampling. Random oversampling duplicates existing minority observations to expand the minority group, while synthetic oversampling generates new synthetic instances for minority group observations. These additional instances offer important data to the minority group and avoid incorrect classification [33]

- **Synthetic minority over-sampling technique (SMOTE)**

  SMOTE is a successful oversampling strategy that creates synthetic samples for the minority class. It has demonstrated considerable success across several applications [34]. This approach was introduced to improve the size of minority groups by creating synthetic observations within feature space. For balancing the dataset, SMOTE begins by randomly selecting an observation $M_a$ from the minority group. Then, it chooses the k nearest neighbors of $M_a$ within this minority group. Next, it selects a second observation $M_b$ from this set of neighbors. $M_a$ and $M_b$ are joined to create a section within the feature space. The novel synthetic instance is created as a convex blend in the middle of $M_a$ and $M_b$. This process continues till achieving the balance between the majority and the minority groups. Due to SMOTE's effectiveness, various extensions of this method have been developed [20] .

Table 1. *Related work*

| *Ref* | *Title* | *Database* | *Method* | *Result* |
|---|---|---|---|---|
| [22] | COVID-19 health analysis and prediction using machine learning algorithms for Mexico and Brazil patients | - 1,129,258 COVID-19 positive patients from Mexico (The dataset used in our study)<br>- 541,746 COVID-19 positive cases from Brazil | - Logistic Regression<br>- Decision Tree<br>- Boosted Random Forest | - Logistic Regression<br>• Mexico dataset:<br>Accuracy = 92.272%<br>F1-score = 53.215%<br>Recall = 46.516%<br>Precision = 62.169%<br>• Brazil dataset:<br>Accuracy = 69.158%<br>F1-score = 50.715%<br>Recall = 38.502%<br>Precision = 74.276% |
| [23] | Personalized predictive models for symptomatic COVID-19 patients using basic preconditions: hospitalizations, mortality, and the need for an ICU or ventilator | - 91,000 infected individuals from Mexico | - Support Vector Machine<br>- Random Forest<br>- Logistic Regression<br>- XGBoost | - XGBoost<br>• Accuracy = 0.8945<br>• F1-score = 0.6237<br>• Recall = 0.5921<br>• Precision = 0.6589 |
| [24] | Development and Validation of a Machine learning prediction model of respiratory failure within 48 hours of patient admission for COVID19 | - Data from COVID-19 patients during first 48 hours of hospitalization | - XGBoost:<br>- Logistic Regression<br>- XGBoost + SMOTEENN | - XGBoost:<br>• accuracy = 91.9%<br>• AUC = 77% |
| [25] | Comparing machine learning algorithms for predicting COVID-19 mortality | - Patient admission data from Ayatollah Taleghani Hospital, Iran | - Naïve Bayes<br>- KNN<br>- Logistic Regression<br>- RF<br>- MLP<br>- J48 Decision tree<br>- XGBoost | - SMOTE-RF<br>• Accuracy = 95.03%<br>• sensitivity = 90.70%<br>• precision = 94.23%<br>• specificity = 95.10%<br>• ROC = 99.02% |
| [26] | Comparing machine learning algorithms for predicting ICU admission and mortality in COVID-19 | - Dataset from the integrated health care system, Mass General Brigham, in New England, USA | - 18 ML models | - All ensemble models with Random Under Sampling technique:<br>• F1-score $\geq$ 0.8 for ICU<br>• F1-score $\geq$ 0.83 for mortality |
| [27] | COVID-19 mortality prediction among patients using epidemiological parameters: an ensemble machine learning approach | - Hospitalized patient data from Mexico | - AdaBoost<br>- CatBoost<br>- XGBoost<br>- LightGBM<br>- Gradient Boost<br>- RF | - XGBoost:<br>• Accuracy = 96%<br>• Precision = 95%<br>• Recall = 95%<br>• F1 = 95%<br>• ROC_AUC = 96%<br>• AP = 99% |

| | | | | |
|---|---|---|---|---|
| [28] | A novel combined dynamic ensemble selection model for imbalanced data to detect COVID-19 from complete blood count | - Whole blood count dataset | - Hybrid DES + SMOTE-ENN + HMCBCG | - HMCBCG + k-nearest oracles eliminate algorithm:<br>• Accuracy = 99.81%<br>• F1-score = 99.86%<br>• G-mean = 99.78%<br>• AUC = 99.81% |
| [29] | Ensemble learning model for diagnosing COVID-19 from routine blood tests | • Regular blood test data | • Extra Trees<br>• RF<br>• LR<br>• XGBoost | - Ensemble model: • Accuracy = 99.88%<br>• Sensitivity = 98.72%<br>• Specificity = 99.99%<br>• AUC = 99.38% |
| [30] | Real-time data of COVID-19 detection with IoT sensor tracking using artificial neural network | Dataset of 5,644 COVID-19 patients from Brazil | - SMOTE + Deep learning | - SMOTE + Deep learning:<br>• Accuracy = 98%<br>• Precision = 97% |
| [20] | Comparison of machine learning techniques to handle imbalanced COVID-19 CBC datasets | - CBC data from COVID-19 patients | - LR, DT, RF, SVM, KNN, MLP, XGBoost, NB<br>- ROS, RUS, SMOTE, ADASYN, SMOTETomek | - SVM, LR, RF |
| [31] | A novel specific artificial intelligence-based method to identify COVID-19 cases using simple blood exams | - 599 hospitalized patients in Brazil | - SVM<br>- SMOTEBoost<br>- Ensembling methods | - Specificity = 85.98%<br>- Sensitivity = 70.25%<br>- NPV = 94.92%<br>- PPV = 44.96%<br>- ROC_AUC = 86.78% |

- **SMOTE Edited Nearest Neighbors (SMOTE-ENN)**

  SMOTE Edited Nearest Neighbors (SMOTE-ENN) is a version of the original SMOTE combined with ENN [35]. It is a two-step method that combines the advantages of oversampling and data cleaning to better handle imbalanced datasets by generating new samples for the minority class in the first step [36]. SMOTE-ENN improves classifier performance on unbalanced data by combining SMOTE with the Edited Nearest Neighbors (ENN) algorithm [37]. SMOTE creates synthetic instances for the minority class, whereas ENN deletes noisy or unclear examples from both the minority and majority classes [38]. ENN works by examining the class labels of the closest neighbors of an instance by using k=3. As long as an instance contains neighbors from both the majority and minority classes, it is deemed to became near to the decision boundary [39] (Figure 1).

- **SMOTETomek**

  SMOTETomek is also an extension of the original SMOTE. It is a hybrid strategy that tries to clean up overlapping data points for each of the classes spread in sample space [41]. It is an undersampling technique that removes noise from the majority class, which shares comparable features and overlaps [42]. SMOTE is coupled with Tomek link to produce enhanced sampling results [43]. Tomek link connects two data points based on a combination of two variables: Two conditions must be met: they must be nearest neighbors and have distinct class labels [44]. A Tomek link is formed when a couple of samples, $E_i$ and $E_j$, are labeled

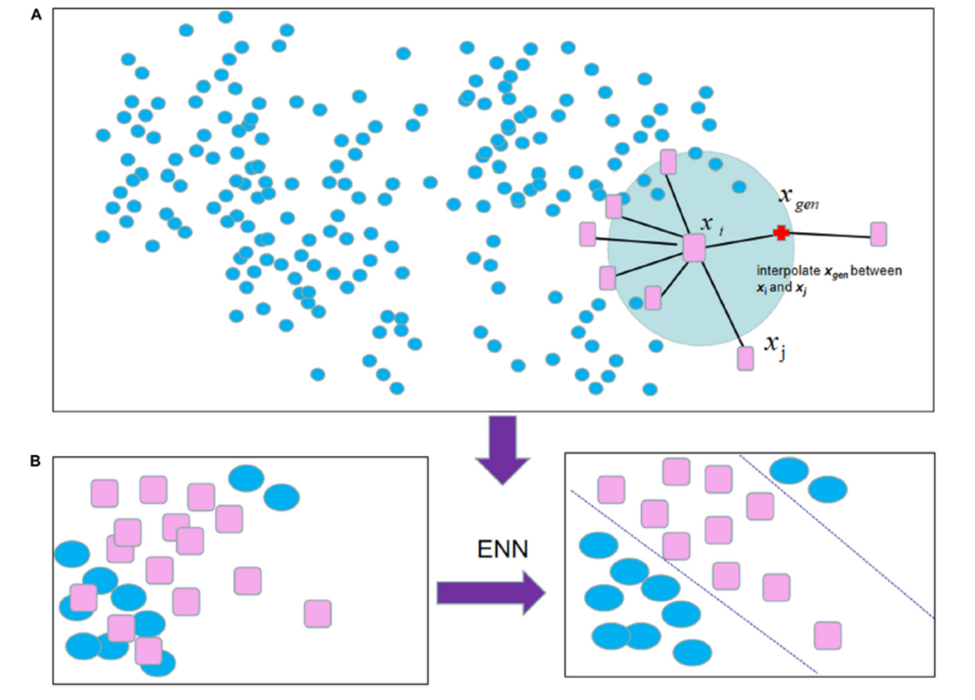

**FIGURE 1 |** The process of SMOTE-ENN algorithm: **(A)** SMOTE selected each sample from the minority samples successively as the root sample for the synthesis of the new sample. **(B)** The following result was obtained by employing ENN to eliminate noise samples when the process of SMOTE is caused.

Figure 1. The process of SMOTE-ENN as proposed in [40]

with different classes and there is no sample $E_k$ so that d($E_i$, $E_k$) ¡ d($E_i$, $E_j$) or d($E_j$, $E_k$) ¡ d($E_i$, $E_j$), d represents the distance function [45].

