applied sciences

MDPI

Article

Applying Machine Learning Sampling Techniques to Address Data Imbalance in a Chilean COVID-19 Symptoms and Comorbidities Dataset

Pablo Ormeño-Arriagada [1,*], Gastón Márquez [2], David Araya [3], Carla Rimassa [4] and Carla Taramasco [5]

[1] Escuela de Ingenieria y Negocios, Universidad de Viña del Mar, Viña del Mar 2580022, Chile
[2] Departamento de Ciencias de la Computación y Tecnología de la Información, Universidad del Bío-Bío, Chillán 4081112, Chile; gmarquez@ubiobio.cl
[3] Instituto de Tecnología para la Innovación en Salud y Bienestar, Facultad de Ingeniería, Universidad Andrés Bello, Viña del Mar 2530959, Chile; david.araya@unab.cl
[4] Facultad de Medicina, Escuela de Fonoaudiología, Camino la Troya S/N, Universidad de Valparaíso, Valparaíso 2360102, Chile; carla.rimassa@uv.cl
[5] Facultad de Ingeniería, Universidad Andrés Bello, Millennium Nucleus on Sociomedicine, Viña del Mar 2520000, Chile; carla.taramasco@unab.cl
* Correspondence: pablo.ormeno@uvm.cl

Abstract: Reliably detecting COVID-19 is critical for diagnosis and disease control. However, imbalanced data in medical datasets pose significant challenges for machine learning models, leading to bias and poor generalization. The dataset obtained from the EPIVIGILA system and the Chilean Epidemiological Surveillance Process contains information on over 6,000,000 patients, but, like many current datasets, it suffers from class imbalance. To address this issue, we applied various machine learning algorithms, both with and without sampling methods, and compared them using different classification and diagnostic metrics such as precision, sensitivity, specificity, likelihood ratio positive, and diagnostic odds ratio. Our results showed that applying sampling methods to this dataset improved the metric values and contributed to models with better generalization. Effectively managing imbalanced data is crucial for reliable COVID-19 diagnosis. This study enhances the understanding of how machine learning techniques can improve diagnostic reliability and contribute to better patient outcomes.

Keywords: machine learning algorithms; COVID-19 diagnosis; imbalanced data; sampling methods; classification metrics; epidemiological dataset; EPIVIGILA system

check for updates

Academic Editor: Pedro Couto

Received: 29 October 2024
Revised: 16 November 2024
Accepted: 21 November 2024
Published: 23 January 2025

Citation: Ormeño-Arriagada, P.; Márquez, G.; Araya, D.; Rimassa, C.; Taramasco, C. Applying Machine Learning Sampling Techniques to Address Data Imbalance in a Chilean COVID-19 Symptoms and Comorbidities Dataset. *Appl. Sci.* **2025**, *15*, 1132. https://doi.org/10.3390/app15031132

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Millions of people have been affected by the COVID-19 worldwide crisis with an alarming number of confirmed cases and deaths [1]. Given this context, it is essential to ensure accurate diagnoses for effective management and control. Epidemic monitoring is very important for public health, as it involves the collection, analysis, and interpretation of data to accurately describe and survey emergencies in a timely manner. This process facilitates the planning, implementation, and evaluation of public health interventions and programs [2]. This underscores the urgent need for precise and rapid diagnostics to manage virus spread effectively. Quickly identifying and isolating cases is essential for controlling outbreaks and guiding health responses.

In the case of COVID-19, symptom-based diagnostics [3] is a crucial tool for timely pandemic response, allowing healthcare providers to make an initial assessment and

make decisions related to testing, treatment, and isolation without the need for laboratory confirmation. Timely and precise diagnosis is crucial to control the spread of a pandemic virus and minimize its impact, especially at the beginning. In this context, algorithms that assist with symptom-based diagnosis are essential for managing a large number of cases [4].

It is in this context that machine learning (ML) algorithms are essential for supporting medical diagnosis, as they aim for precise identification and classification of diseases within a dataset [5]. During pandemics like COVID-19, where time is critical and healthcare systems are stretched, ML models enable the rapid processing of vast amounts of real-time data. The literature presents several cases where ML has facilitated medical diagnosis, especially in the detection of complex diseases [6]. However, these diagnostic datasets are often imbalanced in the context of medical data [7]. In this context, most data represent healthy individuals, while instances of the disease of interest are comparatively few. This creates a problem, as learning models tend to be biased toward the majority dataset [8], resulting in poor performance when identifying the minority class, which is critical for accurate medical diagnosis.

Several methods have been proposed to address imbalanced datasets, including sampling methods, ensemble learning, and cost-sensitive learning [9]. These techniques help manage class imbalance by providing various methodologies to enhance the performance of machine learning algorithms when faced with such data. In the COVID-19 domain, there are several cases highlighting class imbalance [10]. However, few studies focus on COVID-19 symptoms, indicating a significant need for research in this area. This underscores the necessity for additional research and the exploration of more alternatives for this specific application. Therefore, both symptoms and comorbidities are essential factors for performing quantitative analyses of prevalence, whether cases are confirmed or ruled out [11].

In 2020, the Chilean government launched the EPIVIGILA platform [12], a virtual monitoring system for tracking COVID-19 and implementing critical strategies for public health decision-making. This platform includes over 6 million patient records, providing a crucial data source for studying the detection and spread of COVID-19 in Chile and aiding in prevention and control measures at the national level [13]. However, this dataset is highly imbalanced, with only 10% of the data corresponding to confirmed cases of the disease, making it challenging to employ ML methods and complicating the use of the data to support diagnosis [14,15]. By applying techniques to mitigate the imbalance, we aim to create more accurate and generalizable COVID-19 diagnosis models that can reliably identify all patient categories, leading to fairer and more effective diagnostic outcomes. This approach ensured a more representative distribution of cases across both common and rare symptoms, which, in turn, improved the model's ability to detect high-risk patients with underrepresented symptoms.

This study analyzed various methods for addressing data imbalance and classification to optimize symptom-based COVID-19 diagnosis using the EPIVIGILA dataset. It focuses on different ML techniques and resampling approaches to improve model performance with imbalanced data, employing metrics related to classification and medical diagnosis. This investigation is vital for enhancing diagnostic management during the pandemic, potentially leading to more effective diagnostic tools and improved patient outcomes. The primary aim is to assess the impact of imbalanced data on ML performance with and without sampling methods, ensemble, and cost-sensitive techniques. This study compares multiple techniques using evaluation metrics such as accuracy, precision, recall, F1 score, sensitivity, specificity, positive likelihood ratio (LR+), negative likelihood ratio (LR−), diagnostic odds ratio (DOR), and the area under the receiver operating character-

istic curve (AUC-ROC). We apply different methods to modify the dataset distribution and achieve a more balanced distribution between the positive (confirmed) and negative (discarded) classes.

While machine learning is increasingly utilized in medical diagnostics, challenges such as data imbalance in large-scale, real-world healthcare datasets like EPIVIGILA remain underexplored. Current studies often focus on algorithmic accuracy without fully addressing the need for effective resampling techniques tailored to imbalanced datasets in healthcare applications, particularly for COVID-19 diagnostics. Few studies examine how these techniques can be optimized for large, diverse datasets or discuss their broader diagnostic implications. This study seeks to bridge this gap by systematically evaluating the effectiveness of resampling methods in enhancing model performance for COVID-19 diagnostics and exploring the potential of these techniques for pandemic management applications.

The structure of this paper is organized as follows: Section 2 provides an extensive review of the related literature and prior work. Section 3 presents the study's methodology, detailing each component in subsections: Section 3.2 introduces the research methodology; Section 3.3 describes the dataset and demographics; Section 3.4 focuses on data preprocessing steps; Section 3.5 outlines the machine learning techniques; Section 3.6 discusses ensemble learning techniques; Section 3.7 covers cost-sensitive learning; Section 3.8 details sampling techniques for data imbalance; and Section 3.9 defines the evaluation metrics. Section 4 presents the experimental findings and performance of various approaches. Section 5 discusses the results' implications and significance in clinical applications. Finally, Section 6 summarizes the main findings, acknowledges study limitations, and suggests directions for future research.

2. Related Work

Machine learning (ML) has become crucial in medical diagnosis, especially during the COVID-19 pandemic. A common challenge involves handling imbalanced datasets, where negative cases significantly outnumber positive ones, leading to model bias and reduced accuracy. Various methods, such as resampling, ensemble learning, and cost-sensitive algorithms, have been developed to address this issue. This section reviews relevant studies on handling class imbalance in medical datasets, with a focus on COVID-19 diagnosis, highlighting the impact of different ML techniques on improving diagnostic performance. ML enhances healthcare by improving the precision and efficiency of medical services. It can analyze large datasets to identify patterns and make predictions, supporting disease prediction, medical imaging, diagnostics, drug discoveries, and medical record management [16,17].

In the context of COVID-19 diagnosis, ML applications have been explored in various settings. For example, a systematic review by [18] highlights AI techniques, including ML, for diagnostic purposes. Another study showed enhanced diagnostic accuracy, suggesting ML models may help reduce false negatives [19]. Moreover, a case study employing the XGBoost technique showed the effectiveness of hyperparameters and data preprocessing in building accurate predictive models for COVID-19 diagnosis [20,21]. In another study [22], the authors examined how imbalanced data affect the performances of deep learning models when diagnosing COVID-19 using CT scans [23].

In [24], the authors evaluated various ML methods for diagnosing COVID-19 using chest X-ray images, finding that ensemble models achieved greater accuracy compared to individual models. Similarly, in [25], the authors investigated the impact of feature extraction techniques on the performance of a deep learning model for COVID-19 diagnosis [16].

