

Article

COVID-19 Data Analysis: The Impact of Missing Data Imputation on Supervised Learning Model Performance

Jorge Daniel Mello-Román  and Adrián Martínez-Amarilla * 

Faculty of Exact and Technological Sciences, Universidad Nacional de Concepción, Concepción 010123, Paraguay; jdmello@facet-unc.edu.py

* Correspondence: adrian@facet-unc.edu.py

Abstract: The global COVID-19 pandemic has generated extensive datasets, providing opportunities to apply machine learning for diagnostic purposes. This study evaluates the performance of five supervised learning models—Random Forests (RFs), Artificial Neural Networks (ANNs), Support Vector Machines (SVMs), Logistic Regression (LR), and Decision Trees (DTs)—on a hospital-based dataset from the Concepción Department in Paraguay. To address missing data, four imputation methods (Predictive Mean Matching via MICE, RF-based imputation, K-Nearest Neighbor, and XGBoost-based imputation) were tested. Model performance was compared using metrics such as accuracy, AUC, F1-score, and MCC across five levels of missingness. Overall, RF consistently achieved high accuracy and AUC at the highest missingness level, underscoring its robustness. In contrast, SVM often exhibited a trade-off between specificity and sensitivity. ANN and DT showed moderate resilience, yet were more prone to performance shifts under certain imputation approaches. These findings highlight RF's adaptability to different imputation strategies, as well as the importance of selecting methods that minimize sensitivity–specificity trade-offs. By comparing multiple imputation techniques and supervised models, this study provides practical insights for handling missing medical data in resource-constrained settings and underscores the value of robust ensemble methods for reliable COVID-19 diagnostics.



Academic Editors:
Dmytro Chumachenko and
Sergiy Yakovlev

Received: 24 January 2025

Revised: 4 March 2025

Accepted: 6 March 2025

Published: 8 March 2025

Citation: Mello-Román, J.D.; Martínez-Amarilla, A. COVID-19 Data Analysis: The Impact of Missing Data Imputation on Supervised Learning Model Performance. *Computation* **2025**, *13*, 70. <https://doi.org/10.3390/computation13030070>

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/>).

Keywords: supervised learning methods; COVID-19 diagnostics; data imputation; machine learning models

1. Introduction

The global COVID-19 pandemic has generated a vast amount of data related to the spread and impact of the disease worldwide. These data represent an invaluable resource for healthcare professionals, epidemiologists, and researchers, offering critical insights into the dynamics of the disease and its effects on populations [1]. However, working with these datasets poses significant challenges, particularly due to the presence of missing values, which can compromise the accuracy of analyses and predictions [2,3].

Machine learning plays a pivotal role in diverse fields, including healthcare, industry, and scientific research. In healthcare, it has been employed for early disease detection, such as COVID-19 diagnosis, as well as for optimizing treatment strategies and predicting clinical outcomes [4,5]. Across industries, machine learning enhances automation and intelligent decision-making, thereby improving efficiency and productivity [6]. It is also widely used in research to uncover patterns in large datasets and support real-time decision-making in sectors such as finance, marketing, and logistics [7]. This versatile approach provides powerful tools to address complex challenges across various domains.

Supervised learning models, which are trained using labeled data, are particularly effective for classifying patients into confirmed or dismissed COVID-19 cases [8,9]. This research evaluates several supervised learning techniques, including Support Vector Machines (SVMs), Artificial Neural Networks (ANNs), Logistic Regression (LR), Decision Trees (DTs), and Random Forests (RFs), to determine their performance in predicting COVID-19 cases. This study also investigates how these models respond to missing data, focusing on the use of imputation techniques to address such gaps [10].

Recently, advanced data imputation strategies have emerged, leveraging deep learning architectures such as Variational Autoencoders (VAEs) or Generative Adversarial Networks (GANs) [11,12]. While these methods can capture highly complex distributions in medical data, they often require specialized hardware, extensive hyperparameter tuning, and substantial computational resources [13]. In this study, we focus on four well-established imputation approaches—Random Forest (RF), Predictive Mean Matching (PMM) by Multiple Imputation by Chained Equations (MICE), K-Nearest Neighbor (KNN), and eXtreme Gradient Boosting (XGBoost)—chosen for their robust performance, relative ease of implementation, and suitability for large-scale datasets with moderate computational demands [14]. By doing so, we aim to balance methodological rigor with practical applicability in real-world healthcare environments.

This study uses a dataset from the Department of Concepción, Paraguay, which provides a regional perspective of COVID-19 cases in the South American context. Although localized, this dataset provides valuable information on disease dynamics and contributes to the global effort to improve diagnostic methods and data analysis practices.

2. Problem Statement and Research Questions

The effective analysis of COVID-19 data is essential for understanding disease dynamics and guiding public health strategies [15]. However, missing values in medical records remain a significant obstacle, particularly in resource-constrained healthcare systems such as Paraguay's public health sector [16]. These gaps often arise from inconsistencies in record-keeping, incomplete patient-provided information, or inadequate documentation during data collection. Such deficiencies undermine the reliability of machine learning models, which depend on comprehensive and high-quality data to generate accurate predictions. To address this challenge, it is crucial to implement robust data imputation techniques and evaluate their influence on the performance of predictive models. This study seeks to explore these challenges and provide insights through the following research questions:

RQ1. How do supervised learning models perform in classifying COVID-19 cases when trained on imputed datasets?

RQ2. What is the impact of different data imputation techniques on the evaluation metrics of machine learning models?

This study holds significant importance as it combines advanced data imputation techniques with machine learning to tackle a critical challenge in medical data analysis. By focusing on a region with unique socio-economic and cultural characteristics, the research provides valuable insights that can be applied to similar contexts facing comparable public health issues. The findings extend beyond the immediate scope of COVID-19, contributing to broader efforts in leveraging computational methods across diverse fields of knowledge. By offering practical solutions to enhance the quality and utility of incomplete datasets, this study not only addresses current challenges but also sets a foundation for future applications of machine learning in health data analysis and beyond.

3. Theoretical and Methodological Framework

This section outlines the key concepts and methodologies that underpin the study, focusing on machine learning and its applications, the challenges posed by missing data, and the techniques employed to address these issues. By integrating advanced machine learning models and imputation methods, this research seeks to enhance the reliability and accuracy of COVID-19 case predictions.

3.1. Supervised Learning Models

Machine learning (ML) is a branch of Artificial Intelligence focused on creating algorithms capable of learning from data without explicit programming. These algorithms rely on large datasets to train effectively, enabling them to identify patterns and make predictions based on the learned relationships [17]. Among the key paradigms of ML is supervised learning (SL), which is grounded in statistical learning theory and generalization principles [18]. In SL, algorithms are trained on datasets with labeled input features and outputs, equipping them to predict outcomes for new, unseen data [19].

Among the SL models utilized in this study are Artificial Neural Networks (ANNs), Support Vector Machines (SVMs), Decision Trees (DTs), Logistic Regression (LR), and Random Forests (RFs). Each model brings unique strengths and is chosen for its suitability to the dataset and research goals. A concise overview of each model is provided below.

3.1.1. Artificial Neural Networks (ANNs)

ANNs, inspired by biological neural networks, process data through interconnected layers—input, hidden, and output—to predict outcomes. They are particularly effective in handling complex data structures, but their performance heavily depends on data volume and network architecture [20]. ANNs are mathematically modeled by the following structure for a neuron k :

$$U_k = \sum_{j=1}^n (W_{kj} X_j) \quad (1)$$

$$Y_k = f(U_k + b_k)$$

where U_k represents the linear combination of inputs, X_j are the input signals, W_{kj} are the weights associated with neuron k , and b_k is the bias term. The function f is the activation function that governs the output signal, Y_k , of the neuron. For more comprehensive details on this model, refer to [21].

3.1.2. Support Vector Machines (SVMs)

SVMs classify data by mapping them into high-dimensional space and determining an optimal hyperplane for separation. Their success is influenced by the selection of kernel functions and parameters, such as polynomial or radial basis kernels [22].

Consider the nonlinear transformation $\Phi : R^m \rightarrow H$, which maps input vectors into a new feature space $\Phi(x) \in H$. The kernel function measures similarity and is defined as the scalar product between two vectors in the transformed space, $\Phi(u) \cdot \Phi(v) = K(u, v)$ [23].

For the binary classification problem involving N training examples, each example is represented by a tuple (X_i, y_i) , where X corresponds to the set of attributes for example i , and the class label is denoted by $y_i \in \{1, -1\}$. The SVM learning task can be formulated as the following constrained optimization problem [24]:

$$\max L = \sum_{i=1}^N \lambda_i - \frac{1}{2} \sum_{i,j} \lambda_i \lambda_j y_i y_j K(X_i, X_j) \quad (2)$$

$$\text{subject to } \sum_{i=1}^N \lambda_i y_i = 0, \quad \lambda_i \geq 0 \text{ for all } i$$

A test case Z can be classified using the decision function

$$f(z) = \text{sign}\left(\sum_{i=1}^n \lambda_i y_i K(X_i, Z) + b\right) \quad (3)$$

where λ_i represents a Lagrange multiplier, b is a bias term, and K is the kernel function used for mapping the data into the feature space.

3.1.3. Decision Trees (DTs)

Decision Trees is an SL method that uses a recursive partitioning process to split data into distinct groups at each node based on a given criterion. This process continues until the tree is fully constructed [25].

Each node in a decision tree corresponds to an attribute from a set $U = \{A_1, \dots, A_n\}$, with the root node containing all objects in the dataset Ω . The classification process starts at the root and progresses through nodes until it reaches a leaf node. The Chi-squared Automatic Interaction Detector (CHAID) algorithm, introduced by Kass in 1980 [26], uses the Chi-squared test for node splitting. The test statistic is calculated as

$$\chi^2 = \sum_{j=1}^J \sum_{i=1}^I \frac{(n_{ij} - \hat{m}_{ij})^2}{\hat{m}_{ij}} \quad (4)$$

where n_{ij} is the observed cell frequency, $\hat{m}_{ij}$ is the expected frequency, and $p = P(\chi^2 > X^2)$ determines the statistical significance, with χ^2 distributed across $(J - 1)(I - 1)$ degrees of freedom.