- **Adaptive Synthetic Sampling (ADASYN)**

ADASYN is an oversampling approach that effectively improves learning by focusing on data distribution [46]. A density estimation measure is employed to define the number of synthetic observations demanded for every minority group sample. This approach helps to balance the majority and minority groups and generates synthetic observations where the data points are hard to study. The process of generating synthetic data involves several steps: first, determine the number of novel data points required to produce a balanced dataset. Next, the density evaluation is achieved using the k-nearest neighbors for every observation in the minority group Eq. 1 and then normalization Eq. 2. The number of required instances for each sample is subsequently computed Eq. 3, and the new synthetic sample is generated accordingly.[20]

$$r_i = \frac{\Delta_i}{K}, \quad i = 1, \ldots, m_s \tag{1}$$

$$\hat{r}_i = \frac{r_i}{\sum_{i=1}^{m_s} r_i} \tag{2}$$

$$g_i = \hat{r}_i \times G \tag{3}$$

### *3.1.2. Under-sampling approaches* :

Among the most common techniques for handling with class imbalance is the undersampling method, which involves choosing the majority class from samples using a prototype [47]. The process of undersampling involves the removal of samples from the majority class [48]. In the undersampling strategy, the functioning region in the dataset is the majority class, where instances from the majority class are eliminated randomly or using certain technique to balance the classes, and then typical classification techniques are employed to categorize the data [49]. For lower ratios of class imbalance, the undersampling methods work well [50].

- **Random Under Samling approach (RUS)**

RUS is a resampling technique used to address the issue of class imbalance [51]. It is one of the simplest methods, where observations from the majority group are randomly removed to achieve a more balanced class distribution. Thus, RUS helps reduce model bias and can lead to improved overall performance during the learning process [52].

The random undersampling algorithm operates as follows [53]:

1. Initiate by selecting a random batch of data from the majority group.
2. Remove observations from the majority group to balance the dataset.
3. Repeat steps 1 and 2 till the majority and the minority groups are balanced.
4. Train the algorithm on a balanced dataset and evaluate its performance.

## *3.2. Machine Learning approaches*

### *3.2.1. K-Nearest Neighbors algorithm (KNN)* :

The Nearest Neighbors approach is based on the idea that observations within a dataset are close to each other in terms of similar characteristics [54]. Wherever unclassified instance is encountered, it will be labeled following its nearest neighbors. The extended version of this approach, called k-Nearest Neighbors (kNN), introduces the parameter k, which specifies the quantity of neighbors to examine. The classification process is simple: the unclassified data is assigned the label that is most prevalent from its neighbors. The method uses a distance measure to identify the k nearest neighbors. In this study, we employed the Euclidean Distance (Eq. 4):

$$D(x, y) = \sqrt{\sum_{i=1}^{n} |x_i - y_i|^2} \tag{4}$$

### *3.2.2. Decision Tree (DT)* :

DT is a straightforward and effective non-parametric data analysis technique. It is also a controlled learning algorithm employed for regression and classification tasks. It aims to develop an algorithm predicting the dependent variable value by learning fundamental decision information derived from the data attributes.

### *3.2.3. Logistic Regression (LR)* :

LR is an analytical technique belonging to the Generalized Linear Models family. Despite its name, it is employed for classification rather than regression. Log-linear classifier, logit regression, as well as maximum-entropy classification (MaxEnt) are used to denote logistic regression. In this approach, a logistic function is employed to represent a dichotomous variable, which is dependant on any sort of explanatory features.

The equation used in logistic regression is similar to the one used for linear regression, as shown below:

$$y = \frac{e^{(b_0 + b_1 \times x)}}{1 + e^{(b_0 + b_1 \times x)}} \tag{5}$$

Where,

- $b_0$ is the bias,
- $b_1$ refers to the coefficient of the input value ($x$),
- $y$ is the value of the output,
- $x$ refers to the value of the input.

### *3.2.4. Gradient Naïve Bayes (GNB)* :

Naïve Bayes is an elementary supervised learning algorithm founded on the probabilistic technique. It is a classification approach, which employs Bayes theorem. It supposes that each pair of attributes is independent of one another. This assumption simplifies the calculations. Thus, it is named naive. It is referred to as class conditional independence as well. The Gaussian Naïve Bayes is applied in the event of continuous data with normal distribution.

### *3.2.5. Random Forest (RF)* :

RF is a collection of learning approach used for regression and ranking problems. It utilizes several non-pruned decision trees for regression and classification assignments. In a random forest classifier, every decision tree is built using an amount of the data's variables. Once numerous trees are constructed, every tree votes the new data point's class [55]. Since random forest builds every tree using a bootstrap sampling, the minority group could be excluded in these samples. This can lead to trees that perform poorly and exhibit bias toward the majority group [56].

## *3.3. Tuning hyperparameters with Random Search*

A hyperparameter is a key variable in machine learning, that if not set, the default value is utilized [57]. For this reason, tuning hyper-parameters is typically considered an optimization task [58]. Therefore, choosing the right hyper-parameters for a machine learning algorithm needs expertise, intuition, and trial [59]. A random search is used to find candidates for the hyper-parameters, and an experiment is subsequently carried out on the chosen candidates [60]. This approach tries multiple specified combinations, evaluates hyperparameters, and selects the best results [61].

## *3.4. Feature selection*

Feature selection is a critical step in the classification of high-dimensional data [62]. It involves identifying a subset of relevant features that either preserve or improve the performance of predictive models [63]. Feature selection algorithms aim to identify the most informative combination of features that promote the development of models that are not only more accurate and efficient but also easier to interpret [64]. By isolating the features that contribute meaningfully to predicting the target variable, feature selection helps reduce noise and enhance model generalization. Despite its importance, this step is often underestimated or overlooked in practice [65].

### *3.4.1. SHAP* :

SHapley Additive exPlanations (SHAP) is a method used to determine how individual features contribute to the prediction of a dependent variable. The key idea is that the importance of a feature depends not only on its own contribution but also on its interactions with all other features in the dataset [66]. Consequently, SHAP provides valuable insights for clinical practice by identifying which factors should be prioritized [67]. Moreover, SHAP's applicability beyond commonly used machine learning classifiers offers a comprehensive framework for improving both the interpretability and performance of machine learning models [68]. Rooted in game theory, SHAP is designed to explain machine learning model outputs in a transparent and consistent manner [69]. The explanation for a given instance x is as follows [70]:

Where the explanatory model is represented by (Eq. 6):

$$g(z') = \phi_0 + \sum_{j=1}^{M} \phi_j z'_j \tag{6}$$

### *3.5. Evaluation metrics*

Evaluation metrics are quantitative indicators employed to estimate the action and efficacy of a statistical or machine learning model [71]. These metrics provide valuable information on the performance of the model and simplify the comparison of various models or algorithms [72]. Several evaluation metrics used to estimate model performance. The most commonly used for binary classification are accuracy, precision, recall, F1-score, area Under ROC Curve, and average Precision [73].

*3.5.1. Accuracy* :

Accuracy is a metric that indicates how frequently a machine learning model accurately predicts an outcome. Thus, is described as the portion of precisely predicted instances to the totality of instances in the dataset [74]. A high accuracy rate demonstrates that the majority of labels correctly reflect reality. This can dramatically improve machine learning model performance by lowering prediction errors [75]. For a binary classificator, accuracy is generally defined as follows [76]:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{7}$$

TP refers for True Positive, TN for True Negative, FP for False Positive, and FN for False Negative.

*3.5.2. Precision* :

Precision is the portion of accurately divined positive instances among all instances projected as positive [77]. It reveals which fraction of positive predictions were in fact reasonable by calculating the samples correctly predicted as positive (TP) and divided by the total positive predictions, correct or incorrect (TP, FP) [78]. High precision indicates a low rate of false positives [79]. It can be described as follows:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{8}$$

*3.5.3. Recall* :

Recall, frequently cited as sensitivity or the True Positive Rate (TPR), represents the portion of positive samples that are accurately arranged [80]. It quantifies the portion of accurate class predictions in relation to the totality of samples within the respective class [77]. Recall is determined by the ratio of correctly classified positive samples to the total number of samples designated as belonging to the positive class as follows [81]. It is defined as follows:

$$\text{Recall} = \frac{TP}{TP + FN} \tag{9}$$

*3.5.4. F1-score* :

F1 score is an appropriate statistic for measuring classifier performance since it respects both the accuracy and predictive power of a Machine Learning model [82]. It is a synthetic one-dimensional indicator, and it is frequently used to assess the performance of classifiers [83]. Moreover, F1-score balances precision and recall by considering their harmonic mean of these two indices as follows [84] [85]:

$$\text{F1-score} = 2 * \frac{precision * recall}{precision + recall} \tag{10}$$

*3.5.5. Area Under ROC Curve (ROC_AUC)* :

The ROC curve is defined as a plot of the test True-Positive Rate versus the matching False-Positive rate [86]. It depicts the associated False Positive Rate (FPR) and True Positive Rate (TPR) on the x-axis and y-axis, respectively, for binary classification after taking into account all decision thresholds [87]. Therefore, ROC analysis has the advantage of being threshold-agnostic, allowing for the estimation of a predictor's performance without a specified threshold. It also provides criteria for selecting the ideal threshold based on a cost function or objective [88].

Otherwise, the AUC calculation is a diagnostic tool that considers both true-positive and false-positive Rate. It can help identify predictors that are more accurate than others and those that are near to a 50-50 estimate or worse [89].

#### *3.5.6. Average Precision (AP)* :

Recall and precision are trade-offs, and both of them must be taken into account at the same time when comparing and evaluating various prediction methods [90]. Average Precision (AP) is a fundamental parameter for measuring the accuracy of an object detection algorithm [91]. It is a metric that integrates precision and recall in the context of ranked retrieval results [92]. The average precision for a certain information demand is calculated by averaging the precision ratings obtained from each relevant document. It can be calculated like so:

$$AP = \sum_{n} (R_n - R_{n-1}) P_n \tag{11}$$

Where $R_n$ and $P_n$ refer to the recall and the precision at the n$^{\text{th}}$ threshold

## 4. Data and Methods

### *4.1. Dataset description*

The dataset utilized in this investigation was provided by Iwendi et al. [22]. It is a Mexican dataset that includes 1,129,258 records of COVID-19 hospitalized patients. This dataset Contains 1,023,066 recovered cases from COVID-19 infection and 106,192 of deaths (Figure2).

