Malaysian Journal of Computing, 10 (1): 2135-2158, 2025

eISSN: 2600-8238

EXAMINING THE IMPACT OF FEATURE SELECTION TECHNIQUES ON MACHINE AND DEEP LEARNING MODELS FOR THE PREDICTION OF COVID-19

Hafiza Zoya Mojahid[1*], Jasni Mohamad Zain[2], Marina Yusoff[3], Abdul Basit[4], Abdul Kadir Jumaat[5] and Mushtaq Ali[6]

[1,2,3,4,5]College of Computing, Informatics and Mathematics, University Teknologi MARA, 40450 Shah Alam, Selangor, Malaysia*

[2,3,5]Institute for Big Data Analytics and Artificial Intelligence(IBDAAI), Kompleks Al-Khawarizmi, Universiti Teknologi MARA, 40450 Shah Alam, Selangor, Malaysia

[6]Riphah International University, 7th Avenue, Sector G-7/4, 44000 Islamabad, Pakistan

[1*]2022547787@student.uitm.edu.my, [2]jasni67@uitm.edu.my, [3]marina998@uitm.edu.my, [4]2021691374@student.uitm.edu.my, [5]abdulkadir@tmsk.uitm.edu.my, [6]mushtaq.ali@riphah.edu.pk

ABSTRACT

Feature selection is a vital preprocessing step for identifying the most informative features in complex datasets, enhancing the efficiency and accuracy of machine learning models. Its applications extend across various domains, including big data analytics, finance, chemometrics, medical diagnostics, biological research, intrusion detection systems, and renewable energy solutions. In medical contexts, feature selection serves a dual purpose: it reduces dimensionality while simultaneously improving the comprehension of disease etiology. This study delves into key variable selection methods—specifically Recursive Feature Elimination (RFE), Principal Component Analysis (PCA) and Least Absolute Shrinkage and Selection Operator (LASSO). We evaluate the interaction of these methods with Support Vector Machines (SVM), Logistic Regression (LR), and eXtreme Gradient Boosting (XGBoost) for COVID-19 prediction. Key performance metrics, including F1-score, precision, recall, and accuracy. LASSO with SVM performed the best overall in terms of accuracy = 0.7679 and precision=0.8236, but PCA outperformed RFE with XGBoost, underscoring the importance of matching feature selection methods to model types. In addition, we employ a deep learning Feature Selection method based on Extreme Learning Machine (FSELM) and compare its effectiveness against the established feature selection techniques. Our work reveals that Lactate Dehydrogenase (LDH) is the most relevant feature while predicting COVID-19. This research aims to provide insights into the optimal integration of feature selection techniques with advanced machine learning models for accurate prediction of COVID-19 virus.

Keywords: *COVID-19, Deep Learning, Extreme Learning Machine, Feature Selection, Machine Learning Models, Prediction.*

Received for review: 09-01-2025; Accepted: 18-03-2025; Published: 01-04-2025
DOI: 10.24191/mjoc.v10i1.4475

1. Introduction

While specific curative treatments for COVID-19 remain elusive, management strategies such as antivirals, vaccines, and supportive care have significantly mitigated its impact as of January 2024. Mild cases often resolve with home-based care, including adequate rest, hydration, and over-the-counter medications to relieve symptoms such as fever and pain. Severe cases, however, may require hospitalization and more intensive interventions, such as oxygen therapy, antiviral medications, or mechanical ventilation. Early detection through timely screening remains critical to ensuring appropriate treatment and reducing the risk of severe outcomes. The timely and accurate diagnosis of COVID-19 is essential for controlling its spread and mitigating transmission (Saberi-Movahed *et al.,* 2022). While Reverse Transcription Polymerase Chain Reaction (RT-PCR) is recognized by the World Health Organization (WHO) as the gold standard for COVID-19 testing, challenges such as long wait times, limited availability, and accuracy issues (with 15-20% false negatives) highlight the need for faster, more affordable, and accessible diagnostic alternatives (Li *et al.,* 2020; Brinati *et al.,* 2020; Rikan *et al.,* 2021; Alsharif & Qurashi, 2020; Basit *et al.,* 2022).

Expert knowledge is crucial for accurately diagnosing COVID-19 infections. However, the use of machine learning, which learns from patient data and healthcare records, can significantly enhance the speed, accuracy, and outcomes for patients by reducing errors (Basit *et al.,* 2024). In Machine Learning (ML), supervised learning is a dynamic field that involves training predictive models using labeled datasets with known target outputs. The challenges in supervised learning arise in regression, where the goal is to predict continuous outputs, and in classification, where the aim is to predict discrete categories. Classification, a key method in data mining, efficiently categorizes data into distinct classes. This study utilizes machine learning, primarily classification methods such as Support Vector Machines (SVM), Logistic Regression (LR), and eXtreme Gradient Boosting (XGBoost) to predict COVID-19 outcomes. To improve accuracy and efficiency, relevant features were selected using various methods such as filter, wrapper, and embedding which are essential for processing large health datasets and enhancing disease diagnosis. Table 1 outlines the important parameters for evaluating different feature selection methods.

2. Introduction

While specific curative treatments for COVID-19 remain elusive, management strategies such as antivirals, vaccines, and supportive care have significantly mitigated its impact as of January 2024. Mild cases often resolve with home-based care, including adequate rest, hydration, and over-the-counter medications to relieve symptoms such as fever and pain. Severe cases, however, may require hospitalization and more intensive interventions, such as oxygen therapy, antiviral medications, or mechanical ventilation.

Early detection through timely screening remains critical to ensuring appropriate treatment and reducing the risk of severe outcomes. The timely and accurate diagnosis of COVID-19 is essential for controlling its spread and mitigating transmission (Saberi-Movahed *et al.,* 2022). While Reverse Transcription Polymerase Chain Reaction (RT-PCR) is recognized by the World Health Organization (WHO) as the gold standard for COVID-19 testing, challenges such as long wait times, limited availability, and accuracy issues (with 15-20% false negatives) highlight the need for faster, more affordable, and accessible diagnostic alternatives (Li *et al.,* 2020; Brinati *et al.,* 2020; Rikan *et al.,* 2021; Alsharif & Qurashi, 2020; Basit *et al.,* 2022).

Expert knowledge is crucial for accurately diagnosing COVID-19 infections. However, the use of machine learning, which learns from patient data and healthcare records, can significantly enhance the speed, accuracy, and outcomes for patients by reducing errors (Basit *et al.,* 2024). In Machine Learning (ML), supervised learning is a dynamic field that involves training predictive models using labeled datasets with known target outputs. The challenges in supervised learning arise in regression, where the goal is to predict continuous

outputs, and in classification, where the aim is to predict discrete categories. Classification, a key method in data mining, efficiently categorizes data into distinct classes.

This study utilizes machine learning, primarily classification methods such as Support Vector Machines (SVM), Logistic Regression (LR), and eXtreme Gradient Boosting (XGBoost) to predict COVID-19 outcomes. To improve accuracy and efficiency, relevant features were selected using various methods such as filter, wrapper, and embedding which are essential for processing large health datasets and enhancing disease diagnosis. Table 1 outlines the important parameters for evaluating different feature selection methods.

Table 1. Outline of Important Parameters for Feature Selection.

Parameters	Filter	Wrapper	Embedded
Procedure Scrutiny	Statistical test.	Cross-validation.	Cross-validation.
Benchmarks	Feature subset relevance.	Feature subset usefulness.	Feature subset usefulness.
Search	By traversing the features sequentially, either through nested feature subsets or by individually ranking features.	Explore every conceivable feature subset.	The search process is governed by the learning algorithm.
Findings	-Robust against overfitting. -There exists a potential for failure in selecting pertinent features.	-Susceptible to overfitting. -Identifies the most valuable features but entails substantial time complexity	-Exhibits reduced susceptibility to overfitting. -Demonstrates comparatively lower time complexity.

Filter approaches examine the data directly (independently of the model) to identify significant features. These methods operate in two steps: (1) ranking features according to metrics such as distance or correlation and (2) selecting the top-ranked features while discarding the rest. This approach reduces model complexity by eliminating irrelevant features (Akhiat et al., 2018). In this study, we employ Principal Component Analysis (PCA) as a Filter approach. Wrappers, unlike Filter techniques, create and analyse models for all conceivable feature combinations. They select the optimum combination based on its ability to predict outcomes while taking into account feature interactions. Wrappers tend to be less in accurate than Filters, and we employ Recursive Feature Elimination (RFE) as our Wrapper approach in this study. Embedded methods combine Filtering and Wrapping procedures. As the model is being trained, they concurrently select features and improves it (Akhiat *et al.,* 2020). This makes them faster than Wrappers and LASSO (Least Absolute Shrinkage and Selection Operator) for embedded techniques (Akhiat *et al.,* 2019; Bouchlaghem *et al.,* 2022; Dede *et al.,* 2023; Jain and Xu, 2023). Blood testing, according to studies, can help detect COVID-19 outcomes. Specific cell types such as monocytes, basophils, lymphocytes, and neutrophils may provide information on a patient's progression. This is significant because early discovery can lead to more effective therapy, particularly for those at risk of serious health conditions (Ferrari *et al.,* 2020). Predicting the impact of COVID-19 on patients can be made more accurate by using blood testing. Figure 1 highlights the function of each of these methods.

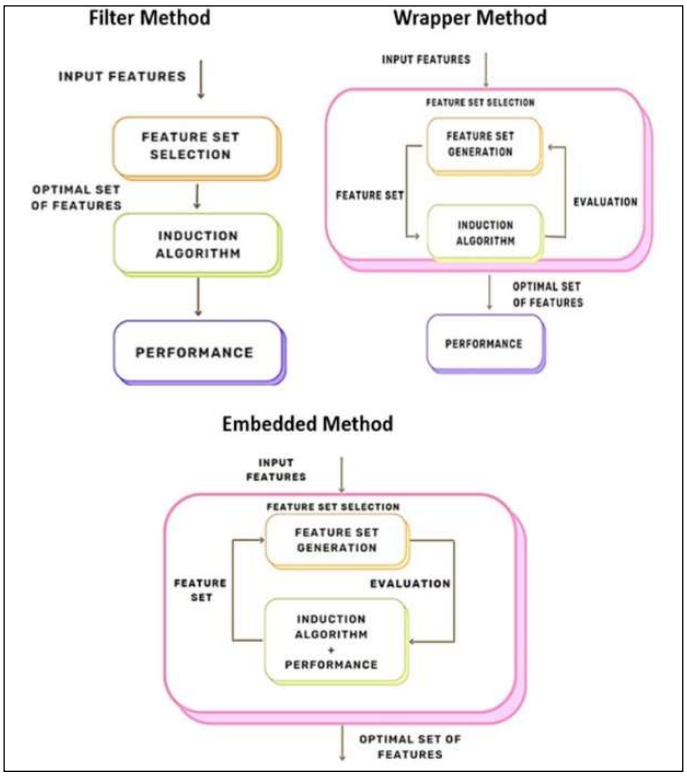

Figure 1. Workflow of Filter, Wrapper and Embedded Method.