3.1.4. Random Forest (RF)

Random Forest is an ensemble learning method that builds on DT to enhance classification performance and generalization. Like DT, RF operates on a dataset Ω , using a set of attributes $U = \{A_1, \dots, A_n\}$. However, instead of constructing a single tree, it generates multiple decision trees by applying the Bootstrap method: repeatedly sampling subsets of Ω with replacement to form training subsets [27,28].

Each tree in the forest independently classifies data points, with the final prediction $\hat{f}(x')$ for an input x' determined by the mode (majority vote) of predictions from all B trees:

$$\hat{f}(x') = \text{Mode}\left(\{f_b(x')\}_{b=1}^B\right) \quad (5)$$

Here, $f_b(x')$ represents the prediction of the b -th tree for input x' , and B is the total number of trees. This aggregation reduces overfitting by combining diverse trees, each trained on different subsets of Ω and U . For classification tasks, Random Forest minimizes a loss function such as cross-entropy to ensure accuracy:

$$H(y, \hat{y}) = - \sum_{i=1}^C y_i \log(\hat{y}_i) \quad (6)$$

where y is the true class label, $\hat{y}_i$ is the predicted probability for class i , and C is the number of classes.

3.1.5. Logistic Regression (LR)

LR is a widely used classification method known for its simplicity and interpretability, though it is limited when applied to nonlinear or highly correlated data. The model

estimates the probability of a binary outcome ($y \in \{0, 1\}$) based on input variables $x_{i1}, \dots, x_{ip}$. The probability function is defined as

$$F(t) = \frac{\exp(t)}{1 + \exp(t)} \quad (7)$$

where $t = h(x_{i1}, \dots, x_{ip}) = \beta_0 + \beta_1 x_{i1} \dots + \beta_p x_{ip}$, also known as the Logit model, and F is the transformed function that maps t to a probability. A more detailed explanation of the LR model can be found in [29].

3.1.6. Evaluation Metrics

Evaluation metrics are critical tools for assessing the performance of predictive models, particularly in fields like epidemiology where accurate classification can significantly impact public health outcomes. These metrics are calculated using a confusion matrix, a table that provides a detailed comparison of the predicted versus actual outcomes, enabling the precise evaluation of model performance [30]. Table 1 summarizes the key metrics derived from the confusion matrix, along with their descriptions and formulas [31].

Table 1. Evaluation metrics derived from the confusion matrix.

Metric	Description	Formula ¹
Sensitivity	Proportion of actual positives correctly identified	$\frac{TP}{TP+FN}$
Specificity	Proportion of actual negatives correctly identified	$\frac{TN}{TN+FP}$
Accuracy	Overall proportion of correct classifications	$\frac{TP+TN}{TP+FN+FP+TN}$
Precision	Proportion of predicted positives that are correct	$\frac{TP}{TP+FP}$
F1-Score	Harmonic mean of precision and recall (sensitivity)	$2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$
MCC	Balanced measure robust to imbalanced datasets	$\frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$

¹ TP: true positives; TN: true negatives; FP: false positives; FN: false negatives. As defined in the confusion matrix.

Among these metrics, sensitivity measures the proportion of actual positives correctly identified by the model, crucial for minimizing missed cases in disease detection. Specificity evaluates the model's ability to correctly identify negatives, reducing false alarms. Accuracy provides an overall correctness measure, while precision highlights the reliability of positive predictions. The F1-score balances precision and recall, making it particularly useful in scenarios with imbalanced datasets. F1-score values range from 0 to 1, where 1 indicates perfect precision and recall, and 0 signifies the worst possible performance. Finally, the Matthews Correlation Coefficient (MCC) offers a comprehensive measure that considers all elements of the confusion matrix, making it robust for datasets with class imbalances. MCC values range from -1 to 1 , where 1 represents a perfect prediction, 0 indicates no better than random guessing, and -1 reflects total disagreement between predictions and observations [32].

Another key evaluation tool is the AUC-ROC (area under the receiver operating characteristic curve). This metric plots sensitivity against $1 - \text{specificity}$ across varying thresholds to assess the trade-off between these measures. The AUC quantifies a model's discrimination ability, with scores ranging from 0.5 (random guessing) to 1.0 (perfect classification). Its versatility and capacity to evaluate model performance independent of decision thresholds make it an indispensable complement to the metrics derived from the confusion matrix, allowing adjustments tailored to specific application goals [33,34].

3.2. Machine Learning in Medicine and the Challenge of Missing Data

Machine learning (ML) has become a cornerstone of modern medicine, providing innovative solutions for diagnosis, treatment planning, and patient management. Its applications range from interpreting medical images to predicting disease outcomes, significantly enhancing clinical decision-making and operational efficiency [35]. For example, ML models have been successfully developed to analyze radiological images, aiding in the detection of conditions such as tumors and fractures. Additionally, predictive analytics powered by ML enables the forecasting of patient outcomes, facilitating proactive interventions. The integration of ML into healthcare seeks to improve patient outcomes, reduce costs, and elevate the overall quality of care [36,37].

However, the performance of ML models heavily relies on the quality and completeness of the data used for training. Missing data are a common challenge in medical datasets, resulting from various issues such as patient non-responses, data entry errors, or equipment malfunctions. These gaps in data can introduce biases, skew analyses, and compromise the reliability of ML predictions [38]. Addressing the problem of incomplete data is therefore crucial to ensuring the accuracy and validity of ML-based insights.

Imputation Techniques

Imputation techniques, which estimate and replace missing values, are widely used to mitigate the effects of missing data. The selection of an appropriate imputation method depends on factors such as the nature of the dataset, the extent of missingness, and the specific goals of the analysis [39]. Missingness is typically classified into three types: MCAR (missing completely at random), MAR (missing at random), and NMAR (not missing at random), each requiring tailored strategies to maintain data integrity and reliability [40].

Understanding these distinctions is crucial for guiding the choice of imputation strategy. In MCAR scenarios, the probability of missingness is unrelated to either observed or unobserved variables, making the missing values essentially random. Under MAR, the probability of missingness depends only on observed information, enabling many conventional imputation methods to yield unbiased estimates. NMAR, however, arises if the missingness depends on unobserved data, thus demanding specialized approaches or additional variables to accurately model the reasons behind missingness [41]. In practice, confirming the exact mechanism can be challenging. Consequently, researchers often adopt MAR as a practical assumption in medical datasets, being aware that an unaddressed NMAR component may introduce systematic bias in the results.

This study employs four imputation techniques: Random Forest (RF), Predictive Mean Matching (PMM) by Multiple Imputation by Chained Equations (MICE), K-Nearest Neighbor (KNN), and an eXtreme Gradient Boosting (XGBoost)-based imputation approach. PMM is a semi-parametric approach within MICE that selects observed values closest to the predicted value from a regression model, preserving the original data distribution and ensuring plausible imputations. RF imputation is a non-parametric method that predicts missing values by building random forest models for each variable and leveraging patterns in other variables within the dataset. This approach does not rely on strict assumptions about data distribution, making it particularly effective for handling complex, nonlinear relationships and interactions. This is especially advantageous in medical datasets, where feature interdependencies are often intricate. Further details about the imputation method are provided in [42].

The MICE imputation method involves creating multiple imputations for missing data through an iterative series of predictive models, allowing for a comprehensive assessment of uncertainty due to missingness. It assumes the missing values are missing at random (MAR). The basic idea behind the algorithm is to treat each variable that has missing values

as a dependent variable in regression and treat the others as independent (predictors). PMM enhances the quality of imputations by avoiding unrealistic values often produced by pure regression. It selects observed values based on proximity to the predicted ones, thereby maintaining the variability and plausibility of imputed data [43].

In addition, we also experimented with two other methods—KNN and XGBoost—focusing on their suitability under MAR assumptions. KNN- and tree-based models (such as XGBoost) can capture complex interactions among variables without imposing restrictive distributional requirements, which is particularly beneficial in medical datasets [44,45]. These methods provide a balance between accuracy and implementation feasibility in resource-constrained settings, thereby supporting robust imputation for MAR data. More advanced deep learning techniques may offer additional advantages in certain missing data scenarios but typically require specialized architectures, greater computational resources, and extensive hyperparameter tuning [46].

Although deep learning-based or hybrid approaches to missing data imputation have shown promise in recent studies [11,12], their practical deployment requires more complex infrastructures and longer training times than many conventional techniques [13]. As a result, simpler but still robust methods, including PMM, RF, KNN, and XGBoost, remain attractive for healthcare scenarios where computational resources or timeframes are constrained. Such approaches strike a balance between accuracy, interpretability, and implementation feasibility, which is essential in clinical contexts.

By utilizing these popular imputation techniques, we can enhance the robustness and accuracy of ML models, addressing the critical challenges associated with missing data in healthcare analytics.

4. Materials and Methods

This study adopts a quantitative research approach, consistent with the experimental design methods prevalent in machine learning research [47]. The methodology emphasizes the use of multiple algorithms to address the research problem, with a focus on feature selection to enhance model performance for predictive tasks [48]. Algorithm efficiency is assessed through metrics derived from the confusion matrix, complemented by an evaluation of computational processing time [49].

The framework centers on analyzing and implementing supervised learning models trained on imputed datasets to evaluate their performance and reliability. This experimental approach provides a robust foundation for addressing the research questions, leveraging established machine learning techniques and quantitative metrics to tackle the challenges posed by missing data and improve prediction accuracy in medical datasets.

4.1. Dataset Description

The dataset analyzed in this study comprises records from 33,028 individuals admitted with suspected COVID-19 to healthcare facilities in the Concepción Department, Paraguay, during the period 2020–2022. These records were provided by the Health Surveillance Directorate of the Ministry of Public Health and were collected to monitor the pandemic's progression specifically within this region. The dataset reflects administrative records and may not include individuals who did not officially access the public healthcare system [50]. However, this omission is estimated to be minimal, with negligible impact on the models' effectiveness.