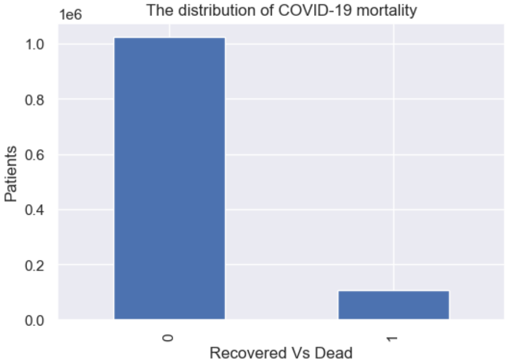

Figure 2. The distribution of COVID-19 mortality

We picked out 19 features for this study, including 18 medical and demographic characteristics, as well as the target feature: Death. The medical and demographic variables encompass Sex, Type of patient, Intubation, Pneumonia, Pregnancy, Diabetes, EPOC, Asthma, Immunosuppression, Hypertension, Other complication, Cardiovascular disease, Obesity, Renal failure, Smoking, ICU admissio, Age, and Other COVID contact. Initially, the dataset encoded the 'Sex' variable with 1 representing 'female', 2 representing 'male', the 'Type of patient' variable used 1 for 'in transit' and 2 for 'in hospital'. The rest of qualitative attributes were coded as 1 for 'positive' and 2 for 'negative'. In the present research, we applied a different coding style. For variable 'Sex', we attributed 0 to males and 1 to females, and the 'Type of patient' variable was adjusted to 0 for 'in transit' and 1 for 'in hospital'. When it comes to the other qualitative features, 0 was used for 'negative' and 1 for 'positive' (Table 2).

Table 2. dataset description

| No. | Feature | Description | Encoding | Data Type |
|---|---|---|---|---|
| 1 | Sex | Indicates the gender of infected individual | 0-male, 1-female | Qualitative |
| 2 | Type of patient | Type of the medical care the patient received | 0-transit, 1-hospital | Qualitative |
| 3 | Intubated | Indicates if the patient necessitates intubation | 0-no, 1-yes | Qualitative |
| 4 | Pneumonia | Indicates if the patient had pneumonia | 0-no, 1-yes | Qualitative |
| 5 | Pregnancy | Denotes if the patient was pregnant | 0-no, 1-yes | Qualitative |
| 6 | Diabetes | Identifies if the patient had diabetes | 0-no, 1-yes | Qualitative |
| 7 | EPOC | Indicates if the patients had Excess Post-exercise oxygen consumption | 0-no, 1-yes | Qualitative |
| 8 | Asthma | Specifies if the patient had asthma | 0-no, 1-yes | Qualitative |
| 9 | Immunosuppression | Indicates if the patient had immunosuppression | 0-no, 1-yes | Qualitative |
| 10 | Hypertension | Identifies if the patient had hypertension | 0-no, 1-yes | Qualitative |
| 11 | Another complication | Presence of other medical complications | 0-no, 1-yes | Qualitative |
| 12 | Cardiovascular | Indicates if the patient had cardiovascular issues | 0-no, 1-yes | Qualitative |
| 13 | Obesity | Specifies if the patient was obese | 0-no, 1-yes | Qualitative |
| 14 | Renal failure | Indicates the presence of kidney failure | 0-no, 1-yes | Qualitative |
| 15 | Smoking | Indicates wether the patient is smoking | 0-no, 1-yes | Qualitative |
| 16 | ICU | Determines if the patient must have intensive care | 0-no, 1-yes | Qualitative |
| 17 | Age | Records the age of the patient | NA | Numeric |
| 18 | Other_Case | Indicates if the patient had contact with another COVID-19 case | 0-no, 1-yes | Qualitative |
| 19 | Outcome (Death) | Indicates if the patient survived or died | 0-survived, 1-dead | Qualitative |

### *4.2. Proposed approach*

The effective prediction of COVID-19 mortality among hospitalized patients is of paramount importance for enabling medical systems to optimize resource allocation, favor cases at high risk of death for careful consideration, and avoid delays when offering necessary health services, especially when the infrastructure is overburdened with individuals. Thereby, this reinforces the robustness of the healthcare sector and supports clinicians in addressing this critical medical crisis. However, the imbalance in COVID-19 datasets hinders the development of an accurate machine learning algorithm. Thus, this study aims to develop an effective machine learning model for predicting the risk of fatality among COVID-19 patients by employing several approaches for balancing the data.

Initially, we started with data preprocessing. All variables are between 0 and 1, apart from the variable 'Age'. We standardized the variable 'Age' using Z-score normalization to avoid the impact of features with various scales. Then, we split the data into training and testing sets, with 75% for the training and 25% for testing. We used the training dataset to build various machine learning models including Gaussian Naive Bayes, Logistic

Regression, K-Nearest Neighbors, Random Forest, and Decision Tree. These machine learning models have been chosen because they are among the most extensively used algorithms for COVID-19 data and they represent a diverse set of methodological approaches. Decision tree and logistic regression are recognized for their high classification performance and model explainability in binary outcome prediction. Gaussian Naïve Bayes can estimate the probability of fatality based on relevant attributes, while K-Nearest Neighbors can be appropriate for predicting severity by identifying similarities in patient characteristics. Furthermore, Random Forest caught attention due to its ensemble nature, which is expected to handle class imbalance more effectively than individual models. Additionally, resampling approaches including, SMOTE, SMOTE-ENN, SMOTETomek, ADASYN, and RUS have been performed to handle the issue of class imbalance. On the other hand, feature selection analysis has been performed using SHapley Additive exPlanations (Shap) technique to determine the most significant features for predicting COVID-19 mortality among patients (Figure 3).

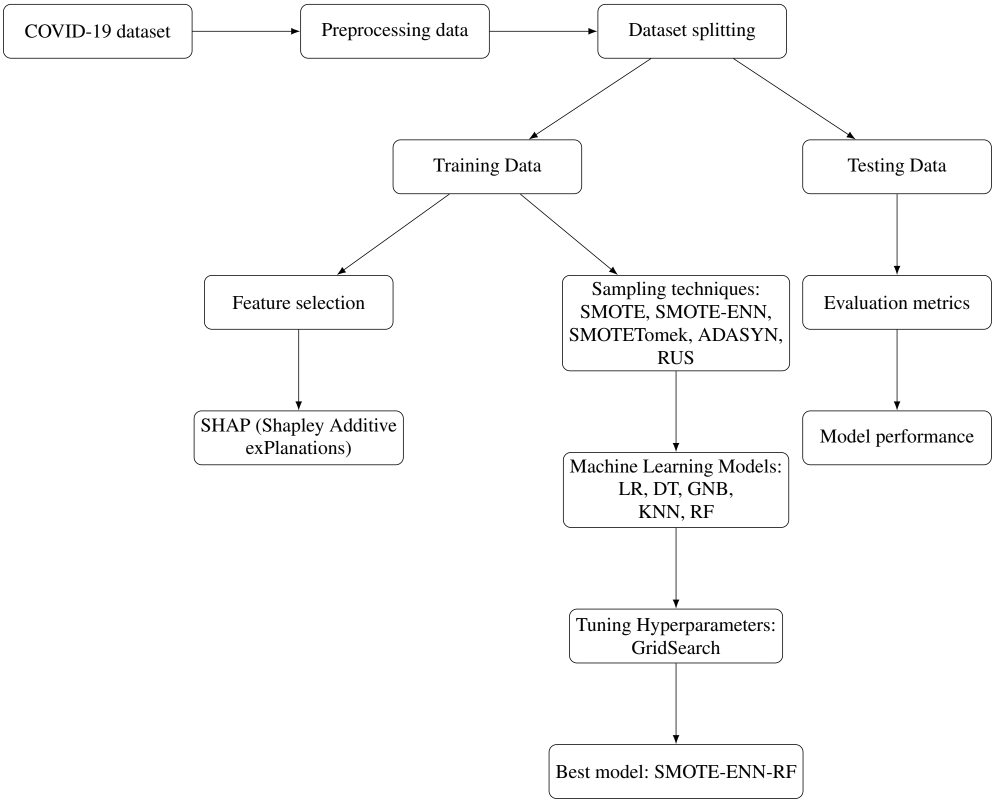
COVID-19 dataset
Preprocessing data
Dataset splitting
Training Data
Testing Data
Feature selection
Sampling techniques: SMOTE, SMOTE-ENN, SMOTETomek, ADASYN, RUS
Evaluation metrics
SHAP (Shapley Additive exPlanations)
Model performance
Machine Learning Models: LR, DT, GNB, KNN, RF
Tuning Hyperparameters: GridSearch
Best model: SMOTE-ENN-RF

Figure 3. Overview of the proposed solution

## 5. Results

We implemented five machine learning models: Decision Tree (DT), Gradient Naïve Bayes (GNB), K-Nearest Neighbors (KNN), Logistic Regression (LR), and Random Forest (RF). Since our dataset was highly imbalanced, we applied five data-balancing techniques: SMOTE, SMOTEENN, SMOTETomek, ADASYN, and RUS to address this issue.

Before applying these balancing techniques, the dataset exhibited significant imbalance, consisting of 767,195 recovered cases and 79,748 deaths. This imbalance is visually shown in Figure 4.

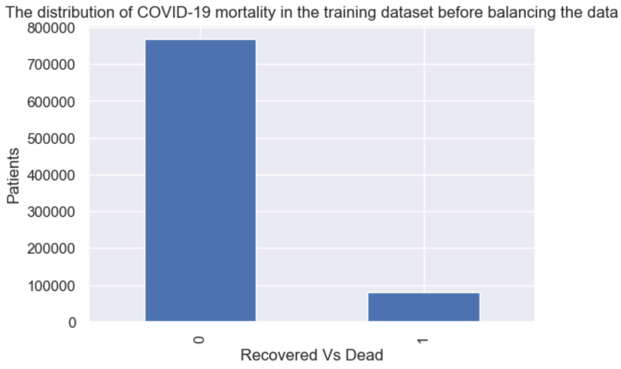

Figure 4. The distribution of COVID-19 mortality in the training set before balancing data

