International Research Journal of Multidisciplinary Scope (IRJMS), 2024; 5(1): 714-728

Original Article | ISSN (O): 2582-631X DOI: 10.47857/irjms.2024.v05i01.0345

Covid-19 Prediction Using Enhanced KNN Imputation for Data Pre-Processing

Hari Priya N*, Rajeswari S

Department of Computer Science, Sree Saraswathi Thyagaraja College, Pollachi, Tamil Nadu, India. *Corresponding Author's Email: haripriyanarasimma@gmail.com

Abstract

In the wake of the recent Coronavirus Disease 2019 (COVID-19) pandemic, health care systems all over the world have been heavily affected. The rapid detection of COVID-19 has emerged as a top priority for global health systems to prevent its spread. Although Reverse Transcription-Polymerase Chain Reaction (RT-PCR) is still the method of choice for COVID-19 detection, the potential of blood test data in predictive modelling is currently being utilized gradually. In this study, we investigated the effectiveness of machine learning models for detecting COVID-19 from blood test data, with a particular emphasis on the pre-processing step involving Enhanced K-Nearest Neighbours (KNN) Imputation. By utilizing Enhanced KNN Imputation, our methodology sought to provide a more robust and precise imputation of missing values in blood test datasets. Support Vector Machine - Recursive Feature Elimination (SVM-RFE) based feature selection has been utilized to identify the most significant features. Then, we trained 5 different machine learning classifiers using both traditionally imputed and Enhanced KNN imputed data. Based on the experimental results, Random Forest model outperformed other classifiers using dataset imputed with Enhanced KNN imputation with an accuracy of 80% with all the features. The same methodology has been carried out with the exclusion of GENDER feature and as a result SVM model achieved an accuracy of 84%. The study suggests that the combination of Enhanced KNN Imputation and machine learning could be a valuable tool for COVID-19 detection, potentially aiding in faster and more accurate diagnosis.

Keywords: COVID-19, Machine learning, KNN imputation, Pre-processing, Random forest, SVM.

Introduction

The respiratory illness COVID-19, which refers to "Coronavirus Disease 2019," is caused by the new coronavirus severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). In late 2019, it initially appeared in Wuhan, Hubei Province, China, and it has since spread over the world, causing a pandemic. Many countries have imposed lockdown to restrict civilians from traveling around excessively. Because of this societal distancing element and activity limits, the health and economy of several countries have been affected (1). An initial cluster of cases was traced back to a Wuhan seafood market, pointing to an animal-to-human transmission. However, as the virus disseminated, human-to-human transmission became the dominant route of propagation (2). This global COVID-19 outbreak has placed a significant strain on healthcare systems. Multiple countries were affected within a few months, raising an international outrage. Maintaining a healthcare system that is both responsive and able to deliver basic services has proven difficult for many countries (3). The World Health Organization, also known as the WHO, provided interim guidelines to countries and it has made significant efforts to coordinate globally on monitoring, disease epidemiology, modelling using mathematics, diagnostics, medical care, and prevention in order to fight this global epidemic (4).

The burden on healthcare systems can be reduced with effective screening that allows for prompt and efficient diagnosis of COVID-19. In an effort to aid medical professionals around the world in assessing patients, forecasting techniques have been developed that use a combination of factors to evaluate the likelihood of infection (5). COVID-19 spreads when an infected individual sneezes, coughs, or speaks. Touching an infected surface or object and then touching your lips, nose, or eyes is another means of contracting the virus. Infected people might experience moderate to severe

(Received 01st December 2023; Accepted 11th January 2024; Published 30th January 2024)

respiratory infections after being infected with COVID-19, and the severity of the main symptoms varies substantially. In most cases, the main signs were high fever, cough, throat discomfort, and pain in the muscles (6). Based on the symptoms that have been diagnosed, patients with illness are given medication. Some preventative measures include keeping a safe distance from sick persons, washing hands often, wearing a mask, and avoiding crowded and public places (7).

The global spread of the COVID-19 pandemic has forced health care and scientific groups to begin searching for novel approaches to diagnosis. In response to this issue, many different diagnostic methods have been targeted by various Machine Learning (ML) models (8). Utilizing computed tomography (CT) images, laboratory blood test data, and patient comorbidities, ML has been employed to develop COVID-19 diagnostic and prognostic predictive models (9). The RT-PCR test has been regarded as the gold standard for diagnosing COVID-19. Nevertheless, RT-PCR assays are laborious to perform and might take up to 6 hours to get findings (10). In addition, some examples of false-negative results in RT-PCR testing have been reported due to low viral loads in individuals with early-stage COVID-19 infections (11, 12).

The potential of using blood-test data for detecting COVID-19 has been highlighted in recent clinical studies. In this study, we have used laboratory blood test data for predicting COVID-19 and special importance has been given to data pre-processing as it is the crucial stage in the machine learning (ML) pipeline. Also, the accuracy of predictive modelling is impacted directly by the quality and structure of the data. Using an enhanced K-Nearest Neighbours (KNN) estimation method in combination with a dynamic K selection algorithm for data pre-processing has been proposed in this work. The main objective of this work includes,

- To compare the predictive performance of various machine learning classifiers on the dataset imputed with basic KNN Imputer against the same dataset imputed using proposed enhanced KNN Imputation with dynamic k selection algorithm
- To identify and rank the most influential features in predicting the target variable using SVM-RFE.
- To evaluate the impact of KNN imputation on the predictive accuracy of machine learning models
- To assess and select the best-performing model(s) based on various evaluation metrics, such as accuracy, recall, precision, and F1 score.

The remaining sections of the paper are organized as follows. Section 2 presents relevant research works pertaining to the problem domain. In Section 3, the utilized materials, including the dataset, data pre-processing methods, feature selection technique, and ML algorithms are described. In Section 4, we wrap up by discussing the outcomes of Data pre-processing and the ML prediction. Section 5 ends with the conclusion respectively.

Related works

Cabitza *et al.* (13) have used hematochemical values from routine blood tests for the study and stated that this could be a faster and less expensive alternative. Three different training data sets of hematochemical values from 1,624 patients admitted to San Raphael Hospital (OSR) were used to build machine learning (ML) models. Multivariate k-nearest neighbours algorithm was used with k=5 for imputation. For selecting features, the recursive feature-elimination method was used. Hyper-parameter optimization was used to find the optimal features. For classification, they tested with five different algorithms on the three different datasets. With OSR dataset, the accuracy of random forest model was 88% and with COVID-specific dataset, KNN and SVM achieved 86% accuracy and finally KNN got an accuracy of 86% using the CBC dataset in the standard version.

Bao *et al.* (14) using data from the Chinese hospitals and Wuhan Union Hospital, applied Random Forest and SVM to analyze 294 blood samples. To evaluate this method, they used a dataset consisting of 86 patients with moderate non-COVID-19 viral pneumonia and 208 patients with moderate COVID-19. Experimental results on the fifteen features selected for analysis show that SVM outperforms random forest classifier with an accuracy of 84%. They concluded by saying that both medical and machine learning theories can account for their findings and their method has the potential to add additional rapid COVID-19 testing

option that can be performed in laboratories that are equipped to perform routine blood tests,

Zhao *et al.* made a logistic regression-based classification model to predict two main outcomes: admission to the intensive care unit (ICU) and death. This was another attempt to figure out how comorbidities affect risks in COVID-19 patients. For the testing dataset, the risk score model was accurate with an Area under the curve (AUC) of 0.74 for predicting ICU entry and 0.83 for predicting death. This model was tested using data from the COVID-19 persons under investigation (PUI) registry of 4997 patients from Stony Brook University Hospital in New York. In addition, the authors discovered that cardiopulmonary measures were the most reliable predictors of mortality (15).