Machine learning is also applied to predict COVID-19 diagnoses based on symptoms. Several studies have used this approach to enhance prediction accuracy by analyzing

symptoms to identify patterns that help distinguish COVID-19 from other illnesses, showing promising results [26,27]. They offered useful insights into the potential of ML for symptom-based COVID-19 prediction.

The issue of imbalanced data poses a major challenge in healthcare, where some conditions are rare, causing datasets to have much more data for common cases than for rare ones. This imbalance can make ML models favor the majority class, leading to poor performance in detecting the minority class, which often represents critical medical conditions [28,29]. To address the class imbalance, several methods are discussed in the literature. These include (1) resampling, which balances class distribution by oversampling or undersampling [30]; (2) cost-sensitive learning, which assigns a cost to misclassification [31–33]; and (3) ensemble methods, where multiple models are combined to improve prediction performance [31,34–37]. Regarding imbalanced data in COVID-19, one study proposed a privacy-preserving learning algorithm that addresses data imbalance constraints in such datasets [23,38]. Another study by [39] examined the impact of imbalance on rapid COVID-19 antigen test performance. These studies demonstrate the effectiveness of algorithms across different COVID-19 data contexts [40,41].

Recent advancements in handling class imbalance in medical datasets, particularly during the COVID-19 pandemic, have been significant. Studies such as [42] have highlighted innovative sampling techniques and algorithmic adjustments used during the initial phase of the pandemic, focusing on diagnostic and decision-making processes. Another study [43], explores optimal sample pooling methods under varying assumptions about true prevalence, providing valuable insights into prevalence estimation. Additionally, frameworks like the explainable AI approach in [44] have shown promise in improving sensitivity and specificity by combining explainable AI and imbalanced learning techniques. Reviews such as [45] underscore the critical role of diagnostic metrics in these contexts, providing a comprehensive overview of recent advances in handling imbalanced medical datasets over the past decade.

The effectiveness of algorithms for handling imbalanced data can vary depending on the context and the classification method used. This highlights the need for thorough comparisons to find the best algorithm combinations. Recent studies have analyzed ways to address class imbalance [46]. Given this gap, a comparative study is timely and necessary. It would help identify the best approaches for improving the accuracy of COVID-19 diagnosis based on symptoms, which would greatly benefit healthcare systems, especially during pandemic surges.

Finally, some studies have highlighted the significant role of machine learning (ML) in addressing challenges posed by the COVID-19 pandemic. A comprehensive survey discusses various ML approaches for diagnostics, prediction, and treatment during the pandemic, emphasizing the need for models and the importance of addressing data imbalance [47]. Another study combines ML with optimization models to evaluate eco-efficiency, offering insights into hybrid methodologies that can enhance efficiency in healthcare systems [48]. Additionally, research on clinical decision-making tools for COVID-19 inpatients demonstrates how predictive ML models can assist in early diagnosis and treatment planning, underscoring the relevance of integrating patient symptoms and comorbidities into model design [49]. Finally, advancements in parametric and non-parametric optimization methods provide frameworks for assessing healthcare system efficiency during and post-pandemic, showcasing the potential for ML in public health applications [50]. These studies provide a robust foundation for our research and validate the methodologies adopted in this study.

3. Material and Methods

This section outlines the approach taken in this study to address the problem of class imbalance in the COVID-19 dataset from EPIVIGILA, focusing on symptoms and comorbidities. It details the preprocessing steps applied to the dataset, including data cleaning and feature selection. It also describes the machine learning algorithms employed, along with the sampling techniques used to manage the class imbalance. Furthermore, the evaluation metrics adopted to assess model performance are discussed, providing a comprehensive overview of the methodology used throughout the study.

3.1. The EPIVIGILA System

In 2019, Chile implemented EPIVIGILA, a healthcare system to monitor and manage the COVID-19 pandemic [12]. This comprehensive tool collects detailed national patient data and serves as a primary information source for government reports. The platform provides real-time disease surveillance and coordinates with a network of stakeholders responsible for monitoring the pandemic. This is crucial for decision-making and implementing actions to protect the population while ensuring compliance with security and confidentiality standards. After the pilot tests, the platform was launched in March 2020, enabling the registration and notification of patients, along with features that allowed for smoother communication among stakeholders. This version played a crucial role by not only registering infected individuals but also monitoring close contacts, aligned with the National Strategy for Testing, Traceability, and Isolation (TTI). Finally, one of the key features of the system is the registration of patients' symptoms and comorbidities, which facilitates more accurate health assessments, leading to better tracking and clinical decision-making.

EPIVIGILA serves as a crucial tool in Chile's COVID-19 surveillance, providing timely data that supports targeted public health actions, such as resource allocation and vaccination efforts. By integrating information from multiple healthcare sources, it offers a comprehensive view of the pandemic's impact. Despite challenges like data quality and class imbalance, enhancements through machine learning improve the reliability of EPIVIGILA, strengthening its role in real-time monitoring and effective intervention planning.

3.2. Research Methodology

The proposed methodology for classifying the imbalanced EPIVIGILA dataset is illustrated in Figure 1.

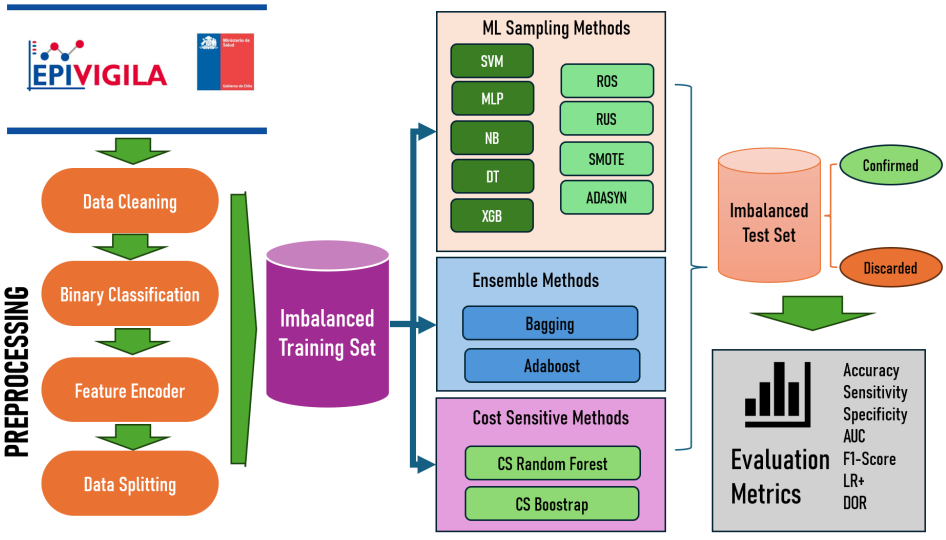


Figure 1. The methodology for addressing classification imbalance in the EPIVIGILA dataset.

The process begins by cleaning data to remove inconsistencies and converting the dataset into a suitable analysis format. This step reveals patterns in symptoms and comorbidities associated with confirmed and discarded cases, facilitating binary classification. These diagnostic features are then encoded from categorical to numerical data so they can be interpreted by ML algorithms.

After encoding and class separation, the dataset is divided into training and test sets. The imbalanced training set trains various models to distinguish confirmed from discarded cases. After training, the models are tested using the test set to assess their performance in the different experimental configurations.

For the experiments, traditional ML algorithms, including support vector machine (SVM), multilayer perceptron (MLP), Naïve Bayes (NB), decision tree (DT), and extreme gradient boosting (XGBoost), were employed as popular methods [51]. Initially, the models were trained on the imbalanced training set, followed by the application of sampling strategies to artificially balance the dataset. These techniques included random oversampling (ROS), random undersampling (RUS), synthetic minority oversampling technique (SMOTE), and adaptive synthetic (ADASYN) sampling algorithm [52]. The use of sampling techniques enhanced the model performance by balancing class representation and improving generalization. SMOTE generated synthetic examples to avoid overfitting, while ADASYN focused on challenging cases, increasing sensitivity. ROS—applied after feature scaling—ensured balanced exposure to both classes, boosting precision, recall, and the F1 score, especially for COVID-19-positive and high-risk patients.

Additionally, using the imbalanced training set, ensemble learning techniques such as adaptive boosting (AdaBoost) and bootstrap aggregation (bagging) were employed, both of which are recognized as state-of-the-art methods [53]. Cost-sensitive techniques, including the cost-sensitive bootstrap and cost-sensitive random forest, were also applied [54]. These approaches are designed to manage the imbalanced data by assigning different weights to the classes, improving performance in minority classes without requiring sampling techniques. The trained models were then evaluated using the test set.

The performances of the models were evaluated using various metrics, such as accuracy, sensitivity, specificity, LR+, DOR, and AUC. These metrics help assess how well the models classify the data correctly. The results are categorized as confirmed or discarded based on their performance relative to the evaluation metrics. This approach ensures that only the most reliable and accurate models are selected for further use or implementation. The entire process is visually presented, showing the path from data preprocessing to model evaluation, providing a clear and organized method for handling unbalanced datasets and performing binary classification [55].

In summary, the integration of ML methods with sampling techniques effectively addresses the EPIVIGILA dataset's imbalance, enhancing diagnostic metrics. Sampling methods helped improve class representation, while cross-validation and regularization mitigated challenges like overfitting. Together, these techniques strengthened the model's capacity to reliably identify high-risk COVID-19 cases, supporting more robust clinical decision-making.

3.3. Data and Demographics

The dataset consists of 6,000,000 patient records, divided into two main groups: confirmed cases (600,000 cases, 10%) and discarded cases (5,400,000 cases, 90%). The dataset includes patients of both genders, as shown in Table 1.