In consideration of the aforementioned constraints and requirements, the primary objectives of this study are outlined below:

1. Examine and contrast machine learning models (SVM, LR, XGBoost) and feature selection techniques (RFE, LASSO, PCA) for COVID-19 prediction. We want to investigate how these strategies combine to increase accuracy in output results.
2. Investigate Deep Learning (DL) model for Feature Selection such as Extreme Learning Machines (FSELM). This approach will be compared to established methods (RFE, LASSO, and PCA) to evaluate its impact on improving model prediction of COVID-19. We believe that this will uncover complicated relationships in the data, allowing for higher accuracy in output results.

The structure of this paper as follows: The Section 1 begins with the Introduction. The Section 2 discusses the literature. The Section 3 covers the conceptual design, resources, and procedures employed for this research. Section 4 provides the results. Section 5 presents a debate and prospective study guidelines. Lastly, Section 6 discusses the conclusions.

3. Literature Review

This section will provide an outline of machine learning models used to diagnose COVID-19 utilizing regular laboratory and/or clinical data. The most widely used models for predictions were focused on RF (Sharma *et al.,* 2021; Aria *et al.,* 2021), Extreme Gradient Boosting (XGBooost), LR (Jawa, 2022), SVM (Mojahid *et al.,* 2024) and ANN (Azeem *et al.,* 2023). Additionally, diverse feature variety techniques combined with ML models using routine blood tests exhibit notable potential in identifying key features influencing the COVID-19's severity and fatality rate.

For instance, Wang *et al.* (2020) proposed a prediction model to predict three clinical outcomes among COVID-19 patients at NYU Langone Health Center (NYULH): ventilation, admission to the intensive care unit, or death. Both XGBoost and Logistic Regression (LR), using the Least Absolute Shrinkage and Selection Operator (LASSO) for feature selection, were

assessed in this study. The XGBoost model achieved significantly better outcomes than LR when trained on a sample of 3,740 patients. Parchure *et al.* (2020) established an Random Forest (RF) model for predicting near-term death (20-48 hours) utilizing time-series inpatient data gathered from Electronic Health Records (EHRs). Data from 567 infected patients who were hospitalized to a New York hospital were included in their study. Instead of using the traditional method that depends on clinical symptoms, Banerjee *et al.* (2020) used complete blood counts as a discriminative tool for identifying COVID-19 infected patients. According to the study, individuals with positive COVID-19 have decreased leukocyte, platelet, and lymphocyte counts.

Using normal blood indicators, Brinati *et al.* (2020) were able to detect COVID-19 cases among 279 COVID-19-infected cases. Using ML techniques, these researchers were able to identify COVID-19 cases with good recall (92%–95%) and accuracy (82%–86%). Vaid *et al.* (2020) used the XGBoost classification model to predict serious incidents and in-hospital mortality at several time frames (three, five, seven, and ten days, respectively) following admission. The model was created and evaluated on EHRs from infected patients hospitalized at the Mount Sinai Health System in New York City, with COVID-19 virus. A study by Thell *et al.* (2021) found that the XGBoost model performed admirably in predicting critical events and mortality. The researchers examined the use of common blood tests and their occurrence rates to distinguish between patients infected with SARS-CoV-2 and those who were not. Researchers investigated the possibility of using models based on machine learning to evaluate common blood tests for the purpose of detecting COVID-19 cases, building on the findings of Yang *et al.* (2020). This strategy could be especially useful in areas where traditional diagnostic techniques, such as the reverse transcriptase assay, are not readily available.

The study also looked into how ML models might be utilized to predict the severity and mortality risks related to the virus. This body of research underscores the significance of ML applications in harnessing information gleaned from routine blood tests for comprehensive virus assessment, particularly in scenarios where traditional screening methods may be limited or unavailable. Based on the significance and classification of XGBoost features, Linden *et al.* (2021) proposed that higher lactate dehydrogenase (LDH), hyperglycemia, acute renal injury, age, C-reactive protein (CRP), and anion gap were important factors in predicting serious incidents and death in infection cases. Rahman *et al.* (2021) predicted the likelihood of COVID-19 infections using easily accessible parameters from complete blood counts (CBCs). An external dataset was used to validate the classification model, which showed strong predicted accuracy. Similarly, Chowdhury *et al.* (2021) investigated response results, clinical representations, and demographic characteristics in an effort to identify important medical and demographic variables. To predict patient death, multiple variables were taken into attention,

Such as age along with data gathered upon hospitalization, such as neutrophils, lymphocytes, hs-CRP, and LDH. The achievement was achieved by employing a multi-tree XGBoost model. Moreover, a nomogram was developed by the researchers to predict the death risk associated with verified COVID-19 cases. First, a combined score that was correlated with the likelihood of the patient's mortality was calculated. Individuals with infection were then separated into three risk categories: high, moderate, and low. Impressive Area Under the Curve (AUC) values of 0.961 and 0.991 were obtained by the training and validation study nomogram, respectively.

In order to predict death prospects in COVID-19 infected cases, Zhu *et al.* (2020) used an approach of 6-layer deep neural network to classify the top five factors from 56 features studied at entrance. For the purpose of training and testing the model, the dataset they used included information from 181 instances that were gathered from a prominent medical care in Wuhan, China. De Terwangne *et al.* (2020) demonstrated the prognostic power of a Bayesian network-based model using five critical clinical parameters (age, acute kidney damage, lymphocytes, activated partial thromboplastin time, and LDH) for COVID-19 severity categorization. Aladağ and Atabey (2020) used coagulopathy indicators to attempt to predict the likelihood of death in sternly infected cases. Zhang *et al.* (2017) examined the independent association among the severity of COVID-19 infection and the starting point stages of four medical criteria: D-dimer, LDH, Computed Tomography score (CT), and Neutrophil-to-

Lymphocyte Ratio (NLR) upon entry. They accomplished this by using logistic regression. Notably, in high-risk populations, an increased level of NLR and LDH showed effectiveness in identifying confirmed COVID-19 cases. Moreover, the sensitivity of the model was improved by the combination of NLR and LDH. In a retrospective research, Huang *et al.* (2020) utilized nine independent health-related risk indicators at patient entrance to calculate risk ratings and categorize the 336 confirmed COVID-19 cases and the 139 controls who were negative.

Wang *et al.* (2020) developed two predictive models that used a mix of laboratory and healthcare information to predict death caused by COVID-19 virus after hospitalization. The laboratory structure demonstrated superior discriminative power when validated using a separate cohort dataset, accomplishing an AUC of 0.88. Features included in the model comprised Aspartate Aminotransferase (AST), lymphocyte and neutrophil count, age, hs-CRP, Peripheral Capillary Oxygen Satuation (SpO2), D-dimer, and Glomerular Filtration Rate (GFR). In order to determine relevant features, such as GFR, White Blood Cell (WBC) count, myoglobin, neutrophil count, and age, Zhang *et al.* (2020) utilized univariate and multivariate logistic regression analyses. These characteristics were then used to progress a grading system that predicted the severity of COVID-19 cases. The model's effectiveness was confirmed by validation on outside data, which included 22 infected cases. In order to create a fatality prognosis model, Aznar-Gimeno *et al.* (2021) carefully selected age, lymphocyte count, LDH, and SpO2, as a subgroup of important characteristics. Following validation on a separate cohort, an excellent AUC of 0.98 was obtained. The authors also presented a nomogram to calculate the death likelihood utilizing their well-established prediction system.

Bolourani *et al.* (2021) developed an XGBoost model that makes use of important variables such as age, respiratory rate, serum lactate, emergency severity index (ESI) level, demographics, and the sort of oxygen delivery system utilized in the emergency room. The XGBoost model outperformed other models in predicting respiratory breakdown for confirmed COVID-19 patients within 48 hours of admission, whose mean accuracy was 0.919 and its AUC was 0.77. In order to predict COVID-19 mortality, Yan *et al.* (2020) argued for an understandable single-tree XGBoost model that made use of the three most relevant characteristics such as LDH, lymphocytes, and hs-CRP. The early potential for prediction of this model were highlighted by its staggering 94% accuracy within three days preceding the individual's outcomes. In light of current works employing ML models to analyse blood biomarkers for predicting COVID-19 outcomes and severity, it is evident that the existing models, while informative, fall short of achieving the requisite accuracy and precision.

The literature review showcases various models leveraging blood biomarkers and classification models, yet there remains a gap in establishing a clinically reliable predictive biomarker for COVID-19 outcomes. Our research, as outlined in the abstract, aims to bridge existing gaps in the literature by employing ensemble techniques and ML models for feature selection. We focus on enhancing the performance of the experiments through the spotlight on hybrid and ensemble approaches. By delving into high-dimensional blood biomarker-based COVID-19 clinical dataset, our study aims to uncover the most vital subcategory of blood biomarkers capable of precisely predicting COVID-19 outcomes, particularly mortality, with enhanced accuracy and precision. This approach is innovative in its comprehensive exploration of feature selection methodologies and classification models, bridging the existing gaps in the literature and providing a more robust framework for preliminary detection of COVID-19 outcomes upon hospital admission. Through meticulous testing and validation, our research aims to contribute significantly to the field by offering a more reliable and accurate predictive model, thereby aiding in the understanding of the relative hazard of death for COVID-19 virus entities during the early stages of their hospitalization.

4. Methods

The study opted a publicly available dataset from a local hospital in Italy. Hence, formal informed permission was not required, and no human subjects were engaged. Table 2 provides the detail description for all variables of the dataset.

Table 2. Features of San Raffaele Health Dataset.

Features	Type
Gender	Categorical
Age	Numerical (discrete)
Leukocytes	Numerical (continuous)
Platelets	Numerical (continuous)
C-reactive Protein (CRP)	Numerical (continuous)
Transaminases (AST)	Numerical (continuous)
Transaminases (ALT)	Numerical (continuous)
Transaminases (ALP)	Numerical (continuous)
Gamma Glutamil Transferasi (GGT)	Numerical (continuous)
Lactate dehydrogenase (LDH)	Numerical (continuous)
Neutrophils	Numerical (continuous)
Lymphocytes	Numerical (continuous)
Monocytes	Numerical (continuous)
Eosinophils	Numerical (continuous)
Basophils	Numerical (continuous)
Swab	Categorical

Following data collection and preparation, three unique types of feature selection approaches were progressively applied: Filter (PCA), Wrapper (RFE), and embedded method as LASSO. This procedure entailed identifying the most pertinent subset of features, which was critical for subsequent model development. Following which, three different ML procedures were used to train models on the modified feature sets. The efficiency of the model and feature selection was then evaluated, using a variety of assessment measures such as accuracy, precision, recall and F1-score that comprehensively assess model performance.

