Journal of Engineering Science and Technology
Vol. 19, No. 2 (2024) 598 - 617

ENHANCING PREDICTIVE PERFORMANCE IN COVID-19 HEALTHCARE DATASETS: A CASE STUDY BASED ON HYPER ADASYN OVER-SAMPLING AND GENETIC FEATURE SELECTION

HAYDER MOHAMMEDQASIM[1,*], ABDULRAHMAN AHMED JASIM[2], ROA'A MOHAMMEDQASEM[3], OGUZ ATA[4]

[1,3]Faculty of Engineering, Istanbul Aydin University, Istanbul, Turkey
[2]Dept. of Electrical and Computer Engineering, Altinbas University, Istanbul, Turkey
[2]Collage of Engineering, Al-Iraqia University, Baghdad, Iraq
[4]Collage of Engineering, Altinbas University, Istanbul, Turkey
*Corresponding Author: hmohammedqasim@aydin.edu.tr

Abstract

Predictive analytics is paramount in the health industry, where it finds its wide application, in that it helps increase the forecast's accuracy level based on big data. Most of the time, there is a tendency toward the imbalance of the datasets in healthcare. In this study, two COVID-19 datasets from Kaggle were used as a case study of dataset imbalance. In such scenarios of imbalanced datasets like COVID-19, conventional sampling methods like ADASYN (Adaptive Synthetic Sampling Approach for Imbalanced Learning) tend to yield only modest accuracy levels. To address another problem like finding the optimal features, this study proposes a novel approach that combines oversampling techniques with genetic feature selection (GFs) using laboratory data. This innovative method aims to construct machine-learning clinical prediction models for the identification of COVID-19 infected patients, leveraging two widely recognized datasets by using hyper ADASYN over-sampling and genetic feature selection, stands out for its unprecedented precision in identifying relevant features crucial for accurate predictions. Unlike the traditional approach, it can solve the class imbalance problem and tune the feature space to bring about a dramatic increase in accuracy, precision, recall, and overall predictive performance by using our hypermodel. Our approach significantly enhanced the performance of the classifier, and the Random Forest (RF) model with "n" trees classifies accurately to the limit of 99%, with precision 99%, recall 99%, and F1-score 99% for each of the datasets. Decision Tree (DT) model achieved 92% with all metrics for Dataset I, and 95% with all metrics for Dataset II. Multilayer Perceptron (MLP) achieved 99% with all metrics, respectively, for both datasets. Gradient Boosting (XGB) achieved 97% for all metrics with dataset I and 98% with all metrics for dataset II. These results underscore the efficacy of our proposed method in balancing COVID-19 datasets and enhancing predictive accuracy.

Keywords: ADASYN, Genetic feature selection, Healthcare, Imbalanced datasets, Oversampling, Predictive analytics.

1. Introduction

1.1. Challenges in healthcare data management

Despite these continuous initiatives, a significant obstacle still exists in the absence of standard medical records that completely capture the complex relationships between people and sickness. The National Prevention Initiative (NPI) focuses on enhancing healthcare outcomes through effective data management and disease prevention strategies [1]. It highlights the most relevant problems in healthcare data, which has to be accurate, balanced, and reliable for informed decision-making and accurate prediction of diseases. These problems are felt most acutely when handling such a complex and imbalanced dataset as COVID-19.

This paper proposes that the proposed research directly meets the above challenge by proposing a balancing approach for the COVID-19 dataset. It proposes a balancing methodology in the dataset and hence contributes to NPI goals. The proposed study, therefore, would bring improved accuracy in predicting and preventing diseases since the utility of the data in line with the objectives of NPI would fill an important gap, as things stand now, in the practice of managing healthcare data. This is very crucial in enhancing health strategies, especially with the current COVID-19 pandemic still existing.

As a subset of AI applications, machine learning (ML) algorithms have the unusual ability to improve performance via experience without the need for explicit programming [2]. The following examples are used to aid with comprehension. Let's say the goal is to develop an AI-powered medical diagnostic tool that can forecast a patient's risk of developing a certain kind of cancer. The computer receives training data comprised of a diverse range of samples to do this. The doctor records any possible symptoms connected to the target cancer type as the first stage in training the computer [3].