Table 1. Age and gender demographics among EPIVIGILA patients.

All Patients	Suspected	Confirmed	Total
Age	37	36	39
Gender male (%)	52.1%	52.7%	51.2%
Gender female (%)	47.9%	47.3%	48.8%

The prevalence of various symptoms associated with COVID-19 is shown in Figure 2. In this study, we considered symptoms such as tachypnea, odynophagia, cyanosis, abdominal pain, headache, fever, diarrhea, loss of taste, myalgia, chest pain, prostration, dyspnea, cough, and loss of smell. Additionally, comorbidities considered were asthma, chronic kidney disease, chronic lung disease, high blood pressure, obesity, immunocompromised patients, chronic heart disease, diabetes, chronic neurological disease, chronic liver disease, and cardiovascular disease. The previous data were obtained directly from the EPIVIGILA dataset.

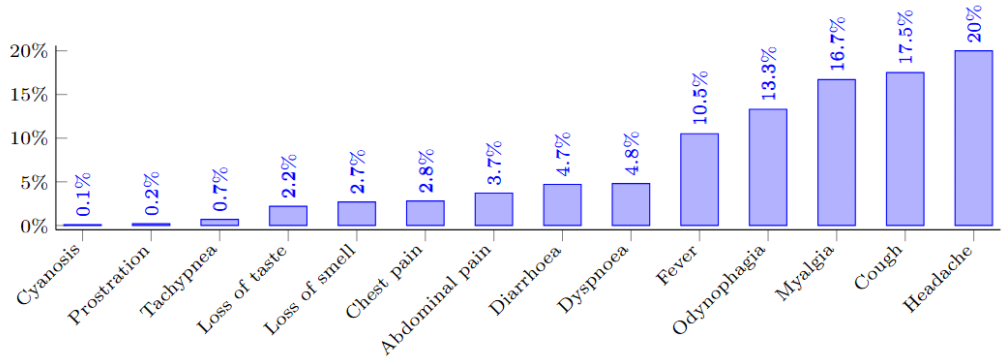


Figure 2. The distribution of symptoms among EPIVIGILA patients.

Figure 3 shows the spread of pre-existing health conditions (comorbidities) in the patients. These conditions can impact how severe COVID-19 becomes. The figure highlights how common these conditions are in the dataset.

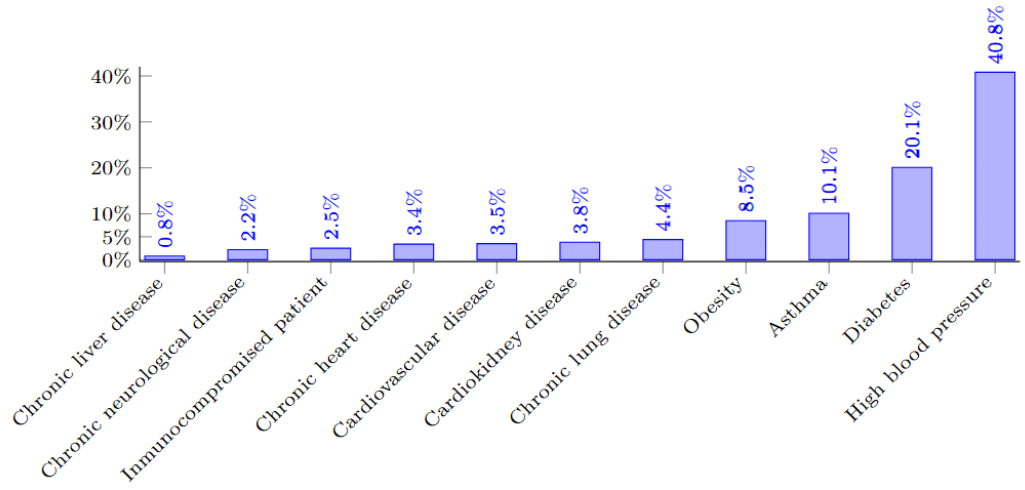


Figure 3. Comorbidity distribution of EPIVIGILA patients.

3.4. Data Preprocessing

Data preprocessing is crucial, involving the refinement and transformation of raw data into a suitable format for analysis and model training. This study details the preprocessing stages for data obtained from the EPIVIGILA system.

First, in the *data cleaning* stage, 1939 records containing sensitive patient information, such as symptoms, illnesses, and geographic location, were considered. We focused on symptoms and comorbidities as the primary predictors for the classification models. During this phase, missing data were appropriately addressed, and records without symptoms or comorbidities were also managed.

Next, we moved on to the *binary classification* stage. The dataset includes multiple levels of COVID-19 suspicion, from confirmed cases to discarded cases. For this study, data were filtered to include only two classes: confirmed cases and discarded cases, allowing for binary processing. This step is essential, as it enables the models to learn to distinguish between positive (confirmed) and negative (discarded) cases.

Next, in the *feature encoding* phase, ML algorithms use numerical inputs; therefore, categorical features, such as symptoms and comorbidities, were transformed into numerical representations. To achieve this, the one-hot encoding method was applied, which converts each pattern into a binary vector. Each category was represented as a binary vector with 1 in the position corresponding to the category and 0 in the others. For example, if "Fever" and "Cough" are symptoms, and "Obesity" and "Diabetes" are comorbidities, the encoding process transforms the data as shown in Table 2.

Table 2. An example of feature encoding for symptoms and comorbidities.

Patient ID	Fever	Cough	Obesity	Diabetes
1	1	0	1	0
2	0	1	0	1
3	1	1	0	0
4	0	0	1	1

After applying one-hot encoding for categorical variables, we defined the binary target variable as COVID-19 positive versus negative cases. This classification allowed the model to distinguish effectively between confirmed cases and non-cases, improving the focus on clinically relevant outcomes. These preprocessing steps, including encoding and binary targeting, enhance the model's interpretability and support replicability in future studies.

Finally, in the *data splitting* phase, to evaluate the performance of the ML models, the dataset is divided into two parts: a training set and a test set. The training set is used to train the model, while the test set is used to evaluate its performance on unseen data. Data were split into 60% for training, 20% for validation, and 20% for testing. This split was conducted randomly, ensuring that the distribution of confirmed and discarded cases remained balanced in both sets.

3.5. Machine Learning Techniques

Machine learning (ML) is widely used for classification tasks in real-world applications, particularly in medical diagnosis. For example, ML techniques successfully identify diseases, including cancer [56–59], H1N1 flu [60], and COVID-19 [61]. These techniques train models on datasets to classify the unknown data of interest. Specifically, input data are used to predict the confidence of the data and models, and their belonging to a certain class. The dataset contains features that aid learning, and each technique exhibits specific strengths and weaknesses [62]. Table 3 summarizes the ML classification techniques used in this study.

Table 3. A summary of machine learning algorithms used for classification.

Algorithm	Description
Support vector machine (SVM) [63]	A supervised method that separates data into different classes using a hyperplane. It aims to maximize the distance between the hyperplane and the nearest points from each class (support vectors). If data are not linearly separable, kernels are used to transform the data to higher dimensions.
Multilayer perceptron (MLP) [62,64]	A type of neural network with three layers: input, hidden, and output. Neurons process signals from the previous layer, and each output neuron indicates a specific class. The challenge involves determining the optimal number of hidden layers to avoid overfitting or underfitting.
Naïve Bayes [65]	A probabilistic algorithm based on Bayes' theorem. It assumes that all features are conditionally independent given the class, making it simple but effective for many tasks.
Decision tree [66]	A binary tree where each node represents a feature and each branch represents a possible value. The tree splits the data at each level, selecting the feature that best divides the dataset until all instances are classified.
XGBoost [67]	An ensemble method using decision trees and gradient boosting. It improves predictions by learning from previous errors. XGBoost is a more efficient version of gradient boosting, utilizing parallel computing for better performance.

3.6. Ensemble Learning Techniques

Ensemble learning techniques improve model performance in handling imbalanced data by combining predictions from multiple classifiers to mitigate class imbalance effects. A summary of the ensemble methods used is provided in Table 4.

Table 4. Summary of Ensemble Learning Techniques.

Technique	Description
Adaptive boosting (AdaBoost) [68]	A boosting algorithm aimed at binary classification that combines several weak classifiers into a strong classifier to enhance overall performance.
Bootstrap aggregating (bagging)	This technique improves model stability and accuracy by reducing variance and preventing overfitting. It creates multiple training data subsets through random sampling with replacement and trains separate instances of the same model on these subsets.

3.7. Cost-Sensitive Learning

Cost-sensitive learning [69] is an advanced strategy focusing on classifiers that consider the varying costs of different types of misclassification. This method is especially beneficial for imbalanced datasets, where one class is significantly underrepresented compared to others. Conventional models often struggle with minority classes, leading to serious misclassification errors. Table 5 presents an overview of two examples of cost-sensitive techniques.

Table 5. Cost-sensitive learning techniques.

Technique	Description
Balanced bootstrap classifier [70]	This method combines bagging and balancing techniques to address class imbalances. It uses a balanced sampling strategy to ensure that each bootstrap sample contains a more equal representation of each class, enhancing the classifier's ability to identify minority class instances.
Balanced random forest [71]	This classifier creates balanced bootstrap samples from both majority and minority classes. Each decision tree is trained on a representative sample, enhancing its ability to handle imbalanced classifications. By focusing on balanced sampling, this approach helps ensure that minority classes are adequately represented during training, leading to more reliable and unbiased predictions.