AlJame *et al.* (16) proposed a prediction model for COVID-19 that can be used with existing medical data to produce an accurate diagnosis. They have used a new ensemble-based approach known as deep forest (DF) that uses multiple classifiers in multiple layers in order to enhance performance. Layer-by-layer processing is utilized at the cascade level, which is built from three distinct classifiers Extra trees, XGBoost, and Light GBM. Two open-source datasets were used for both training and testing the prediction model. The proposed DF model achieves 99.5% accuracy, 99.96% specificity and 95.28% sensitivity in experiments. They concluded that DF model can serve as a rapid screening tool for COVID-19 patients in areas where testing is scarce

Chadaga *et al.* (17) used standard blood tests and AI to diagnose and predict COVID-19. The dataset had been collected from Patients with COVID-19 who were admitted to Brazil's Israelita Albert Einstein Hospital. Prediction was performed using several different classifiers, including random forest, k nearest neighbours, logistic regression and XG Boost. Oversampling was performed using the SMOTE method to handle the class imbalance issue in the dataset. They have identified that leukocytes, platelets, eosinophils, and monocytes were the most significant indicators and the highest accuracy of 92% was achieved by Random Forest model.

Yao *et al.* (18) conducted research on the detection of severity level in COVID-19 patients by using clinical information in conjunction with data from blood and urine tests. 137 COVID-19 patients from Tongji Hospital in China were included in the dataset as clinically proven cases. The authors used five machine learning models, including Adaboost, LR, SVM, RF and KNN and feature selection was carried out using conservative recursive feature elimination (cRFE) strategy to identify the most important features. SVM achieved an overall accuracy of 81.48 % on the independent test dataset and an accuracy of 99% on the validation dataset by using 28 features making it the most accurate model.

Over 2,670,000 COVID-19 patients from 146 different nations with 307,382 labelled samples were used by Pourhomayoun and Shakibi (19). They proposed an AI-based model for prioritizing patients in need of immediate attention or hospitalization. They have utilized both filter and wrapper feature selection methods in this study to find the most relevant features in the dataset. The findings indicate that the model has an accuracy of 89.98% overall in predicting the mortality rate with the 10-fold cross-validation of the neural network model when compared with the other machine learning models.

Authors in (20) explained the importance of rigorous evaluation methodologies supporting the Enhanced KNN Imputation approach. The decision to use a comprehensive assessment strategy which includes calculating Mean Absolute Error (MAE) and training a ML Classifier is based on the idea that evaluating imputation accuracy and downstream task performance provides a more holistic view of the method's effectiveness. The utilization of the weighted averaging technique in enhanced KNN is based on the assumption that assigning weights to instances that are more similar results in imputations that more accurately depict the fundamental patterns in the data. The incorporation of similarity measurements in the imputation process has been emphasized in (21) and the authors suggested that weighted estimation has significant impact in increasing the efficacy of imputation process (22).

In the study (23), the authors have used Multivariate Imputation by Chained Equation (MICE) technique to mitigate the issue of missing data. The choice between MICE and K-Nearest Neighbours (KNN) imputation for medical data depends on various factors, including the characteristics of the data, the nature of the missingness, and the specific goals of the research

problem. MICE is flexible and can handle a mix of continuous and categorical variables, which is often the case in medical datasets. It can capture complex relationships between variables and provides imputations that preserve observed patterns in the data. However, KNN is a non-parametric method and makes fewer assumptions about the underlying distribution of the data. The underlying principle of KNN is that it imputes missing values based on the values of nearest neighbours, which can be intuitive and effective, especially when missingness is related to similarity between instances. According to our research problem, the missing data mechanism may not always be fully explained by other observed variables. Hence, we have considered using KNN based Imputation method for imputing missing values. To make KNN Imputation more sophisticated and potentially robust imputation technique, we have introduced the combination of dynamic k selection, weighted average, and additional evaluation steps in our methodology. Compared to basic KNN imputation, the proposed enhanced KNN imputation method provides enhancements in adaptability, efficiency, imputation accuracy, and downstream task performance. Another remarkable aspect of our work when compared with the existing study is that, with the enhanced KNN Imputation with Dynamic K Selection Algorithm we have achieved higher accuracy in predicting COVID-19 with the same datasets.

Methodology

In the field of medical research, the presence of missing values creates significant obstacles in generating accurate and insightful conclusions. Our research intends to address these gaps by employing and comparing the efficacy of basic KNN Imputation and proposed method. The main aim of this study is to develop an enhanced KNN Imputation method for data pre-processing. The imputed dataset obtained using basic KNN and the proposed KNN will be used as an input to predict COVID-19 by training various machine learning models. The proposed methodology of the entire of the entire work flow is shown in Figure 1.

Data sources

The data set utilized for this study was obtained from the Zenodo website (22), which was used by (23), and the authors stated that this dataset was made available by IRCCS Ospedale San Raffaele.

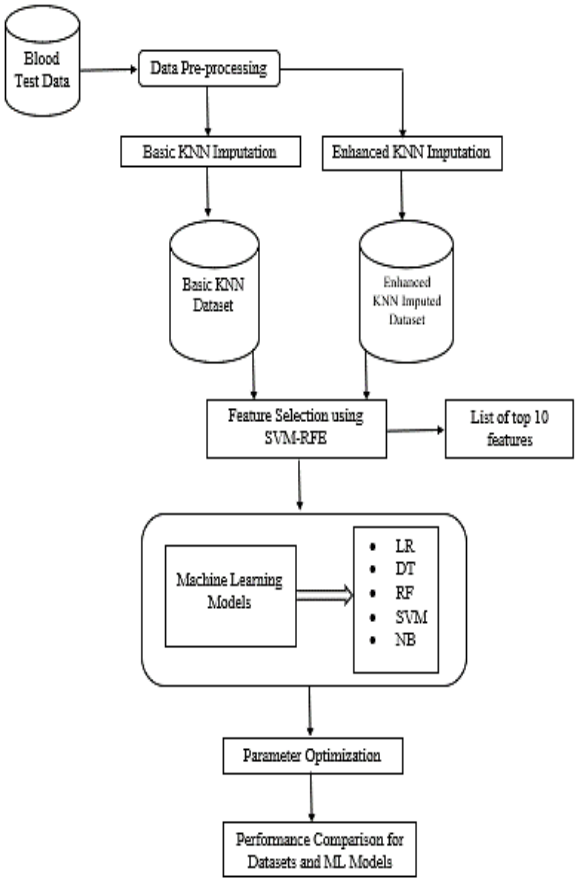

Figure 1: Proposed methodology

The dataset has originally 279 COVID-19 cases and 16 features, including 2 demographic feature (age and gender), 13 blood test indicators, and an RT-PCR result from a nasopharyngeal swab test as the target variable. The dependent variable SWAB has been renamed as “target” and Basophil column has been dropped as it has 0 as the majority value in the dataset. The dependent variable "target" is a binary variable which represents 1 for positive COVID-19 case and 0 for negative COVID-19 case. The list of all features used in the dataset has been shown Table 1.