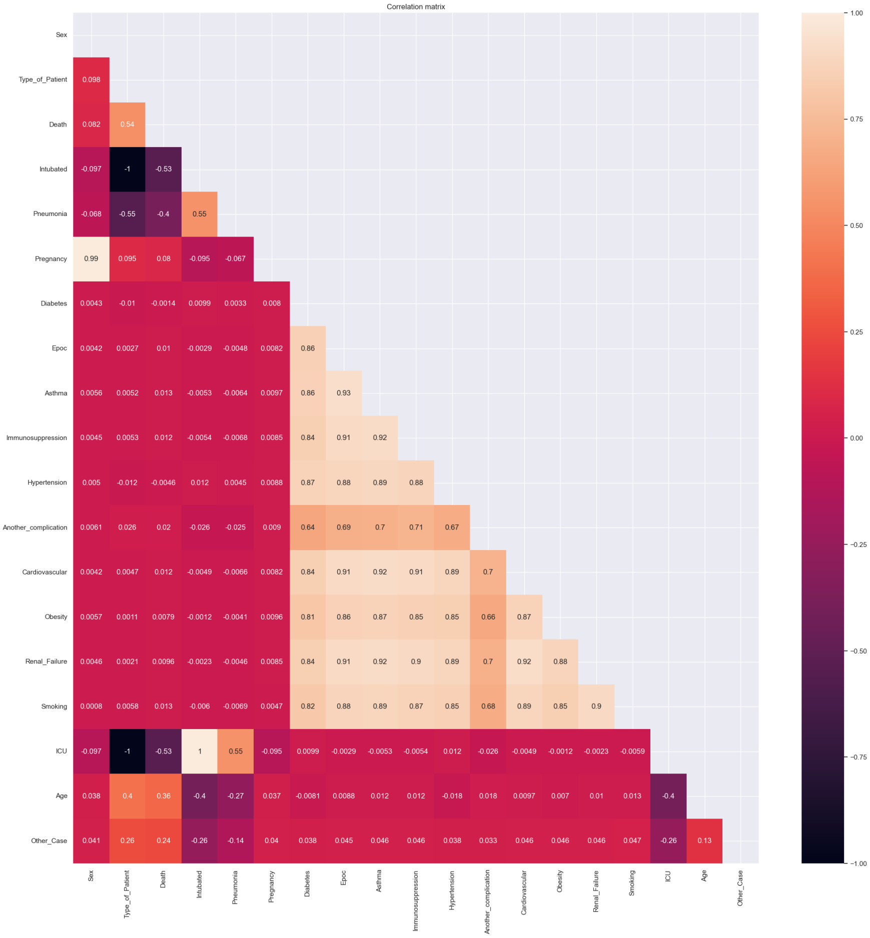

Figure 5. The distribution of COVID-19 mortality in the training set before balancing data

### 5.1. *Correlation analysis*

In this investigation, correlation analysis was used to identify the relashionship between different attributes in the dataset as well as the relashionship between the dependent variable (death) and the independent features. It has been noticed that the independent variables ‘Type of patient’, ’Age’ and ‘Other_case’ reveal a moderate positive correlation with the target variable ‘Death’. Whereas, the features ‘ICU’, ‘Intubated’ and ‘Pneumonia’ present a moderate negative correlation with the dependent variable ‘Death’ (Figure 5).

### *5.2. Balancing data*

*5.2.1. Balancing data with ADASYN* :
After applying the ADASYN technique, the dataset was balanced and contained 767,195 recovered cases and 763,687 deaths (Figure6).

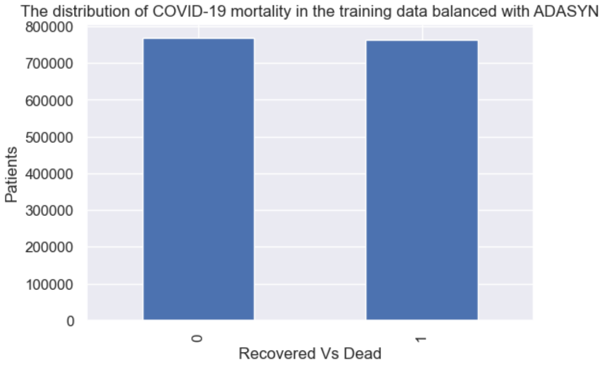

Figure 6. The distribution of COVID-19 mortality after balancing data with ADASYN

Table 3. ADASYN Random search results

| Algorithm | Hyper parameter | Value |
|---|---|---|
| GNB | priors<br>var_smoothing | None<br>4.056727410148507e-09 |
| DT | criterion<br>max_depth<br>max_features<br>min_samples_leaf<br>min_samples_split<br>splitter | entropy<br>19<br>None<br>17<br>18<br>best |
| RF | bootstrap<br>max_depth<br>max_features<br>min_samples_leaf<br>min_samples_split<br>n_estimators | False<br>None<br>sqrt<br>1<br>17<br>111 |
| LR | solver<br>penalty<br>C | liblinear<br>l2<br>21.544346900318867 |
| KNN | n_neighbors<br>p<br>weights | 15<br>1<br>distance |

To further enhance model performance, we employed the Random Search method to tune hyperparameters for the machine learning models. The optimal hyperparameters are presented in Table 3.

Following Random Search optimization, we developed the machine learning models using the ADASYN technique with tuned hyper parameters. The results indicate that Logistic Regression achieved the best performance among all classifiers with 86.92% accuracy, 57.03% F1-score, 92.66% Recall, 41.19% precision, 38.85% average precision, and 89.49% ROC_AUC. The other models, Decision Tree (DT), Random Forest (RF), K-Nearest Neighbors (KNN), and Gaussian Naïve Bayes (GNB) followed in terms of performance, as summarized in Table 4.

Table 4. Comparison of machine learning models after tuning hyperparameters and balancing data with ADASYN

| **Model** | **Train_Score** (%) | **Test_accuracy** (%) | **f1score** (%) | **Recall** (%) | **Precision** (%) | **AP** (%) | **roc_auc** (%) |
|---|---|---|---|---|---|---|---|
| GNB | 84.88 | 85.75 | 53.97 | 89.18 | 38.69 | 35.52 | 87.29 |
| DT | 90.61 | 86.79 | 56.30 | 90.84 | 40.79 | 37.91 | 88.61 |
| RF | 92.80 | 87.82 | 56.56 | 84.61 | 42.48 | 37.38 | 86.38 |
| LR | 86.78 | 86.92 | 57.03 | 92.66 | 41.19 | 38.85 | 89.49 |
| KNN | 80.05 | 91.68 | 56.76 | 58.30 | 55.30 | 36.15 | 76.72 |

### *5.2.2. Balancing data with RUS Technique* :

After the application of RUS method, we achieved a balanced dataset with 79,748 for both Recovered patients and dead individuals (Figure 7)

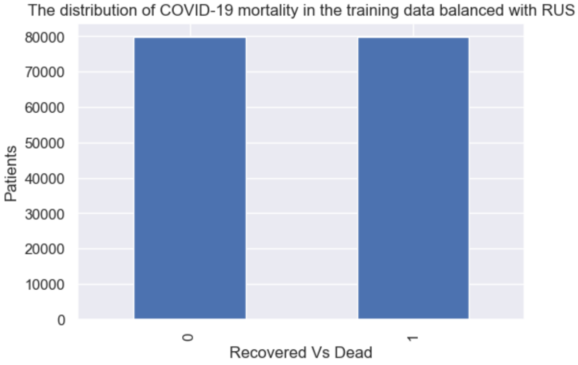

Figure 7. The distribution of COVID-19 mortality after balancing data with RUS

We utilized the Random Search technique to optimize hyperparameters for the machine learning models. The best-performing hyperparameters are detailed in Table 5

A comparative analysis of the machine learning models using the Random Under-Sampling (RUS) approach with tuned hyper parameters, revealed a slight improvement in overall performance. Among all classifiers, Logistic Regression (LR) achieved the highest performance with an accuracy of 87.83%, F1-score of 58.21%, recall of 90.53%, precision of 42.90%, average precision (AP) of 39.73%, and ROC-AUC of 89.04%. Following LR, K-Nearest Neighbors (KNN), the Random Forest (RF), Decision Tree (DT), and Gaussian Naïve Bayes (GNB) also demonstrated enhanced results (Table 6).

### *5.2.3. Balancing data with SMOTE Technique* :

After applying the SMOTE method to balance the data, we obtained an equal distribution of 767,195 cases for both recovered and deceased patients (Figure 8).

Table 5. RUS Random search results

| Algorithm | Hyper parameter | Value |
|---|---|---|
| GNB | priors<br>var_smoothing | None<br>1.4140637435040791e-08 |
| DT | criterion<br>max_depth<br>max_features<br>min_samples_leaf<br>min_samples_split<br>splitter | gini<br>9<br>None<br>15<br>17<br>best |
| RF | n_estimators<br>min_samples_split<br>min_samples_leaf<br>max_features<br>max_depth<br>bootstrap | 192<br>8<br>5<br>sqrt<br>15<br>True |
| LR | solver<br>penalty<br>C<br>random_state | liblinear<br>l2<br>37.649358067924716<br>1 |
| KNN | n_neighbors<br>p<br>weights | 12<br>1<br>uniform |

Table 6. Comparison of machine learning models after tuning hyperparameters with RUS balancing

| Model | Train_Score (%) | Test_accuracy (%) | f1score (%) | Recall (%) | Precision (%) | AP (%) | roc_auc (%) |
|---|---|---|---|---|---|---|---|
| GNB | 87.30 | 85.73 | 53.92 | 89.12 | 38.65 | 35.47 | 87.25 |
| DT | 90.36 | 86.76 | 57.21 | 94.45 | 41.03 | 39.27 | 90.21 |
| RF | 90.58 | 86.75 | 57.27 | 94.80 | 41.03 | 39.38 | 90.35 |
| LR | 89.15 | 87.83 | 58.21 | 90.53 | 42.90 | 39.73 | 89.04 |
| KNN | 87.52 | 89.04 | 58.78 | 83.39 | 45.38 | 39.40 | 86.51 |

Using balanced data with SMOTE, we employed the Random Search method to determine the best hyperparameters of the classifiers and enhance performance metrics. The identified optimal hyperparameters are exhibited in Table 7.

After tuning hyper parameters using Random Search, the performance of ML classifiers employing balanced data with SMOTE method has been generally enhanced. Logistic regression persist the most performing classifier compared to the other models with 87.85% accuracy, 58.26% F1-score, 90.52% Recall, 42.95% precision, 39.77% average precision, and and 89.05% ROC_AUC. Followed by DT, RF, KNN, and GNB (Table 8).

### *5.2.4. Balancing data with SMOTETomek Technique* :

After balancing data with SMOTETomek method, we obtained 763,567 for both recovered patients and deaths (Figure 9).