4.1 Data Description

For this study, a publicly accessible dataset was utilized from the IRCCS: Scientific Institute for Research, Hospitalization and Healthcare (Kaggle, 2020), comprises records from 279 patients hospitalized to San Raffaele Hospital in Milan, Italy, across late February 2020 to mid-March 2020. The dataset has 16 columns, namely: GENDER, AGE, Monocytes, WBC, AST, Platelets, Lymphocytes, CRP, Eosinophils, Basophils, ALT, ALP, GGT, Neutrophils, LDH, and SWAB. Of the 279 patients, 177 tested positive for COVID-19, with the "Swab" variable being the target variable denoting the diagnosis. The negative COVID-19 test results occurred 102 times (37%), while the positive COVID-19 test results occurred 177 times (63%). The dataset contains missing values, especially in columns Neutrophils, Lymphocytes, Monocytes, Eosinophils, Basophils, and several others. The study has been separated into three broad phases illustrated as shown in Figure 2.

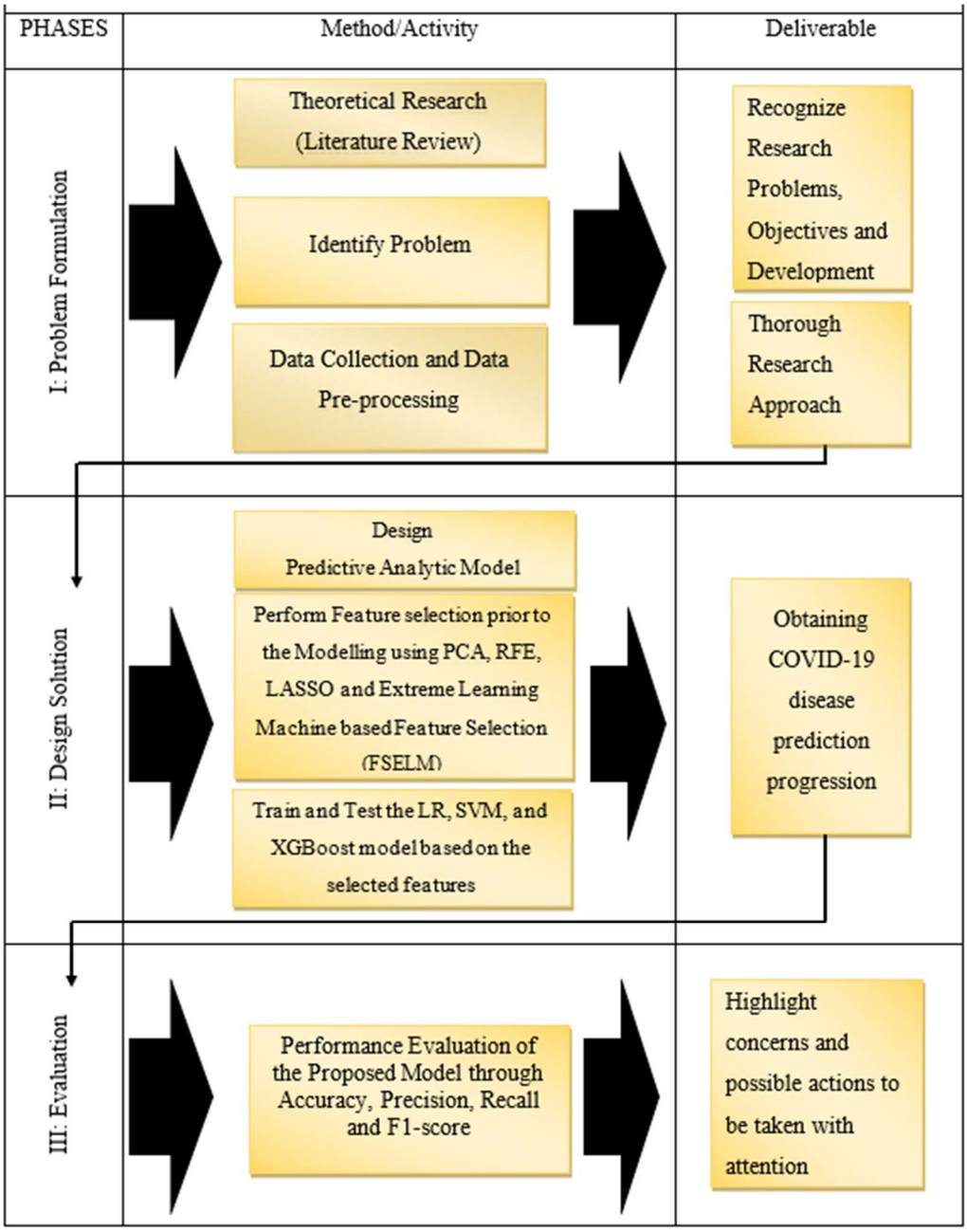

Figure 2. Phases of Research Methodology.

4.2 Missing Values

Before applying the imputation method, a Missing Completely at Random (MCAR) test was conducted to determine whether the missing data occurred randomly or followed a specific pattern. 'Chi-square': 0.0 indicates that the calculated chi-square statistic is 0.0. This indicates that the occurrence of missing data is unrelated to the observed values. Secondly, 'Degrees of freedom' in this test are '1' because there is one degree of freedom for a 2x2 contingency table. Further, the p-value is 1.0, indicating that under the null hypothesis (the information is missing completely at random), there is no evidence to discard the null hypothesis. This suggests that the missing values in the dataset is indeed completely at random according to this test. A heatmap was generated to illustrate the correlation among all features within the dataset. Figures 3(a) and 3(b) depict the heatmap and visualizations of missing data, respectively.

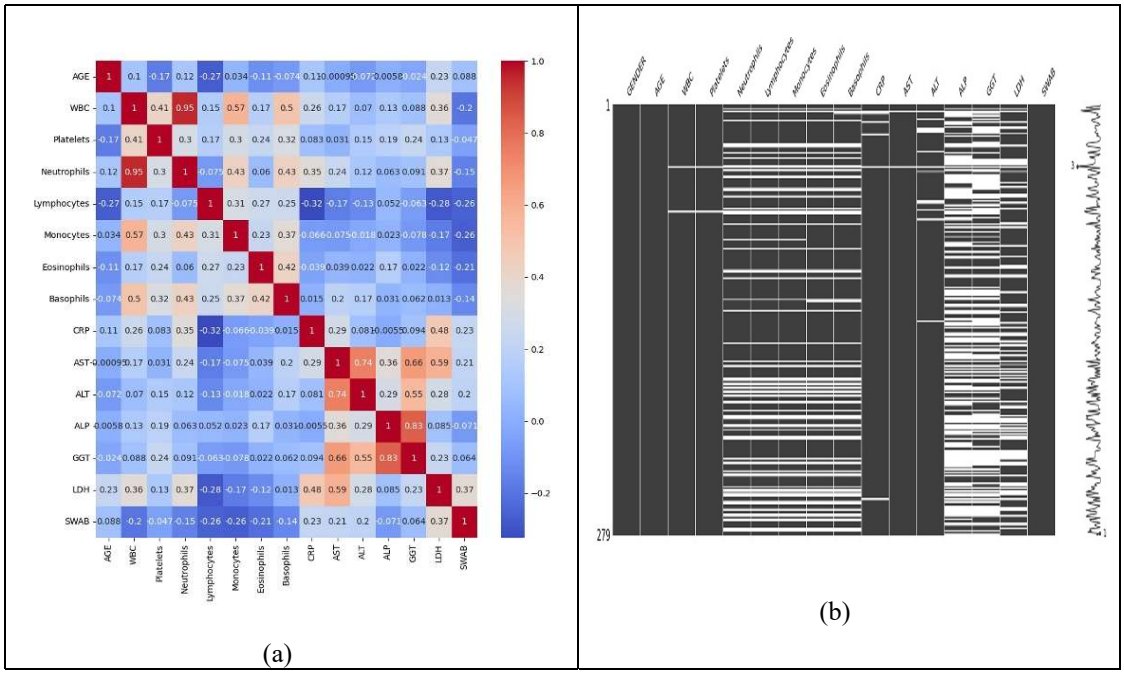

Figure 3. (a) Heatmap and (b) Missing values.

4.3 Data Pre-processing

During preprocessing, the dataset, confirmed to have missing values occurring at random through missing data plots and the MCAR test, underwent Expectation-Maximization (EM) imputation that helps to estimate and fill in missing values based on the observed data. Subsequently, the 'GENDER' column was transformed into numeric values using 'label encoder.' The 'StandardScaler' from scikit-learn python library was then applied to standardize the feature variables. This standardization, aimed at achieving zero mean and unit variance in input features, facilitates enhanced performance of ML models. It is important to mention that both the 'GENDER' column, being categorical, and the target variable 'SWAB' were excluded from standardization, as standardization is typically applied to numerical features only. Once the preprocessing is completed, the dataset is then split into 80% training and 20% testing set using train-test split.

An important consideration during preprocessing is ensuring that the integrity of the dataset remains intact, particularly when handling categorical data and imputation technique. The use of EM imputation not only addresses missing values but also maintains the dataset's underlying statistical structure, which is crucial for building reliable predictive models. Furthermore, the decision to exclude categorical variables like 'GENDER' and the target variable from standardization aligns with best practices, as improper standardization of such columns could distort their meaningful contribution to the model. By splitting the data into training and testing sets, the model ensures that the ML models are evaluated on unseen data, providing a robust measure of their generalization performance. These steps collectively form the foundation for development of an accurate and scalable predictive model for COVID-19 diagnosis. Framework of Diagnostic Prediction of COVID-19 disease is shown in Figure 4.