Data Pre-processing

The occurrence of missing values in datasets may cause substantial difficulties in the fields of data analysis and machine learning. Simple methods like mean, median, and mode imputation are frequently used in conventional approaches (24). This is problematic since they may generate biases, especially if the missingness is not completely random. This issue has been addressed

Table 1: List of all features

S. No	Attribute	Expansion	Missing Values
1	AGE	Age	NIL
2	GENDER	Gender	NIL
3	WBC	White Blood Cells	2 (0.72%)
4	PLT	Platelets	2 (0.72%)
5	NEU	Neutrophils	70 (25.09%)
6	LY	Lymphocytes	70 (25.09%)
7	MO	Monocytes	70 (25.09%)
8	EO	Eosinophils	70 (25.09%)
9	CRP	C-Reactive Protein	6 (2.15%)
10	AST	Aspartate aminotransferase	2 (0.72%)
11	ALT	Alanine aminotransferase	13 (4.66%)
12	ALP	Alkaline phosphatase	148 (53.05%)
13	GGT	Gamma glutamyltransferase	143 (51.25%)
14	LDH	Lactate dehydrogenase	85 (30.47%)
15	target	Covid-19 (Positive/Negative)	NIL

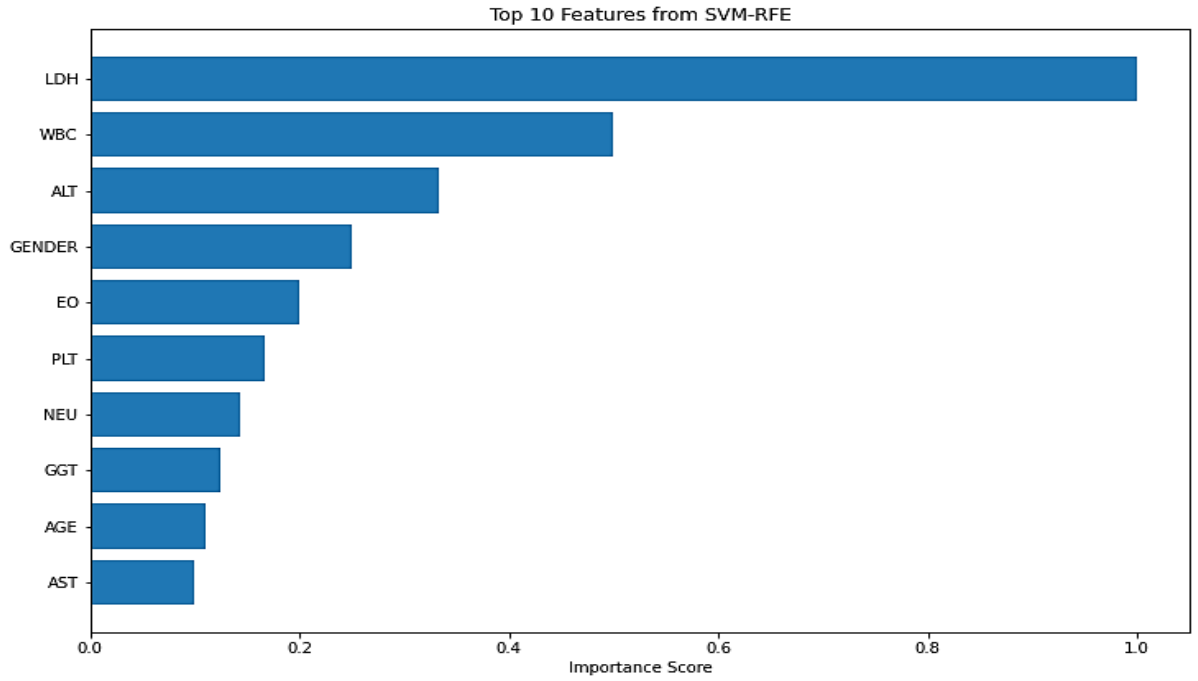

Figure 2: Feature selection using SVM - RFE

in this work and a technique called "Enhanced KNN Imputation with Dynamic K Selection Algorithm" has been proposed. This approach strives to obtain more precise imputation by making use of K-Nearest Neighbours (KNN) and optimizing it with a dynamic selection of the 'K' parameter. Combining the stability of K-Nearest Neighbours (KNN) with a dynamic 'K' optimization, data normalization, and a distance weighting methodology, this method ensures a fine-grained and accurate imputation procedure.

Proposed enhanced KNN imputation with dynamic k selection algorithm

Input

- Data set with missing values (D)
- Test set with known true values (Test_set)

Output

- Data set D′ with imputed values using Enhanced KNN Imputation algorithm

Steps

1. Initialize Data Sets:
   - Partition D into Dm (instances with at least one missing value) and Dc (Complete samples).
   - Standardize Dc and Dm to yield datasets Dcscaled and Dmscaled.
2. Dynamic k Selection for Each Sample x in Dm_{scaled}:
   - Compute distances between x and all samples in using Dcscaled observed features.
   - Determine Koptimal by incrementing k until average distance of top k neighbours is below a threshold or reaches kmax.
3. KNN Imputation:

For each sample x in Dm_{scaled}:

   - Compute weighted average of the variable from the koptimal nearest neighbours to estimate missing values in x.
4. Finalization:
   - Merge imputed Dm with Dc to create D′
5. Evaluation:
   - Calculate Mean Absolute Error (MAE) between imputed values in D' and true values in the Test_set for instances present in both sets.
   - Train a Random Forest Classifier on D′ and evaluate its accuracy.
6. Return D′

The proposed Enhanced KNN Imputation with Dynamic K Selection Algorithm is designed to impute missing values in a dataset (D) using a K-nearest neighbours (KNN) approach. The first step is to partition the original dataset D into two subsets known as Dm (instances with at least one missing value) and Dc (complete samples without missing values). Then standardizing of these two datasets Dc and Dm has been carried out to yield datasets Dcscaled and Dmscaled. Standardization involves scaling the features to have a mean of 0 and a standard deviation of 1. In order to select Dynamic k for each sample x in Dmscaled, distances between x and all samples in Dcscaled using observed features has been computed. Then the optimal value of k (Koptimal) is determined by incrementing k until the average distance of the top k neighbours is below a certain threshold or reaches a maximum value (kmax). This step dynamically adapts the value of k for each sample. For performing KNN Imputation, for each sample x in Dmscaled, the weighted average of the variable from the Koptimal nearest neighbours is computed to estimate missing values in x. This involves assigning weights to the neighbours based on their distance to the target sample. Finally, the imputed dataset Dm is imputed with Dc to create a complete dataset D′ without any missing values.

Calculation of mean absolute error (MAE):

Mean Absolute Error (MAE) is a metric used to assess the accuracy of the imputation process. In regression problems, MAE is often used to compute the average absolute differences between the predicted and observed values. In this instance:

Observed values: These values refer to the actual values found in the Test_set, which is a subset of the original dataset (D) for which the exact values for specific instances are known.

Predicted values: The predicted values are the values that were filled in with values from the imputed dataset (D), which was made with our imputation technique.