This dataset was collected by the Health Surveillance Directorate, which consolidates patient information from healthcare facilities throughout Concepción—both private and public hospitals. During the COVID-19 pandemic, the Paraguayan government mandated the consistent documentation of all suspected COVID-19 cases, facilitating the creation

of comprehensive clinical records. This effort sought to inform epidemiological models for real-time monitoring and policy decisions. Additionally, regular public reporting of case counts promoted transparency and accountability, further supporting the dataset's reliability. As a result, the likelihood of excluding individuals who did not engage with official healthcare services is presumed low, making the dataset broadly representative of the country.

The dataset's population is nearly evenly distributed by gender, with 48.9% male (16,137 individuals) and 51.1% female (16,891 individuals). This balanced representation ensures that the results are not unduly influenced by gender-specific characteristics, enhancing the robustness of the analysis. The dependent variable, "Final Classification", categorizes the individuals into four diagnostic outcomes: Confirmed, Discarded, Inconclusive, and Suspected. These classifications are based on medical evaluations, laboratory tests (PCR and/or antigen), or other diagnostic criteria. The cases were classified as follows:

- 41.8% (13,819 individuals) were classified as Confirmed, indicating SARS-CoV-2 infection;
- 29.8% (9837 individuals) were classified as Discarded, indicating a negative diagnosis for COVID-19 despite initial suspicion;
- 26.7% (8833 individuals) were classified as Inconclusive, where the available evidence was insufficient to determine infection status;
- 1.62% (535 individuals) were classified as Suspected, requiring further testing or evaluation for a definitive diagnosis;
- A single case (0.003%) was categorized as Not Applicable, likely due to issues with record registration or processing.

A temporal analysis of the data reveals that the highest proportion of cases (53%) occurred in 2021, followed by 33.6% in 2022 and 13.5% in 2020. The lower case count in 2020 likely reflects the impact of strict quarantine measures during the early phase of the pandemic [51].

4.2. Data Preprocessing Stages

This study focuses on hospital records with a "Final Classification" of Confirmed or Discarded, based on PCR and/or antigen test results. Records classified as Suspected or Inconclusive, or those diagnosed using alternative methods, were excluded, refining the dataset to only include cases with definitive diagnostic outcomes. Preprocessing was conducted in six stages, resulting in datasets with varying levels of completeness and missingness handling.

- *Stage 1:* Columns with missing values exceeding thresholds of 5%, 10%, 20%, or 40% were removed to form datasets labeled as Level 1, Level 2, Level 3, and Level 4, respectively. This step ensures that variables with high missingness do not undermine the robustness of the statistical models.
- *Stage 2:* Retained variables were encoded to suit supervised learning models. Most variables were dichotomous, nominal, or ordinal, with some numeric features such as Age and Week of Notification included for predictive analysis.
- *Stage 3:* Rows with missing values exceeding thresholds of 5%, 10%, 20%, or 40% were excluded, corresponding to the creation of datasets Level 1 through Level 4. This step reduces the risk of bias from incomplete records while balancing data retention and quality.
- *Stage 4:* Each dataset was split into two subsets: 70% for training and 30% for testing. This partitioning ensures that the imputation process only affects the training set, preserving the integrity of the test set, which remains untouched, preserving its natural missingness.

- *Stage 5*: PMM, RF, KNN, and XGBoost imputation were applied to training datasets at all levels to estimate missing values based on patterns within the data.
- *Stage 6*: a dataset with no missing values (Level 0) was created by removing all rows containing any missing values, providing a baseline for comparison with the imputed datasets.

After defining these six stages, we obtained five final dataset variants (Levels 0 to 4), each reflecting different thresholds of missing data removal. Table 2 summarizes the number of variables and records in each dataset level, and how many Confirmed and Discarded cases remained at each level. A detailed description of the variables is provided in Appendix A.

Table 2. Number of retained variables and records at each dataset level.

Dataset Level	Thresholds (% Column, % Row)	Variables Retained	Records Retained	Confirmed Cases	Discarded Cases
Level 0	0	50	2534	1668	866
Level 1	$\leq 5, \leq 5$	44	8251	4703	3548
Level 2	$\leq 10, \leq 10$	45	10,455	5798	4657
Level 3	$\leq 20, \leq 20$	47	14,209	7638	6571
Level 4	$\leq 40, \leq 40$	49	16,945	9670	7275

To investigate whether the missing values were missing completely at random (MCAR) or missing at random (MAR), we performed a basic statistical assessment using Little’s MCAR test and exploratory logistic regressions on key variables. Little’s test ($p < 0.05$) rejected the MCAR assumption, indicating that the absence of data depends on other observed features in the dataset. Supplementary logistic models further suggested a MAR pattern, as missingness in specific variables was significantly associated with values of other variables. We repeated these checks across all datasets (Levels 1 to 4) and observed consistent results, supporting the use of MAR-appropriate imputation methods. While these tests do not entirely exclude the possibility of MNAR, they provide reasonable evidence that MAR is the dominant mechanism in our data [52], thus justifying the imputation strategies adopted in this study.

Hyperparameter Configuration and Data Partitioning

The configuration and selection of hyperparameters for the supervised learning models were conducted for the complete dataset (Level 0) as a preliminary step to optimize performance and ensure robust implementation. An exhaustive internal search procedure—balancing accuracy and computational feasibility—was carried out. For each model, we used a grid search with 5-fold cross-validation to explore multiple parameter combinations. In SVM, we tested a range of cost (1–100) and gamma (1×10^{-4} – 1×10^{-2}) values; in ANN, different hidden layer sizes (1–3) and decay factors (0.01–0.1) were examined; for DT, cp was varied between 0.0005 and 0.01; and for RF, the number of trees ranged from 100 to 1000 alongside different mtry values. Once the best settings were identified, we applied them consistently to all imputation levels (Levels 1 to 4) to maintain comparability across experiments.

Based on this tuning, the final hyperparameters used for each supervised learning model were as follows: ANN with a hidden layer size of 1 and a decay of 0.1, SVM configured with a cost of 15 and gamma of 0.001, DT employing a complexity parameter (cp) of 0.0015, and RF optimized with 500 trees and an mtry value of 6. Logistic Regression (LR) served as a baseline without hyperparameter tuning.

After selecting these hyperparameters, we proceeded with the final partitioning step. For each dataset (Levels 0–4), we performed 10 runs, each time randomly selecting 70% of the data for training and 30% for testing. To preserve the class distribution of Confirmed vs. Discarded cases, the split was stratified, ensuring that each class was proportionally represented in both training and testing subsets. By repeating this process, we obtained performance metrics as an average of the 10 runs, thus reducing the impact of any single random partition. While cross-validation provides a more exhaustive approach, the size of our dataset—combined with time and resource constraints—made it less practical in this setting. Moreover, a consistent 70/30 split is often preferred in large-scale medical studies to reflect a real-world scenario: a stable training set for model development and an untouched test set for final evaluation [53].

For datasets with missing values (Levels 1 to 4), only the training subsets were imputed. The test subsets remained unaltered, with missing values handled by listwise deletion, thereby emulating real-world conditions where incomplete data often appear at prediction time [54]. This approach allowed models trained on imputed data to be evaluated on raw, incomplete test data, ensuring a realistic assessment of robustness and generalization.

Regarding hyperparameter settings for the imputation techniques, we employed 50 iterations for PMM and 10 trees for the RF method. For KNN, we set $k = 5$ and used Euclidean distance, while for XGBoost-based imputation, we used a maximum depth of 3, an eta of 0.1, and 100 boosting rounds. Preliminary tests indicated stable performance for these configurations, which were also guided by relevant research [55].

The computational framework for this study was developed in R. For supervised learning model training, we used the libraries `e1071` (SVM), `randomForest` (RF), `rpart` (DT), and `nnet` (ANN). For the imputation procedures, we employed `mice` (PMM and RF), the `VIM` package (KNN), and `xgboost` (XGBoost-based imputation). All experiments were conducted on an HP ZBook 15v G5 (Intel® Core™ i5-8300H CPU @ 2.30 GHz, 16 GB RAM, Windows 11 Pro 64-bit). In terms of relative complexity, SVM generally required the longest runtime, while Random Forest and ANN showed moderate processing times, and Decision Trees (DTs) and Logistic Regression (LR) were consistently faster. Among the imputation methods, PMM- and RF-based approaches generally incurred higher overhead than XGBoost and KNN, yet all remained computationally feasible in our moderate-resource environment. This qualitative comparison provides practical insights for real-world applications, particularly in resource-constrained scenarios such as the one addressed in this study.

5. Results

The supervised models (SVM, RF, DT, LR, and ANN), widely recognized for their efficacy in medical diagnostics, have been extensively applied to epidemiological research, particularly in the context of COVID-19 [15,56]. The results are organized into two primary sections. Section 5.1. details the performance of the supervised learning models on the imputed datasets. Section 5.2. discusses the impact of the data imputation techniques on the evaluation metrics.

5.1. Supervised Learning Models' Performance on Imputed Datasets

The performance of the models across various imputation levels (Levels 0 to 4) was evaluated using key metrics, including accuracy, sensitivity, specificity, F1-score, MCC, and AUC. Table 3 summarizes the results achieved using the PMM imputation method, highlighting how each model's performance evolved as the level of missingness increased.

Table 3. Supervised leaning models' performance with PMM imputation.