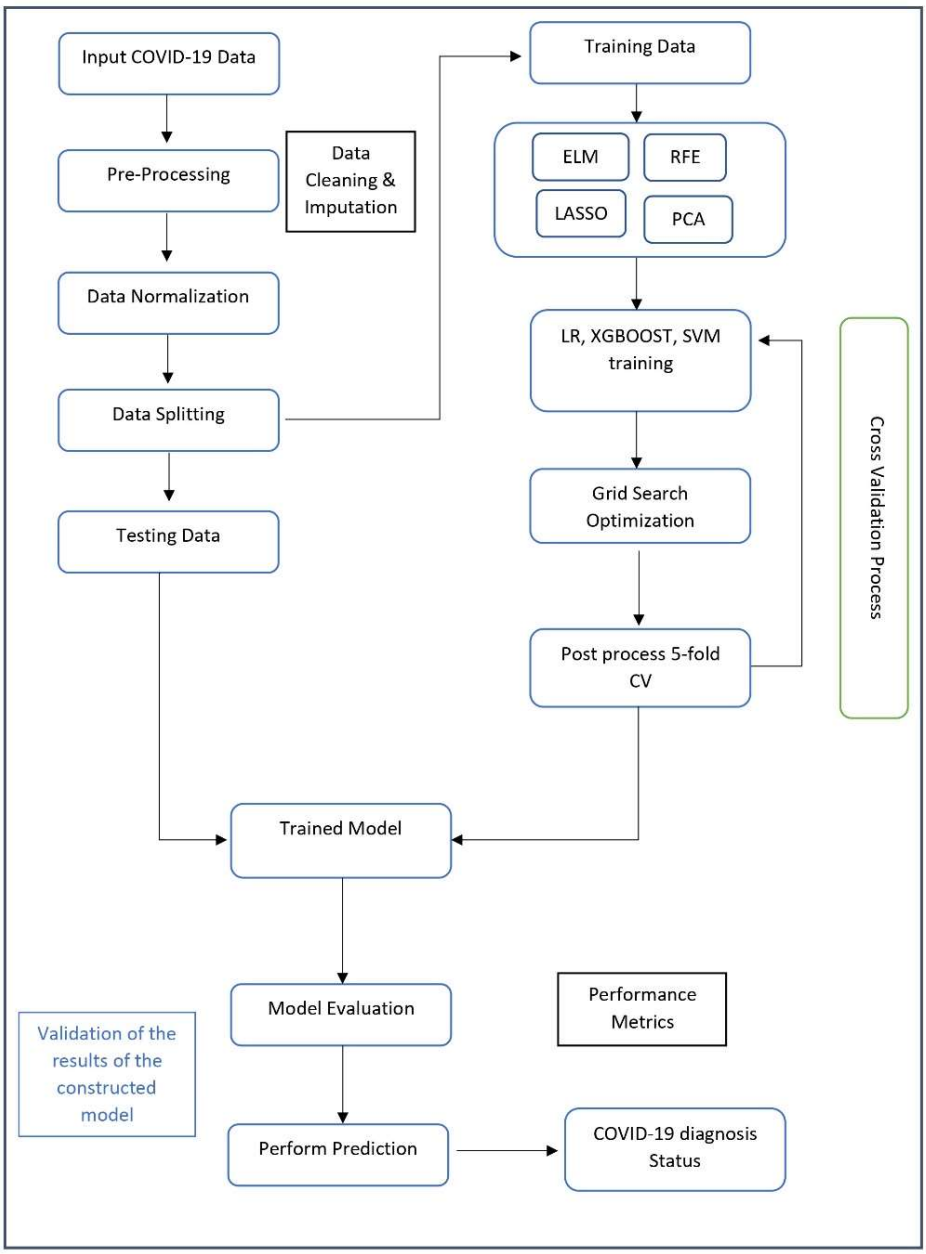

Figure 4. Framework of Diagnostic Prediction of COVID-19 Disease.

4.4 Feature Selection Techniques

As described, the feature selection techniques utilized in this research are: PCA, RFE, and LASSO. PCA feature exclusion generates new characteristics through linear combinations of original instances, maximizing dataset variance while excluding already accounted variance. The resulting PCA components capture reduced variance compared to the initial features. Widely used in machine learning, PCA effectively reduces dimensionality and eliminates redundant records within selected features (Lang *et al.,* 2019). RFE optimizes feature combinations by recursively omitting specific features, building models on the remaining data, and cleaning the optimal blend constructed on modelling outcomes (Zhou *et al.,* 2022). RFE finds widespread use in various fields such as landslide vulnerability evaluation, biological data recognition, and land utilization categorization because of its benefits and practicality (Lin *et al.,* 2017; Mostafiz *et al.,* 2020). LASSO is a type of regularization method that can be used to prevent overfitting in statistical models (Friedrich *et al.,* 2022). In contrast to other techniques for identifying variables like Classification and Regression Tree (CART) and Random Forest,

LASSO is particularly effective in correctly identifying relevant variables in psychiatric data. Table 3 provides the comparison of these techniques.

Table 3. Comparison of Feature Selection Techniques.

Parameters	PCA	RFE	LASSO
Procedure Scrutiny	Dimensionality reduction technique	Feature elimination method	Feature selection method with regularization
Benchmarks	Captures maximum variance in data	Selects features based on importance	Shrinks coefficients, selects features
Search	Transformation of original features into orthogonal components	Iteratively removes features based on model performance	Penalizes coefficients towards zero, performing feature selection
Findings	Reduced dimensionality, orthogonal feature space	Reduced feature set, preserved predictive power	Sparse coefficient vector, feature selection

4.5 Machine Learning Models

This study employed three models trained utilizing machine learning to forecast COVID-19. The Support Vector Machine (SVM) is excellent at identifying trends and classifying data, separating classes using high-dimensional dividing lines. Logistic Regression (LR), a prominent statistical method, determines the probability that a specific instance is a member of a particular class, making it appropriate to handle binary and multi-class issues. Lastly, Extreme Gradient Boosting (XGBoost) utilizes a sequential learning technique, constructing a series of progressively accurate models to improve overall prediction performance. Each model offers distinct strengths to the task of predicting COVID-19.

4.6 Evaluation Metrics

To determine how well the model predicts, it is tested using a test set of unobserved data. The percentage of the test set predictions that the model correctly identified is known as accuracy. In simple terms, it shows the model's overall ability to classify data that is received. The quality of positive predictions—that is, the degree to which the researcher can have faith that an identified positive case is, in fact, positive—is the main emphasis of precision as Equation (1) states:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

The accuracy of positive predictions, that is, how well the model finds every genuinely positive case in the data, is the main emphasis of recall as Equation (2) states:

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive} \tag{2}$$

Recall and precision are combined into one parameter called the F1-Score. It creates a balance between these two metrics, making sure a model cannot get a high score by being exceptionally good at one thing, as Equation (3):

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative} \tag{3}$$

Because of this, the F1-Score serves as a helpful gauge of a model's general efficiency in detecting infected patients, provided by Equation (4):

$$F1\ score = \frac{2 * Recall * Precision}{Recall + Precision} \quad (4)$$

5. Results

5.1 Experiment 1: Exclusion of Grid Search CV

This study examined the dataset of 279 patients hospitalized in IRCCS Ospedale San Raffaele Hospital in Milan, Italy, from late February to mid-March 2020. The information contains demographics, blood test results, and COVID-19 test status (positive or negative). EM imputation was opted to handle missing values for blood tests. To identify COVID-19 risk factors, the data was subjected to a variety of feature extraction methods (such as RFE, LASSO, and PCA) and ML models such as SVM, LR, and XGBoost. Table 4 highlights the number of features each technique selects.

Table 4. Selected Features for Each Technique.

Methods	FS Technique	Selected Features	Types of Features
Filter	PCA	6	Eosinophils, Lymphocytes, Basophils, Neutrophils, GGT, and LDH
Wrapper	RFE	5	GENDER, WBC, AST, LDH, and CRP
Embedded	LASSO	8	GENDER, WBC, Eosinophils, CRP, AST, ALT, ALP, LDH

The PCA analysis demonstrates the importance of dimensionality reduction by selecting six principal components that encapsulate the most significant variations in the dataset while avoiding redundancy. RFE (Recursive Feature Elimination) emphasizes the utility of wrapper methods by selecting five key features based on iterative model performance evaluation, which includes both clinical and demographic variables. Embedded methods like LASSO prioritize sparsity by enforcing feature selection through regularization, highlighting the eight most relevant predictors, including both blood biomarkers and demographic data. The balance achieved in each technique underscores the importance of tailoring feature selection to the specific algorithm and dataset properties. Moreover, the selected features align well with established clinical markers of COVID-19 severity, enhancing model interpretability and real-world application potential. These methodologies collectively optimize the trade-off between computational efficiency and predictive accuracy, offering robust frameworks for diagnostic model development.

The choice of features significantly impacts the model's ability to generalize and make accurate predictions, especially in medical datasets with high-dimensional data. PCA's ability to retain key information while reducing dimensionality highlights its utility in preprocessing steps, particularly for algorithms sensitive to multicollinearity. RFE, as a wrapper method, ensures the inclusion of features that contribute directly to model performance, making it ideal for datasets with moderate complexity. LASSO, by penalizing less relevant features, not only simplifies the model but also helps in avoiding overfitting, a common issue in predictive modeling. The integration of these feature selection techniques ensures a comprehensive approach to capturing both linear and nonlinear relationships within the data. By employing these varied methodologies, the framework enhances the robustness of diagnostic models, making them adaptable to diverse clinical datasets and scenarios. Figure 5(a) and Figure 5(b) displays the relevant results.

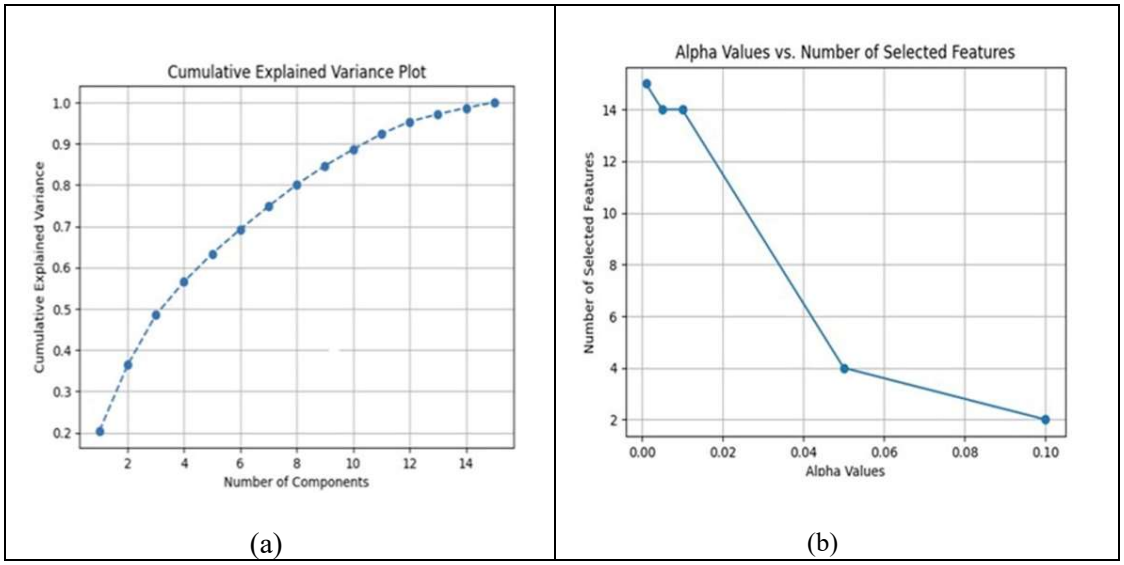

Figure 5. (a) Cumulative Explained Variance vs. Number of Components (b) Alpha Values vs. Selected Features.