The MAE is calculated using the formula:

$$\mathbf{MAE = (1/n)\ \Sigma\ (i=1\ to\ n)\ |y_i - \hat{y}_i|}$$

where n is the number of instances, y_i is the true value, and $\hat{y}_i$ is the imputed/predicted value.

Comparison is done between the imputed values in D' to the true values in the Test_set. Specifically, is calculated as this step assesses how well the imputed values align with the known true values.
To Train a Random Forest Classifier, imputed dataset D' has been used which includes the imputed values and known true values from the Test_set. This step aims to evaluate how well the imputed values contribute to the accuracy of a downstream task (classification). It assesses the impact of imputation on the performance of a machine learning model. This algorithm is designed for not only imputing missing values in the original dataset but also evaluate the imputation performance using both the MAE and the accuracy of a Random Forest Classifier trained on the imputed dataset. The inclusion of a separate test set (Test_set) with known true values enhances the robustness of the evaluation. The algorithm measures the average absolute difference between the true and imputed values. A lower MAE indicates more accurate imputation. This metric gives a simple and interpretable measure of how closely imputed values match actual values, making it an appropriate choice for evaluating the performance of imputation algorithms.

Feature selection

In high-dimensional datasets, where the number of features can be substantial, prioritizing the most informative features becomes essential. This can improve the model's performance by removing unnecessary or redundant features, and also improves the model's interpretability. For our analysis, we utilized the Support Vector Machine Recursive Feature Elimination (SVM-RFE) technique, a well-known method for ranking features for classification tasks. SVM-RFE operates by recursively fitting a linear SVM to the data set and ranking features based on the SVM's weight magnitude (25, 26). Upon applying SVM-RFE to our dataset, we identified a subset of features that were most indicative of covid-19 prediction and out of all, the top 3 features were LDH, WBC, ALT which is shown in Figure 2.
LDH is an enzyme present in numerous body tissues, such as the liver, heart, and lungs. Increased LDH levels may indicate tissue damage. Multiple studies have demonstrated that patients with severe COVID-19 infection have

MAE for instances present in both D' and Test_set

elevated LDH levels. Having a high white blood cell count may indicate an acute infection or inflammation, whereas a decrease in white blood cell count (leukopenia) may indicate an impaired immune system or certain viral infections. Some COVID-19 patients have been observed to have leukopenia, specifically a reduction in lymphocytes (a subset of WBCs), particularly in the early stages of the disease. ALT is an enzyme that resides primarily in the liver. Increased ALT levels can indicate liver damage. Some COVID-19 patients have elevated ALT levels, indicating possible liver damage. This may be caused by the direct viral effect, systemic inflammation, drug-induced liver damage, or a combination of these factors. Patients receiving treatments that may be harmful to the liver should have their ALT and other liver enzymes monitored closely.

Descriptive analytics

Comprehensive descriptive analytics will include the following components.

- Summary Statistics
- Distribution of features
- Target Distribution
- Correlation Matrix

Summary statistics

Each of the 279 samples in the dataset has 15 attributes including the target variable associated with it. The count verifies the number of non-missing data points. The average number is given by the mean, which is a measure of the central tendency. The dispersion or variance in the feature values is represented by the standard deviation. The range of the data for a certain feature is shown by the range between the minimum and maximum numbers. In order to quickly grasp the breadth and probable skewness of the data, the 25th, 50th (median), and 75th percentiles can be used as a rapid reference. For example, the “AGE" feature has an average value of 61.34 years and a standard deviation of 18.49 years, which suggests that the age range in the dataset is fairly large. Figure 3 shown below provides a quick overview of the dataset by emphasizing central tendencies, standard deviations, and potential outliers.

Distribution of features

Histograms are useful for describing distributions by highlighting patterns and outliers and elucidating the shape of the distribution as a whole. Features like WBC, EO, CRP, AST, ALT, ALP, GGT, and LDH exhibit a right-skewed distribution, indicating that most of the data points are clustered on the left side, with a tail stretching to the right. AGE, although with a peak around 60-70, appears relatively uniform compared to other features. Some features, like WBC and CRP have potential outliers with values significantly distant from the central cluster. For visualizing the features using histograms all the 13 features except the target variable and it is depicted in Figure 4.

	count	mean	std	min	25%	50%	75%	max
GENDER	279.0	0.326165	0.469651	0.0	0.000000	0.000000	1.000000	1.0
AGE	279.0	61.336918	18.491523	0.0	49.000000	64.000000	76.000000	98.0
WBC	279.0	8.550179	4.838849	1.1	5.100000	7.100000	10.700000	29.2
PLT	279.0	226.252330	100.866326	20.0	163.500000	204.000000	271.000000	620.0
NEU	279.0	6.303226	3.940475	0.5	3.800000	5.200000	7.450000	26.4
LY	279.0	1.230645	0.760323	0.2	0.700000	1.083333	1.500000	7.2
MO	279.0	0.633035	0.386080	0.0	0.400000	0.583333	0.800000	3.2
EO	279.0	0.062425	0.120860	0.0	0.000000	0.000000	0.100000	1.3
CRP	279.0	90.620669	93.479409	0.1	21.800000	54.500000	128.600000	478.0
AST	279.0	54.081840	57.425851	11.0	27.000000	36.000000	60.000000	550.0
ALT	279.0	45.335723	44.841109	9.0	22.000000	32.000000	47.000000	335.0
ALP	279.0	82.665472	63.241052	34.0	60.166667	72.000000	85.916667	838.0
GGT	279.0	74.798088	99.127133	10.0	28.166667	48.000000	83.083333	839.0
LDH	279.0	373.146953	173.800344	98.0	253.000000	328.000000	437.000000	1195.0
target	279.0	0.634409	0.482461	0.0	0.000000	1.000000	1.000000	1.0

Figure 3: Summary Statistics of the dataset

Target distribution

The distribution of the target variable in the dataset is shown in Figure 5 which indicates the number of positive and negative COVID -19 cases.

Correlation matrix

The correlation matrix is shown as a heatmap in Figure 6, which shows the pairwise correlation coefficients between the features and the target variable. There is a strong positive correlation between the WBC count and Neutrophil count. This is expected since neutrophils are a type of white blood cell. These two liver enzymes AST and ALT also show a strong positive correlation. Features like LDH, WBC, CRP, and NEU have relatively higher positive correlations with the target variable. It suggests that as these feature values increase, the likelihood of the target being 1 also increases. Conversely, PLT (Platelet count) has a negative correlation with the target, suggesting that higher platelet counts might be associated with a target value of 0.