Subsequently, data on these symptoms are extracted from each medical report during the training data preparation phase (i.e., the feature retrieval process). A sophisticated machine learning method is then employed to train the system using the amassed dataset. This algorithm incorporates information regarding recorded conditions extracted from each medical record, along with the associated outcomes (i.e., whether the patient received a cancer diagnosis or not Another issue in machine learning is feature selection, which is difficult since the set of derived functions has a big influence on the algorithm's output [4]. This issue is crucial to correctly collecting the key elements of the input data.

1.2. Machine learning in healthcare

The ML technique quickly builds the connection between the retrieved characteristics and the required output after the best features have been found. When the extracted characteristics, like in the previous example, can be manually identified and the expert system can choose the function, this machine learning technique is helpful (i.e., the list of diseases). The most common issue with most medical data sets is imbalanced data. Several open-source methods and packages [5] have been created to aid analysts in training and testing classifiers for this disruption of binary naming entries, numerous classes, and various nomenclatures. Among the training strategies that integrate data imbalance, some utilize the data balance method for under-testing or monitoring, others alter the loss function to

learn the algorithm to overcome imbalances, and yet others employ math sets. However, when evaluating the efficacy of classifiers, the problem is to develop measures with low precision that are easy to compute and comprehend; for example, some of the indicators discovered might misrepresent findings. Others, for example, fail to offer comprehensive performance reports, such as the resolution of accuracy or macro calls, when utilized separately.

Feature selection is another problem in machine learning as it better reflects the most relevant and applicable properties of the input data, which is critical, although the set of derived functions has a significant impact on the algorithm output [6]. Numerous studies examining the impact of feature selection on classification algorithms have shown that feature extraction not only reduces the dimensionality of the data but also improves the model's quality, Feature extraction in machine learning may be categorized as a filtering, wrapper, or metaheuristic [7].

1.3. Metaheuristic algorithms for feature selection in imbalanced healthcare datasets

In this work, a genetic feature selection algorithm is one of the metaheuristic algorithms; metaheuristic algorithms have taken great prominence in all domains, including engineering, biomedicine, and economics, as powerful tools for solving complex real-world problems [8]. Diversity and intensification are the two most important ingredients of these algorithms. Both must be controlled at an adequate level for tackling real-world problems. Most of these algorithms find motivation in biological evolution and swarming behavior.

Generally, these algorithms can be divided into two categories: the meta-population-based and the single-solution algorithms. In single-solution metaheuristic algorithms, the journey usually starts with a candidate solution, and therefore, local searching is done. However, there's a potential drawback wherein a single-solution approach, when employing a metaheuristic, may become ensnared in a local optimum [9]. While searching, population-based metaheuristics consider a wide range of potential answers.

These metaheuristics actively fight against population homogeneity and keep solutions from being stuck in local optima. Population-based metaheuristic algorithms are notable for genetic algorithms (GA), ant-colony optimization (ACO), particle swarm optimization (PSO) [10], etc. In this study, imbalanced datasets are considered to be one of the most critical challenges in health, viewed from the perspective of COVID-19 data. Such an imbalance, by its very nature of disproportionate representation of cases in the prediction datasets, has been shown to impair precision and predictive model reliability. The current work will detail an attempt to develop a method that can be effective for balancing, thereby increasing the precision and usefulness of the machine learning tool within the domain of healthcare analytics.

2. Related works

In order to categorize COVID-19 patients more effectively and gain trustworthy information about the disease's causes, researchers used a variety of algorithms and techniques. Predicting clinical outcomes is challenging because of a variety of confounding variables, including uneven data, irrelevant attributes, and training

duration, all of which can impair patient categorization. Healthcare organizations might be able to define and explain medical outcomes more precisely by addressing these problems.

Mondal et al. [11] developed a brand-new COVID-19 prediction model with a high feature rating to streamline the Hospital Albert Einstein dataset by deleting pointless features. Multilayer Perception (MLP) and Logistic Regression (LR), with accuracy rates of 93% and 92%, respectively, were the two best-performing evaluation models. For the purpose of comparing various machine learning (ML) models, the COVID dataset was divided into training and testing subsets.

Samuel et al. [12] provided views into the evolution of the coronavirus through textual analysis supported by statistical statistics. Machine learning was used to classify tweets bearing text with different duration regarding the coronavirus. In-depthness and logic are hence applied using the logistic regression classification to the Naive Bayes technique, which offers an accuracy of 91% and 74% for these kinds of tweets, respectively.