The Table 5 presents a comparative analysis of various feature selection techniques integrated with ML models for predicting outcomes. Techniques such as PCA, RFE, and LASSO were paired with classifiers like SVC, LR and XGBoost to evaluate their performance across key metrics: accuracy, precision, recall, and F1-score. The results highlight the compatibility between feature selection techniques and ML models, showcasing how feature selection impacts predictive accuracy and model's efficiency.

Table 5. Evaluation of ML Models with Integrated Feature Selection.

Technique	Accuracy	Precision	Recall	F1-Score
PCA + SVC	0.7143	0.6863	1.0000	0.8140
PCA + LR	0.7321	0.7500	0.8571	0.7999
PCA + XGBoost	0.7321	0.7941	0.7714	0.7826
RFE + SVC	0.7143	0.7317	0.8571	0.7895
RFE + LR	0.7143	0.7209	0.8857	0.7949
RFE + XGBoost	0.7143	0.7317	0.8571	0.7895
LASSO + SVC	0.7679	0.7619	0.9143	0.8312
LASSO + LR	0.7679	0.7619	0.9143	0.8312
LASSO + XGBoost	0.7679	0.7750	0.8857	0.8267

The accuracy achieved among the feature selection techniques, LASSO consistently outperformed PCA and RFE, achieving the highest accuracy of 0.7679 across all paired ML models (SVC, LR, and XGBoost). This indicates that LASSO's ability to refine features by penalizing irrelevant variables significantly enhances model performance. In contrast, PCA and RFE achieved similar accuracy levels, with PCA slightly outperforming RFE when paired with Logistic Regression and XGBoost (both achieving 0.7321) compared to RFE's accuracy of 0.7143. Interestingly, SVC consistently performed slightly lower in accuracy, irrespective of the feature selection technique used.

This comparative analysis on accuracy underscores the importance of selecting the appropriate feature selection techniques and ML models integration. While LASSO proved superior in this case, the choice of technique might vary based on dataset characteristics and the problem's complexity. Additionally, understanding the compatibility between feature selection techniques and models is crucial, as suboptimal integration can lead to decreased model performance. This emphasizes the need for exploratory analysis to determine the best configuration for specific use cases. Figure 6 demonstrates the accuracy results obtained, highlighting the variation across different combinations.

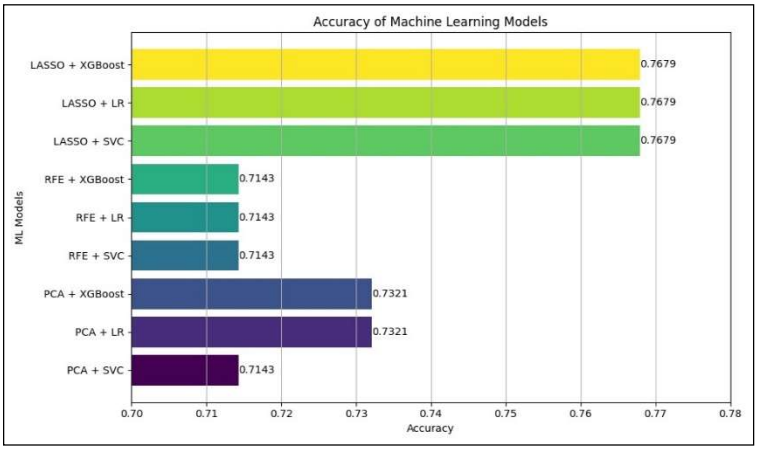

Figure 6. Accuracy of Integrated Models (exclusion of Grid Search CV).

The recision scores among the combinations, PCA with XGBoost achieved the highest precision of 0.7941, indicating its effectiveness in minimizing false positives while correctly identifying positive cases. This is closely followed by LASSO with XGBoost, which recorded a precision of 0.7750, demonstrating a similarly robust performance. LASSO generally performed well across all ML models, maintaining precision scores above 0.76, showcasing its ability to select features that enhance the discriminative capability of the models. Comparatively, RFE showed moderate performance, with precision values ranging from 0.7209 to 0.7317, slightly trailing PCA in terms of precision. The analysis highlights the varying impact of feature selection techniques on precision. While PCA and LASSO proved more effective in ensuring higher precision, RFE exhibited less consistent results. This emphasizes the importance of aligning feature selection techniques with specific ML models to achieve optimal precision, depending on the dataset and problem context. Figure 7 highlights the precision results obtained by the experimentations.

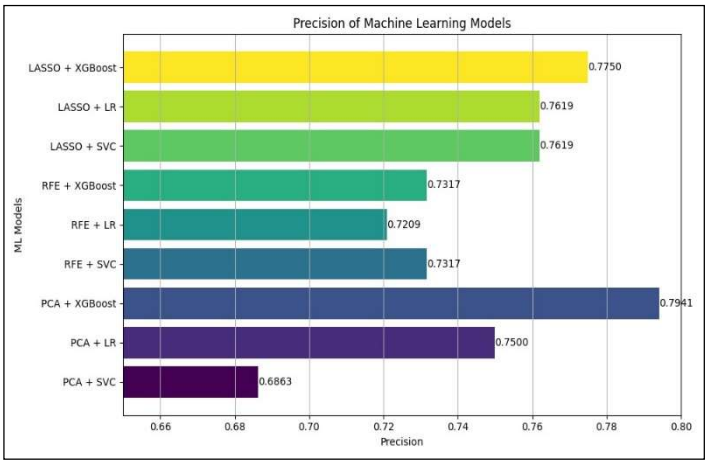

Figure 7. Precision of Integrated Models (Exclusion of Grid Search CV).

For the recall analysis, PCA with SVC combination stands out with a perfect recall of 1.0000, indicating it identified all relevant cases without missing any. However, this might come at the cost of increased false positives, which needs further examination alongside other metrics. LASSO-based combinations demonstrated consistently high recall values, with LASSO + SVC and LASSO + LR both achieving 0.9143, followed closely by LASSO + XGBoost at 0.8857. This indicates that LASSO is adept at selecting features that enhance sensitivity, making it suitable for applications where capturing all positive cases is critical. In contrast, PCA with XGBoost reported the lowest recall at 0.7714, suggesting a trade-off between its precision and recall capabilities. Meanwhile, RFE-based combinations showed balanced performance, with

recall values ranging from 0.8571 to 0.8857, highlighting its effectiveness in moderately sensitive scenarios. For scenarios prioritizing sensitivity, PCA with SVC or LASSO-based combinations emerge as optimal choices, while models like PCA with XGBoost may be better suited for precision-focused tasks. Figure 8 highlights the recall analysis.

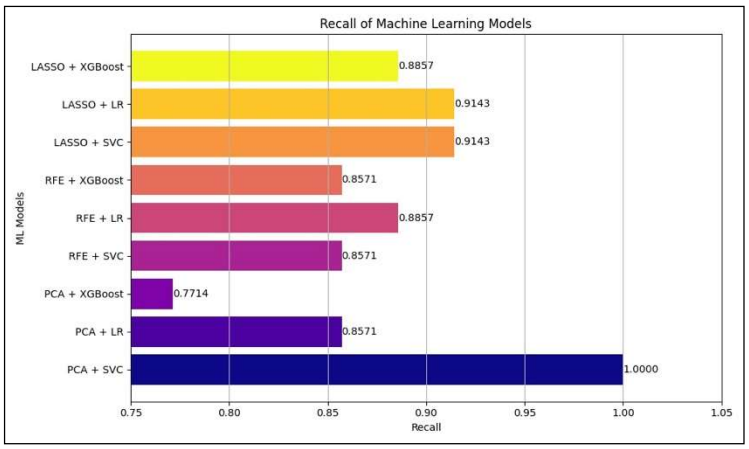

Figure 8. Recall Analysis of Integrated Models (Exclusion of Grid Search CV).

In terms of F1-score, the results show how well the models balance both precision and recall, considering the harmonized mean of these two metrics.

1. LASSO-based techniques (LASSO + SVC, LASSO + LR, LASSO + XGBoost) outperform other methods with the highest F1-scores of 0.8312 for both SVC and LR, and 0.8267 for XGBoost. These high scores suggest that LASSO feature selection combined with machine learning models is particularly effective at providing a good balance between precision and recall.

2. PCA and RFE techniques follow closely behind, with PCA + SVC yielding an F1-score of 0.8140, the highest among the PCA methods. PCA + LR and PCA + XGBoost both show slightly lower scores at 0.7999 and 0.7826, respectively.

3. RFE + LR has an F1-score of 0.7949, slightly outperforming RFE + XGBoost and RFE + SVC, both of which return an F1-score of 0.7895. Overall, LASSO techniques show consistent top performance across multiple machine learning models, while PCA and RFE also perform well, with PCA showing slightly better results in some cases. This analysis highlights LASSO as the most reliable feature selection method in terms of achieving a strong balance between precision and recall. Figure 9 highlights the F1-score analysis.

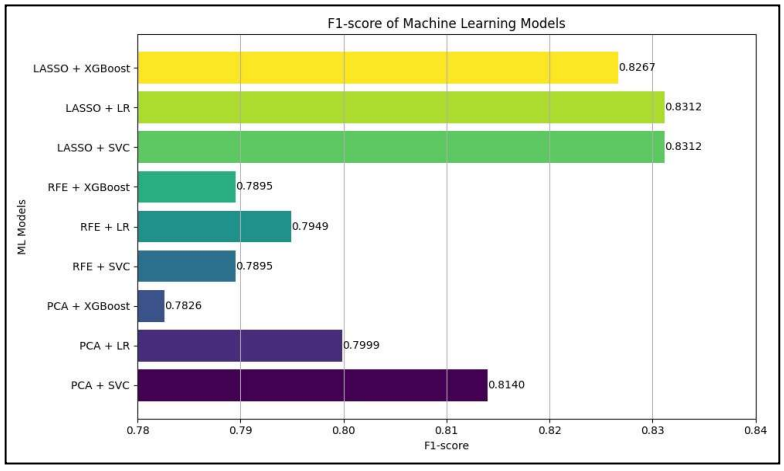

Figure 9. F1-score of Integrated Models (Exclusion of Grid Search CV).