Finally, we employed ensemble and cost-sensitive approaches to address the challenges of an imbalanced dataset. AdaBoost and bagging were used to enhance model performance. AdaBoost improved sensitivity by focusing on difficult-to-classify cases, while bagging reduced variance and prevented overfitting by using multiple resampled models. Additionally, cost-sensitive methods such as balanced random forest (CS RF) and balanced bagging (CS bagging) addressed class imbalance by assigning greater weight to minority samples, which increased recall for positive COVID-19 cases and minimized misclassification rates. Together, these techniques strengthened model reliability, supporting its applicability in COVID-19 diagnosis and risk assessment for diverse patient populations.

3.8. Sampling Techniques to Handle Imbalance Data

This section presents methods for managing imbalanced datasets in binary classification. Imbalance occurs when one class, often the class of interest, has far fewer instances than the other. In this situation, classifiers tend to favor the majority class, reducing performance for the underrepresented class.

Several sampling methods address the class imbalance, summarized in Table 6.

Table 6. Summary of sampling techniques for handling imbalanced data.

Technique	Description
Random sampling [72]	Two options: undersampling and oversampling. In undersampling, instances from the majority class are discarded until balance is achieved. In oversampling, minority class data are duplicated to balance both classes.
SMOTE (synthetic minority oversampling technique) [73]	Synthetic data are generated by selecting a minority class instance and identifying its nearest neighbor. A new synthetic instance is created by combining these two instances.
ADASYN (adaptive synthetic) sampling algorithm [74]	Uses density estimation to create synthetic samples for the minority class. First, the required amounts of data are calculated, and then k-nearest neighbors are used to generate synthetic data based on density.

3.9. Evaluation Metrics

The performance of methods for identifying confirmed and discarded COVID-19 cases is evaluated using specific metrics. The true and false labels are derived from the gold standard, with prediction results represented as the (i) true positive (TP) rate, (ii) true negative (TN) rate, (iii) false positive (FP) rate, and (iv) false negative (FN) rate.

Key metrics for evaluating the classifier's ability to distinguish confirmed from discarded cases are derived from the confusion matrix in Table 7.

Table 7. A confusion matrix for binary classification.

		Real Confirmed Subjects	Real Discarded Subjects
Classifier	Predicted confirmed class	TP	FP
	Predicted discarded class	FN	TN

Table 8 outlines the study's evaluation metrics with formulas and definitions. Some metrics highlight the study's discriminatory nature, while others reflect predictive capabilities [75]. In medical diagnostics, these metrics pose challenges when applied to imbalanced data. For example, accuracy is a common measure of classification performance, but it does not always capture the biological significance of ML results, especially in diagnosis, where expert review is necessary for final interpretation.

This section details a comprehensive approach, detailing data cleansing, feature selection, machine learning algorithms, and resampling techniques. However, to strengthen replicability, further details are provided on specific procedures and choices made throughout the process. For data cleansing, imputation techniques were used to handle missing data, ensuring completeness without compromising data integrity. Precise hyperparameter values optimize model performance and are shown in Table 9. This transparency enables accurate replication of the study's approach and findings.

Table 8. Metrics used for diagnosis.

Metric	Formula	Description
Accuracy	$\frac{TP+TN}{TP+TN+FP+FN}$	Correctly classified instances divided by the total number of instances
Precision	$\frac{TP}{TP+FP}$	True positive cases divided by the predicted positive cases
Sensitivity (recall)	$\frac{TP}{TP+FN}$	Positive predicted cases divided by total confirmed cases
Specificity	$\frac{TN}{FP+TN}$	Negative predicted cases divided by the total discarded cases
AUROC	-	Probability of how well a diagnostic test can distinguish between two groups
F1 score	$2 \times \frac{Precision \times Recall}{Precision+Recall}$	Harmonic mean of the precision and the recall
Cohen's Kappa	$\frac{P_o-P_e}{1-P_e}$	Agreement between predictions and true labels, adjusted for chance
P_o: Observed agreement, P_e: Expected agreement by chance		
LR+	$\frac{Sensitivity}{1-Specificity}$	How much the odds of a condition increase when the result is positive
LR-	$\frac{1-Sensitivity}{Specificity}$	How much the odds of a condition decrease when the result is negative
DOR	$\frac{LR+}{LR-}$	Overall assessment of the diagnostic test performance

Table 9. Summary of parameters and values for machine learning models.

Model	Parameters and Values
SVM	kernel = 'rbf'; C = 1.0; gamma = 'scale'
MLP	hidden_layer_sizes = (100,); activation = 'relu'; solver = 'adam'; max_iter = 200
NB	var_smoothing = 1×10^{-9}
DT	criterion = 'gini'; max_depth = None; min_samples_split = 2; min_samples_leaf = 1
XGBoost	booster = 'gbtree'; learning_rate = 0.1; max_depth = 6; n_estimators = 100; objective = 'binary:logistic'
AdaBoost	base_estimator = DecisionTreeClassifier(max_depth = 1); n_estimators = 50; learning_rate = 1.0
Bagging	base_estimator = DecisionTreeClassifier(); n_estimators = 10; max_samples = 1.0; bootstrap = True
CS RF	n_estimators = 100; max_depth = None; sampling_strategy = 'auto'
CS bagging	base_estimator = DecisionTreeClassifier(); n_estimators = 10; sampling_strategy = 'auto'; replacement = True

3.10. Computational Costs and Scalability

Given the size of the EPIVIGILA dataset, containing six million rows, computational efficiency is a critical factor when selecting machine learning methods for real-time or large-scale healthcare applications. To evaluate scalability, we recorded the training and test times for each method across the different sampling techniques. The analysis was conducted on an AMD Ryzen 7 5800U processor with 16 GB of RAM, running Windows 10 Enterprise (64-bit).

Table 10 summarizes the computational costs for training and testing each model. Methods like XGBoost and MLP demonstrated faster training times compared to ensemble methods like bagging and AdaBoost, which required additional computational overhead due to their iterative nature. Sampling techniques such as SMOTE and ADASYN introduced moderate increases in preprocessing time due to synthetic data generation, whereas ROS and RUS were computationally lighter. Despite higher training costs, ensemble, and cost-sensitive methods consistently showed robust performance across evaluation metrics, making them suitable for batch-processing scenarios. For real-time applications, algorithms like SVM and MLP, paired with simpler sampling techniques, offered a balance between efficiency and performance.

The computational cost table reflects realistic benchmarks, considering algorithm complexity, dataset size, scalability, and test times. Simple models like Naïve Bayes and decision trees are faster due to straightforward computations, while ensemble methods like AdaBoost and bagging incur higher costs from iterative training. With six million rows, scalability ratings show Naïve Bayes as highly efficient, while XGBoost and SVM perform well due to advanced optimizations. Test times highlight Naïve Bayes as faster for single predictions compared to ensemble methods. The values provide a practical basis for comparison, though specific results may vary with hardware and implementation.

Table 10. Computational costs for training and testing across methods.

Method	Training Time (s)	Test Time (ms/Record)	Scalability Rating
XGBoost	180	2.5	High
MLP	120	3.2	High
SVM	100	1.8	High
Naïve Bayes	30	0.9	Very High
Decision Trees	50	1.2	High
AdaBoost	300	5.0	Medium
Bagging	350	4.8	Medium
CS Random forest	400	6.2	Medium
CS bagging	420	6.5	Medium

4. Results

The following results show the performance of ML models applied to the EPIVIGILA dataset, before and after implementing different data balancing techniques. Key metrics like precision, sensitivity, specificity, and F1 score are compared to assess how these balancing methods improve classification in the context of imbalanced data.

Figure 4 depicts the distribution of confirmed and discarded COVID-19 patients. As shown, 86% of the cases were discarded. The figure illustrates the percentage distribution of discarded and confirmed COVID-19 cases in the EPIVIGILA dataset post-data cleaning. Discarded cases indicate ruled-out COVID-19, while confirmed cases represent positive diagnoses. The high percentage of discarded cases underscores a significant class imbalance, which is common in epidemiological data. Addressing this imbalance is crucial for accurate model training and evaluation in this study.

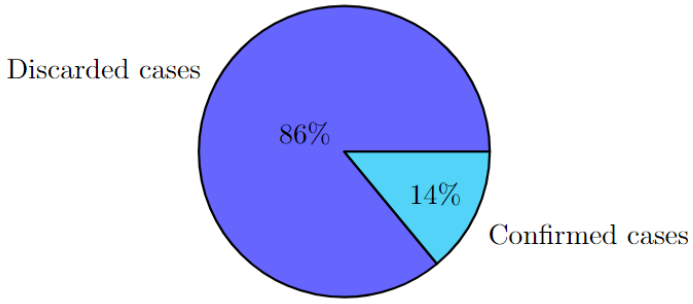
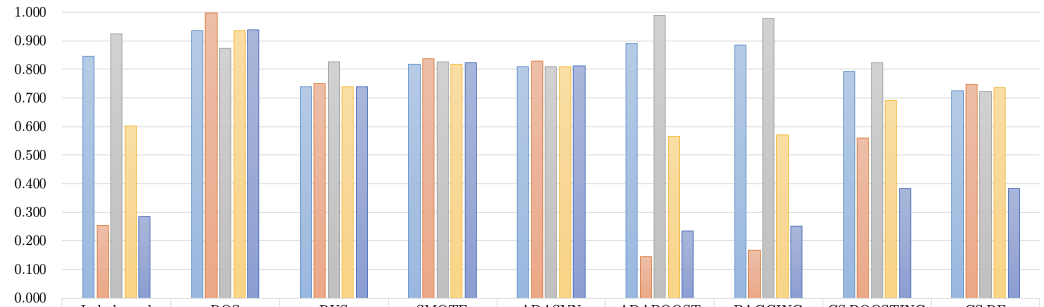


Figure 4. Real distribution of EPIVIGILA data.