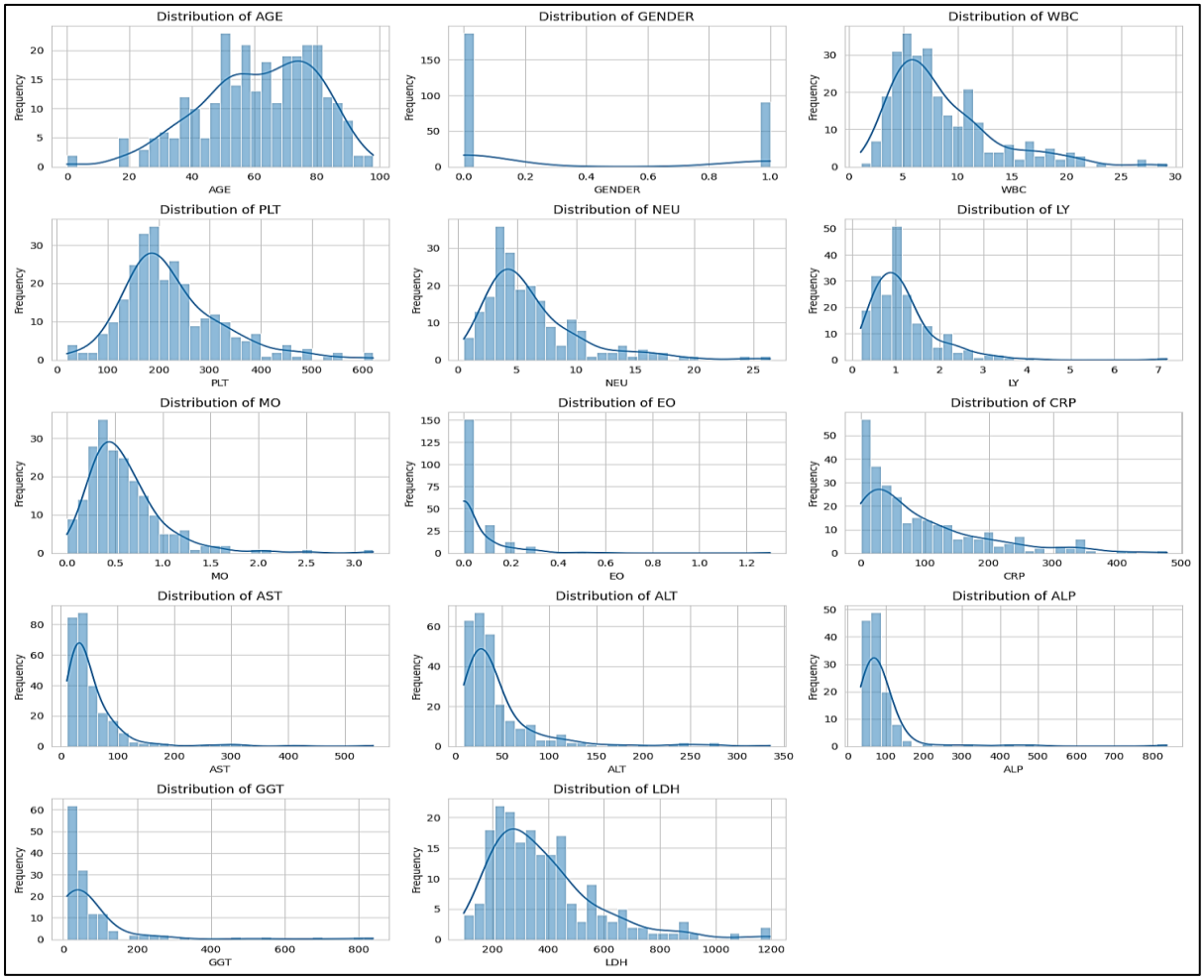

Figure 4: Distribution of features using histogram

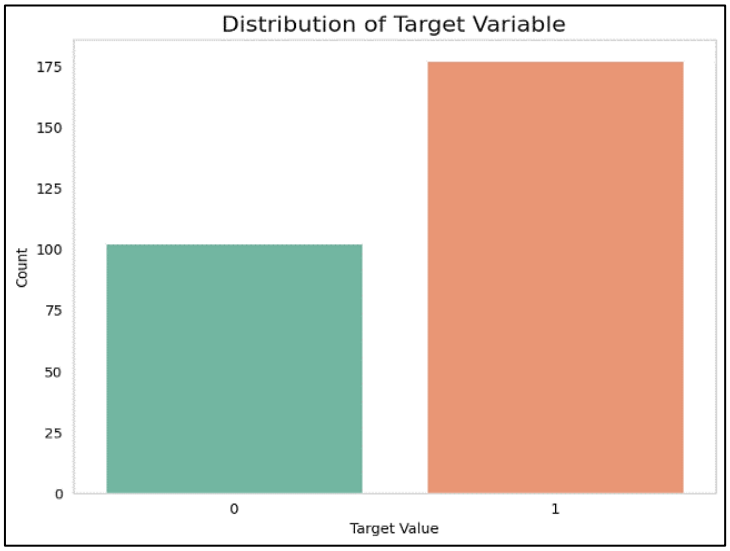

Figure 5: Distribution of target variable

Machine learning algorithms

After applying feature selection technique using SVM-RFE, we used various machine learning algorithms to build a predictive model. We compared various classification techniques, specifically we evaluated the following classifier models:

- Logistic Regression (LR)
- Decision Tree (DT)
- Random Forest (RF)
- Support Vector Machine (SVM)
- Naïve Bayes (NB)

The selection of these classifiers was based on their prevalence in the field of machine learning and the diversity of their underlying algorithms, providing a comprehensive view of the influence of imputation techniques. Each classifier has its own set of hyperparameters that can significantly impact its performance. We used Grid Search CV to ensure that each model's predictive ability was maximized. This utility uses cross-validation to systematically explore all possible combinations of hyperparameter values and select the one with the best validation performance. For each classifier, two models were trained. One containing the dataset pre-processed using Basic KNN Imputation. Another dataset using the Enhanced KNN Imputation. After tuning hyperparameters and training the model, predictions were made on the respective test sets.

Results and Discussions

Results of ynamic KNN imputation

We used dynamic K-nearest neighbours (KNN) estimation to fill in missing data in our dataset. We experimented with several values of k to determine the optimal value for neighbours and imputation precision. Through our experimentation, the best k value for KNN imputation was determined to be 2. This value yielded the lowest mean absolute error (MAE) in our imputation results. The different metrics observed based on the proposed imputation have been shown in Table 2.

Table 2: Performance metrics of proposed KNN imputation method

S.NO	Measures	Result
1	Best k value for KNN Imputation	2
2	Mean Absolute Error	4.0522
3	Accuracy on Basic Imputed Data	66.07 %
4	Accuracy on Imputed Data with best k (2)	82.14 %