5.2 Experiment 2: Inclusion of Grid Search CV

Further, in this study, we opted Grid Search with Cross Validation (CV) to optimize model performance as shown in Table 6. This method enabled us to find the most effective combinations of feature selection approaches and algorithms for predicting COVID-19 risk, establishing a solid foundation for constructing more accurate and reliable prediction models.

Table 6. Evaluation of ML Models with Integrated Feature Selection (Grid Search CV).

Technique	Accuracy	Precision	Recall	F1-Score
PCA + SVC	0.7143	0.7111	0.9143	0.8000
PCA + LR	0.7321	0.7500	0.8571	0.7999
PCA + XGBoost	0.7500	0.7692	0.8571	0.8108
RFE + SVC	0.7143	0.7568	0.8000	0.7777
RFE + LR	0.7143	0.7111	0.9143	0.8000
RFE + XGBoost	0.6964	0.725	0.8286	0.7733
LASSO + SVC	0.7679	0.8236	0.8000	0.8116
LASSO + LR	0.7679	0.7619	0.9143	0.8312
LASSO + XGBoost	0.7679	0.7619	0.9143	0.8312

The evaluation of machine learning models using Grid Search CV reveals that LASSO-based models consistently outperform others across all performance metrics. Both LASSO + LR and LASSO + XGBoost achieve the highest F1-score of 0.8312, demonstrating an excellent balance between precision and recall, alongside the highest accuracy of 0.7679. These results highlight LASSO's strength in selecting meaningful features and minimizing false positives effectively, with LASSO + SVC achieving the highest precision of 0.8236. Similarly, PCA-based models, particularly PCA + XGBoost, show robust performance with an F1-score of 0.8108, accuracy of 0.7500, and a recall of 0.8571, making them a competitive alternative for feature selection tasks. While RFE-based models exhibit slightly lower accuracy and F1-scores—ranging from 0.7733 to 0.8000—they still offer a reliable solution for capturing essential features. Overall, LASSO-based approaches emerge as the most effective, providing superior performance across metrics, followed by PCA as a strong contender. RFE, while slightly behind, remains a viable option for applications with less stringent performance requirements. Figure 10 explores the experimentation results achieved

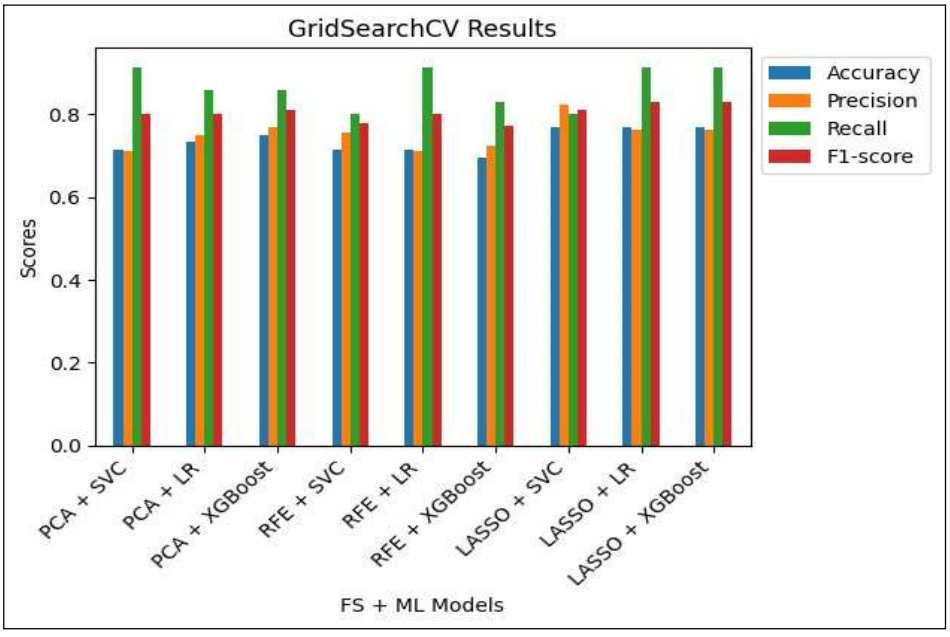

Figure 10. Evaluation of Machine Learning Models with Grid Search CV.

5.3 Experiment 3: Extreme Learning Machine

The Extreme Learning Machine approach was used to determine feature weights indirectly by computing correlation significance coefficients, correlation indices, absolute influence coefficients, and feature weights (Rajpal *et al.,* 2022). Here is an outline into the way their approach was applied.

1. Training the ELM Classifier: We built an ELM Classifier object and trained the model.
2. Extraction of Beta (Coefficient) Matrix: After training the ELM model, we acquired the optimum beta matrix. This beta matrix represents the coefficients associated with the connections between the ELM model's input and hidden layers.
3. Initializing Random Weight Matrix: Then we created a random weight matrix based on the amount of characteristics in the dataset and nodes in the ELM model's hidden layer.
4. Calculate the Correlation Significance Coefficient : This function uses the weight matrix, the beta matrix, and the index as input. It iterates across the concealed nodes, calculating the correlation significant coefficient using the ELM model parameters.
5. Calculating Correlation Index: The next function calculates the correlation index using the correlation significant coefficient. This index measures the strength of the link between the input features and the target variable, as described by the ELM.
6. Calculating Absolute Influence Coefficient: The function uses the calculating correlation index and pre-calculated R values to get the absolute influence coefficient. This coefficient indicates the relative value of each parameter for forecasting the target variable.
7. Feature Weight: This calculates the feature weight by aggregating the absolute influence coefficients over all features. This weight embodies the grade of significance of each feature in the classification task.

We employed this Extreme Learning Machine model-based feature selection (FSELM) technique to find key features for predicting COVID-19, namely, Age, Platelets, Neutrophils, Lymphocytes, Eosinophils, Basophils, CRP, AST, ALT, and LDH. Weights for every feature were determined by dissecting the model's architecture through the learning procedure.

The weights represent how much each feature contributes to the forecast. Then, using these weights, we prioritize or choose the most informative features for our classification models (SVM, LR, XGBoost). Table 8 highlights the evaluation of the FSELM approach without applying Grid Search CV provides key insights into the performance of different machine learning models. Table 7 highlights the experimental results

Table 7. Evaluation of Feature Selection Exteme Learning Machine.

Technique	Accuracy	Precision	Recall	F1-Score
FSELM + SVM	0.6964	0.6875	0.9429	0.7952
FSELM + LR	0.6964	0.7143	0.8571	0.7792
FSELM + XGBoost	0.6786	0.7179	0.8000	0.7568

Among the techniques, FSELM + SVM demonstrates the highest recall value of 0.9429, indicating a strong ability to identify positive cases. However, its accuracy of 0.6964 and precision of 0.6875 suggest a relatively higher occurrence of false positives, slightly affecting its overall reliability. The F1-score of 0.7952 highlights its balanced trade-off between precision and recall. On the other hand, FSELM + LR exhibits consistent performance with an accuracy of 0.6964 and a slightly higher precision of 0.7143 compared to FSELM + SVM. Its recall of 0.8571 and F1-score of 0.7792 indicate that it effectively balances both sensitivity and specificity, making it a reliable choice for general applications. FSELM + XGBoost, while achieving the lowest accuracy (0.6786) among the techniques, still maintains reasonable precision (0.7179) and recall (0.8000). Its F1-score of 0.7568 suggests that the technique prioritizes precision over sensitivity compared to the other two models. Overall, FSELM + SVM excels in recall, making it highly suitable for tasks requiring accurate identification of

positive cases. However, FSELM + LR provides more balanced performance across all metrics, while FSELM + XGBoost demonstrates moderate effectiveness, with room for improvement in both accuracy and recall. These results provide a baseline for comparison with the models fine-tuned using Grid Search CV. The evaluation of FSELM techniques with Grid Search CV shows noticeable improvements in performance metrics compared to the previous results without optimization as shown in Table 8.

Table 8. Evaluation of Feature Selection Exteme Learning Machine with Grid Search CV.

Technique	Accuracy	Precision	Recall	F1-Score
FSELM + SVC	0.6786	0.6889	0.8857	0.775
FSELM + LR	0.6964	0.7143	0.8571	0.7792
FSELM + XGBoost	0.7321	0.7500	0.8571	0.8000

Among the models, FSELM + XGBoost emerges as the most robust, achieving the highest accuracy of 0.7321, precision of 0.7500, and an F1-score of 0.8000. This indicates that Grid Search CV has effectively optimized the model's hyperparameters, allowing it to deliver a balanced performance across all metrics while maintaining a strong recall value of 0.8571. FSELM + LR displays consistent results, with an accuracy of 0.6964, precision of 0.7143, and a recall of 0.8571. The F1-score of 0.7792 signifies balanced performance, although its overall improvement is less pronounced than that of FSELM + XGBoost.

FSELM + SVC achieves a recall of 0.8857, which is the highest among the three models, highlighting its ability to identify positive cases effectively. However, its lower accuracy (0.6786) and precision (0.6889) suggest some trade-offs in false positives. Its F1-score of 0.775 reflects its moderately balanced performance, but it trails behind the other techniques in overall accuracy. In conclusion, FSELM + XGBoost demonstrates the most significant enhancement through Grid Search CV, delivering the best overall performance. While FSELM + LR provides reliable and consistent results, FSELM + SVC is particularly strong in recall, making it suitable for tasks where sensitivity is a priority. These findings underscore the effectiveness of Grid Search CV in optimizing model performance for feature selection and machine learning integration. Figure 11 highlights the results of our experiment.

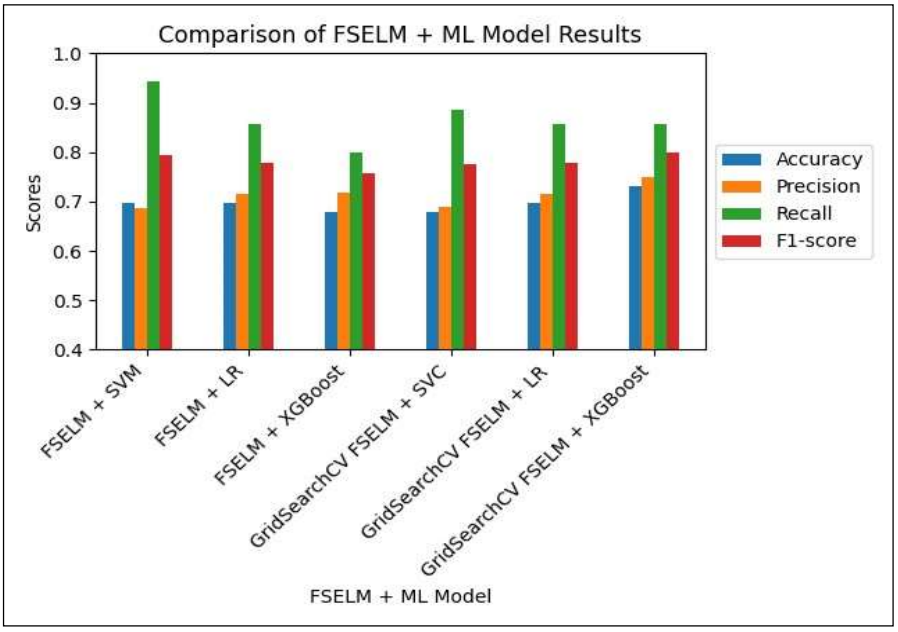

Figure 11. Evaluation of Machine Learning Models with Grid Search CV.