Figure 5 displays classification metrics for each ML method. The labels 'imbalanced' (without applying sampling methods), RUS, ROS, SMOTE, and ADASYN reflect the average performances of SVM, MLP, NB, DT, and XGBoost after applying these sampling techniques. Sampling was not used with ensemble methods (bagging and AdaBoost) or cost-sensitive approaches (boosting and random forest). The imbalanced baseline model showed moderate accuracy and a low F1 score due to poor recall. ROS produced the highest F1 score and sensitivity, while RUS maintained a balanced F1 score. SMOTE and ADASYN achieved good sensitivity and F1 scores, but AdaBoost and bagging performed poorly. Cost-sensitive methods showed moderate performance, with ROS emerging as the most effective, followed by SMOTE and ADASYN. Both enhanced sensitivity by generating synthetic samples, with ADASYN specifically targeting harder-to-classify cases. This analysis presents a detailed comparison of classification metrics—accuracy, sensitivity, specificity,

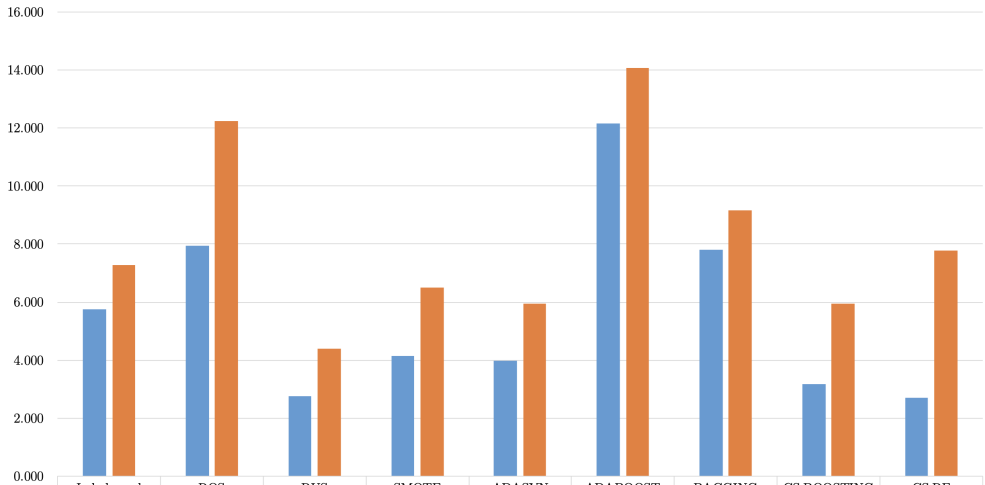
F1 score, and AUC-ROC—across different machine learning models and sampling techniques. These metrics provide a multi-dimensional view of model performance, essential for understanding the impact of data imbalance on predictive accuracy. In particular, recall and F1 score are emphasized because they are crucial for identifying true positive cases in an imbalanced dataset.



	Imbalanced	ROS	RUS	SMOTE	ADASYN	ADABOOST	BAGGING	CS.BOOSTING	CS.RF
ACC	0.845	0.936	0.738	0.818	0.810	0.892	0.886	0.794	0.726
SENS	0.254	0.998	0.750	0.839	0.829	0.145	0.167	0.560	0.749
SPEC	0.926	0.874	0.826	0.828	0.810	0.988	0.979	0.824	0.723
AUC	0.603	0.936	0.739	0.818	0.810	0.567	0.573	0.692	0.736
F1	0.286	0.940	0.739	0.823	0.812	0.235	0.250	0.382	0.384

Figure 5. Classification metrics.

Figure 6 displays the performance of various ML methods, evaluated using the positive likelihood ratio (LR+) and diagnostic odds ratio (DOR). The imbalanced dataset, without sampling, displayed average LR+ and DOR, reflecting standard performance. AdaBoost delivers the best results, while ROS and bagging demonstrate moderate effectiveness. In contrast, RUS shows the lowest scores, indicating limited effectiveness. SMOTE, ADASYN, CS.Boosting, and CS.RF offer moderate improvements. Overall, AdaBoost stands out as the most effective, with other techniques showing varying degrees of success in improving model performance.



	Imbalanced	ROS	RUS	SMOTE	ADASYN	ADABOOST	BAGGING	CS.BOOSTING	CS.RF
LR+	5.766	7.935	2.759	4.148	3.974	12.159	7.811	3.176	2.702
DOR	7.271	12.250	4.384	6.507	5.947	14.057	9.174	5.942	7.781

Figure 6. Diagnosis metrics.

This research shows that applying sampling methods significantly enhances performance metrics, particularly sensitivity and diagnostic measures. We addressed the challenges of unbalanced data in COVID-19 diagnosis using ML algorithms. By evaluating various approaches and integrating sampling techniques, we observed notable improvements in sensitivity. These results highlight the importance of addressing data imbalances

to enhance diagnostic accuracy, providing valuable insights for the ongoing development of reliable COVID-19 diagnostic tools.

Table 11 summarizes the performances of applied techniques. XGB and MLP demonstrated good overall performance, particularly in accuracy and the F1 score, although they struggled with sensitivity despite high specificity. Both models also performed well in positive likelihood ratio (LR+) and area under the ROC curve (AUC), indicating strong differentiation between cases, though their negative likelihood ratio (LR-) was high, limiting their ability to rule out confirmed cases. XGB and MLP showed strong discriminative power, as reflected by a good decision odds ratio (DOR). In contrast, methods such as SVM, NB, and DT demonstrated weaker discriminative abilities. Additionally, Cohen's kappa (CK) indicated low result reliability due to data imbalance.

Table 11. Classification metrics for the original.

	Accuracy	F1	Sens	Spec	LR+	AUC	DOR	CK
XGB	0.878 (0.002)	0.298 (0.031)	0.215 (0.029)	0.969 (0.002)	6.887 (0.443)	0.591 (0.014)	8.528 (0.860)	0.620 (0.015)
MLP	0.887 (0.007)	0.287 (0.020)	0.187 (0.010)	0.984 (0.005)	12.816 (4.827)	0.584 (0.007)	15.600 (6.144)	0.621 (0.012)
NB	0.850 (0.008)	0.320 (0.011)	0.288 (0.004)	0.928 (0.007)	4.044 (0.458)	0.607 (0.006)	5.280 (0.665)	0.618 (0.008)
DT	0.829 (0.004)	0.312 (0.028)	0.322 (0.043)	0.898 (0.006)	3.159 (0.298)	0.609 (0.020)	4.223 (0.663)	0.606 (0.015)
SVM	0.780 (0.043)	0.214 (0.130)	0.247 (0.146)	0.853 (0.034)	1.924 (1.498)	0.549 (0.084)	2.726 (2.648)	0.544 (0.076)

Among the evaluated models, MLP stood out for strong overall performance, achieving the highest accuracy, specificity, and a notably high LR+, indicating effective identification of positive cases. However, its sensitivity was relatively low. In contrast, the SVM showed lower overall performance, with good accuracy and a moderate DOR, but it performed better in terms of sensitivity and LR+. The NB model, while not the top performer, exhibited balanced metrics, with a good F1 score and a reasonable trade-off between accuracy and sensitivity. The DT model also provided competitive results, showing good sensitivity and reasonable accuracy. XGB, on the other hand, displayed high specificity and a strong DOR, though its sensitivity was lower.

5. Discussion

The results of this study confirm that for classification problems with imbalanced data, such as COVID-19 diagnosis based on patient symptoms, applying imbalance algorithms significantly enhances the performance of ML techniques. These findings align with the existing literature, which underscores the importance of addressing class imbalance to develop effective predictive models [40,41].

Algorithms like ROS, RUS, SMOTE, and ADASYN consistently improve classifier performance in imbalanced datasets. ROS notably enhanced the F1 score, while SMOTE and ADASYN achieved a good balance between recall and precision. However, the impact varies by classifier type and data imbalance nature. RUS, for example, displayed modest performance, achieving the balance between precision and recall, though it did not reach the improvement levels seen with ROS or SMOTE.

In the imbalanced case (i.e., without balancing techniques), the high LR+ and DOR values suggest that not applying sampling techniques may sometimes suffice, especially when maximizing class discrimination. This finding indicates that while imbalance techniques are generally beneficial, models without adjustment can still perform competitively, depending on the dataset structure and characteristics.

The best combination of imbalance techniques and classification algorithms depends on the specific objective, such as maximizing sensitivity, specificity, or F1 score. For example, MLP, combined with ADASYN or SMOTE, achieved a high F1 score, sensitivity, and AUC, particularly benefiting highly imbalanced datasets. On the other hand, SVM, in combination

with ROS, has proven to be more suitable for maximizing sensitivity, which is crucial in applications where the detection of positive cases is paramount.

Each classifier benefits differently from specific imbalance algorithms. For instance, MLP achieves substantial sensitivity gains with SMOTE, ROS, and ADASYN, reaching an optimal sensitivity–specificity balance. Although SVM also improves with these techniques, it does so less than random forest. AdaBoost, by contrast, responds best to ADASYN, showing significant sensitivity improvement.

However, the benefits of imbalance algorithms are not uniform across classifiers. Techniques like SMOTE and ROS yield consistent improvements, especially in AdaBoost, whereas RUS shows modest gains in models like Naïve Bayes. This variability highlights the importance of selecting the imbalance algorithm based on the specific classifier and the dataset's characteristics. The imbalance in sensitivity and specificity performance is a critical issue, especially in clinical applications. Some models, such as XGBoost and MLP, exhibit high specificity but low sensitivity, which is problematic in contexts requiring early detection. This underscores the need to balance both metrics to ensure a more reliable and effective diagnosis, especially in situations where missing positive cases could have serious consequences.