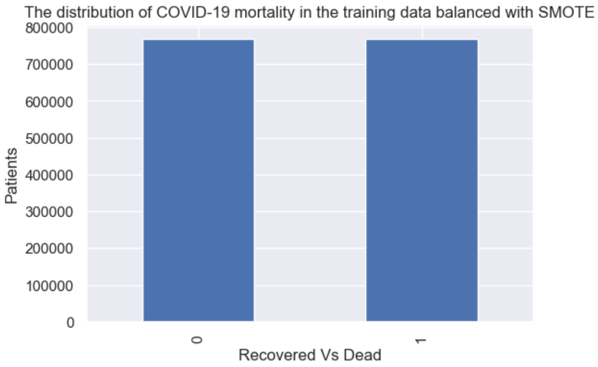

Figure 8. The distribution of COVID-19 mortality after balancing data with SMOTE

Table 7. SMOTE Random search results

| Algorithm | Hyper parameter | Value |
|---|---|---|
| GNB | priors<br>var_smoothing | None<br>1.2943857387580748e-08 |
| DT | criterion<br>max_depth<br>max_features<br>min_samples_leaf<br>min_samples_split<br>splitter | entropy<br>18<br>None<br>14<br>17<br>best |
| RF | n_estimators<br>min_samples_split<br>min_samples_leaf<br>max_features<br>max_depth<br>bootstrap | 111<br>17<br>1<br>sqrt<br>None<br>False |
| LR | solver<br>penalty<br>C<br>random_state | liblinear<br>l1<br>45.34878508128591<br>1 |
| KNN | n_neighbors<br>p<br>weights | 15<br>1<br>distance |

Using SMOTE-Tomek balanced data, we utilized the Random Search method to identify the optimal hyperparameters for the classifiers, aiming to improve performance metrics. The selected best hyperparameters are presented in Table 9

After optimizing hyperparameters with Random Search, the performance of ML classifiers using SMOTE-Tomek balanced data has generally improved. Logistic Regression remains the top-performing model, achieving an accuracy of 87.90%, an F1-score of 58.36%, a recall of 90.49%, a precision of 43.06%, an average precision of 39.86%, and a ROC AUC of 89.06%. Followed by Decision Tree (DT), Random Forest (RF), K-Nearest Neighbors (KNN), and Gaussian Naïve Bayes (GNB) (Table 10).

Table 8. Comparison of machine learning models after tuning hyperparameters and balancing data with SMOTE

| **Model** | **Train Score** | **Test Accuracy** | **F1-Score** | **Recall** | **Precision** | **AP** | **ROC AUC** |
|---|---|---|---|---|---|---|---|
| GNB | 87.28 | 85.82 | 54.07 | 89.09 | 38.81 | 35.60 | 87.29 |
| DT | 91.92 | 88.11 | 58.35 | 88.87 | 43.43 | 39.64 | 88.45 |
| RF | 93.68 | 88.70 | 57.81 | 82.66 | 44.44 | 38.36 | 85.99 |
| LR | 89.15 | 87.85 | 58.26 | 90.52 | 42.95 | 39.77 | 89.05 |
| KNN | 85.75 | 91.84 | 56.89 | 57.49 | 56.30 | 36.35 | 76.44 |

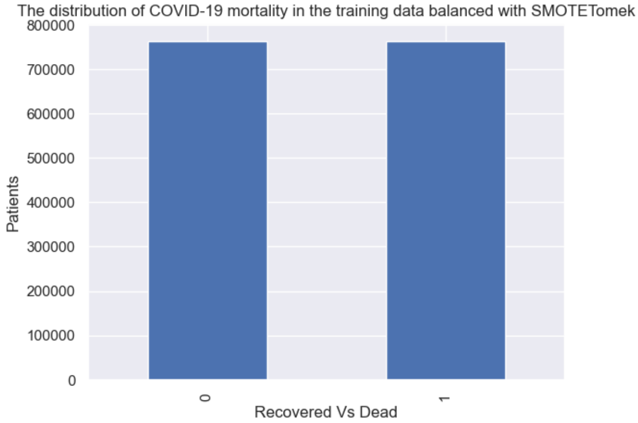

Figure 9. The distribution of COVID-19 mortality after balancing data with SMOTETomek

#### 5.2.5. *Balancing data with SMOTEENN Technique* :

After balancing the training dataset with SMOTEENN, the dataset contained 685,529 recovered patients and 392,480 of deaths (Figure 10)

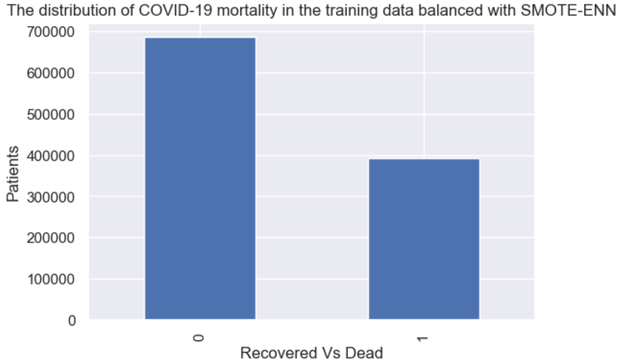

Figure 10. The distribution of COVID-19 mortality after balancing data with SMOTEENN

Using the balanced dataset obtained through SMOTE-ENN, we employed the Random Search technique to determine the optimal hyperparameters for the classifiers with the goal of enhancing performance metrics. The resulting best hyperparameter configurations are summarized in Table 11.

After hyperparameter optimization using Random Search, the performance of the machine learning classifiers on the SMOTE-ENN balanced dataset showed a general improvement. Among the models, Random Forest achieved the best results, with an accuracy of 89.44%, F1-score of 60.92%, recall of 87.88%, precision of 46.62%, average

Table 9. SMOTE-Tomek Random search results

| Algorithm | Hyper parameter | Value |
|---|---|---|
| GNB | priors<br>var_smoothing | None<br>5.946741356463224e-09 |
| DT | criterion<br>max_depth<br>max_features<br>min_samples_leaf<br>min_samples_split<br>splitter | gini<br>17<br>None<br>16<br>7<br>best |
| RF | n_estimators<br>min_samples_split<br>min_samples_leaf<br>max_features<br>max_depth<br>bootstrap | 111<br>17<br>1<br>sqrt<br>None<br>False |
| LR | solver<br>penalty<br>C<br>random_state | liblinear<br>l2<br>21.544346900318867<br>1 |
| KNN | n_neighbors<br>p<br>weights | 15<br>1<br>distance |

Table 10. Comparison of machine learning models after tuning hyperparameters and balancing data with SMOTETomek

| Model | Train Score | Test Accuracy | F1-Score | Recall | Precision | AP | ROC AUC |
|---|---|---|---|---|---|---|---|
| GNB | 87.46 | 85.80 | 54.04 | 89.11 | 38.78 | 35.57 | 87.28 |
| DT | 91.89 | 87.91 | 58.28 | 90.14 | 43.06 | 39.74 | 88.91 |
| RF | 93.80 | 88.70 | 58.17 | 83.38 | 44.53 | 38.85 | 86.52 |
| LR | 89.36 | 87.90 | 58.36 | 90.49 | 43.06 | 39.86 | 89.06 |
| KNN | 85.68 | 91.75 | 57.30 | 59.08 | 55.61 | 36.69 | 77.10 |

precision of 42.10%, and ROC AUC of 88.94%. It was followed in performance by Logistic Regression (LR), Decision Tree (DT), K-Nearest Neighbors (KNN), and Gaussian Naïve Bayes (GNB), as detailed in Table 12.

### *5.3. Feature selection using SHAP Technique*

To assess the influence of individual variables on outcome prediction in machine learning models, we conducted a SHAP analysis. The results indicate that Type_of_Patient, Age, Pneumonia, Other_Case, and Intubated are strongly associated with higher risk of mortality. Consequently, these variables emerge as the most critical features for predicting COVID-19 mortality in hospitalized patients (Figure 11).

Table 11. SMOTE-ENN Random search results

| Algorithm | Hyper parameter | Value |
|---|---|---|
| GNB | priors | None |
| | var_smoothing | 1.3347189009806692e-08 |
| DT | criterion | gini |
| | max_depth | 19 |
| | max_features | None |
| | min_samples_leaf | 16 |
| | min_samples_split | 17 |
| | splitter | best |
| RF | n_estimators | 130 |
| | min_samples_split | 4 |
| | min_samples_leaf | 1 |
| | max_features | log2 |
| | max_depth | 14 |
| | bootstrap | True |
| LR | solver | liblinear |
| | penalty | l2 |
| | C | 79.24828983539186 |
| | random_state | 1 |
| KNN | n_neighbors | 15 |
| | p | 1 |
| | weights | distance |

Table 12. Comparison of machine learning models after tuning hyperparameters and balancing data with SMOTE-ENN

| Model | Train Score | Test Accuracy | F1-Score | Recall | Precision | AP | ROC AUC |
|---|---|---|---|---|---|---|---|
| GNB | 91.75 | 85.63 | 53.78 | 89.23 | 38.49 | 35.35 | 87.25 |
| DT | 98.10 | 90.85 | 59.73 | 72.41 | 50.83 | 39.39 | 82.58 |
| SMOTE_ENN_RF | 96.61 | 89.44 | 60.92 | 87.88 | 46.62 | 42.10 | 88.74 |
| LR | 94.39 | 88.24 | 58.85 | 89.79 | 43.77 | 40.26 | 88.94 |
| KNN | 99.86 | 89.48 | 56.02 | 71.55 | 46.03 | 35.60 | 81.44 |

## 6. Discussion

The findings of this study illustrate the strong predictive capabilities of Random Forest algorithm trained on balanced data using SMOTEENN resampling technique (SMOTE-ENN-RF). This model outperformed the other models with 89.44% accuracy, 60.92% F1-score, 87.88% Recall, 46.62% precision, 42.10% AP, and 88.74% ROC_AUC.