Model	Level	Accuracy	Sensitivity	Specificity	F1-Score	MCC	AUC
ANN	0	0.764	0.787	0.719	0.817	0.491	0.753
	1	0.709	0.709	0.709	0.731	0.415	0.709
	2	0.718	0.695	0.746	0.729	0.439	0.721
	3	0.758	0.667	0.869	0.752	0.540	0.768
	4	0.746	0.676	0.872	0.774	0.526	0.774
SVM	0	0.809	0.80	0.828	0.851	0.598	0.814
	1	0.720	0.620	0.846	0.712	0.470	0.733
	2	0.712	0.539	0.922	0.672	0.487	0.730
	3	0.747	0.570	0.963	0.712	0.564	0.766
	4	0.740	0.633	0.930	0.757	0.546	0.782
DT	0	0.801	0.807	0.798	0.735	0.584	0.874
	1	0.731	0.766	0.704	0.716	0.467	0.814
	2	0.716	0.845	0.610	0.730	0.462	0.801
	3	0.755	0.833	0.691	0.754	0.523	0.849
	4	0.753	0.719	0.771	0.676	0.480	0.839
RF	0	0.826	0.854	0.773	0.866	0.620	0.902
	1	0.745	0.772	0.711	0.772	0.484	0.842
	2	0.747	0.727	0.772	0.759	0.497	0.836
	3	0.780	0.742	0.826	0.787	0.565	0.873
	4	0.773	0.796	0.731	0.818	0.518	0.858
LR	0	0.805	0.802	0.811	0.848	0.586	0.890
	1	0.714	0.701	0.732	0.732	0.429	0.716
	2	0.721	0.695	0.752	0.731	0.444	0.723
	3	0.760	0.662	0.878	0.751	0.545	0.770
	4	0.741	0.677	0.856	0.770	0.511	0.766

From Table 3, RF emerged as the top-performing model, achieving the highest accuracy (0.826) and AUC (0.902) at Level 0. Across all levels, RF demonstrated minimal performance degradation, maintaining an accuracy of 0.773 and an AUC of 0.858 at Level 4. Its metrics, including accuracy, sensitivity, specificity, and F1-score, remained relatively stable, highlighting its resilience to imputation variability. SVM excelled in specificity, reaching 0.930 at Level 4, surpassing its specificity at Level 0 (0.828). However, its sensitivity declined significantly to 0.633 at Level 4, indicating a tendency to favor negative predictions. This improvement in specificity with imputation suggests that PMM may introduce a bias toward negative classification outcomes. ANN displayed marked sensitivity to missing data, with a substantial decline in accuracy and F1-score between Level 0 and Level 1. While its metrics showed slight recovery at higher levels, its overall performance was notably affected by the presence of missing data.

Similar results were obtained with RF imputation, as shown in Table 4. Across all levels of imputation using the RF method, it demonstrated minimal performance degradation, maintaining an accuracy of 0.760 and an AUC of 0.857 at Level 4. Its metrics, including sensitivity, specificity, and F1-score, remained consistently high, reflecting its robustness and adaptability to imputation variability. SVM excelled in specificity, reaching 0.950 at Level 3 and 0.918 at Level 4, surpassing its Level 0 specificity of 0.828. However, sensitivity declined significantly, indicating a trade-off that favored negative classifications at higher levels of imputation. ANN exhibited sensitivity to missing data, showing less stability across imputation levels.

Table 4. Supervised learning models’ performance with RF imputation.

Model	Level	Accuracy	Sensitivity	Specificity	F1-Score	MCC	AUC
ANN	0	0.764	0.787	0.719	0.817	0.491	0.753
	1	0.719	0.701	0.743	0.740	0.440	0.722
	2	0.718	0.676	0.770	0.727	0.444	0.723
	3	0.767	0.700	0.846	0.765	0.547	0.773
	4	0.736	0.665	0.874	0.769	0.511	0.769
SVM	0	0.809	0.800	0.828	0.851	0.598	0.814
	1	0.741	0.654	0.855	0.742	0.509	0.755
	2	0.721	0.544	0.944	0.685	0.516	0.744
	3	0.750	0.581	0.950	0.715	0.559	0.765
	4	0.727	0.629	0.918	0.752	0.521	0.773
DT	0	0.801	0.807	0.798	0.735	0.584	0.874
	1	0.728	0.731	0.726	0.698	0.453	0.816
	2	0.727	0.712	0.739	0.698	0.450	0.815
	3	0.768	0.804	0.739	0.761	0.541	0.854
	4	0.755	0.807	0.729	0.692	0.510	0.839
RF	0	0.826	0.854	0.773	0.866	0.620	0.902
	1	0.747	0.782	0.701	0.779	0.484	0.843
	2	0.751	0.729	0.778	0.765	0.504	0.841
	3	0.785	0.757	0.818	0.792	0.573	0.874
	4	0.760	0.769	0.742	0.809	0.493	0.857
LR	0	0.805	0.802	0.811	0.848	0.586	0.890
	1	0.722	0.704	0.746	0.743	0.446	0.725
	2	0.720	0.671	0.782	0.728	0.451	0.727
	3	0.765	0.691	0.852	0.761	0.545	0.772
	4	0.736	0.666	0.874	0.769	0.512	0.770

From Table 5, we see that RF remains quite stable under KNN-based imputation, maintaining an accuracy of 0.772 and an AUC of 0.859 at Level 4. These figures underscore RF’s robustness, even when missing values are replaced via KNN. By contrast, SVM achieves a standout specificity of 0.962 at Level 3—surpassing its Level 0 value of 0.828—but does so at the expense of sensitivity, which declines to 0.576. ANN records moderate metrics overall, with accuracy diminishing from 0.764 at Level 0 to 0.743 at Level 4, indicating that it remains somewhat sensitive to the nuances introduced by imputation. DT sustains a reasonably balanced profile, particularly at Levels 3 and 4, where accuracy hovers around 0.756–0.750. Lastly, LR maintains consistent performance but seldom outperforms the ensemble approaches (RF, DT) or SVM on the principal metrics.

Turning to Table 6, which shows the models’ outcomes with XGBoost-based imputation, RF again leads the pack. Although its accuracy dips slightly at some intermediate levels, the model finishes strong with a 0.769 accuracy and 0.860 AUC at Level 4, underscoring its adaptability across different imputation techniques. SVM once more leverages high specificity—peaking at 0.953 at Level 3—yet sees a corresponding drop in sensitivity (to 0.569), reinforcing the trade-off observed in other imputation settings. ANN presents respectable but modest results, with accuracy falling from 0.764 at Level 0 to 0.748 at Level 4. DT remains fairly steady, moving from a 0.801 accuracy at Level 0 to roughly 0.750–0.754 at higher imputation levels—less stable than RF, but still competitive. LR stays consistent, though it rarely eclipses the ensemble techniques or SVM in accuracy or F1-score.

Table 5. Supervised learning models' performance with KNN imputation.

Model	Level	Accuracy	Sensitivity	Specificity	F1-Score	MCC	AUC
ANN	0	0.764	0.787	0.719	0.817	0.491	0.753
	1	0.711	0.709	0.713	0.732	0.419	0.711
	2	0.716	0.694	0.743	0.728	0.435	0.719
	3	0.760	0.668	0.873	0.754	0.544	0.770
	4	0.743	0.672	0.870	0.770	0.520	0.771
SVM	0	0.809	0.80	0.828	0.851	0.598	0.814
	1	0.719	0.619	0.845	0.711	0.468	0.732
	2	0.713	0.539	0.923	0.672	0.489	0.731
	3	0.750	0.576	0.962	0.717	0.568	0.769
	4	0.735	0.620	0.942	0.750	0.546	0.781
DT	0	0.801	0.807	0.798	0.735	0.584	0.874
	1	0.731	0.704	0.766	0.745	0.467	0.814
	2	0.717	0.610	0.845	0.702	0.462	0.801
	3	0.756	0.704	0.820	0.760	0.523	0.849
	4	0.750	0.754	0.742	0.794	0.481	0.841
RF	0	0.826	0.854	0.773	0.866	0.620	0.902
	1	0.746	0.768	0.718	0.771	0.485	0.842
	2	0.747	0.729	0.770	0.759	0.496	0.836
	3	0.778	0.743	0.822	0.786	0.562	0.873
	4	0.772	0.799	0.724	0.818	0.514	0.859
LR	0	0.805	0.802	0.811	0.848	0.586	0.890
	1	0.714	0.701	0.731	0.732	0.429	0.716
	2	0.721	0.695	0.753	0.731	0.445	0.724
	3	0.761	0.665	0.877	0.753	0.547	0.771
	4	0.745	0.680	0.861	0.773	0.519	0.770

Overall, Random Forest (RF) performed consistently well across all four imputation methods (PMM, RF-based, KNN, and XGBoost). Even at higher levels of missingness (Level 4), it generally retained an accuracy above 0.76 and AUC above 0.85, indicating relatively low performance degradation compared to the baseline. In contrast, SVM often improved its specificity—sometimes exceeding 0.95—although this typically coincided with lower sensitivity, suggesting a tendency toward negative classifications. Meanwhile, ANN showed moderate resilience but tended to lose 1–2% in accuracy at higher missingness levels, implying some sensitivity to the chosen imputation strategy. Decision Trees (DTs) maintained balanced results, with accuracy values usually fluctuating between 0.74 and 0.80, while Logistic Regression (LR) remained consistent but rarely matched the top accuracy or F1-score results presented by the ensemble- or margin-based models.

When comparing the four imputation methods, PMM- and RF-based approaches offered slightly higher accuracy and AUC for certain models, notably RF itself. However, KNN and XGBoost also produced competitive outcomes: KNN generally preserved RF's high accuracy (near 0.77) and AUC (around 0.86) at Level 4, whereas XGBoost kept RF above a 0.76 accuracy and 0.86 AUC under comparable conditions. For SVM, ANN, and DT, the main patterns held across KNN and XGBoost, suggesting that the models' performance differences stem more from specificity–sensitivity trade-offs rather than large shifts in overall accuracy. In general, the ensemble-based methods—particularly RF—adapted well to missing data imputation, while other algorithms still achieved favorable metrics under certain imputation scenarios.

Table 6. Supervised learning models’ performance with XGBoost-based imputation.