Iwendi et al. [13] used ML categories to determine the virus's severity. Grid search optimization was used to fine-tune the AdaBoost with Random Forest's hyperparameters after analysing numerous methods and deciding that AdaBoost-RF was the best option (AdaBoost-RF). This model is a remotely deployable model for the detection and forecasting of patient severity in COVID-19 outbreaks with geographic and demographic-based data.

Another study that the authors had included employed a novel artificial neural network (ANN) with a deep extreme learning machine (DELM) architecture [14]. The DELM model presents a potential number of coronavirus epidemics in isolated locations and movements that would be helpful in performing the fundamental movements. In order to analyse clinical data for COVID-19 and adjust DELM settings for best performance, the study investigated several alternatives and activation aspects. Four deep learning models (DLM) and two hyper-deep learning models were introduced by Alakus et al. [15] These models used a dataset of 600 patients, 18 clinical characteristics, and trial-and-error hyperparameter tweaks to improve overall performance. With the greatest accuracy of 92% and recall rate of 94%, the CNN LSTM model stood out.

Li et al. [16] created an artificial intelligence system that uses deep learning to identify COVID-19's characteristic lung patterns and gauge the severity of the illness through the examination of dense chest CT images. Using artificial intelligence, this technology assists radiologists in the diagnosis and treatment of COVID-19 patients, as confirmed by medical radiology reports from bioimaging manufacturers. In addressing the imbalanced data issue in order to produce synthetic positive samples.

Sowjanya and Mrudula [17] used two SMOTE (Synthetic Minority Over-sampling Technique) to improve distance and instance-based sampling. When these techniques were evaluated with several classifiers, Distance-based SMOTE (D-SMOTE) and Bi-phasic SMOTE (BP-SMOTE) showed outperformed the original SMOTE in terms of accuracy. With the evolution of advanced machine-learning models for COVID-19, it remains a gaping challenge that cripples its effectiveness in the imbalanced nature of healthcare datasets. No prior literature has ever been leveraged to offer the synergy of oversampling methods and feature

selection to enhance model performance. The work, therefore, bridges this gap with the novel introduction of the integration of hyper-ADASYN oversampling with genetic algorithm-based feature selection, forming a wholesome solution for high predictive accuracy improvement in a clinical setting.

From the above review, it can be clear that the datasets applied are very diversified, including individual patient demographics, clinical features, and methods by which data is collected. Such diversity, while it does enrich the research landscape, carries along the difficulty of direct comparisons between predictive performance. We acknowledge this variability, and thus, our research aims to standardize the preprocessing of the dataset to an extent that would help reduce the impact of heterogeneity in data on predictive accuracy in COVID-19 outcomes. Table 1 shows a brief review of different studies related to COVID-19 prediction carried out over various datasets.

Table 1. Related work on COVID-19 disease prediction using different datasets.

Authors	Novel Approach	ACC	Dataset
Mondala et al. [11]	Multilayer perceptron (MLP), XGBoost and logistic regression	91%	Kaggle COVID-19 diseases dataset (5644 samples, 111 attributes) provided from Hospital Israelita Albert Einstein of Brazil
Iwendi et al. [13]	Fine-tuned Random Forest model boosted by the AdaBoost algorithm	94%	Kaggle as "Novel Corona Virus 2019 Dataset" (1086 samples ,17 attributes)
Khan et al. [14]	Artificial Neural Network (ANN) with Deep Extreme Learning Machine (DELM)	97.59%	Kaggle 2019 novel Coronavirus outbreak (1719 samples,9 attributes)
Alakus et al. [15]	Deep learning	86.66%	GitHub COVID- 19- Clinical (5644 samples,111 attributes)
Li et al. [16]	Deep learning	97%	COVID-19 dataset (538 CT)
Sowjanya et al. [17]	Machine learning, Deep Learning, Ensemble algorithms, BP-SMOTE and D-SMOTE	96%	Kaggle Framingham dataset (16 variables and 4240 observations),

3. Implementation and model design

In this section, we are proposing the mitigation strategies on the challenges found in the medical data set. Basic strategies used in developing this methodology are to develop a highly effective machine learning system that will be able to support unbalanced data sets and other related problems. The proposed technique encompasses the following six key steps:

Step 1: Pre-processing of the COVID-19 dataset.