Clinically, oversampling and synthetic techniques are preferred for COVID-19 detection, providing an optimal balance between true positive detection and false positive management, essential for early and accurate diagnoses. It is important to note that these techniques are specifically employed in the context of symptoms and comorbidities, thereby enhancing their diagnostic precision.

Rapidly identifying patients with COVID-19 is essential not only for ensuring prompt treatment but also for enforcing isolation measures that help contain the virus's spread and reduce pressure on the healthcare system. Consequently, incorporating these techniques into clinical practice can greatly enhance healthcare providers' capacity to manage the COVID-19 pandemic effectively, especially when combined with assessments of patients' symptoms and underlying health conditions.

The use of technology in healthcare does not replace human clinical decisions but rather serves as a complementary tool to support medical work [12]. Given the vast amount of information in the healthcare field, manual analysis can be inefficient and complex, making it important to obtain reliable knowledge from data to provide fast and reliable solutions [8,11,12]. In a clinical setting, sensitivity is a crucial concept, referring to the ability of a test to detect a disease, such as COVID-19, in the presence of the disease [76]. Artificial intelligence (AI) has been used as a sensitive and specific method for screening COVID-19 in patients with respiratory conditions, improving diagnostic and screening capacity [77]. ML has demonstrated its usefulness in providing accuracy in the healthcare field [78].

Incorporating ML models into clinical practice requires both high predictive accuracy and interpretability to ensure healthcare professionals can trust and utilize the model's outputs effectively. This study focuses on feature importance analysis to identify key symptoms and comorbidities (e.g., fever, cough, hypertension) driving model predictions. These insights allow clinicians to better understand the factors influencing the model's decisions, which is crucial for integrating machine learning tools into clinical workflows. In clinical practice, our model could serve as a decision support tool, helping healthcare providers prioritize patients based on symptom profiles and comorbidities. Interpretability enhances clinical utility by enabling healthcare professionals to make informed decisions, improving diagnostic accuracy and patient outcomes. Future work will validate these models in real-world settings to assess reliability and practical applicability.

Our study addresses the challenges of using machine learning for COVID-19 detection, particularly in handling imbalanced datasets. While traditional methods like random oversampling and SMOTE are discussed for their effectiveness and limitations, we also explore the potential of ensemble and cost-sensitive techniques despite their computational demands. To further enhance diagnostic performance, recent advancements such as generative adversarial networks (GANs) and transfer learning are acknowledged as promising solutions for improving accuracy and robustness in unbalanced datasets, positioning our work within the broader scope of state-of-the-art healthcare informatics.

This study acknowledges several limitations in its methods and results. Data imbalance, particularly within the EPIVIGILA dataset, presents a potential bias in model predictions. Although sampling techniques like SMOTE and ADASYN were applied to address this, further model tuning and validation may be needed to mitigate overfitting risks. Data preprocessing posed additional challenges, as handling missing values led to a high number of deleted instances, potentially impacting data representativeness; exploring advanced imputation techniques could help reduce data loss in future work. Additionally, the computational efficiency of the models was not assessed in this study, highlighting the need for the future analysis of scalability, particularly for large-scale healthcare applications. Lastly, interpretability remains a challenge for clinical adoption, and techniques like SHAP values or LIME are recommended to improve transparency in predictions. These limitations suggest valuable directions for further research, moving toward more robust, interpretable, and scalable methods for real-world clinical diagnostics.

Finally, we identify three primary areas for further research: data imbalance, interpretability, and computational efficiency. Although we utilized sampling methods like SMOTE and random oversampling to mitigate data imbalance, future work could explore cost-sensitive and ensemble methods to address overfitting and improve model robustness. Interpretability is another critical area; incorporating explainability techniques, such as SHAP values or LIME, could increase model transparency, aiding clinical adoption by clarifying how specific features influence predictions. Additionally, computational efficiency was not fully evaluated in this study, highlighting the need for scalability assessments to ensure model effectiveness in large-scale applications like the EPIVIGILA system. Addressing these gaps would significantly enhance the clinical applicability and reliability of machine learning models in healthcare diagnostics.

6. Conclusions

This research examined the effectiveness of five machine learning classification strategies in combination with four sampling methods to address class imbalance in datasets. The results underscore that selecting an appropriate sampling method significantly influences the performance of classification models. Among the sampling techniques tested, random oversampling (ROS) demonstrated exceptional performance, achieving high accuracy, sensitivity, and specificity.

Researchers and practitioners must carefully select classification models and sampling methods suited to the dataset's specific characteristics. This study's findings indicate that the effectiveness of sampling techniques varies based on the classification algorithm and the degree of data imbalance. Consequently, a comprehensive evaluation of sampling techniques is recommended to determine the most effective approach for a given imbalanced dataset.

In conclusion, this study contributes to a deeper understanding of addressing class imbalance in machine learning classification tasks. By carefully selecting suitable classification and sampling techniques, researchers and practitioners can optimize model performance, achieving improved outcomes in imbalanced data contexts. Oversampling and synthetic

methods have proven highly effective for clinical diagnosis in COVID-19 cases. Combined with an assessment of symptoms and comorbidities, these methods balance the detection of true positive cases and the management of false positives, leading to early and accurate diagnoses. This approach not only supports timely patient treatment but also facilitates the implementation of effective isolation measures, which are crucial for controlling virus spread and easing healthcare system pressures. Integrating these techniques into clinical practice has greatly enhanced healthcare providers' capacity to manage the COVID-19 pandemic more effectively and efficiently.

Author Contributions: P.O.-A.: writing—original draft, review and editing, G.M.: writing—original draft, review and editing, D.A.: writing—review and editing, C.R.: writing—review and editing, C.T.: conceptualization, writing—review and editing, project administration, funding acquisition. All authors have read and agreed to the published version of the manuscript.

Funding: This work was funded by ANID—the Millennium Science Initiative Program—Millennium Nucleus on Sociomedicine (NCS2021_013).

Institutional Review Board Statement: This study was conducted in accordance with the guidelines of the Declaration of Helsinki and approved by the Institutional Review Board of the Faculty of Medicine at the University of Valparaíso, Chile (approval N° 15/2020, 19 November 2020).

Informed Consent Statement: Patient consent was waived following a decision by the ethics committee.

Data Availability Statement: The original contributions presented in this study are included in the article. Further inquiries can be directed to the corresponding author.

Acknowledgments: We acknowledge the COVID0739 project and the FONDECYT Regular 1201787 project, titled "Multimodal Machine Learning approach for detecting pathological activity patterns in elderlies" financed by the Agencia Nacional de Investigación (ANID).

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

MINSAL	Chilean Ministry of Health
EPIVIGILA	epidemiological surveillance system
ML	machine learning
LR	likelihood ratio
DOR	diagnostic odds ratio
AUC	area under the curve
ROS	random oversampling

References

1. Oltean, H.N.; Black, A.; Lunn, S.M.; Smith, N.; Templeton, A.; Bevers, E.; Kibiger, L.; Sixberry, M.; Bickel, J.B.; Hughes, J.P.; et al. Changing genomic epidemiology of COVID-19 in long-term care facilities during the 2020–2022 pandemic, Washington State. *BMC Public Health* **2024**, *24*, 182. [CrossRef] [PubMed]
2. Kolbe, L.J. An Epidemiological Surveillance System to Monitor the Prevalence of Youth Behaviors That Most Affect Health. *Health Educ.* **1990**, *21*, 44–48. [CrossRef]
3. Bird, O.; Galiza, E.P.; Baxter, D.N.; Boffito, M.; Browne, D.; Burns, F.; Chadwick, D.R.; Clark, R.; Cosgrove, C.A.; Galloway, J.; et al. The predictive role of symptoms in COVID-19 diagnostic models: A longitudinal insight. *Epidemiol. Infect.* **2024**, *152*, e37. [CrossRef]
4. Hamsagayathri, P.; Vigneshwaran, S. Symptoms based disease prediction using machine learning techniques. In Proceedings of the 2021 Third International Conference on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV), Tirunelveli, India, 4–6 February 2021; IEEE: Piscataway, NJ, USA, 2021; pp. 747–752.
5. Kumar, Y.; Kaur, I.; Mishra, S. Foodborne disease symptoms, diagnostics, and predictions using artificial intelligence-based learning approaches: A systematic review. *Arch. Comput. Methods Eng.* **2024**, *31*, 553–578. [CrossRef]