A comparative analysis with previous studies, which deal with the prediction of COVID-19 mortality in imbalanced data sets, reveals that our proposed SMOTE-ENN-RF model consistently outperforms several existing approaches. Compared to the logistic regression model presented in[23], our approach shows superior performance in all metrics reported. Similarly, while the XGBoost model from [24] achieved a slightly higher accuracy, our model substantially outperforms it in terms of recall, ROC AUC, and AP, critical metrics in identifying high-risk patients who are more likely to die from the disease. As for the work presented in [22], which exhibits logistic

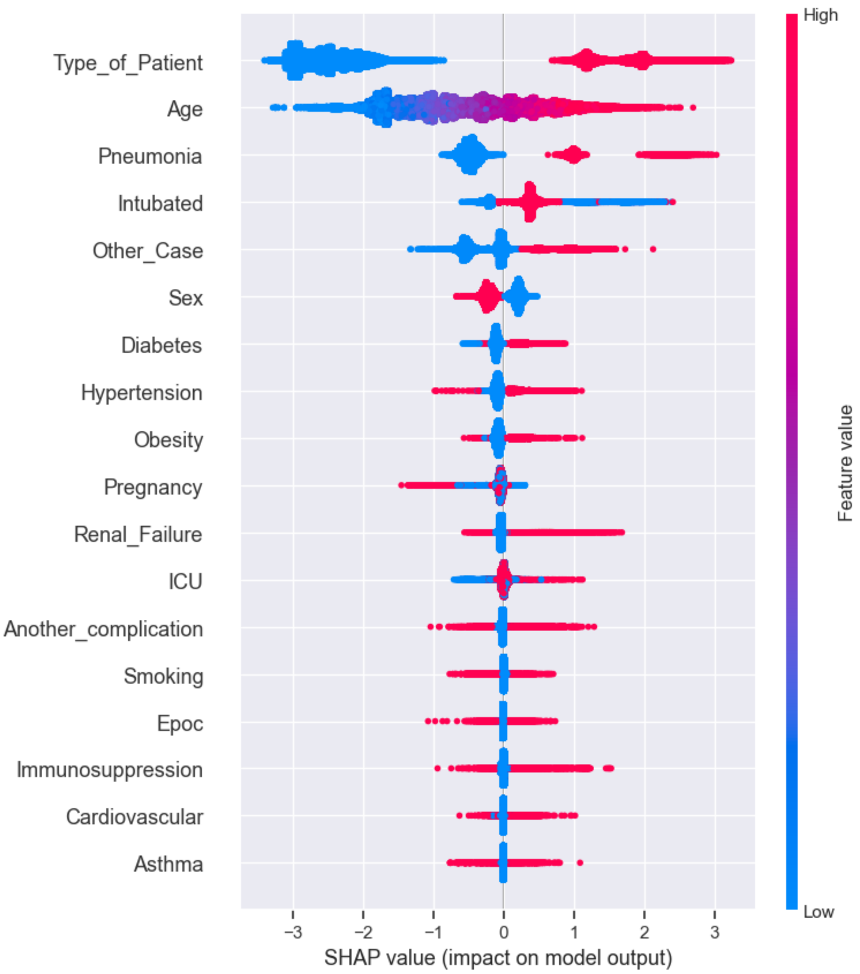

Figure 11. SHAP

regression as the best algorithm using the same dataset as us, our model achieves markedly higher recall and F1-score, although it is slightly outperformed in terms of accuracy and precision. This suggests that our model offers a more balanced and practical trade-off between precision and recall that is an essential attribute when dealing with real-world imbalanced datasets.

It should be noted that while our model delivers competitive performance, it is still outperformed by some ensemble-based models found in [26], as well as by the SMOTE-RF and SMOTE-XGBoost models presented in [25] and [27] respectively. These results underscore the continued relevance of ensemble methods and advanced boosting techniques when tuned appropriately for the data distribution.

Additionally, when comparing our results to studies focused on COVID-19 detection rather than mortality prediction, such as those found in [31], [29], and [28], the SMOTE-ENN-RF model still holds its ground. It outperforms the ER-CoV model in [31] in terms of recall and ROC AUC, which highlights its strength in correctly identifying positive cases. However, it is surpassed by the best-performing models in [28] and [29], which may benefit from more complex architectures or larger datasets (Table 13).

Accordingly, our findings illustrate the effectiveness of combining Random Forest with SMOTE-ENN to address data imbalance in COVID-19 mortality prediction. The results not only outperform several key benchmarks from the literature but also offer a balanced and practical approach suitable for real-world deployment. This underscores the value of resampling techniques in enhancing machine learning performance for medical prognosis, ultimately

Table 13. Results comparison

| Category | Ref. | Model | Accuracy (%) | Precision (%) | Recall (%) | F1-score(%) | ROC_AUC (%) | AP (%) |
|---|---|---|---|---|---|---|---|---|
| **Covid-19 mortality prediction** | [23] | LR: After visiting a medical facility | 76 | - | - | - | 74 | - |
| | [24] | XGBoost | 91.9 | 52.1 | 5.1 | - | 77 | 26 |
| | [22] | Logistic Regression | 92.272 | 62.169 | 46.516 | 53.215 | - | - |
| | [26] | Ensemble based models: ICU Prediction | - | - | - | $\geq 0.8$ | - | - |
| | | Ensemble based models: mortality Prediction | - | - | - | $\geq 0.83$ | - | - |
| | [25] | SMOTE-RF | 95.03 | 94.23 | 90.70 | - | 99.02 | - |
| | [27] | SMOTE-XGBoost | 96 | 95 | 95 | 95 | 96 | 99 |
| | **Our study** | **SMOTE-ENN-RF** | **89.44** | **46.62** | **87.88** | **60.92** | **88.74** | **42.10** |
| **Covid-19 detection** | [31] | ER-CoV | - | 44.96 | 70.25 | - | 86.78 | - |
| | [29] | Ensemble model(ERLX) | 99.88 | - | 98.72 | - | 99.38 | - |
| | [28] | HMCBCG + k-nearest oracles eliminate | 99.81 | - | - | 99.86 | 99.81 | - |

supporting healthcare professionals in making timely and informed decisions for patients at higher risk of COVID-19-related mortality. Moreover, the findings of this study hold promising implications for clinical practice. By accurately identifying patients at elevated risk of mortality, the proposed machine learning approach can aid in early diagnosis, guide risk stratification protocols, and contribute to personalized treatment planning. These applications can enhance clinical decision-making and support more efficient resource allocation during health crises. However, real-world implementation requires addressing key challenges, such as limited data availability, integration with existing healthcare systems, and the need for external validation to ensure robustness across diverse clinical settings.

Our investigation is not without limitations. First, it is important to note that our dataset exclusively represents the population of one country. Therefore, for comprehensive validation, data from diverse geographical locations and healthcare systems are essential to assure the generalizability of our findings. Second, this investigation does not account for the features of unstructured datasets. While the majority of input features employed in predicting COVID-19 mortality are numerical, The current approach may face few challenges when integrating unstructured data. Third, certain potential risk factors influencing COVID-19 severity, such as viral load and body mass index, were not considered in this study. The inclusion of such factors in future research would contribute to a more comprehensive understanding of predictive variables. Finally, our models are based on available epidemiological and clinical data, which may not fully capture the impact of emerging variants or long-term complications. Continuous updates and adaptation to evolving pandemic conditions are necessary for maintaining predictive accuracy.

## 7. Conclusion

The present work investigates the use of Artificial Intelligence approaches to predict COVID-19 mortality within hospitalized patients. Given the prevalence of imbalanced datasets in COVID-19 research, the creation of effective predictive models remains a significant challenge. To address this challenge, this analysis explores various balancing data techniques to improve classification performance and develop an effective model. The outcome of this study demonstrates that Random Forest algorithm developed with balanced data using SMOTEENN approach (SMOTE-ENN-RF) attained the highest performance with 89.44% accuracy, 60.92% F1-score, 87.88% Recall, 46.62% precision, 42.10% AP, and 88.74% ROC AUC. Besides, the present work identifies the variables: Type of Patient, Age, Pneumonia, Intubation, having contact with other COVID-19 patient as the most important features for predicting COVID-19 mortality.

As perspective, additional resampling methods such as cost-sensitive learning strategies could be explored to adress the issue of imbalanced dataset. Furthermore, deep learning approaches can be applied to predict efficiently COVID-19 fatality. Additionally, integrating multi-modal data sources, including genomic data, electronic health records, and real-time wearable sensor data could enhance the prediction of COVID-19 severity. Finally, delving into proteomic analysis methods may offer accurate insights into predicting COVID-19 severity.