Step 2:Application of oversampling techniques

Step 3:Employing genetic feature selection methods.

Step 4:Partition the dataset into training and testing sets.

Step 5:Applying various machine learning techniques.

Step 6:Evaluation the effectiveness of machine learning on diverse performance metrics.

Figure 1 shows the proposed approach. Each aspect of the proposed approach is detailed in the subsections that follow.

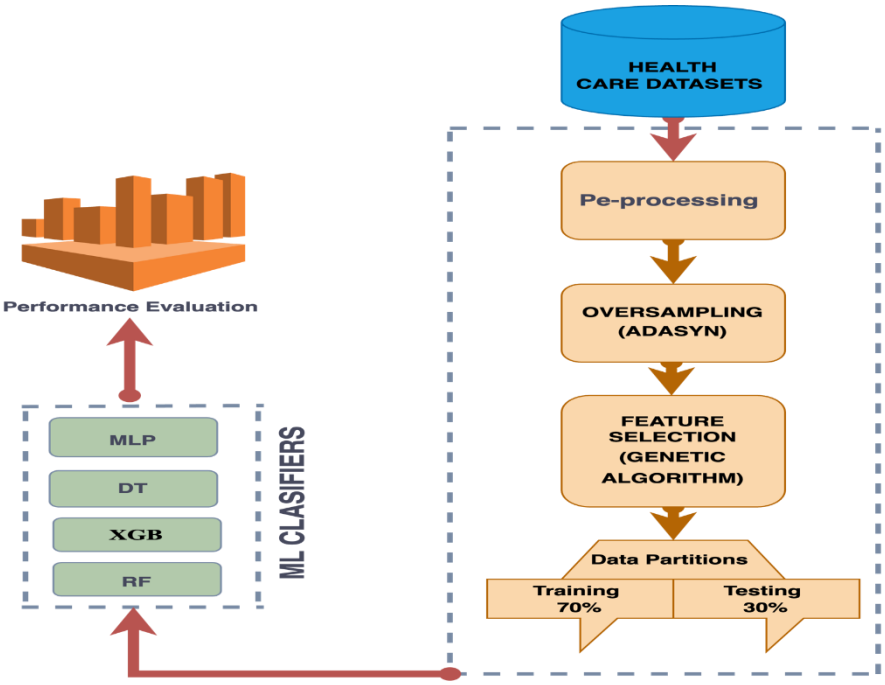

Fig. 1. Basic learning process for developing predictive models.

3.1. Data description

In this research, we utilized two distinct COVID-19 datasets, **Dataset I**, Albert Einstein Hospital in Brazil generously contributed COVID-19 data, encompassing a total of 5,644 patients. Prior to any normalization procedures, the clinical data were meticulously anonymized following the industry's best practices and standards [18]. This will involve all the blood test results of the patients, both positive and negative, with reference to SARS-CoV-2. The dataset used is for all attributes related to the 5,644 individuals with medical tests.

It is important to mention that this data set has a striking level of class imbalance, where the number of positive cases is far less than the large number of negative cases. To tackle this class imbalance issue, we employed the ADASYN oversampling method for both Dataset I and Dataset II. Detailed differences between the negative and positive classes for the original Dataset I and its over-sampled form are given in Table 2.

Table 2. Comparison of COVID-19 dataset I for non-patients and patient's samples.

Dataset	Positive	Negative	Total
Original Data	99 samples	501 samples	600 samples
ADASYN Data	501 samples	501 samples	1002 samples

For **Dataset II,** the source of our data was Kaggle, "Novel Coronavirus 2019 Dataset." From this dataset, we collected a number of instances of pneumonia cases found in Wuhan. Despite the fact that the data set includes as few as 17 features and 1085 samples, the model samples are quite numerous, and they really have an evident relationship in the scope of their influence on the target value. At that point, it should be noted that this data set is characterized by quite a serious class imbalance problem: the overwhelming number of cases is for negative patients, and only a small part is positive. The oversampling technique has been applied in order to balance such a class imbalance. Table 3 shows in detail the original view of Dataset II, and, on the other hand, the view of the latter after the application of the oversampling technique.

Table 3. Comparison of COVID-19 dataset II for non-patients and patient's samples.