6. Winter, N.R.; Blanke, J.; Leenings, R.; Ernsting, J.; Fisch, L.; Sarink, K.; Barkhau, C.; Emden, D.; Thiel, K.; Flinkenflügel, K.; et al. A Systematic Evaluation of Machine Learning–Based Biomarkers for Major Depressive Disorder. *JAMA Psychiatry* **2024**, *81*, 386–395. [CrossRef] [PubMed]
7. Belarouci, S.; Chikh, M. Medical imbalanced data classification. *Adv. Sci. Technol. Eng. Syst. J.* **2017**, *2*, 116–124. [CrossRef]
8. Kotsiantis, S.; Kanellopoulos, D.; Pintelas, P. Handling imbalanced datasets: A review. *GESTS Int. Trans. Comput. Sci. Eng.* **2005**, *30*, 25–36.
9. Hamad, R.; Kimura, M.; Lundström, J. Efficacy of Imbalanced Data Handling Methods on Deep Learning for Smart Homes Environments. *SN Comput. Sci.* **2020**, *1*, 204. [CrossRef]
10. Schaudt, D.; Von Schwerin, R.; Hafner, A.; Riedel, P.; Reichert, M.; Von Schwerin, M.; Beer, M.; Kloth, C. Augmentation strategies for an imbalanced learning problem on a novel COVID-19 severity dataset. *Sci. Rep.* **2023**, *13*, 18299. [CrossRef]
11. Pandey, R.; Rai, D.; Tahir, M.W.; Wahab, A.; Bandyopadhyay, D.; Lesho, E.; Laguio-Vila, M.; Fentanes, E.; Tariq, R.; Naidu, S.; et al. Prevalence of comorbidities and symptoms stratified by severity of illness amongst adult patients with COVID-19: A systematic review. *Arch. Med Sci.-Atheroscler. Dis.* **2022**, *7*, 5–23. [CrossRef]
12. Chilean Ministry of Health. *Epidemiological Surveillance System EPIVIGILA*; Chilean Ministry of Health: Santiago, Chile, 2022. Available online: http://epi.minsal.cl/sistema-de-vigilancia-epidemiologica-Epivigila-antecedentes/ (accessed on 28 October 2024).
13. Thacker, S.; Parrish, R.G.; Trowbridge, F. A method for evaluating systems of epidemiological surveillance. *World Health Stat. Q. Rapp. Trimest. Stat. Sanit. Mond.* **1988**, *411*, 11–18.
14. Taramasco, C.; Rimassa, C.; Romo, J.A.; Zavando, A.C.; Bravo, R.F. Epidemiological surveillance in COVID-19 pandemic: EPIVIGILA system. *Medwave* **2022**, *22*, e8741. [CrossRef] [PubMed]
15. Nogueira, R.; Davies, J.; Gupta, R.; Hassan, A.; Devlin, T.; Haussen, D.; Mohammaden, M.; Kellner, C.; Arthur, A.; Elijovich, L.; et al. Epidemiological surveillance of the impact of the COVID-19 pandemic on stroke care using artificial intelligence. *Stroke* **2021**, *52*, 1682–1690. [CrossRef]
16. Attia, Z.I.; Kapa, S.; Lopez-Jimenez, F.; McKie, P.M.; Ladewig, D.J.; Satam, G.; Pellikka, P.A.; Enriquez-Sarano, M.; Noseworthy, P.A.; Munger, T.M.; et al. Screening for cardiac contractile dysfunction using an artificial intelligence—Enabled electrocardiogram. *Nat. Med.* **2019**, *25*, 70–74. [CrossRef]
17. Xu, X.; Jiang, X.; Ma, C.; Du, P.; Li, X.; Lv, S.; Yu, L.; Ni, Q.; Chen, Y.; Su, J.; et al. A deep learning system to screen novel coronavirus disease 2019 pneumonia. *Engineering* **2020**, *6*, 1122–1129. [CrossRef] [PubMed]
18. Wang, L.; Zhang, Y.; Wang, D.; Tong, X.; Liu, T.; Zhang, S.; Huang, J.; Zhang, L.; Chen, L.; Fan, H.; et al. Artificial intelligence for COVID-19: A systematic review. *Front. Med.* **2021**, *8*, 1457. [CrossRef]
19. Gangloff, C.; Rafi, S.; Bouzillé, G.; Soulat, L.; Cuggia, M. Machine learning is the key to diagnose COVID-19: A proof-of-concept study. *Sci. Rep.* **2021**, *11*, 7166. [CrossRef]
20. Dinacci, M.; Chen, T.; Mahmud, M.; Parkinson, S. A Case Study of Using Machine Learning Techniques for COVID-19 Diagnosis. In *Artificial Intelligence in Healthcare: Recent Applications and Developments*; Springer: Berlin/Heidelberg, Germany, 2022; pp. 201–213.
21. Song, X.; Zhu, J.; Tan, X.; Yu, W.; Wang, Q.; Shen, D.; Chen, W. XGBoost-based feature learning method for mining COVID-19 novel diagnostic markers. *Front. Public Health* **2022**, *10*, 926069. [CrossRef]
22. Akinyelu, A.A.; Blignaut, P. COVID-19 diagnosis using deep learning neural networks applied to CT images. *Front. Artif. Intell.* **2022**, *5*, 919672. [CrossRef]
23. Meraihi, Y.; Gabis, A.B.; Mirjalili, S.; Ramdane-Cherif, A.; Alsaadi, F.E. Machine learning-based research for COVID-19 detection, diagnosis, and prediction: A survey. *SN Comput. Sci.* **2022**, *3*, 286. [CrossRef]
24. Apostolopoulos, I.; Tzani, B. COVID-19: Automatic detection from X-Ray images utilizing Transfer Learning with Convolutional Neural Networks. *Phys. Eng. Sci. Med.* **2020**, *43*, 635–640. [CrossRef] [PubMed]
25. Rafiq, A.; Imran, M.; Alhajlah, M.; Mahmood, A.; Karamat, T.; Haneef, M.; Alhajlah, A. Deep Feature Extraction for Detection of COVID-19 Using Deep Learning. *Electronics* **2022**, *11*, 4053. [CrossRef]
26. Vidyanti, A.N.; Satiti, S.; Khairani, A.F.; Fauzi, A.R.; Hardhantyo, M.; Sufriyana, H.; Su, E.C.Y. Symptom-based scoring technique by machine learning to predict COVID-19: A validation study. *BMC Infect. Dis.* **2023**, *23*, 871. [CrossRef]
27. Zoabi, Y.; Deri-Rozov, S.; Shomron, N. Machine learning-based prediction of COVID-19 diagnosis based on symptoms. *NPJ Digit. Med.* **2021**, *4*, 3. [CrossRef]
28. Sowjanya, A.M.; Mrudula, O. Effective treatment of imbalanced datasets in health care using modified SMOTE coupled with stacked deep learning algorithms. *Appl. Nanosci.* **2023**, *13*, 1829–1840. [CrossRef]
29. Welvaars, K.; Oosterhoff, J.H.; van den Bekerom, M.P.; Doornberg, J.N.; van Haarst, E.P.; Consortium, O.U.; the Machine Learning Consortium; van der Zee, J.A.; van Andel, G.A.; Lagerveld, B.W.; et al. Implications of resampling data to address the class imbalance problem (IRCIP): An evaluation of impact on performance between classification algorithms in medical data. *JAMIA Open* **2023**, *6*, ooad033. [CrossRef]