Model	Level	Accuracy	Sensitivity	Specificity	F1-Score	MCC	AUC
ANN	0	0.764	0.787	0.719	0.817	0.491	0.753
	1	0.710	0.710	0.710	0.732	0.417	0.710
	2	0.715	0.694	0.742	0.727	0.434	0.718
	3	0.761	0.671	0.870	0.755	0.545	0.770
	4	0.748	0.683	0.863	0.777	0.524	0.773
SVM	0	0.809	0.80	0.828	0.851	0.598	0.814
	1	0.719	0.622	0.842	0.712	0.467	0.732
	2	0.714	0.541	0.923	0.674	0.490	0.732
	3	0.747	0.569	0.953	0.710	0.565	0.766
	4	0.742	0.642	0.921	0.762	0.544	0.782
DT	0	0.801	0.807	0.798	0.735	0.584	0.874
	1	0.730	0.688	0.784	0.740	0.469	0.814
	2	0.716	0.607	0.847	0.701	0.461	0.800
	3	0.754	0.706	0.812	0.759	0.517	0.849
	4	0.750	0.798	0.664	0.803	0.459	0.836
RF	0	0.826	0.854	0.773	0.866	0.620	0.902
	1	0.748	0.770	0.719	0.773	0.489	0.842
	2	0.752	0.736	0.772	0.765	0.506	0.836
	3	0.778	0.741	0.823	0.785	0.561	0.873
	4	0.769	0.808	0.701	0.818	0.504	0.860
LR	0	0.805	0.802	0.811	0.848	0.586	0.890
	1	0.714	0.700	0.733	0.732	0.430	0.716
	2	0.718	0.693	0.749	0.729	0.439	0.721
	3	0.760	0.663	0.879	0.752	0.546	0.771
	4	0.746	0.686	0.854	0.776	0.518	0.770

5.2. Impact of Data Imputation Techniques on Evaluation Metrics

This section evaluates how the four imputation methods (PMM, RF-based, KNN, and XGBoost) impact model performance across increasing missingness levels (0–20%). We first analyze global trends in accuracy and AUC, then dissect granular trade-offs in F1-score, MCC, sensitivity, and specificity. A focused comparison of ANN and SVM—the most imputation-sensitive models—highlights critical performance fluctuations at extreme missingness (Levels 0 vs. 4). Finally, we synthesize recommendations for method–model pairing based on metric priorities.

Examining accuracy and AUC first, Random Forest (RF) consistently retains higher values relative to other models. For instance, with PMM, RF’s accuracy declines modestly from 0.826 at Level 0 to 0.773 at Level 4, while AUC similarly falls from 0.902 to 0.858. A comparable pattern emerges with RF-based imputation, where accuracy starts at 0.826 and ends at 0.760, and AUC remains above 0.85. KNN and XGBoost also produce competitive results—under KNN, for example, RF’s accuracy sits around 0.772 and its AUC at 0.859 by Level 4, and with XGBoost imputation, the final accuracy is near 0.769 and AUC near 0.860. In contrast, algorithms like ANN and SVM show more noticeable drops in accuracy and AUC when missingness rises, indicating that ensemble-based methods are generally more robust to how data are filled.

Turning to the F1-score and MCC, RF again demonstrates stability. Typically, it retains F1-scores above 0.77 and MCC values above 0.49–0.50 across all imputation levels. Meanwhile, ANN sees steeper dips, especially at lower imputation levels under PMM,

where the F1-score can drop by about 8–10 percentage points between Levels 0 and 1. SVM is somewhat more volatile: it benefits from higher specificity but often loses ground in sensitivity, which in turn lowers its F1-score and MCC. Notably, RF-based imputation appears to smooth out these fluctuations for tree-based algorithms like DT, which remains relatively stable in its F1-scores (roughly 0.69–0.76) and MCC values (0.45–0.54) across different levels of missingness.

To further quantify these trade-offs, Table 7 contrasts ANN and SVM performance at Level 0 and Level 4 across imputation methods.

Table 7. ANN and SVM performance at Level 0 vs. Level 4 across imputation methods.

Model	Method	Level	Accuracy	Sensitivity	Specificity	F1-Score
ANN	PMM	0	0.764	0.787	0.719	0.817
		4	0.746	0.676 (−14%)	0.872 (+21%)	0.774 (−5%)
	RF	0	0.764	0.787	0.719	0.817
		4	0.736	0.665 (−15%)	0.874 (+22%)	0.769 (−6%)
	KNN	0	0.764	0.787	0.719	0.817
		4	0.743	0.672 (−15%)	0.870 (+21%)	0.770 (−6%)
	XGBoost	0	0.764	0.787	0.719	0.817
		4	0.748	0.683 (−13%)	0.863 (+20%)	0.777 (−5%)
	PMM	0	0.809	0.8	0.828	0.851
		4	0.74	0.633 (−21%)	0.930 (+12%)	0.757 (−11%)
SVM	RF	0	0.809	0.8	0.828	0.851
		4	0.727	0.629 (−21%)	0.918 (+11%)	0.752 (−12%)
	KNN	0	0.809	0.8	0.828	0.851
		4	0.735	0.620 (−23%)	0.942 (+14%)	0.750 (−12%)
	XGBoost	0	0.809	0.8	0.828	0.851
		4	0.742	0.642 (−20%)	0.921 (+11%)	0.762 (−10%)

Table 7 highlights two critical patterns. For ANN, sensitivity declines by 13–15% at Level 4, while specificity improves by 20–22%, suggesting a bias toward negative classifications under missingness. XGBoost mitigates this effect slightly, with the smallest sensitivity drop (−13%). For SVM, specificity spikes (e.g., +14% with KNN), but sensitivity plummets by 20–23%, indicating a conservative avoidance of false positives. XGBoost strikes the best balance for SVM, minimizing losses in sensitivity (−20%) and F1-score (−10%).

Regarding sensitivity (recall) and specificity, SVM often spikes in specificity as missingness increases—reaching values above 0.90 with KNN or XGBoost—but its sensitivity can concurrently drop below 0.60. ANN, though not as extreme, also experiences moderate changes in sensitivity (around 5–10% differences), suggesting that margin-based and neural models are particularly sensitive to shifts in data quality post-imputation. In contrast, RF consistently achieves balanced sensitivity and specificity (frequently both above 0.70), underscoring its resilience to the choice of imputation method.

Overall, these findings indicate that PMM- and RF-based approaches often preserve a higher accuracy, AUC, and F1-score for models such as RF and DT. KNN and XGBoost also

produce comparably strong results, though they can introduce more pronounced trade-offs in sensitivity versus specificity for models like SVM or ANN. As Table 7 demonstrates, these trade-offs are most acute at Level 4, where KNN maximizes SVM’s specificity (0.942) but at a steep cost to sensitivity (0.620). Consequently, while most methods maintain respectable performance at lower missingness levels, the degree of metric fluctuation at higher levels (e.g., Level 3 or 4) can help determine which combination of approach and model is best—particularly when certain metrics (e.g., sensitivity or AUC) are given priority.

Following these observations, Figure 1 provides a visual overview of how each imputation method influences the two principal metrics, accuracy and AUC, as missingness increases from Level 0 to Level 4. In Figure 1a, the four approaches—PMM, RF-based, KNN, and XGBoost—are plotted against the average accuracy for all models at each level of missingness. Meanwhile, Figure 1b compares these same methods in terms of their average AUC. Together, these visuals reinforce the conclusion that an RF-based imputation method tends to sustain higher accuracy and AUC across most models and levels of missingness.

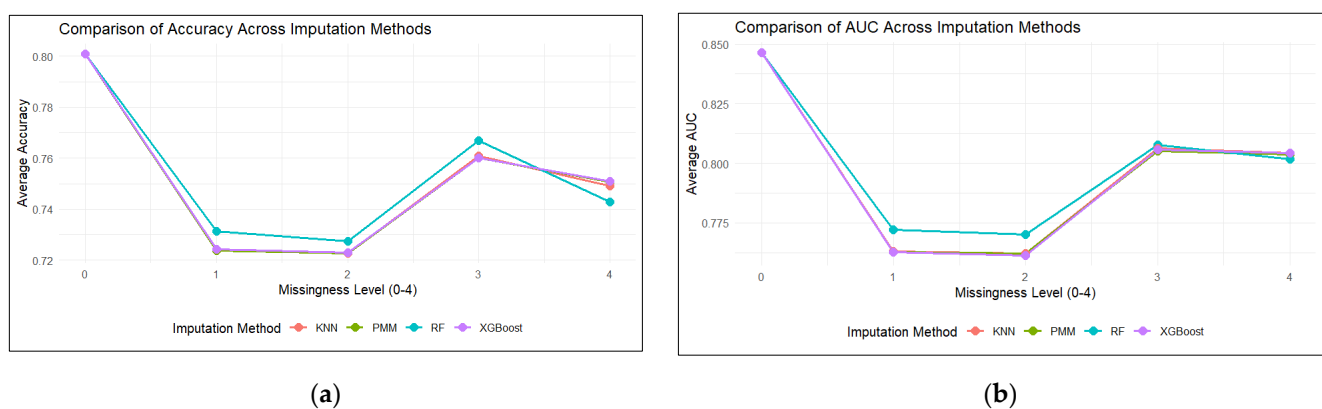


Figure 1. (a) Comparison of average accuracy across the four imputation methods at each missingness level; (b) comparison of average AUC across the four imputation methods at each missingness level.

6. Discussion

The findings of this study confirm the effectiveness of RF models in imputing missing medical data under various conditions of absence. As demonstrated in [57], RF outperforms methods such as mean imputation or nearest neighbors in scenarios with missing data at random (MCAR and MAR), owing to its ability to handle nonlinear relationships and high-dimensional data. However, the study also noted RF’s limitations in handling missing not at random (MNAR) data. Similarly, [58] further emphasized RF’s utility for mixed medical data, highlighting its advantages over parametric methods in avoiding distributional assumptions and efficiently managing complex interactions. Nevertheless, [59] cautioned against its use in specific scenarios.

Feng et al. [2] identified notable stability of RF in contexts with moderate proportions of missing data (less than 20%), aligning with the findings of this study. However, they also highlighted the superiority of multiple imputation methods in more complex scenarios or with higher levels of missingness. In [60], RF’s potential as an adaptable and precise method for imputing missing medical data was reinforced, with opportunities for enhanced predictive capacity through specific optimizations.

Regarding the impact on evaluation metrics, [14] stressed the importance of selecting appropriate imputation methods to minimize bias in sensitivity and specificity, noting RF’s solid balance between precision and flexibility, consistent with this study’s results. For instance, [61] also confirmed RF’s robust coverage under MAR data and its superiority over PMM in nonlinear scenarios due to its ability to manage complex interactions.

However, both [14,61] observed that ANN is more sensitive to imputation methods, further reinforcing RF's advantage in clinical contexts.