## REFERENCES

1. A. U. M. Shah, S. N. A. Safri, R. Thevadas, N. K. Noordin, A. Abd Rahman, Z. Sekawi, A. Ideris, and M. T. H. Sultan, *COVID-19 outbreak in Malaysia: Actions taken by the Malaysian government*, International Journal of Infectious Diseases, vol. 97, pp. 108–116, 2020.
2. C. Huang, Y. Wang, X. Li, L. Ren, J. Zhao, Y. Hu, L. Zhang, G. Fan, J. Xu, X. Gu, et al., *Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China*, The Lancet, vol. 395, no. 10223, pp. 497–506, 2020.
3. Q. Li, X. Guan, P. Wu, X. Wang, L. Zhou, Y. Tong, R. Ren, K. S. M. Leung, E. H. Y. Lau, J. Y. Wong, et al., *Early transmission dynamics in Wuhan, China, of novel coronavirus–infected pneumonia*, New England Journal of Medicine, vol. 382, no. 13, pp. 1199–1207, 2020.
4. D. Wolff, S. Nee, N. S. Hickey, and M. Marschollek, *Risk factors for Covid-19 severity and fatality: a structured literature review*, Infection, vol. 49, pp. 15–28, 2021.
5. M. Jawad Hashim, A. R. Alsuwaidi, and G. Khan, *Population risk factors for COVID-19 mortality in 93 countries*, Journal of Epidemiology and Global Health, vol. 10, no. 3, pp. 204–208, 2020.
6. S. El Khamlichi, A. Maurady, and A. Sedqui, *Comparative study of COVID-19 situation between lower-middle-income countries in the eastern Mediterranean region*, Journal of Oral Biology and Craniofacial Research, vol. 12, no. 1, pp. 165–176, 2022.
7. A. H. M. Antommaria, T. S. Gibb, A. L. McGuire, P. R. Wolpe, M. K. Wynia, M. K. Applewhite, A. Caplan, D. S. Diekema, D. M. Hester, L. S. Lehmann, *et al.*, *Ventilator triage policies during the COVID-19 pandemic at US hospitals associated with members of the association of bioethics program directors*, Annals of Internal Medicine, vol. 173, no. 3, pp. 188–194, 2020.
8. J. Silberzweig, T. A. Ikizler, H. Kramer, P. M. Palevsky, J. Vassalotti, and A. S. Kliger, *Rationing scarce resources: The potential impact of COVID-19 on patients with chronic kidney disease*, Journal of the American Society of Nephrology, vol. 31, no. 9, pp. 1926–1928, 2020.
9. N. Salari, H. Khazaie, A. Hosseinian-Far, H. Ghasemi, M. Mohammadi, S. Shohaimi, A. Daneshkhah, B. Khaledi-Paveh, and M. Hosseinian-Far, *The prevalence of sleep disturbances among physicians and nurses facing the COVID-19 patients: a systematic review and meta-analysis*, Globalization and Health, vol. 16, pp. 1–14, 2020.
10. N. Salari, H. Khazaie, A. Hosseinian-Far, B. Khaledi-Paveh, M. Kazeminia, M. Mohammadi, S. Shohaimi, A. Daneshkhah, and S. Eskandari, *The prevalence of stress, anxiety and depression within front-line healthcare workers caring for COVID-19 patients: a systematic review and meta-regression*, Human Resources for Health, vol. 18, pp. 1–14, 2020.
11. J.-L. Vincent and J. Creteur, *Ethical aspects of the COVID-19 crisis: How to deal with an overwhelming shortage of acute beds*, European Heart Journal: Acute Cardiovascular Care, vol. 9, no. 3, pp. 248–252, 2020.
12. Y. E. I. El-Bouzaidi and O. Abdoun, *Advances in artificial intelligence for accurate and timely diagnosis of COVID-19: a comprehensive review of medical imaging analysis*, Scientific African, vol. 22, p. e01961, 2023.
13. L. Taidi and S. El Khamlichi, *A comprehensive review of artificial intelligence techniques for timely and accurate prediction of Down syndrome*, Agile Security in the Digital Era, pp. 179–194, 2024.
14. S. El Khamlichi, I. B. A. Ouahab, M. Bouhorma, F. Elouaï, A. Sedqui, and A. Maurady, *Intelligent Systems and Applications in Engineering*.
15. Y. E. I. El-Bouzaidi and O. Abdoun, *Artificial Intelligence for Sustainable Dermatology in Smart Green Cities: Exploring Deep Learning Models for Accurate Skin Lesion Recognition*, Procedia Computer Science, vol. 236, pp. 233–240, 2024.
16. J. L. Leevy, T. M. Khoshgoftaar, R. A. Bauder, and N. Seliya, *A survey on addressing high-class imbalance in big data*, Journal of Big Data, vol. 5, no. 1, pp. 1–30, 2018.
17. J. M. Johnson and T. M. Khoshgoftaar, *Survey on deep learning with class imbalance*, Journal of Big Data, vol. 6, no. 1, pp. 1–54, 2019.
18. V. López, A. Fernández, S. García, V. Palade, and F. Herrera, *An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics*, Information Sciences, vol. 250, pp. 113–141, 2013.
19. Y. Zhao, Z. S.-Y. Wong, and K. L. Tsui, *A framework of rebalancing imbalanced healthcare data for rare events' classification: a case of look-alike sound-alike mix-up incident detection*, Journal of Healthcare Engineering, vol. 2018, no. 1, p. 6275435, 2018.
20. M. Dorn, B. I. Grisci, P. H. Narloch, B. C. Feltes, E. Avila, A. Kahmann, and C. S. Alho, *Comparison of machine learning techniques to handle imbalanced COVID-19 CBC datasets*, PeerJ Computer Science, vol. 7, p. e670, 2021.

21. J. Kim and O. Kwon, *A model for rapid selection and COVID-19 prediction with dynamic and imbalanced data*, Sustainability, vol. 13, no. 6, p. 3099, 2021.
22. C. Iwendi, C. G. Y. Huescas, C. Chakraborty, and S. Mohan, *COVID-19 health analysis and prediction using machine learning algorithms for Mexico and Brazil patients*, Journal of Experimental & Theoretical Artificial Intelligence, vol. 36, no. 3, pp. 315–335, 2024.
23. S. Wollenstein-Betech, C. G. Cassandras, and I. Ch. Paschalidis, *Personalized predictive models for symptomatic COVID-19 patients using basic preconditions: hospitalizations, mortality, and the need for an ICU or ventilator*, International Journal of Medical Informatics, vol. 142, pp. 104258, 2020.
24. S. Bolourani, M. Brenner, P. Wang, T. McGinn, J. Hirsch, D. Barnaby, and T. Zanos, *Development and Validation of a Machine learning prediction model of respiratory failure within 48 hours of patient admission for COVID-19*, Journal of Medical Internet Research, 2021.
25. K. Moulaei, M. Shanbehzadeh, Z. Mohammadi-Taghiabad, and H. Kazemi-Arpanahi, *Comparing machine learning algorithms for predicting COVID-19 mortality*, BMC Medical Informatics and Decision Making, vol. 22, no. 1, p. 2, 2022.
26. S. Subudhi, A. Verma, A. B. Patel, C. C. Hardin, M. J. Khandekar, H. Lee, D. McEvoy, T. Stylianopoulos, L. L. Munn, S. Dutta, *et al.*, *Comparing machine learning algorithms for predicting ICU admission and mortality in COVID-19*, NPJ Digital Medicine, vol. 4, no. 1, p. 87, 2021.
27. K. Chadaga, S. Prabhu, S. Umakanth, N. Sampathila, and R. Chadaga, *COVID-19 mortality prediction among patients using epidemiological parameters: an ensemble machine learning approach*, Engineered Science, vol. 16, no. 10, pp. 221–233, 2021.
28. J. Wu, J. Shen, M. Xu, and M. Shao, *A novel combined dynamic ensemble selection model for imbalanced data to detect COVID-19 from complete blood count*, Computer Methods and Programs in Biomedicine, vol. 211, pp. 106444, 2021.
29. M. AlJame, I. Ahmad, A. Imtiaz, and A. Mohammed, *Ensemble learning model for diagnosing COVID-19 from routine blood tests*, Informatics in Medicine Unlocked, vol. 21, pp. 100449, 2020.
30. H. Mohammedqasim and O. Ata, *Real-time data of COVID-19 detection with IoT sensor tracking using artificial neural network*, Computers and Electrical Engineering, vol. 100, pp. 107971, 2022.
31. F. Soares, A. Villavicencio, F. S. Fogliatto, M. H. P. Rigatto, M. J. Anzanello, M. A. P. Idiart, and M. Stevenson, *A novel specific artificial intelligence-based method to identify COVID-19 cases using simple blood exams*, MedRxiv, pp. 2020–04, 2020.
32. N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, *SMOTE: Synthetic minority over-sampling technique*, Journal of Artificial Intelligence Research, vol. 16, pp. 321–357, 2002.
33. M. S. Shelke, P. R. Deshmukh, and V. K. Shandilya, *A review on imbalanced data handling using undersampling and oversampling technique*, International Journal of Recent Trends in Engineering and Research, vol. 3, no. 4, pp. 444–449, 2017.
34. B. Das, N. C. Krishnan, and D. J. Cook, *RACOG and wRACOG: Two probabilistic oversampling techniques*, IEEE Transactions on Knowledge and Data Engineering, vol. 27, no. 1, pp. 222–234, 2014.
35. E. Parkinson de Castro, *An examination of the SMOTE and other SMOTE-based techniques that use synthetic data to oversample the minority class in the context of credit-card fraud classification*, Technological University Dublin, 2020.
36. G. Husain, D. Nasef, R. Jose, J. Mayer, M. Bekbolatova, T. Devine, and M. Toma, *SMOTE vs. SMOTEENN: A Study on the Performance of Resampling Algorithms for Addressing Class Imbalance in Regression Models*, Algorithms, vol. 18, no. 1, p. 37, 2025.
37. R. Bounab, K. Zarour, B. Guelib, and N. Khlifa, *Enhancing Medicare Fraud Detection Through Machine Learning: Addressing Class Imbalance With SMOTE-ENN*, IEEE Access, 2024.
38. K. Swain, T. K. Tak, K. Upreti, P. R. Kshirsagar, S. Bala, R. C. P. Krishnan, S. R. Nayak, and M. N. Mohanty, *Enhancing Stroke Prediction Using LightGBM With SMOTE-ENN and Fine-Tuning: A Comprehensive Analysis*, Cureus, vol. 16, no. 12, 2024.
39. A. X. Wang, S. S. Chukova, and B. P. Nguyen, *Synthetic minority oversampling using edited displacement-based k-nearest neighbors*, Applied Soft Computing, vol. 148, p. 110895, 2023.
40. Y. Shang, T. Liu, Y. Wei, J. Li, L. Shao, M. Liu, Y. Zhang, Z. Zhao, H. Xu, Z. Peng *et al.*, *Scoring systems for predicting mortality for severe patients with COVID-19*, EClinicalMedicine, vol. 24, Elsevier, 2020.
41. S. Satpathy, *SMOTE for Imbalanced Classification with Python*, Anal. Vidhya, 2020.
42. H. Hairani, A. Anggrawan, and D. Priyanto, *Improvement performance of the random forest method on unbalanced diabetes data classification using Smote-Tomek Link*, JOIV: International Journal on Informatics Visualization, vol. 7, no. 1, pp. 258–264, 2023.
43. R. Manoharan, M. S. Stalin, G. B. Loganathan, and others, *Ensemble Model for Educational Data Mining Based on Synthetic Minority Oversampling Technique*, 2023.
44. I. Tomek, *Two modifications of CNN*, 1976.
45. E. de la Cal, J. R. Villar, P. Vergara, J. Sedano, and Á. Herrero, *A SMOTE extension for balancing multivariate epilepsy-related time series datasets*, in *Proceedings of the International Joint Conference SOCO'17-CISIS'17-ICEUTE'17*, León, Spain, Sep. 6–8, 2017, pp. 439–448, Springer, 2018.
46. A. Amin, S. Anwar, A. Adnan, M. Nawaz, N. Howard, J. Qadir, A. Hawalah, and A. Hussain, *Comparing oversampling techniques to handle the class imbalance problem: A customer churn prediction case study*, IEEE Access, vol. 4, pp. 7940–7957, 2016.
47. Z. Sun, W. Ying, W. Zhang, and S. Gong, *Undersampling method based on minority class density for imbalanced data*, Expert Systems with Applications, vol. 249, p. 123328, 2024.
48. H. He and Y. Ma, *Imbalanced learning: foundations, algorithms, and applications*, John Wiley & Sons, 2013.
49. P. Kaur and A. Gosain, *Comparing the behavior of oversampling and undersampling approach of class imbalance learning by combining class imbalance problem with noise*, in *ICT Based Innovations: Proceedings of CSI 2015*, Springer, 2018, pp. 23–30.
50. D. Devi, S. K. Biswas, and B. Purkayastha, *A review on solution to class imbalance problem: Undersampling approaches*, in *2020 International Conference on Computational Performance Evaluation (ComPE)*, IEEE, 2020, pp. 626–631.
51. M. Saripuddin, A. Suliman, S. S. Sameon, and B. N. Jorgensen, *Random undersampling on imbalance time series data for anomaly detection*, In *Proceedings of the 2021 4th International Conference on Machine Learning and Machine Intelligence*, pp. 151–156, 2021.