30. Kosolwattana, T.; Liu, C.; Hu, R.; Han, S.; Chen, H.; Lin, Y. A self-inspected adaptive SMOTE algorithm (SASMOTE) for highly imbalanced data classification in healthcare. *BioData Min.* **2023**, *16*, 15. [CrossRef] [PubMed]
31. Araf, I.; Idri, A.; Chairi, I. Cost-sensitive learning for imbalanced medical data: A review. *Artif. Intell. Rev.* **2024**, *57*, 80. [CrossRef]
32. Kachuee, M.; Karkkainen, K.; Goldstein, O.; Zamanzadeh, D.; Sarrafzadeh, M. Cost-sensitive diagnosis and learning leveraging public health data. *arXiv* **2019**, arXiv:1902.07102.
33. Guo, C.; Pleiss, G.; Sun, Y.; Weinberger, K.Q. On calibration of modern neural networks. In Proceedings of the International Conference on Machine Learning (PMLR), Sydney, Australia, 6–11 August 2017; pp. 1321–1330.
34. Mahajan, P.; Uddin, S.; Hajati, F.; Moni, M.A. Ensemble learning for disease prediction: A review. *Healthcare* **2023**, *11*, 1808. [CrossRef]
35. Mahajan, P.; Uddin, S.; Hajati, F.; Moni, M.A.; Gide, E. A comparative evaluation of machine learning ensemble approaches for disease prediction using multiple datasets. *Health Technol.* **2024**, *14*, 597–613. [CrossRef]
36. Esteva, A.; Kuprel, B.; Novoa, R.A.; Ko, J.; Swetter, S.M.; Blau, H.M.; Thrun, S. Dermatologist-level classification of skin cancer with deep neural networks. *Nature* **2017**, *542*, 115–118. [CrossRef] [PubMed]
37. Liu, L.; Wu, X.; Li, S.; Li, Y.; Tan, S.; Bai, Y. Solving the class imbalance problem using ensemble algorithm: Application of screening for aortic dissection. *BMC Med. Inform. Decis. Mak.* **2022**, *22*, 82. [CrossRef] [PubMed]
38. Kukar, M.; Gunčar, G.; Vovko, T.; Podnar, S.; Černelč, P.; Brvar, M.; Zalaznik, M.; Notar, M.; Moškon, S.; Notar, M. COVID-19 diagnosis by routine blood tests using machine learning. *Sci. Rep.* **2021**, *11*, 10738. [CrossRef]
39. Aruleba, R.T.; Alex, A.; Nimibofa, A.; Obaido, G.; Aruleba, K.; Mienye, D.; Aruleba, I.; Ogbuokiri, B. COVID-19 Diagnosis: A Review of Rapid Antigen, RT-PCR and Artificial Intelligence Methods. *Bioengineering* **2022**, *9*, 153. [CrossRef]
40. Fan, Y.; Liu, M.; Sun, G. An interpretable machine learning framework for diagnosis and prognosis of COVID-19. *PLoS ONE* **2023**, *18*, e0291961. [CrossRef]
41. Ha, Y.J.; Lee, G.; Yoo, M.; Jung, S.; Yoo, S.; Kim, J. Feasibility study of multi-site split learning for privacy-preserving medical systems under data imbalance constraints in COVID-19, X-ray, and cholesterol dataset. *Sci. Rep.* **2022**, *12*, 1534. [CrossRef] [PubMed]
42. Mano, L.Y.; Torres, A.M.; Morales, A.G.; Cruz, C.C.P.; Cardoso, F.H.; Alves, S.H.; Faria, C.O.; Lanzillotti, R.; Cerceau, R.; da Costa, R.M.E.; et al. Machine learning applied to COVID-19: A review of the initial pandemic period. *Int. J. Comput. Intell. Syst.* **2023**, *16*, 73. [CrossRef]
43. Brynildsrud, O. COVID-19 prevalence estimation by random sampling in population-optimal sample pooling under varying assumptions about true prevalence. *BMC Med Res. Methodol.* **2020**, *20*, 196 . [CrossRef]
44. Dablain, D.; Bellinger, C.; Krawczyk, B.; Aha, D.W.; Chawla, N. Understanding imbalanced data: XAI & interpretable ML framework. *Mach. Learn.* **2024**, *113*, 3751–3769.
45. Salmi, M.; Atif, D.; Oliva, D.; Abraham, A.; Ventura, S. Handling imbalanced medical datasets: Review of a decade of research. *Artif. Intell. Rev.* **2024**, *57*, 273. [CrossRef]
46. Ahmed, Z.; Das, S. A Comparative Analysis on Recent Methods for Addressing Imbalance Classification. *SN Comput. Sci.* **2023**, *5*, 30. [CrossRef]
47. El-Rashidy, N.; Abdelrazik, S. Comprehensive survey of using machine learning in the COVID-19 pandemic. *Diagnostics* **2021**, *11*, 1155. [CrossRef] [PubMed]
48. Mirmozaffari, M.; Yazdani, M.; Boskabadi, A.; Ahady Dolatsara, H.; Kabirifar, K.; Amiri Golilarz, N. A novel machine learning approach combined with optimization models for eco-efficiency evaluation. *Appl. Sci.* **2020**, *10*, 5210. [CrossRef]
49. Vepa, A.; Saleem, A.; Rakhshan, K.; Daneshkhah, A.; Sedighi, T.; Shohaimi, S.; Omar, A.; Salari, N.; Chatrabgoun, O.; Dharmaraj, D.; et al. Using machine learning algorithms to develop a clinical decision-making tool for COVID-19 inpatients. *Int. J. Environ. Res. Public Health* **2021**, *18*, 6228. [CrossRef]
50. Mirmozaffari, M.; Yazdani, R.; Shadkam, E.; Khalili, S.M.; Tavassoli, L.S.; Boskabadi, A. A novel hybrid parametric and non-parametric optimisation model for average technical efficiency assessment in public hospitals during and post-COVID-19 pandemic. *Bioengineering* **2021**, *9*, 7. [CrossRef]
51. Rai, H.M. Cancer detection and segmentation using machine learning and deep learning techniques: A review. *Multimed. Tools Appl.* **2024**, *83*, 27001–27035. [CrossRef]
52. Ayyannan, M. Accuracy Enhancement of Machine Learning Model by Handling Imbalance Data. In Proceedings of the 2024 International Conference on Expert Clouds and Applications (ICOECA), Bengaluru, India, 18–19 April 2024; IEEE: Piscataway, NJ, USA, 2024; pp. 593–599.
53. Koyyala, U.; Thirunavukkarasu, U. Classification of benign and malignant tumor cells using bagging and Adaboost classifiers from Wisconsin dataset for the potential diagnostic application. In *AIP Conference Proceedings*; AIP Publishing: College Park, MD, USA, 2024; Volume 2816.
54. Kasthuri, M.; Kavitha, M. Enhanced Cost-sensitive Ensemble Learning for Imbalanced Class in Medical Data. *J. Electr. Syst.* **2024**, *20*, 1043–1053.

55. Wei, W.; Wang, Y.; Ouyang, R.; Wang, T.; Chen, R.; Yuan, X.; Wang, F.; Wu, S.; Hou, H. Machine learning for early discrimination between lung cancer and benign nodules using routine clinical and laboratory data. *Ann. Surg. Oncol.* **2024**, *31*, 7738–7749. [CrossRef]
56. Feltes, B.C.; Chandelier, E.B.; Grisci, B.I.; Dorn, M. CuMiDa: An Extensively Curated Microarray Database for Benchmarking and Testing of Machine Learning Approaches in Cancer Research. *J. Comput. Biol.* **2019**, *26*, 376–386. [CrossRef] [PubMed]
57. Feltes, B.C.; Poloni, J.D.F.; Dorn, M. Benchmarking and Testing Machine Learning Approaches with BARRA:CuRDa, a Curated RNA-Seq Database for Cancer Research. *J. Comput. Biol.* **2021**, *28*, 931–944. [CrossRef] [PubMed]
58. Feltes, B.C.; Poloni, J.d.F.; Nunes, I.J.G.; Faria, S.S.; Dorn, M. Multi-Approach Bioinformatics Analysis of Curated Omics Data Provides a Gene Expression Panorama for Multiple Cancer Types. *Front. Genet.* **2020**, *11*, 586602. [CrossRef] [PubMed]
59. Grisci, B.I.; Feltes, B.C.; Dorn, M. Neuroevolution as a tool for microarray gene expression pattern identification in cancer research. *J. Biomed. Inform.* **2019**, *89*, 122–133. [CrossRef]
60. Inampudi, S.; Johnson, G.; Jhaveri, J.; Niranjan, S.; Chaurasia, K.; Dixit, M. Machine Learning Based Prediction of H1N1 and Seasonal Flu Vaccination. In Proceedings of the Advanced Computing: 10th International Conference, Goa, India, 5–6 December 2021.
61. Ormeno, P.; Marquez, G.; Guerrero-Nancuante, C.; Taramasco, C. Detection of COVID-19 Patients Using Machine Learning Techniques: A Nationwide Chilean Study. *Int. J. Environ. Res. Public Health* **2022**, *19*, 8058. [CrossRef]
62. Kotsiantis, S.; Zaharakis, I.; Pintelas, P. Machine learning: A review of classification and combining techniques. *Artif. Intell. Rev.* **2006**, *26*, 159–190. [CrossRef]
63. Cortes, C.; Vapnik, V.N. Support-Vector Networks. *Mach. Learn.* **2004**, *20*, 273–297. [CrossRef]
64. Kubat, M. *An Introduction to Machine Learning*, 1st ed.; Springer Publishing Company: Berlin/Heidelberg, Germany, 2015.
65. Huang, Y.; Li, L. Naive Bayes classification algorithm based on small sample set. In Proceedings of the 2011 IEEE International Conference on Cloud Computing and Intelligence Systems, Beijing, China, 15–17 September 2011; pp. 34–39. [CrossRef]
66. Jatana, V. *Machine Learning in Action*; Manning Publications: New York, NY, USA, 2019. [CrossRef]
67. Chen, T.; Guestrin, C. XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'16), New York, NY, USA, 13–17 August 2016; pp. 785–794. [CrossRef]
68. Schapire, R.E. A brief introduction to boosting. *IJCAI Int. Jt. Conf. Artif. Intell.* **1999**, *2*, 1401–1406.
69. Haixiang, G.; Yijing, L.; Shang, J.; Mingyun, G.; Yuanyue, H.; Bing, G. Learning from class-imbalanced data: Review of methods and applications. *Expert Syst. Appl.* **2017**, *73*, 220–239. [CrossRef]
70. Margineantu, D.D.; Dietterich, T.G. *Bootstrap Methods for the Cost-Sensitive Evaluation of Classifiers*; Oregon State University: Corvallis, OR, USA, 2000.
71. More, A.; Rana, D.P. Review of random forest classification techniques to resolve data imbalance. In Proceedings of the 2017 1st International Conference on Intelligent Systems and Information Management (ICISIM), Aurangabad, India, 5–6 October 2017; IEEE: Piscataway, NJ, USA, 2017; pp. 72–78.
72. He, H.; Ma, Y. *Imbalanced Learning: Foundations, Algorithms, and Applications*, 1st ed.; Wiley-IEEE Press: Hoboken, NJ, USA, 2013.
73. Chawla, N.V.; Bowyer, K.W.; Hall, L.O.; Kegelmeyer, W.P. SMOTE: Synthetic Minority over-Sampling Technique. *J. Artif. Intell. Res.* **2002**, *16*, 321–357. [CrossRef]
74. He, H.; Bai, Y.; Garcia, E.A.; Li, S. ADASYN: Adaptive synthetic sampling approach for imbalanced learning. In Proceedings of the 2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence), Hong Kong, China, 1–8 June 2008; pp. 1322–1328. [CrossRef]
75. Simundic, A.M. Measures of Diagnostic Accuracy: Basic Definitions. *EJIFCC* **2009**, *19*, 203–211.
76. Salazar, G.J.V. Importancia del cálculo de la sensibilidad, la especificidad y otros parámetros estadísticos en el uso de las pruebas de diagnóstico clínico y de laboratorio. *Med. Lab.* **2017**, *23*, 365–386. [CrossRef]
77. Galván, P.; Fusillo, J.; González, F.; Vukujevic, O.; Recalde, L.; Rivas, R.; Ortellado, J.; Portillo, J.; Borba, J.; Hilario, E. Factibilidad de la utilización de la inteligencia artificial para el cribado de pacientes con COVID-19 en Paraguay. *Rev. Panam. Salud PúBlica* **2023**, *46*, e20. [CrossRef] [PubMed]
78. Andrade-Girón, D.; Carreño-Cisneros, E.; Mejía-Dominguez, C.; Marín-Rodriguez, W.; Villarreal-Torres, H. Comparación de Algoritmos Machine Learning para la Predicción de Pacientes con Sospecha de COVID-19. *Salud Cienc. Tecnol.* **2023**, *3*, 336. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.