In conclusion, this study highlights the effectiveness of supervised learning models, particularly Random Forest (RF), in classifying COVID-19 cases using imputed datasets, even under significant levels of missing data. RF emerged as the most robust model, maintaining high values for accuracy, area under the curve (AUC), and F1-score across various levels of imputation and across PMM, KNN, RF-based, and XGboost-based imputation methods. These results align with previous research that emphasizes RF's adaptability in imputing complex medical datasets, especially in MAR (missing at random) scenarios.

Addressing *RQ1*, supervised learning models exhibited variable performance depending on the imputation method and the level of missing data. RF proved to be the most consistent model, achieving the highest metrics for accuracy (0.826) and AUC (0.902) in datasets without missing data (Level 0) and showing minimal degradation at higher levels of imputation (accuracy of 0.760 and AUC of 0.857 at Level 4 with RF-based imputation). This demonstrates its capacity to adapt to changes introduced by imputation methods. Conversely, ANN and SVM were more sensitive to the quality of imputed data: ANN showed a pronounced decline in F1-score and accuracy at lower levels (Level 1 and Level 2), while SVM improved specificity (0.950 at Level 3 with RF imputation) at the expense of sensitivity.

Importantly, when alternative imputation strategies—such as KNN or XGBoost—were employed, RF generally retained its leading position, whereas SVM and ANN exhibited larger shifts in sensitivity or accuracy, reinforcing the view that ensemble-based methods are more resilient to different missing data handling approaches.

Although a comprehensive feature-importance analysis was beyond this study's primary scope, a preliminary Random Forest assessment on the complete dataset (Level 0) highlighted the specimen collection technique (e.g., PCR vs. antigen test) as the most influential variable. This finding aligns with clinical research indicating that test sensitivity and reliability vary based on sampling methods [62]. Additional key predictors included the epidemiological week, patient age, and the timing of symptom onset, echoing prior evidence that temporal factors and demographics profoundly affect diagnostic outcomes. Finally, fever and sore throat also emerged as relevant indicators, suggesting that timely documentation of symptoms can enhance case detection [63,64]. These insights underscore the practical importance of systematically capturing both logistical (e.g., test availability) and clinical variables (e.g., patient age, symptom onset) to improve COVID-19 diagnostics in real-world public health contexts.

These results are consistent with findings in high-dimensional or clinical datasets [65,66]. ANN and SVM are powerful machine learning models, but their performance can be significantly affected by hyperparameter settings and data quality, particularly in the presence of outliers or noise introduced by imputation. ANNs require large, well-preprocessed datasets to ensure stable training dynamics, while SVMs are sensitive to feature scaling and parameter tuning.

Regarding *RQ2*, the choice of imputation technique significantly affected model performance. Although RF-based imputation delivered more consistent and stable results compared to PMM, we also observed that KNN and XGBoost produced competitive outcomes for ensemble algorithms (RF, DT), while occasionally inducing greater specificity–sensitivity trade-offs for margin-based methods (SVM). While RF experienced slight decreases in accuracy and AUC (from 0.826 to 0.760 and 0.902 to 0.857, respectively), it maintained stability in key metrics such as the F1-score and MCC across all imputation levels. In contrast, PMM enhanced specificity for models like SVM but introduced greater variability in sensitivity and F1-score. These findings underscore the robustness of RF-based approaches, which

preserve predictive performance even in scenarios with high levels of missing data, while showing that KNN and XGBoost can also yield strong results, particularly for RF.

Regarding practical implications and future applications, the comparative analysis of several models under varying levels of missingness offers valuable guidance for researchers dealing with real-world clinical data. Transferring these findings to other regions with similar data characteristics would primarily involve adjusting thresholds for missingness and reviewing local epidemiological contexts [67]. Nonetheless, the approach outlined here—comparing multiple imputation techniques across various supervised methods—could be replicated to adapt to different disease profiles or healthcare systems.

Additionally, it is essential to acknowledge that each imputation method introduces a degree of uncertainty, which can propagate through subsequent model training and predictions. This “imputation error” may disproportionately affect algorithms sensitive to noise or outliers, as exemplified by the variability observed in ANN and SVM. Although our study did not quantify error propagation in depth, references such as [68] discuss frameworks for measuring imputation-induced variance and its effects on model estimates. Future work incorporating such techniques could provide a more rigorous understanding of how imputation uncertainty impacts critical metrics like the F1-score and AUC, ultimately leading to more robust, interpretable predictions in clinical contexts.

Beyond confirming the effectiveness of established imputation methods, this study advances the understanding of missing data handling in resource-constrained medical settings through a systematic comparison of four widely used techniques (PMM, RF-based, KNN, and XGBoost) across varying missingness levels. Three key insights emerge: First, RF-based imputation paired with RF classifiers maintains robust performance even at high missingness levels (e.g., accuracy: 0.760 at 40%), outperforming ANN and SVM, which exhibit sensitivity to imputation-induced noise. This challenges the assumption that complex models like ANN inherently dominate in clinical predictions, instead highlighting RF’s unique adaptability to incomplete datasets. Second, imputation strategies introduce distinct specificity–sensitivity trade-offs; while PMM enhanced specificity for SVM (0.950 at Level 3), it concurrently reduced sensitivity—a critical concern in high-risk applications like COVID-19 diagnosis where false negatives carry severe consequences. This underscores the necessity of context-aware imputation selection aligned with clinical priorities. Third, XGBoost-based imputation, though less explored in medical contexts, performed comparably to RF-based methods, expanding the toolkit of viable alternatives for practitioners.

These findings invite targeted future research directions. Hybrid frameworks integrating imputation strategies with model architectures optimized for specific missing data patterns (e.g., combining RF with MNAR-adjusted weighting) could address current limitations. Additionally, quantifying the long-term impact of imputation uncertainty on clinical decision-making—such as through MNAR-aware sensitivity analyses—would further refine predictive robustness. By validating the practicality of established techniques under real-world constraints, this work provides clinicians and researchers with a scalable, interpretable framework for medical datasets, prioritizing methods like RF and XGBoost over resource-intensive alternatives in low-infrastructure settings.

Among the limitations of this study is the exclusive use of data from the Concepción Department in Paraguay, which limits the generalizability of findings to other regions with different epidemiological characteristics. Additionally, while the selected imputation methods were effective, the conclusions cannot be generalized to other methods not considered in this work. As a future direction, expanding the geographical and epidemiological scope is recommended to validate findings across different contexts. Moreover, incorporating hybrid imputation methods and exploring model stability in MNAR (missing not at random)

could further enhance their applicability to more complex scenarios. It should also be noted that details regarding computational resource usage, processing times, and model scalability were not provided, which is a limitation for practical implementation.

Furthermore, adopting more advanced imputation strategies—such as deep learning approaches (GANs or Variational Autoencoders)—goes beyond the present scope due to their higher computational and design complexity, even though they may prove valuable, especially in MNAR scenarios. Broadening this study to other regions or countries would likewise require ensuring comparable data quality and standardized recording protocols, a step that involves substantial coordination and data collection efforts. In addition, carrying out more exhaustive outlier analysis and employing thorough validation methods (e.g., repeated k-fold cross-validation) could further refine the results but entail more extensive methodological requirements. Finally, a deep exploration of MNAR-specific models (e.g., selection models or sensitivity analyses) would require additional datasets and assumptions, making it a natural extension of the work presented here.

Although supervised learning predominates in clinical diagnostics due to the availability of well-defined labels, unsupervised and semi-supervised methods offer potential advantages for handling missing data by leveraging partially labeled or unlabeled information. These approaches can sometimes uncover latent structures or subgroups not apparent in strictly supervised frameworks. Nevertheless, they also entail more complex modeling assumptions and may be difficult to align with standard clinical thresholds or outcome measures. Given our dataset's clear positive/negative labels, we focused on supervised techniques for this study. Looking ahead, future work might explore semi-supervised or unsupervised paradigms, particularly when labels are sparse or when broader exploratory insights are desired in resource-constrained healthcare settings.

Author Contributions: J.D.M.-R.: conceptualization, literature review, and methodology; J.D.M.-R. and A.M.-A.: writing and interpretation of data and results; A.M.-A.: data curation; J.D.M.-R. formal analysis. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The data used in this study were third-party data. Restrictions apply to the availability of these data. The data were obtained from the Ministry of Public Health and Social Welfare of Paraguay and are available through the authors upon reasonable request, subject to the approval of the cited Ministry.

Acknowledgments: The authors sincerely thank the First Health Region of the Ministry of Public Health and Social Welfare of Paraguay for kindly providing access to the data for academic and research purposes. Their invaluable support greatly contributed to the success of this study.

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

Appendix A

Table A1. Variable descriptions.

N°	Variable	Description	Type
1	Year	Year of the reported case.	Numerical
2	Week of notification	Epidemiological week when the case was reported.	Numerical
3	Age	Age of the individual at the time of diagnosis.	Numerical
4	Age measure	Measure used for age (days, months, years).	Numerical

Table A1. Cont.