52. J. M. Johnson and T. M. Khoshgoftaar, *Survey on deep learning with class imbalance*, *Journal of Big Data*, vol. 6, no. 1, pp. 1–54, 2019.
53. M. Bach, A. Werner, and M. Palt, *The proposal of undersampling method for learning from imbalanced datasets*, Procedia Computer Science, vol. 159, pp. 125–134, 2019.
54. S. B. Kotsiantis, I. D. Zaharakis, and P. E. Pintelas, *Machine learning: A review of classification and combining techniques*, Artificial Intelligence Review, vol. 26, pp. 159–190, 2006.
55. L. Breiman, *Random forests*, Machine Learning, vol. 45, pp. 5–32, 2001.
56. C. Chen, A. Liaw, and L. Breiman, *Using random forest to learn imbalanced data*, University of California Berkeley, Berkeley, CA, USA, p. 12, 2004.
57. D. A. Anggoro and S. S. Mukti, *Performance comparison of grid search and random search methods for hyperparameter tuning in extreme gradient boosting algorithm to predict chronic kidney failure*, *International Journal of Intelligent Engineering & Systems*, vol. 14, no. 6, 2021.
58. F. Hutter, H. H. Hoos, and T. Stützle, *Automatic algorithm configuration based on local search*, in *Proceedings of the AAAI Conference*, vol. 7, pp. 1152-1157, 2007.
59. R. G. Mantovani, A. L. D. Rossi, J. Vanschoren, B. Bischl, and A. C. P. L. F. De Carvalho, *Effectiveness of random search in SVM hyper-parameter tuning*, in *Proceedings of the 2015 International Joint Conference on Neural Networks (IJCNN)*, pp. 1-8, IEEE, 2015.
60. S. Hanifi, A. Cammarono, and H. Zare-Behtash, *Advanced hyperparameter optimization of deep learning models for wind power prediction*, *Renewable Energy*, vol. 221, p. 119700, Elsevier, 2024.
61. L. Zhang and C. Zhan, *Machine learning in rock facies classification: An application of XGBoost*, in *Proceedings of the International Geophysical Conference, Qingdao, China, 17-20 April 2017*, pp. 1371-1374, Society of Exploration Geophysicists and Chinese Petroleum Society, 2017.
62. P. Yang, W. Liu, B. B. Zhou, S. Chawla, and A. Y. Zomaya, *Ensemble-based wrapper methods for feature selection and class imbalance learning*, in *Advances in Knowledge Discovery and Data Mining: 17th Pacific-Asia Conference, PAKDD 2013, Gold Coast, Australia, April 14-17, 2013, Proceedings, Part I 17*, Springer, pp. 544–555, 2013.
63. K. Gao, T. M. Khoshgoftaar, and A. Napolitano, *Combining Feature Subset Selection and Data Sampling for Coping with Highly Imbalanced Software Data*, in *Proceedings of the International Conference on Software Engineering and Knowledge Engineering (SEKE)*, pp. 439–444, 2015.
64. I. Guyon, S. Gunn, M. Nikravesh, and L. A. Zadeh, *Feature extraction: foundations and applications*, Springer, vol. 207, 2008.
65. H. Guo, W. Shi, and Y. Deng, *Evaluating sensor reliability in classification problems based on evidence theory*, *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 36, no. 5, pp. 970–981, 2006.
66. D. GhoshRoy, P. A. Alvi, and K. C. Santosh, *Unboxing industry-standard AI models for male fertility prediction with SHAP*, *Healthcare*, vol. 11, no. 7, p. 929, 2023.
67. M. Yap, R. L. Johnston, H. Foley, S. MacDonald, O. Kondrashova, K. A. Tran, K. Nones, L. T. Koufariotis, C. Bean, J. V. Pearson, and others, *Verifying explainability of a deep learning tissue classifier trained on RNA-seq data*, *Sci. Rep.*, vol. 11, pp. 1–12, 2021.
68. U. Ahmed, Z. Jiangbin, A. Almogren, M. Sadiq, A. U. Rehman, M. T. Sadiq, and J. Choi, *Hybrid bagging and boosting with SHAP based feature selection for enhanced predictive modeling in intrusion detection systems*, *Scientific Reports*, vol. 14, no. 1, p. 30532, 2024.
69. M. Li, H. Sun, Y. Huang, and H. Chen, *Shapley value: from cooperative game to explainable artificial intelligence*, *Autonomous Intelligent Systems*, vol. 4, no. 1, p. 2, 2024.
70. C. Molnar, *Interpretable machine learning: A guide for making black box models explainable*, Leanpub, 2020.
71. N. Tian, B. Shao, G. Bian, H. Zeng, X. Li, W. Zhao, *Application of forecasting strategies and techniques to natural gas consumption: A comprehensive review and comparative study*, Engineering Applications of Artificial Intelligence, vol. 129, pp. 107644, 2024, Elsevier.
72. T. Schlosser, M. Friedrich, T. Meyer, D. Kowerko, *A consolidated overview of evaluation and performance metrics for machine learning and computer vision*, Tobias Schlosser, Michael Friedrich, Trixy Meyer, and Danny Kowerko–Junior Professorship of Media Computing, Chemnitz University of Technology, vol. 9107, 2024.
73. O. Rainio, J. Teuho, R. Klén, *Evaluation metrics and statistical tests for machine learning*, Scientific Reports, vol. 14, no. 1, pp. 6086, 2024, Nature Publishing Group UK London.
74. V. Agarwal, R. Raman, *A cognitive system for lip identification using convolution neural networks*, in *Cognitive Systems and Signal Processing in Image Processing*, pp. 83–99, Elsevier, 2022.
75. D. Gobov, O. Solovei, *Approaches to Improving the Accuracy of Machine Learning Models in Requirements Elicitation Techniques Selection*, in *International Conference on Computer Science, Engineering and Education Applications*, pp. 574–584, Springer, 2023.
76. D. M. W. Powers, *Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation*, arXiv preprint arXiv:2010.16061, 2020.
77. P. Fränti, R. Mariescu-Istodor, *Soft precision and recall*, Pattern Recognition Letters, vol. 167, pp. 115–121, Elsevier, 2023.
78. K. M. Ting, *Precision and Recall*, in *Encyclopedia of Machine Learning*, vol. 781, 2010.
79. E. J. Michaud, Z. Liu, M. Tegmark, *Precision machine learning*, Entropy, vol. 25, no. 1, pp. 175, MDPI, 2023.
80. A. A. R. Al-chikh Omar, B. Soudan, and others, *A comprehensive survey on detection of sinkhole attack in routing over low power and Lossy network for internet of things*, Internet of Things, vol. 22, pp. 100750, Elsevier, 2023.
81. S. A. Hicks, I. Strümke, V. Thambawita, M. Hammou, M. A. Riegler, P. Halvorsen, and S. Parasa, *On evaluation metrics for medical applications of artificial intelligence*, Scientific Reports, vol. 12, no. 1, pp. 5979, Nature Publishing Group UK London, 2022.
82. N. Ghasemian Sorboni, J. Wang, and M. R. Najafi, *Fusion of Google Street View, LiDAR, and Orthophoto Classifications Using Ranking Classes Based on F1 Score for Building Land-Use Type Detection*, Remote Sensing, vol. 16, no. 11, p. 2011, MDPI, 2024.
83. A. Fujino, H. Isozaki, and J. Suzuki, *Multi-label text categorization with model combination based on F1-score maximization*, Proceedings of the Third International Joint Conference on Natural Language Processing: Volume-II, 2008.

84. M. Açıkkar and S. Tokgöz, *Improving multi-class classification: scaled extensions of harmonic mean-based adaptive k-nearest neighbors*, Applied Intelligence, vol. 55, no. 2, pp. 1-25, Springer, 2025.
85. A. Lavie, K. Sagae, and S. Jayaraman, *The significance of recall in automatic metrics for MT evaluation*, Machine Translation: From Real Users to Research: 6th Conference of the Association for Machine Translation in the Americas, AMTA 2004, Washington, DC, USA, September 28-October 2, 2004. Proceedings 6, pp. 134-143, Springer, 2004.
86. N. J. Wald and J. P. Bestwick, *Is the area under an ROC curve a valid measure of the performance of a screening or diagnostic test?*, Journal of Medical Screening, vol. 21, no. 1, pp. 51-56, SAGE Publications, 2014.
87. J. Li, *Area under the ROC Curve has the most consistent evaluation for binary classification*, PLoS One, vol. 19, no. 12, p. e0316019, Public Library of Science, 2024.
88. J. Muschelli III, *ROC and AUC with a binary predictor: a potentially misleading metric*, Journal of Classification, vol. 37, no. 3, pp. 696-708, Springer, 2020.
89. A. J. Bowers and X. Zhou, *Receiver operating characteristic (ROC) area under the curve (AUC): A diagnostic measure for evaluating the accuracy of predictors of education outcomes*, Journal of Education for Students Placed at Risk (JESPAR), vol. 24, no. 1, pp. 20-46, Taylor & Francis, 2019.
90. M. Zhu, *Recall, precision and average precision*, Department of Statistics and Actuarial Science, University of Waterloo, vol. 2, no. 30, p. 6, 2004.
91. J. Tian, Q. Jin, Y. Wang, J. Yang, S. Zhang, and D. Sun, *Performance analysis of deep learning-based object detection algorithms on COCO benchmark: a comparative study*, *Journal of Engineering and Applied Science*, vol. 71, no. 1, p. 76, Springer, 2024.
92. R. Padilla, S. L. Netto, and E. A. B. Da Silva, *A survey on performance metrics for object-detection algorithms*, in *Proceedings of the 2020 International Conference on Systems, Signals and Image Processing (IWSSIP)*, pp. 237-242, IEEE, 2020.