6. Discussion

Our findings provide significant new insights into the development of COVID-19 risk prediction models and feature selection techniques. Three distinct feature selection methods—LASSO, Recursive Feature Elimination (RFE), and Principal Component Analysis (PCA)—were thoroughly investigated. PCA demonstrated a strong ability to balance maintaining essential features while reducing data complexity. LASSO effectively identified the most relevant attributes and minimized the influence of less significant ones. The results further emphasize the importance of evaluating models using multiple performance metrics.

For instance, LASSO combined with all three machine learning (ML) models delivered superior accuracy overall, and when integrated with Grid Search CV, LASSO with Support Vector Classifier (SVC) achieved the highest precision among all algorithms. For recall, PCA with SVC outperformed other combinations both with and without Grid Search CV. Notably, for the F1-score, LASSO with Logistic Regression (LR) delivered the best overall results. Additionally, PCA consistently outperformed RFE across all three ML algorithms, underscoring that no single feature selection method can universally optimize all performance metrics. We also explored the potential of the Extreme Learning Machine (ELM) technique to identify critical features for COVID-19 risk prediction. This approach allowed us to recover the beta matrix, which contains the connection weights between the input and hidden layers of the ELM classifier. By calculating coefficients such as correlation significance, correlation index, and absolute effect coefficient, we inferred feature weights indirectly. Using these feature weights, we prioritized the most relevant features for ML models like SVM, LR, and XGBoost. The results demonstrated that ELM-based feature selection significantly enhanced the predictive capabilities of these models. Specifically, ELM with SVC achieved a recall of 0.9429 prior to applying Grid Search CV and retained superior performance even after optimization, outperforming LR and XGBoost in this regard.

While this study provides valuable insights into feature selection and ML model development for COVID-19 risk prediction, several limitations should be addressed in future research. Firstly, we utilized a single dataset, which may limit the generalizability of our findings incorporating data from multiple sources could yield more robust and widely applicable results. Secondly, our analysis focused on seven ML algorithms. Expanding this scope to include deep learning approaches could further enhance prediction accuracy. Thirdly, we concentrated solely on clinical data. Including lifestyle factors (e.g., physical activity and smoking habits) and socioeconomic variables could provide a more holistic perspective on COVID-19 risk. Moreover, integrating unstructured data types, such as medical images and ECG signals, could uncover additional insights. Lastly, the limited size of the dataset poses a restriction on the broad applicability of our conclusions. Future studies with larger and more diverse datasets may enhance the generalizability and impact of this research.

7. Conclusion

This study highlights the critical role of feature selection techniques in enhancing the performance of ML models for COVID-19 risk prediction. By evaluating PCA, RFE, and LASSO alongside various ML algorithms, we identified LASSO with SVC as the best performer under Grid Search CV, achieving high accuracy (0.7679) and precision (0.8236). Other combinations, such as LASSO with LR and XGBoost, excelled in recall (0.9143) and F1-score (0.8312), emphasizing the importance of aligning feature selection methods with model types. Additionally, the ELM-based feature selection approach proved effective in identifying key predictors and improving classification performance, particularly with SVM and XGBoost after optimization. Notably, LDH emerged as the most critical feature for COVID-19 prediction, followed by Eosinophils, CRP, and AST. In conclusion, our findings underscore the value of robust feature selection and optimization techniques in developing reliable, interpretable ML models for medical applications.

Data availability

You can access publicly available datasets by visiting the following link https://zenodo.org/records/3886927#.Yc6feGiOmUk.

Acknowledgement

The authors would like to articulate our wholehearted appreciation to the Institute of Big Data Analytics and Artificial Intelligence (IBDAAI), Kompleks Al-Khawarizmi and Universiti Teknologi MARA (UiTM) for their high gratitude support and assistance.

Funding

The author(s) received no specific funding for this work.

Author Contribution

Author1 Developed the research methodology and designed the experimental framework for feature selection and machine learning techniques. Author2 Supervised the research process, provided guidance on model optimization techniques, and finalized the manuscript for submission. Author3 Performed data preprocessing, statistical analysis, and visualization of results, contributing to the interpretation of key findings. Author4 Provided domain expertise, reviewed the manuscript critically for intellectual content, and ensured the clinical relevance of the study. Author5 Coordinated collaborative efforts with international contributors, integrated feedback, and ensured compliance with journal guidelines. Author6 Conducted the literature review and drafted the manuscript, ensuring clarity and logic in the writing process.

Conflict of Interest

The authors have no conflicts of interest to declare.

References

Akhiat, Y., Asnaoui, Y., Chahhou, M., & Zinedine, A. (2020). A new graph feature selection approach. 2020 6th IEEE Congress on Information Science and Technology (CiSt), 156–161. https://doi.org/10.1109/cist49399.2021.9357067.

Akhiat, Y., Chahhou, M., & Zinedine, A. (2018). Feature selection based on graph representation. 2018 IEEE 5th International Congress on Information Science and Technology (CiSt), 3, 232–237. https://doi.org/10.1109/cist.2018.8596467.

Akhiat, Y., Chahhou, M., & Zinedine, A. (2019). Ensemble Feature Selection Algorithm. International Journal of Intelligent Systems and Applications, 11(1), 24–31. https://doi.org/10.5815/ijisa.2019.01.03.

Aladağ, N., & Atabey, R. D. (2020). The role of concomitant cardiovascular diseases and cardiac biomarkers for predicting mortality in critical COVID-19 patients. Acta Cardiologica, 76(2), 132–139. https://doi.org/10.1080/00015385.2020.1810914.

Alsharif, W., & Qurashi, A. (2020). Effectiveness of COVID-19 diagnosis and management tools: A review. Radiography, 27(2), 682–687.

https://doi.org/10.1016/j.radi.2020.09.010.

Aria, M., Cuccurullo, C., & Gnasso, A. (2021). A comparison among interpretative proposals for Random Forests. Machine Learning With Applications, 6, 100094. https://doi.org/10.1016/j.mlwa.2021.100094.

Azeem, M., Javaid, S., Khalil, R., Fahim, H., Althobaiti, T., Alsharif, N., & Saeed, N. (2023). Neural networks for the detection of COVID-19 and other diseases: Prospects and challenges. Bioengineering, 10(7), 850. https://doi.org/10.3390/bioengineering10070850.

Aznar-Gimeno, R., Esteban, L. M., Labata-Lezaun, G., Del-Hoyo-Alonso, R., Abadia-Gallego, D., Paño-Pardo, J. R., Esquillor-Rodrigo, M. J., Lanas, Á., & Serrano, M. T. (2021). A Clinical Decision Web to Predict ICU Admission or Death for Patients Hospitalised with COVID-19 Using Machine Learning Algorithms. International Journal of Environmental Research and Public Health, 18(16), 8677. https://doi.org/10.3390/ijerph18168677.

Banerjee, A., Ray, S., Vorselaars, B., Kitson, J., Mamalakis, M., Weeks, S., Baker, M., & Mackenzie, L. S. (2020). Use of Machine Learning and Artificial Intelligence to predict SARS-CoV-2 infection from Full Blood Counts in a population. International Immunopharmacology, 86, 106705. https://doi.org/10.1016/j.intimp.2020.106705.

Basit, A, Zain, J. M, Jumaat, A. K, Hamdan, N. and Mojahid, H.Z. (2024). PREDICTING COVID-19 TRENDS: A DEEP DIVE INTO TIME- DEPENDENT SIRSD WITH DEEP-LEARNING TECHNIQUE, Malaysian Journal of Computing, vol. 9, no. 2, p. (1955–1978), Oct. 2024, doi: 10.24191/mjoc.v9i2.27425

Basit, A., Zain, J. M., Mojahid, H. Z., & Ali, M. (2022). Genesis of monkeypox. Journal of Pure and Applied Microbiology, 16(suppl 1), 3192–3197. https://doi.org/10.22207/jpam.16.spl1.19.

Bolourani, S., Brenner, M., Wang, P., McGinn, T., Hirsch, J. S., Barnaby, D., & Zanos, T. P. (2021). A Machine Learning Prediction Model of respiratory failure within 48 hours of patient admission for COVID-19: Model development and validation. Journal of Medical Internet Research, 23(2), e24246. https://doi.org/10.2196/24246.

Bouchlaghem, Y., Akhiat, Y., & Amjad, S. (2022). Feature Selection: a review and comparative study. E3S Web of Conferences, 351, 01046. https://doi.org/10.1051/e3sconf/202235101046.

Brinati, D., Campagner, A., Ferrari, D., Locatelli, M., Banfi, G., & Cabitza, F. (2020). Detection of COVID-19 Infection from Routine Blood Exams with Machine Learning: A Feasibility Study. Journal of Medical Systems, 44(8). https://doi.org/10.1007/s10916-020-01597-4.

Chowdhury, M. E. H., Rahman, T., Khandakar, A., Al-Madeed, S., Zughaier, S. M., Doi, S. a. R., Hassen, H., & Islam, M. T. (2021). An early warning tool for predicting mortality risk of COVID-19 patients using machine learning. Cognitive Computation, 16(4), 1778–1793. https://doi.org/10.1007/s12559-020-09812-7.

De Terwangne, C., Laouni, J., Jouffe, L., Lechien, J., Bouillon, V., Place, S., Capulzini, L., Machayekhi, S., Ceccarelli, A., Saussez, S., & Sorgente, A. (2020). Predictive Accuracy of COVID-19 World Health Organization (WHO) Severity Classification and Comparison with a Bayesian-Method-Based Severity Score (EPI-SCORE). Pathogens, 9(11), 880. https://doi.org/10.3390/pathogens9110880.

Dede, G., Filiopoulou, E., Paroni, D., Michalakelis, C., & Kamalakis, T. (2023). Analysis and Evaluation of major COVID-19 features: A pairwise Comparison approach. Operations Research Forum, 4(1). https://doi.org/10.1007/s43069-023-00201-y.