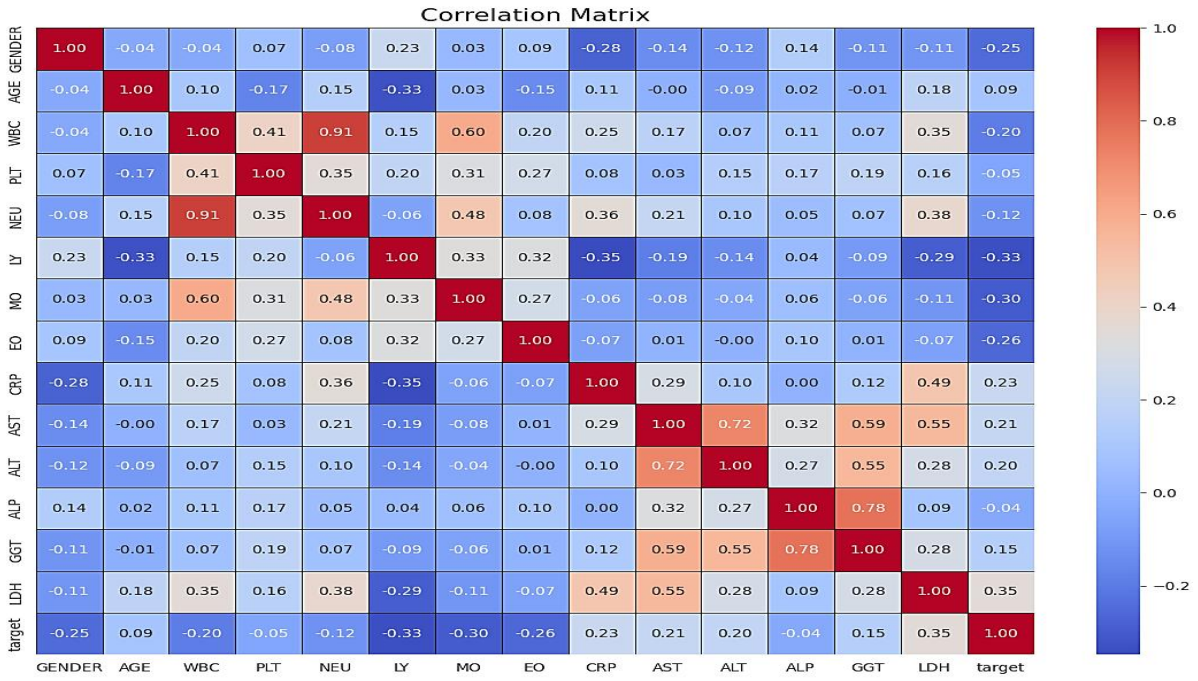

Figure 6: Correlation matrix

Performance metrics

Mean absolute error (MAE)

For the KNN imputation with k=2, the Mean Absolute Error was 4.0522. This measure shows the average size of the differences between the imputed values and the real values. Lower values indicate a more accurate imputation.

Accuracy on basic imputed data

Before we optimized the size of k, our model (RF) was 66.07% accurate when the missing values are inputted using basic KNN Imputation technique. This provides a starting point from which results can be compared with the proposed method.

Accuracy on optimally imputed data

After using KNN imputation with the best k number of 2, our model's (RF) accuracy has been increased to 82.14%. This highlights the significance of selecting an appropriate k value for imputation, which can have a major effect on the final model's efficacy.

These models have been trained and tested using the same dataset that was pre-processed in 2 ways. First, the dataset was pre-processed using basic KNN Imputation technique and the various evaluation metrics such as accuracy, precision, recall, F1 score has been computed.

The same dataset was pre-processed using proposed enhanced KNN Imputation with dynamic k algorithm and the performance metrics has been recorded. The comparison has been carried out using the accuracy score, and it has been proved that dataset imputed with the proposed KNN Imputation method performed better in terms of all evaluation metrics and it is shown in Table 3 and 4 and Figure 7.

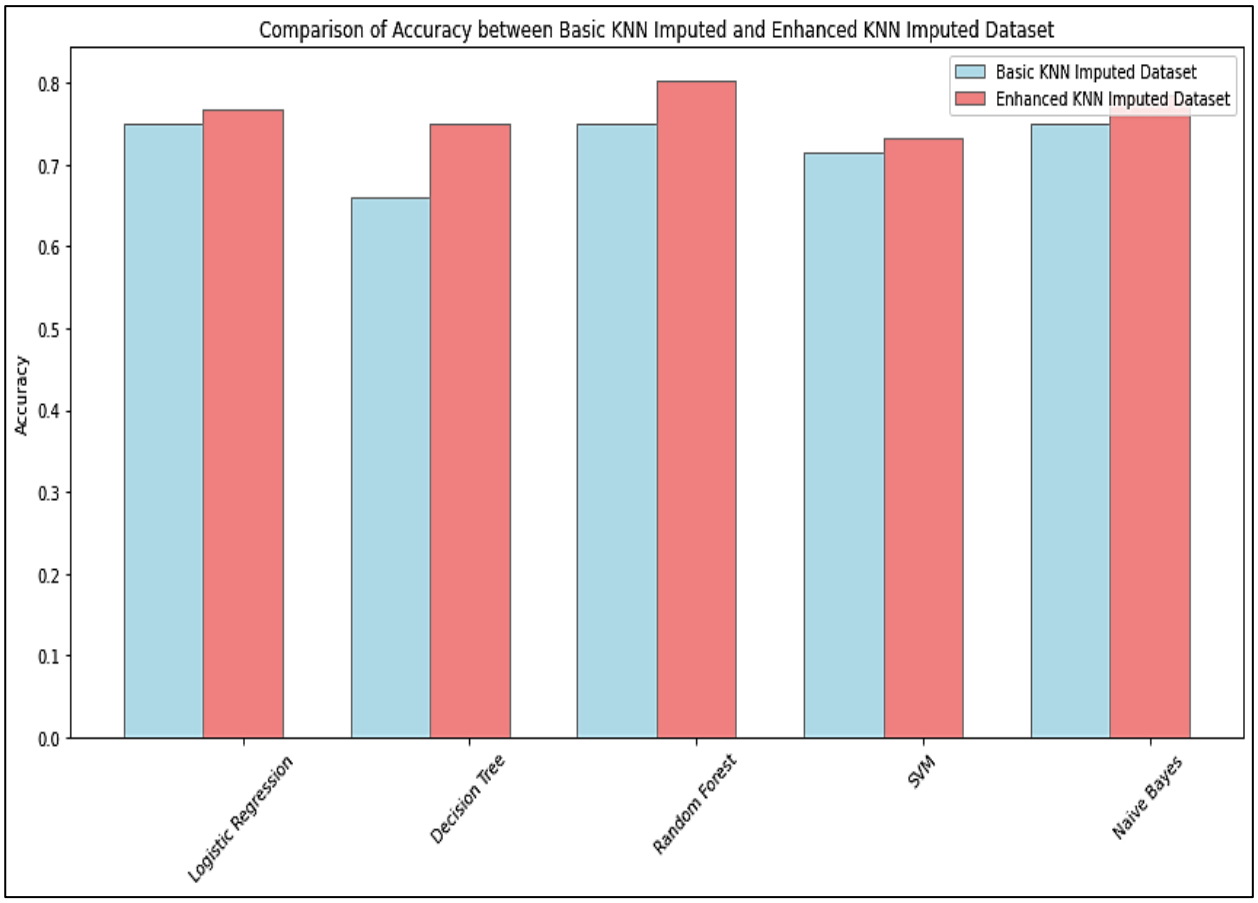

Figure 7: Comparison of accuracy between basic and enhanced KNN imputed dataset (With gender attribute)

Table 3: Performance metrics of ML models using basic KNN imputation dataset (With gender)

S. No	Algorithms	Accuracy	Precision	Recall	F1-Score
1	Logistic Regression	0.75	0.74	0.91	0.82
2	Decision Tree	0.66	0.74	0.71	0.72
3	Random Forest	0.75	0.77	0.86	0.81
4	SVM	0.71	0.72	0.89	0.79
5	Naïve Bayes	0.75	0.78	0.83	0.81

Table 4: Performance metrics of ML models using enhanced KNN imputation dataset (With gender)

S. No	Algorithms	Accuracy	Precision	Recall	F1-Score
1	Logistic Regression	0.77	0.75	0.94	0.84
2	Decision Tree	0.75	0.84	0.74	0.79
3	Random Forest	0.80	0.83	0.86	0.85
4	SVM	0.73	0.73	0.91	0.81
5	Naïve Bayes	0.79	0.81	0.86	0.83

From Tables 3 and 4, the accuracy of various machine learning algorithms computed using dataset that is imputed with Enhanced KNN Imputation is higher. Among all the prediction models considered for the study, the accuracy of RF is higher (80%) when the dataset of proposed method is used. Also, in case of pre-processing using basic KNN imputation, LR, RF and NB achieved equal accuracy rate of 75%.