N°	Variable	Description	Type
5	Age group	Age category classification.	Numerical
6	Sex	Biological sex of the individual (male, female).	Dichotomous
7	Zone	Geographical zone of the case (urban, rural).	Numerical
8	Fever	Presence of fever in the individual.	Dichotomous
9	Week fever/symptoms onset	Week when the symptoms, including fever, started.	Numerical
10	Week first consultation	Week when the individual first consulted for symptoms.	Numerical
11	Mechanical ventilation	Whether the individual required mechanical ventilation.	Dichotomous
12	Hospitalized	Whether the individual was hospitalized.	Dichotomous
13	Died	Whether the individual died.	Dichotomous
14	Signs/symptoms	Presence of reported symptoms (e.g., fever, cough, headache).	Dichotomous
15	Referred fever	Report of fever by the individual.	Dichotomous
16	Temperature > 38 °C	Whether the individual's body temperature exceeded 38 °C.	Dichotomous
17	Coryza, Rhinorrhea	Presence of nasal discharge or a runny nose.	Dichotomous
18	Nasal congestion	Presence of nasal obstruction.	Dichotomous
19	Cough	Presence of coughing.	Dichotomous
20	Difficulty breathing	Presence of respiratory difficulties.	Dichotomous
21	Irritability/confusion	Presence of mental status changes, such as confusion or irritability.	Dichotomous
22	Headache	Presence of headaches.	Dichotomous
23	Conjunctival injection	Presence of redness in the eyes.	Dichotomous
24	Dyspnea, Tachypnea	Difficulty or rapid breathing.	Dichotomous
25	Nausea/vomiting	Presence of nausea or vomiting.	Dichotomous
26	Abdominal pain	Presence of abdominal pain.	Dichotomous
27	Seizures	Presence of seizure episodes.	Dichotomous
28	Abnormal pulmonary auscultation	Detection of abnormal lung sounds during a physical examination.	Dichotomous
29	Ear pain	Presence of pain in the ear.	Dichotomous
30	Sore throat	Presence of throat pain or discomfort.	Dichotomous
31	Myalgia	Presence of muscle pain.	Dichotomous
32	Prostration	Presence of extreme physical weakness or collapse.	Dichotomous
33	Diarrhea	Presence of loose or frequent bowel movements.	Dichotomous
34	Risk factor	Presence of predisposing conditions or exposures.	Dichotomous
35	Chronic heart disease	History of chronic heart disease.	Dichotomous
36	Asthma	History of asthma.	Dichotomous
37	Chronic lung disease	History of chronic respiratory illnesses.	Dichotomous
38	Diabetes	History of diabetes.	Dichotomous
39	Chronic kidney disease	History of chronic kidney disease.	Dichotomous
40	Chronic liver disease	History of chronic liver disease.	Dichotomous
41	Immunodeficiency, disease, treatment	Presence of immunosuppressive conditions or treatments.	Dichotomous

Table A1. Cont.

N°	Variable	Description	Type
42	Chronic neurological disease	History of chronic neurological disorders.	Dichotomous
43	Down's Syndrome	Presence of Down's Syndrome.	Dichotomous
44	Obesity	Presence of obesity as a condition.	Dichotomous
45	Pregnant	Pregnancy status of the individual.	Dichotomous
46	Traveled/resides	Recent travel history or place of residence.	Dichotomous
47	Contact with people	Contact with people in specific scenarios (e.g., crowded spaces).	Dichotomous
48	Contact with infected	Contact with confirmed or suspected COVID-19 cases.	Dichotomous
49	Specimen collection technique	Technique used to collect diagnostic samples (nasopharyngeal swab).	Dichotomous
50	Final SARS CoV-2 classification	Final classification of the case (confirmed, discarded).	Dichotomous

References

- Di Serio, C.; Malgaroli, A.; Ferrari, P.; Kenett, R.S. The reproducibility of COVID-19 data analysis: Paradoxes, pitfalls, and future challenges. *PNAS Nexus* **2022**, *1*, pgac125. [\[CrossRef\]](#) [\[PubMed\]](#)
- Feng, S.; Hategeka, C.; Grépin, K.A. Addressing missing values in routine health information system data: An evaluation of imputation methods using data from the Democratic Republic of the Congo during the COVID-19 pandemic. *Popul. Health Metr.* **2021**, *19*, 44. [\[CrossRef\]](#)
- Pathak, A.; Batra, S.; Sharma, V. An Assessment of the Missing Data Imputation Techniques for COVID-19 Data. In Proceedings of the 3rd International Conference on Machine Learning, Advances in Computing, Renewable Energy and Communication, Ghaziabad, India, 10–11 December 2021; Springer: Singapore, 2022; pp. 701–706. [\[CrossRef\]](#)
- Mondal, M. Diagnosis of COVID-19 Using Machine Learning and Deep Learning: A Review. *Curr. Med. Imaging Rev.* **2021**, *17*, 1403–1418. [\[CrossRef\]](#)
- Montazeri, M.; ZahediNasab, R.; Farahani, A.; Mohseni, H.; Ghasemian, F. Machine Learning Models for Image-Based Diagnosis and Prognosis of COVID-19: Systematic Review. *JMIR Med. Inform.* **2021**, *9*, e25181. [\[CrossRef\]](#) [\[PubMed\]](#)
- Sysoev, A.; Klyavin, V.; Dvurechenskaya, A.; Mamedov, A.; Shushunov, V. Applying Machine Learning Methods and Models to Explore the Structure of Traffic Accident Data. *Computation* **2022**, *10*, 57. [\[CrossRef\]](#)
- Zheng, L.; He, X.; Ding, T.; Li, Y.; Xiao, Z. Analysis of the Accident Propensity of Chinese Bus Drivers: The Influence of Poor Driving Records and Demographic Factors. *Mathematics* **2022**, *10*, 4354. [\[CrossRef\]](#)
- Heidari, A.; Navimipour, N.J.; Unal, M.; Toumaj, S. Machine learning applications for COVID-19 outbreak management. *Neural Comput. Appl.* **2022**, *34*, 15313–15348. [\[CrossRef\]](#)
- 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*, 704256. [\[CrossRef\]](#)
- Bihri, H.; Hsaini, S.; Nejari, R.; Azzouzi, S.; Charaf, M.E.H. Missing Data Analysis in the Healthcare Field: COVID-19 Case Study. In *Networking, Intelligent Systems and Security, Proceedings of the NISS 2021, Kenitra, Morocco, 1–2 April 2021*; Springer: Singapore, 2021; pp. 873–884. [\[CrossRef\]](#)
- Blackthorn, N.; Mahyari, A.A.; Srinivasan, A. Training Variational Autoencoders for Population Synthesis in Public Health with Missing Data. In Proceedings of the 2024 IEEE International Conference on Big Data (BigData), Washington, DC, USA, 15–18 December 2024; pp. 4969–4973. [\[CrossRef\]](#)
- Akpinar, M.H.; Sengur, A.; Salvi, M.; Seoni, S.; Faust, O.; Mir, H.; Molinari, F.; Acharya, U.R. Synthetic Data Generation via Generative Adversarial Networks in Healthcare: A Systematic Review of Image- and Signal-Based Studies. *IEEE Open J. Eng. Med. Biol.* **2024**, *6*, 183–192. [\[CrossRef\]](#)
- Friedrich, P.; Frisch, Y.; Cattin, P.C. Deep Generative Models for 3D Medical Image Synthesis. *arXiv* **2024**. [\[CrossRef\]](#)
- Chebli, A.; Daas, S.; Hafs, T. Evaluating the Impact of Data Imputation on Model Precision in Machine Learning. *Stud. Eng. Exact Sci.* **2024**, *5*, e8310. [\[CrossRef\]](#)
- Podder, P.; Bharati, S.; Mondal, M.R.H.; Kose, U. Application of machine learning for the diagnosis of COVID-19. In *Data Science for COVID-19*; Academic Press: Cambridge, MA, USA, 2021; pp. 175–194. [\[CrossRef\]](#)

16. Mello-Román, J.C.; Gómez-Guerrero, S.; García-Torres, M. Predictive Models for the Medical Diagnosis of Dengue: A Case Study in Paraguay. *Comput. Math. Methods Med.* **2019**, *2019*, 1–7. [\[CrossRef\]](#)
17. Mahesh, B. Machine Learning Algorithms—A Review. *Int. J. Sci. Res. (IJSR)* **2020**, *9*, 381–386. [\[CrossRef\]](#)
18. Vapnik, V.N. *The Nature of Statistical Learning Theory*; Springer: New York, NY, USA, 2000. [\[CrossRef\]](#)
19. Grillo, S.A.; Roman, J.C.M.; Mello-Roman, J.D.; Noguera, J.L.V.; Garcia-Torres, M.; Divina, F.; Sotomayor, P.E.G. Adjacent Inputs With Different Labels and Hardness in Supervised Learning. *IEEE Access* **2021**, *9*, 162487–162498. [\[CrossRef\]](#)
20. Ghorbani, M.A.; Zadeh, H.A.; Isazadeh, M.; Terzi, O. A comparative study of artificial neural network (MLP, RBF) and support vector machine models for river flow prediction. *Environ. Earth Sci.* **2016**, *75*, 476. [\[CrossRef\]](#)
21. Wilusz, T. Neural networks—A comprehensive foundation. *Neurocomputing* **1995**, *8*, 359–360. [\[CrossRef\]](#)
22. Lee, Y.W.; Choi, J.W.; Shin, E.-H. Machine learning model for predicting malaria using clinical information. *Comput. Biol. Med.* **2021**, *129*, 104151. [\[CrossRef\]](#)
23. Segura, M.; Mello, J.; Hernández, A. Machine Learning Prediction of University Student Dropout: Does Preference Play a Key Role? *Mathematics* **2022**, *10*, 3359. [\[CrossRef\]](#)
24. Cristianini, N.; Shawe-Taylor, J. Support Vector and Kernel Methods. In *Intelligent Data Analysis*; Springer: Berlin/Heidelberg, Germany, 2007; pp. 169–197. [\[CrossRef\]](#)
25. Román, J.D.M.; Estrada, A.H. Un estudio sobre el rendimiento académico en Matemáticas. *Rev. Electron. Investig. Educ.* **2019**, *21*, e29. [\[CrossRef\]](#)
26. Kass, G.V. An Exploratory Technique for Investigating Large Quantities of Categorical Data. *Appl. Stat.* **1980**, *29*, 119–127. [\[CrossRef\]](#)
27. Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32. [\[CrossRef\]](#)
28. Cutler, A.; Cutler, D.R.; Stevens, J.R. Random Forests. In *Ensemble Machine Learning*; Springer: New York, NY, USA, 2012; pp. 157–175. [\[CrossRef\]](#)
29. Wendler, T.; Gröttrup, S. *Data Mining with SPSS Modeler*; Springer International Publishing: Dordrecht, The Netherlands, 2021. [\[CrossRef\]](#)
30. Khan, A.R.; Akter, J.; Ahammad, I.; Ejaz, S.; Khan, T.J. Dengue outbreaks prediction in Bangladesh perspective using distinct multilayer perceptron NN and decision tree. *Health Inf. Sci. Syst.* **2022**, *10*, 32. [\[CrossRef\]](#) [\[PubMed\]](#)
31. Trigka, M.; Dritsas, E. Predicting the Occurrence of Metabolic Syndrome Using Machine Learning Models. *Computation* **2023**, *11*, 170. [\[CrossRef\]](#)
32. Chicco, D.; Jurman, G. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genom.* **2020**, *21*, 6. [\[CrossRef\]](#)
33. Riya, N.J.; Chakraborty, M.; Khan, R. Artificial Intelligence-Based Early Detection of Dengue Using CBC Data. *IEEE Access* **2024**, *12*, 112355–112367. [\[CrossRef\]](#)
34. Aljameel, S.S. A Proactive Explainable Artificial Neural Network Model for the Early Diagnosis of Thyroid Cancer. *Computation* **2022**, *10*, 183. [\[CrossRef\]](#)
35. 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.M.; et al. Machine Learning Applied to COVID-19: A Review of the Initial Pandemic Period. *Int. J. Comput. Intell. Syst.* **2023**, *16*, 73. [\[CrossRef\]](#)
36. 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. [\[CrossRef\]](#)
37. Martins, M.V.; Baptista, L.; Luís, H.; Assunção, V.; Araújo, M.-R.; Realinho, V. Machine Learning in X-ray Diagnosis for Oral Health: A Review of Recent Progress. *Computation* **2023**, *11*, 115. [\[CrossRef\]](#)
38. Nguyen, C.D.; Strazdins, L.; Nicholson, J.M.; Cooklin, A.R. Impact of missing data strategies in studies of parental employment and health: Missing items, missing waves, and missing mothers. *Soc. Sci. Med.* **2018**, *209*, 160–168. [\[CrossRef\]](#)
39. Schmitt, P.; Mandel, J.; Guedj, M. A Comparison of Six Methods for Missing Data Imputation. *J. Biom. Biostat.* **2015**, *6*, 1. [\[CrossRef\]](#)
40. Little, R.; Rubin, D. *Statistical Analysis with Missing Data*, 3rd ed.; Wiley Series in Probability and Statistics; Wiley: Hoboken, NJ, USA, 2019. [\[CrossRef\]](#)
41. Kawabata, E.; Major-Smith, D.; Clayton, G.L.; Shapland, C.Y.; Morris, T.P.; Carter, A.R.; Fernández-Sanlés, A.; Borges, M.C.; Tilling, K.; Griffith, G.J.; et al. Accounting for Bias Due to Outcome Data Missing Not at Random: Comparison and Illustration of Two Approaches to Probabilistic Bias Analysis: A Simulation Study. *BMC Med. Res. Methodol.* **2024**, *24*, 278. [\[CrossRef\]](#)
42. Pereira, R.C.; Abreu, P.H.; Rodrigues, P.P.; Figueiredo, M.A. Imputation of data Missing Not at Random: Artificial generation and benchmark analysis. *Expert Syst. Appl.* **2024**, *249*, 123654. [\[CrossRef\]](#)
43. van Buuren, S.; Groothuis-Oudshoorn, K. mice: Multivariate Imputation by Chained Equations in R. *J. Stat. Softw.* **2011**, *45*, 1–67. [\[CrossRef\]](#)
44. Li, J.; Guo, S.; Ma, R.; He, J.; Zhang, X.; Rui, D.; Ding, Y.; Li, Y.; Jian, L.; Cheng, J.; et al. Comparison of the effects of imputation methods for missing data in predictive modelling of cohort study datasets. *BMC Med. Res. Methodol.* **2024**, *24*, 41. [\[CrossRef\]](#)