Ferrari, D., Motta, A., Strollo, M., Banfi, G., & Locatelli, M. (2020). Routine blood tests as a potential diagnostic tool for COVID-19. Clinical Chemistry and Laboratory Medicine (CCLM), 58(7), 1095–1099. https://doi.org/10.1515/cclm-2020-0398.

Friedrich, S., Groll, A., Ickstadt, K., Kneib, T., Pauly, M., Rahnenführer, J., & Friede, T. (2022). Regularization approaches in clinical biostatistics: A review of methods and their applications. Statistical Methods in Medical Research, 32(2), 425–440. https://doi.org/10.1177/09622802221133557.

Huang, D., Wang, T., Chen, Z., Yang, H., Yao, R., & Liang, Z. (2020). A novel risk score to predict diagnosis with coronavirus disease 2019 (COVID-19) in suspected patients: A retrospective, multicenter, and observational study. Journal of Medical Virology, 92(11), 2709–2717. https://doi.org/10.1002/jmv.26143.

Jain, R., & Xu, W. (2023). Artificial Intelligence based wrapper for high dimensional feature selection. BMC Bioinformatics, 24(1). https://doi.org/10.1186/s12859-023-05502-x.

Jawa, T. M. (2022). Logistic regression analysis for studying the impact of home quarantine on psychological health during COVID-19 in Saudi Arabia. Alexandria Engineering Journal, 61(10), 7995–8005. https://doi.org/10.1016/j.aej.2022.01.047.

Lang, S., Bravo-Marquez, F., Beckham, C., Hall, M., & Frank, E. (2019). WekaDeeplearning4j: A deep learning package for Weka based on Deeplearning4j. Knowledge-Based Systems, 178, 48–50. https://doi.org/10.1016/j.knosys.2019.04.013.

Li, Z., Yi, Y., Luo, X., Xiong, N., Liu, Y., Li, S., Sun, R., Wang, Y., Hu, B., Chen, W., Zhang, Y., Wang, J., Huang, B., Lin, Y., Yang, J., Cai, W., Wang, X., Cheng, J., Chen, Z., . . . Ye, F. (2020). Development and clinical application of a rapid IgM-IgG combined antibody test for SARS-CoV-2 infection diagnosis. Journal of Medical Virology, 92(9), 1518–1524. https://doi.org/10.1002/jmv.25727.

Lin, X., Li, C., Zhang, Y., Su, B., Fan, M., & Wei, H. (2017). Selecting Feature Subsets Based on SVM-RFE and the Overlapping Ratio with Applications in Bioinformatics. Molecules, 23(1), 52. https://doi.org/10.3390/molecules23010052.

Linden, T., Hanses, F., Domingo-Fernández, D., DeLong, L. N., Kodamullil, A. T., Schneider, J., Vehreschild, M. J., Lanznaster, J., Ruethrich, M. M., Borgmann, S., Hower, M., Wille, K., Feldt, T., Rieg, S., Hertenstein, B., Wyen, C., Roemmele, C., Vehreschild, J. J., Jakob, C. E., . . . Fröhlich, H. (2021). Machine learning based prediction of COVID-19 mortality suggests repositioning of anticancer drug for treating severe cases. Artificial Intelligence in the Life Sciences, 1, 100020. https://doi.org/10.1016/j.ailsci.2021.100020.

Mojahid, H. Z., Zain, N. J. M., Basit, N. A., Ali, N. M., & Yusoff, N. M. (2024). A REVIEW ON EXTENSIVELY USED MACHINE LEARNING TECHNIQUES FOR THE PREDICTION OF COVID-19. Suranaree Journal of Science and Technology, 31(1), 030167(1-13). https://doi.org/10.55766/sujst-2024-01-e01334.

Mostafiz, R., Uddin, M. S., Alam, N., Reza, M. M., & Rahman, M. M. (2020). Covid-19 detection in chest X-ray through random forest classifier using a hybridization of deep CNN and DWT optimized features. Journal of King Saud University - Computer and Information Sciences, 34(6), 3226–3235. https://doi.org/10.1016/j.jksuci.2020.12.010.

Parchure, P., Joshi, H., Dharmarajan, K., Freeman, R., Reich, D. L., Mazumdar, M., Timsina, P., & Kia, A. (2020). Development and validation of a machine learning-based prediction model for near-term in-hospital mortality among patients with COVID-19. BMJ Supportive & Palliative Care, 12(e3), e424–e431. https://doi.org/10.1136/bmjspcare-2020-002602.

Rahman, T., Al-Ishaq, F. A., Al-Mohannadi, F. S., Mubarak, R. S., Al-Hitmi, M. H., Islam, K. R., Khandakar, A., Hssain, A. A., Al-Madeed, S., Zughaier, S. M., & Chowdhury, M. E. H. (2021). Mortality prediction utilizing blood biomarkers to predict the severity of COVID-19 using machine learning technique. Diagnostics, 11(9), 1582. https://doi.org/10.3390/diagnostics11091582.

Rajpal, S., Agarwal, M., Rajpal, A., Lakhyani, N., Saggar, A., & Kumar, N. (2022). COV-ELM classifier: An extreme learning machine based identification of COVID-19 using chest X-ray images. *Intelligent Decision Technologies*, *16*(1), 193–203. https://doi.org/10.3233/idt-210055.

Rikan, S. B., Azar, A. S., Ghafari, A., Mohasefi, J. B., & Pirnejad, H. (2021). COVID-19 diagnosis from routine blood tests using artificial intelligence techniques. Biomedical Signal Processing and Control, 72, 103263. https://doi.org/10.1016/j.bspc.2021.103263.

Saberi-Movahed, F., Mohammadifard, M., Mehrpooya, A., Rezaei-Ravari, M., Berahmand, K., Rostami, M., Karami, S., Najafzadeh, M., Hajinezhad, D., Jamshidi, M., Abedi, F., Mohammadifard, M., Farbod, E., Safavi, F., Dorvash, M., Mottaghi-Dastjerdi, N., Vahedi, S., Eftekhari, M., Saberi-Movahed, F., . . . Tavassoly, I. (2022). Decoding clinical biomarker space of COVID-19: Exploring matrix factorization-based feature selection methods. Computers in Biology and Medicine, 146, 105426. https://doi.org/10.1016/j.compbiomed.2022.105426.

Sharma, M., Dhiman, N., Vandana, N., & Mishra, V. N. (2021). Mediative fuzzy logic mathematical model: A contradictory management prediction in COVID-19 pandemic. Applied Soft Computing, 105, 107285. https://doi.org/10.1016/j.asoc.2021.107285.

Thell, R., Zimmermann, J., Szell, M., Tomez, S., Eisenburger, P., Haugk, M., Kreil, A., Spiel, A., Blaschke, A., Klicpera, A., Janata, O., Krugluger, W., Sebesta, C., Herkner, H., & Laky, B. (2021). Standard blood laboratory values as a clinical support tool to distinguish between SARS-CoV-2 positive and negative patients. Scientific Reports, 11(1). https://doi.org/10.1038/s41598-021-88844-x.

Vaid, A., Somani, S., Russak, A. J., De Freitas, J. K., Chaudhry, F. F., Paranjpe, I., Johnson, K. W., Lee, S. J., Miotto, R., Richter, F., Zhao, S., Beckmann, N. D., Naik, N., Kia, A., Timsina, P., Lala, A., Paranjpe, M., Golden, E., Danieletto, M., . . . Glicksberg, B. S. (2020). Machine learning to predict mortality and critical events in a cohort of patients with COVID-19 in New York City: model development and validation. Journal of Medical Internet Research, 22(11), e24018. https://doi.org/10.2196/24018.

Wang, J. M., Liu, W., Chen, X., McRae, M. P., McDevitt, J. T., & Fenyö, D. (2020). Predictive modeling of morbidity and mortality in COVID-19 hospitalized patients and its clinical implications. medRxiv (Cold Spring Harbor Laboratory). https://doi.org/10.1101/2020.12.02.20235879.

Wang, K., Zuo, P., Liu, Y., Zhang, M., Zhao, X., Xie, S., Zhang, H., Chen, X., & Liu, C. (2020). Clinical and laboratory predictors of in-hospital mortality in patients with coronavirus disease-2019: a cohort study in Wuhan, China. Clinical Infectious Diseases, 71(16), 2079–2088. https://doi.org/10.1093/cid/ciaa538.

Yan, L., Zhang, H., Goncalves, J., Xiao, Y., Wang, M., Guo, Y., Sun, C., Tang, X., Jing, L., Zhang, M., Huang, X., Xiao, Y., Cao, H., Chen, Y., Ren, T., Wang, F., Xiao, Y., Huang, S., Tan, X., . . . Yuan, Y. (2020). An interpretable mortality prediction model for COVID-19 patients. Nature Machine Intelligence, 2(5), 283–288. https://doi.org/10.1038/s42256-020-0180-7.

Yang, H. S., Hou, Y., Vasovic, L. V., Steel, P. a. D., Chadburn, A., Racine-Brzostek, S. E., Velu, P., Cushing, M. M., Loda, M., Kaushal, R., Zhao, Z., & Wang, F. (2020). Routine laboratory blood tests predict SARS-COV-2 infection using machine learning. Clinical Chemistry, 66(11), 1396–1404. https://doi.org/10.1093/clinchem/hvaa200.

Zhang, C., Qin, L., Li, K., Wang, Q., Zhao, Y., Xu, B., Liang, L., Dai, Y., Feng, Y., Sun, J., Li, X., Hu, Z., Xiang, H., Dong, T., Jin, R., & Zhang, Y. (2020). A novel scoring system for prediction of disease severity in COVID-19. Frontiers in Cellular and Infection Microbiology, 10. https://doi.org/10.3389/fcimb.2020.00318.

Zhang, Y., Wei, C., Guo, C., Bi, R., Xie, J., Guan, D., Yang, C., & Jiang, Y. (2017). Prognostic value of microRNAs in hepatocellular carcinoma: a meta-analysis. Oncotarget, 8(63), 107237–107257. https://doi.org/10.18632/oncotarget.20883.

Zhou, Y., Tian, S., Chen, J., Liu, Y., & Li, C. (2022). Research on Classification of Open-Pit Mineral Exploiting Information based on OOB RFE Feature Optimization. Sensors, 22(5), 1948. https://doi.org/10.3390/s22051948.

Zhu, J. S., Ge, P., Jiang, C., Zhang, Y., Li, X., Zhao, Z., Zhang, L., & Duong, T. Q. (2020). Deep-learning artificial intelligence analysis of clinical variables predicts mortality in COVID-19 patients. Journal of the American College of Emergency Physicians Open, 1(6), 1364–1373. https://doi.org/10.1002/emp2.12205.

Kaggle. (2020, March 27). Diagnosis of COVID-19 and its clinical spectrum. https://www.kaggle.com/datasets/einsteindata4u/covid19.