This study has been carried out in another way in order to find whether GENDER feature has any connection with either increasing or decreasing accuracy of the prediction model. So, in order to find that we have dropped this feature and performed the pre-processing using both methods followed by feature selection and prediction. Based on the experimental results, we found that dropping the GENDER column has impact on predicting the target variable in terms of all performance evaluation metrics and it is depicted in Table 5 and 6 and Figure 8.

From Tables 5 and 6, the accuracy of various machine learning algorithms computed using dataset that is imputed with Enhanced KNN Imputation is higher. Among all the prediction models considered for the study, the accuracy of SVM is higher (84%) when the dataset of proposed method is used. This is an attempt to ensure whether the GENDER feature has any impact on affecting the accuracy of the prediction model. Random Forest model trained using Enhanced KNN Imputation dataset achieved an equal accuracy rate of 80% in both the cases as specified in Table 4 and 6. Without the GENDER feature on Enhanced KNN Imputation dataset, the accuracy of SVM model has been increased to 84 % from 73% which has been identified from Table 4 and 6.

Table 5: Performance metrics of ML models using basic KNN imputation dataset (Without gender)

S. No	Algorithms	Accuracy	Precision	Recall	F1-Score
1	Logistic Regression	0.66	0.69	0.83	0.75
2	Decision Tree	0.62	0.71	0.69	0.70
3	Random Forest	0.71	0.74	0.83	0.78
4	SVM	0.71	0.71	0.91	0.80
5	Naïve Bayes	0.73	0.75	0.86	0.80

Table 6: Performance metrics of ML models using enhanced KNN imputation dataset (Without gender)

S. No	Algorithms	Accuracy	Precision	Recall	F1-Score
1	Logistic Regression	0.73	0.74	0.89	0.81
2	Decision Tree	0.75	0.74	0.91	0.82
3	Random Forest	0.80	0.83	0.86	0.85
4	SVM	0.84	0.84	0.91	0.88
5	Naïve Bayes	0.77	0.79	0.86	0.82

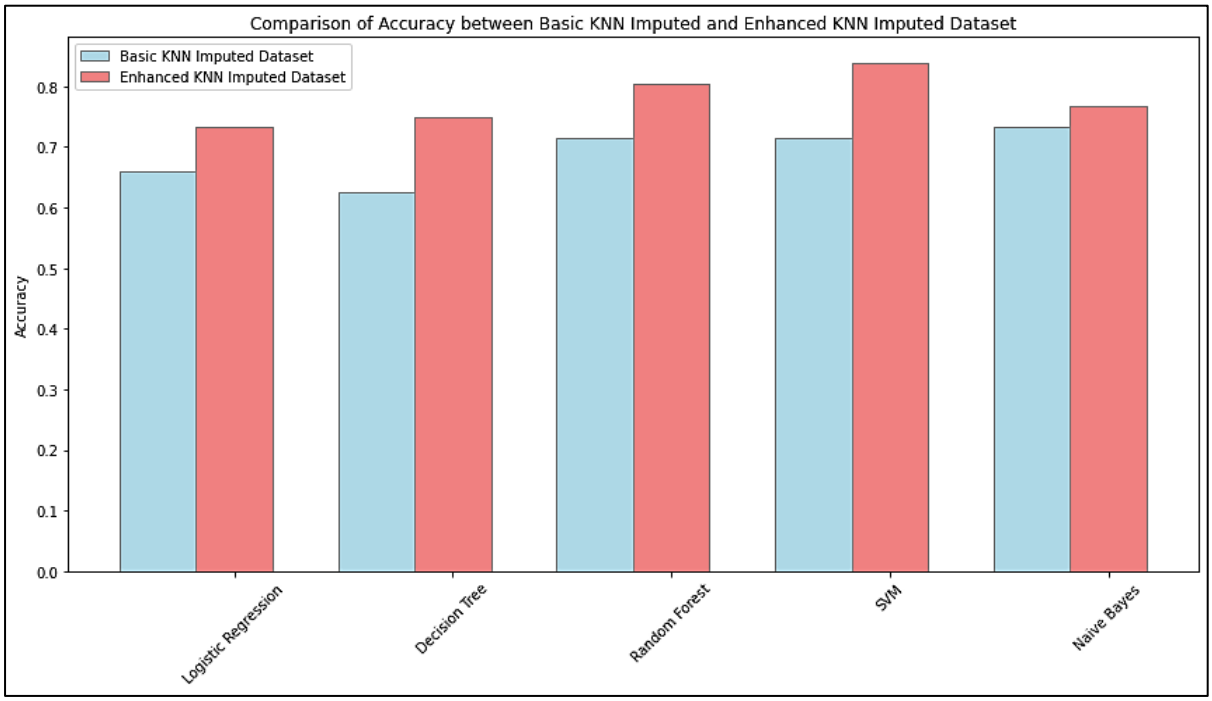

Figure 8: Comparison of accuracy between basic and enhanced KNN imputed dataset (Without gender)

The performance and benefits of algorithms are often evaluated empirically using real-world or benchmark datasets. Here are some potential benefits or advancements that the algorithm could provide.

Dynamic K selection

The dynamic selection of k for each sample based on the average distance to neighbours allows the algorithm to adapt to local patterns in the data. This is useful when working with datasets where the density of neighbours differs between instances.

Imputation using weighted KNN

The algorithm calculates the weighted average of the variable based on the $k_{optimal}$ nearest neighbours. Weighting by distance helps to give more weight to nearby neighbours, potentially capturing more accurate local patterns.

Downstream task integration (Random forest classifier)

Training and testing a Random Forest Classifier on the imputed dataset provides insight into the effect of imputation on downstream tasks. This comprehensive evaluation determines whether the imputed values improve the performance of a machine learning model.

Separate test set for evaluation

The inclusion of a separate test set (Test set) with known true values improves the evaluation's robustness. By using the same data for both training and evaluation, this approach assures that the assessment is not biased, resulting in a more realistic estimate of imputation performance.

Also, several restrictions and difficulties were observed during the execution of our proposed enhanced KNN imputation technique for data pre-processing in machine learning-based COVID-19 prediction. It is crucial to deal with these issues to make sure that our results were accurate and reliable. While applying the proposed methodology on the dataset, the challenge we have faced was that, since Dm contains instances with missing values it might not be suitable to directly calculate the MAE between Dm and the imputed dataset D′, as the true values for the missing entries in Dm are not available. Hence, we have created a separate test set with known true values. This set is the subset of the dataset (perhaps by removing some values and treating them as a validation set). Then, we have applied our proposed Enhanced KNN Imputation technique to impute the missing values and compared the imputed values with the true values in the test set and calculated MAE.

Conclusion

The detection of COVID-19 in a timely and accurate manner is crucial for managing its spread and impact. This study investigated the viability of blood test data as a predictive tool for COVID-19 detection, with a focus on data pre-processing using Enhanced KNN Imputation. Through rigorous testing, we discovered that the Enhanced KNN Imputation method significantly enhanced the performance of machine learning classifiers in comparison to conventional imputation techniques. Based on the experimental results, Random Forest model outperformed other classifiers using dataset imputed with Enhanced KNN method with an accuracy of 80% with all the features. The same methodology has been carried out with the exclusion of GENDER feature and as a result SVM model achieved an accuracy of 84%.