45. Deng, Y.; Lumley, T. Multiple Imputation Through XGBoost. *J. Comput. Graph. Stat.* **2023**, *33*, 352–363. [\[CrossRef\]](#)
46. Ramteke, M.; Raut, S. Enhancing Disease Diagnosis: Leveraging Machine Learning Algorithms for Healthcare Data Analysis. *IETE J. Res.* **2024**, 1–22. [\[CrossRef\]](#)
47. Kamiri, J.; Mariga, G. Research Methods in Machine Learning: A Content Analysis. *Int. J. Comput. Inf. Technol.* **2021**, *10*, 78–91. [\[CrossRef\]](#)
48. Mello-Román, J.D.; Gómez-Chacón, I.M. Creencias y rendimiento académico en matemáticas en el ingreso a carreras de ingeniería. *Aula Abierta* **2022**, *51*, 407–415. [\[CrossRef\]](#)
49. Kröger, H. Predictive machine learning approaches—Possibilities and limitations for the future of life course research. In *Handbook of Health Inequalities Across the Life Course*; Edward Elgar Publishing: Cheltenham, UK, 2023; pp. 112–127. [\[CrossRef\]](#)
50. Rios-González, C.M. Knowledge, Attitudes, and Practices towards COVID-19 in Paraguayans During the Outbreak Period: A Quick Online Survey. *Rev. Salud Publica Parag.* **2020**, *10*, 17–22. [\[CrossRef\]](#)
51. Ramos, P.; Silva, E.; Canese, J.; Velázquez, G. Epidemiología de los casos de COVID-19 diagnosticados en albergues sanitarios del gran Asunción, Paraguay (2020). *Mem. Inst. Investig. Cienc. Salud* **2021**, *19*, 69–77. [\[CrossRef\]](#)
52. Zhou, Y.; Aryal, S.; Bouadjenek, M.R. Review for Handling Missing Data with Special Missing Mechanism. *arXiv* **2024**. [\[CrossRef\]](#)
53. Aguilar-Ruiz, J.S.; Michalak, M. Classification Performance Assessment for Imbalanced Multiclass Data. *Sci. Rep.* **2024**, *14*, 10759. [\[CrossRef\]](#) [\[PubMed\]](#)
54. Li, J.; Wang, Z.; Wu, L.; Qiu, S.; Zhao, H.; Lin, F.; Zhang, K. Method for Incomplete and Imbalanced Data Based on Multivariate Imputation by Chained Equations and Ensemble Learning. *IEEE J. Biomed. Health Inform.* **2024**, *28*, 3102–3113. [\[CrossRef\]](#) [\[PubMed\]](#)
55. Gono, D.N.; Napitupulu, H.; Firdaniza. Silver Price Forecasting Using Extreme Gradient Boosting (XGBoost) Method. *Mathematics* **2023**, *11*, 3813. [\[CrossRef\]](#)
56. Akhtar, A.; Akhtar, S.; Bakhtawar, B.; Kashif, A.A.; Aziz, N.; Javeid, M.S. COVID-19 Detection from CBC using Machine Learning Techniques. *Int. J. Technol. Innov. Manag. (IJTIM)* **2021**, *1*, 65–78. [\[CrossRef\]](#)
57. Kokla, M.; Virtanen, J.; Kolehmainen, M.; Paananen, J.; Hanhineva, K. Random forest-based imputation outperforms other methods for imputing LC-MS metabolomics data: A comparative study. *BMC Bioinform.* **2019**, *20*, 492. [\[CrossRef\]](#)
58. Stekhoven, D.J.; Bühlmann, P. MissForest—Non-parametric missing value imputation for mixed-type data. *Bioinformatics* **2011**, *28*, 112–118. [\[CrossRef\]](#)
59. Hong, S.; Lynn, H.S. Accuracy of random-forest-based imputation of missing data in the presence of non-normality, non-linearity, and interaction. *BMC Med. Res. Methodol.* **2020**, *20*, 199. [\[CrossRef\]](#)
60. Jegadeeswari, K.; Ragunath, R.; Rathipriya, R. A Prediction Model with Multi-Pattern Missing Data Imputation for Medical Dataset. In *Advanced Network Technologies and Intelligent Computing*; Springer Nature: Cham, Switzerland, 2023; pp. 538–553. [\[CrossRef\]](#)
61. Bräm, D.S.; Nahum, U.; Atkinson, A.; Koch, G.; Pfister, M. Evaluation of machine learning methods for covariate data imputation in pharmacometrics. *CPT Pharmacomet. Syst. Pharmacol.* **2022**, *11*, 1638–1648. [\[CrossRef\]](#)
62. Alonaizan, F.; AlHumaid, J.; AlJindan, R.; Bedi, S.; Dardas, H.; Abdulfattah, D.; Ashour, H.; AlShahrani, M.; Omar, O. Sensitivity and Specificity of Rapid SARS-CoV-2 Antigen Detection Using Different Sampling Methods: A Clinical Unicentral Study. *Int. J. Environ. Res. Public Health* **2022**, *19*, 6836. [\[CrossRef\]](#) [\[PubMed\]](#)
63. Mancilla-Galindo, J.; Kammar-García, A.; Martínez-Esteban, A.; Meza-Comparán, H.D.; Mancilla-Ramírez, J.; Galindo-Sevilla, N. COVID-19 patients with increasing age experience differential time to initial medical care and severity of symptoms. *Epidemiology Infect.* **2021**, *149*, e230. [\[CrossRef\]](#) [\[PubMed\]](#)
64. Ding, F.-M.; Feng, Y.; Han, L.; Zhou, Y.; Ji, Y.; Hao, H.-J.; Xue, Y.-S.; Yin, D.-N.; Xu, Z.-C.; Luo, S.; et al. Early Fever Is Associated With Clinical Outcomes in Patients With Coronavirus Disease. *Front. Public Health* **2021**, *9*, 712190. [\[CrossRef\]](#) [\[PubMed\]](#)
65. Guo, J.; Li, W.; Hu, J. Robust Support Vector Machine Based on Sample Screening. In Proceedings of the 2024 14th International Conference on Information Science and Technology (ICIST), Chengdu, China, 6–9 December 2024; pp. 539–546. [\[CrossRef\]](#)
66. Wang, J.; Wu, Z.; Lu, M.; Ai, J. An Empirical Study on the Effect of Training Data Perturbations on Neural Network Robustness. *Sensors* **2024**, *24*, 4874. [\[CrossRef\]](#)
67. Akter, M.S.; Islam, R.; Khan, A.R.; Juthi, S. Big Data Analytics In Healthcare: Tools, Techniques, And Applications—A Systematic Review. *Innov. Eng. J.* **2025**, *2*, 29–47. [\[CrossRef\]](#)
68. Atoum, I. The critical role of evaluation metrics in handling missing data in machine learning. *Int. J. Adv. Appl. Sci.* **2025**, *12*, 112–124. [\[CrossRef\]](#)

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.