In addition, the results obtained by combining Enhanced KNN Imputation with classifiers based on machine learning demonstrate the significance of robust data pre-processing in predictive modelling. In the context of the pandemic, where rapid and accurate diagnostic tools are imperative, our methodology offers a promising approach. It suggests that blood test data can be a valuable diagnostic tool against COVID-19 if properly pre-processed. As we progress, integrating diagnostic tools based on machine learning with conventional methods can pave the way for a more responsive and effective healthcare system.

Abbreviation

Nil

Acknowledgement

We would like to thank Sree Saraswathi Thyagaraja College, Pollachi for encouraging us in doing research pertaining to real time scenario which created an interest in this topic.

Author's contributions

All authors are equally contributed.

Conflict of interest

The authors declare that they have no conflicts of interest to report regarding the present study.

Ethics approval

Not applicable

Funding

No

References

1. Sultana J, Singha AK, Siddiqui ST, Nagalaxmi G, Sriram AK, Pathak N. COVID-19 Pandemic Prediction and Forecasting Using Machine Learning Classifiers. Intelligent Automation & Soft Computing. 2022 May 1;32(2).
2. Malki Z, Atlam ES, Ewis A, Dagnew G, Ghoneim OA, Mohamed AA, Abdel-Daim MM, Gad I. The COVID-19 pandemic: prediction study based on machine learning models. Environmental science and pollution research. 2021 Aug;28:40496-506.
3. Saadatmand S, Salimifard K, Mohammadi R, Kuiper A, Marzban M, Farhadi A. Using machine learning in prediction of ICU admission, mortality, and length of stay in the early stage of admission of COVID-19 patients. Annals of Operations Research. 2023 Sep;328(1):1043-71.
4. Huang J, Zhang L, Liu X, Wei Y, Liu C, Lian X, Huang Z, Chou J, Liu X, Li X, Yang K. Global prediction system for COVID-19 pandemic. Science bulletin. 2020 Nov 11;65(22):1884.
5. Zoabi Y, Deri-Rozov S, Shomron N. Machine learning-based prediction of COVID-19 diagnosis based on symptoms. npj digital medicine. 2021 Jan 4;4(1):3.
6. Painuli D, Mishra D, Bhardwaj S, Aggarwal M. Forecast and prediction of COVID-19 using machine learning. InData Science for COVID-19 2021 Jan 1 (pp. 381-397). Academic Press.
7. Gothai E, Thamilselvan R, Rajalaxmi RR, Sadana RM, Ragavi A, Sakthivel R. Prediction of COVID-19 growth and trend using machine learning approach. Materials Today: Proceedings. 2023 Jan 1;81:597-601.
8. Kwekha-Rashid AS, Abduljabbar HN, Alhayani B. Coronavirus disease (COVID-19) cases analysis using machine-learning applications. Applied Nanoscience. 2023 Mar;13(3):2013-25.
9. Ustebay S, Sarmis A, Kaya GK, Sujan M. A comparison of machine learning algorithms in predicting COVID-19 prognostics. Internal and Emergency Medicine. 2023 Jan;18(1):229-39.
10. Tahamtan A, Ardebili A. Real-time RT-PCR in COVID-19 detection: issues affecting the results. Expert review of molecular diagnostics. 2020 May 3;20(5):453-4.
11. Sun NN, Yang Y, Tang LL, Dai YN, Gao HN, Pan HY, Ju B. A prediction model based on machine learning for diagnosing the early COVID-19 patients. MedRxiv. 2020 Jun 4:2020-06.
12. Ferrari D, Motta A, Strollo M, Banfi G, Locatelli M. Routine blood tests as a potential diagnostic tool for COVID-19. Clinical chemistry and laboratory medicine (CCLM). 2020 Jun 25;58(7):1095-9.
13. Cabitza F, Campagner A, Ferrari D, Di Resta C, Ceriotti D, Sabetta E, Colombini A, De Vecchi E, Banfi G, Locatelli M, Carobene A. Development, evaluation, and validation of machine learning models for COVID-19 detection based on routine blood tests. Clinical Chemistry and Laboratory Medicine (CCLM). 2021 Feb 23;59(2):421-31.
14. Bao FS, He Y, Liu J, Chen Y, Li Q, Zhang CR, Han L, Zhu B, Ge Y, Chen S, Xu M. Triaging moderate COVID-19 and other viral pneumonias from routine blood tests. arXiv preprint arXiv:2005.06546. 2020 May 13.
15. Zhao Z, Chen A, Hou W, Graham JM, Li H, Richman PS, Thode HC, Singer AJ, Duong TQ. Prediction model and risk scores of ICU admission and mortality in COVID-19. PloS one. 2020 Jul 30;15(7):e0236618.
16. AlJame M, Imtiaz A, Ahmad I, Mohammed A. Deep forest model for diagnosing COVID-19 from routine blood tests. Scientific reports. 2021 Aug 17;11(1):16682.
17. Chadaga K, Prabhu S, Vivekananda Bhat K, Umakanth S, Sampathila N. Medical diagnosis of COVID-19 using blood tests and machine learning. InJournal of Physics: Conference Series 2022 Jan 1 (Vol. 2161, No. 1, p. 012017). IOP Publishing.
18. Yao H, Zhang N, Zhang R, Duan M, Xie T, Pan J, Peng E, Huang J, Zhang Y, Xu X, Xu H. Severity detection for the coronavirus disease 2019 (COVID-19) patients using a machine learning model based on the blood and urine tests. Frontiers in cell and developmental biology. 2020:683.
19. Pourhomayoun M, Shakibi M. Predicting mortality risk in patients with COVID-19 using machine learning to help medical decision-making. Smart health. 2021 Apr 1;20:100178.
20. Kumar R, Kumar P, Kumar Y. Time series data prediction using IoT and machine learning technique. Procedia computer science. 2020 Jan 1;167:373-81.
21. Karrar AE. The Effect of Using Data Pre-Processing by Imputations in Handling Missing Values. Indonesian Journal of Electrical Engineering and Informatics (IJEEI). 2022 Apr 30;10(2):375-84.
22. https://zenodo.org/record/3886927
23. Brinati D, Campagner A, Ferrari D, Locatelli M, Banfi G, Cabitza F. Detection of COVID-19 infection from routine blood exams with machine learning: a feasibility study. Journal of medical systems. 2020 Aug;44:1-2.
24. Salem DA, Hashim EM. Impact of data pre-processing on covid-19 diagnosis using machine learning algorithms. International Journal of Intelligent Systems and Applications in Engineering. 2023 Jan 16;11(1s):164-71.
25. Albashish D, Hammouri AI, Braik M, Atwan J, Sahran S. Binary biogeography-based optimization based SVM-RFE for feature selection. Applied Soft Computing. 2021 Mar 1;101:107026.
26. El Kafrawy P, Fathi H, Qaraad M, Kelany AK, Chen X. An efficient SVM-based feature selection model for cancer classification using high-dimensional microarray data. IEEE Access. 2021 Oct 26;9:155353-69.