Dataset	Positive	Negative	Total
Original Data	63 samples	1022 samples	1085 samples
ADASYN Data	1022 samples	1022 samples	2044 samples

3.2. Dataset pre-processing

Data pre-processing includes all the various techniques applied to datasets, either to prepare them for analysis or to structure them [19]. Data preprocessing, in general, can be summarized in the following three basic steps: data cleaning, over-sampling, and feature selection. Data cleaning is therefore the first and most important step that helps detect errors and skewed data points so that they can be eliminated from the dataset. This is a very crucial step in making sure the data are cleaned with the quality and integrity of the complete dataset. Our procedure for data cleaning was very stringent: it included the removal of duplicates, correction of inconsistencies, and imputation of missing values handled through sound imputation techniques in statistics.

The cleaned datasets are checked for outliers and then normalized in such a way that they all form a strong foundation for all the measuring scales in them to be uniform in the subsequent analysis phases. The oversampling in the preprocessing of data will be done through the ADASYN method, hence overcoming this intimidating problem of imbalanced datasets. ADASYN achieves this by generating new data samples randomly, inspiring them from nearby minority data points and their associated counterparts.

It works effectively against imbalanced data. Another machine-learning technique applied is feature selection, whereby the process of determining the required input attributes from the dataset is reduced to only those features that carry out model building effectively. The strategy of normalization enforces the presentation of the stronger, existing relationship in the data. It's important in that it normalizes all the values of the data to have a certain distribution; hence, it enforces the consistency of the data, making it possible for the analysis to have meaning.

3.3. Oversampling technique

The ADASYN algorithm, used in the package of the SMOTE family version 1.3.1, synthesizes positive cases. Here, in order to make the algorithm run, the following steps are applied to customize the main three parameters of the ADAS () function:

i. **Target:** Target is a vector representing the target class attribute that aligns with a given dataset X.
ii. **X:** X stands for the dataset itself, which can be in the form of a data frame or a matrix and should consist of numeric attributes.
iii. **K:** K represents the number of nearest neighbor considered during the sampling process.

The ADASYN algorithm operates on the principle that the density of instances in the minority class is inversely related to the density of synthetic samples it generates. In essence, ADASYN adaptively generates minority data samples based on their underlying distributions. It produces more synthetic data for minority class samples that are deemed harder to learn, in contrast to those that are simpler to learn [20].

In this study, ADASYN has been selected for oversampling due to its adaptive nature in generating synthetic samples. This method overcomes the class imbalance issue; on the other hand, it presents samples that lie close to the border of the decision region, whose largest region is affected by class imbalance. Such an approach is much needed in the case of complex health datasets, such as COVID-19 since they carry a critical minority class value in predictive analyses.

3.4. Genetic algorithm

A method of optimization inspired by natural selection is the genetic algorithm (GA). The survival of the fittest principle is used in this algorithm, which is based on population search [9]. By repeatedly applying genetic elements to members of an existing population, new populations are generated. The start of the population and the random selection of a predetermined number of chromosomes are the first steps in the GA process. Then, each chromosome that the population produces is calculated using fitness functions.

Two chromosomes are chosen using a selection procedure that takes fitness into account [8]. The most crucial stage of GA is crossover, in which the two chromosomes (parents) created by the selection procedure are mated according to the genes they carry to produce new children. The mutation will then be applied at random to some genes in the progeny to maintain population variety. The new population that results from a mutation is represented by the new offspring. The GA methods are as follows:

I. Selection techniques

Selection process: Based on their high fitness, specific chromosomes are chosen at this crucial GA stage to participate in the reproductive process or not. Another name for the selecting process is the reproduction process. Several well-known techniques are employed throughout the selection process. After that, the wheel is rotated at random to find solutions that will be used in the development of the following generation [21].

Nevertheless, there are several disadvantages, such as algorithmic unpredictability leading to inaccurate results. Brindle and De Jong devised the roulette wheel selection approach by using the concept of determinism.

Rank-Selection is the modified version of the roulette wheel. Instead of using fitness values, it uses rankings. They are assigned ranks based on their fitness levels, and each person has a chance to be chosen based on their ranks. The

probability of early convergence of the solution to the local minimum decreases with the rank-selection approach [22].

II. Crossover operators

By combining certain genetic components from one or more parental sources, crossover operators create children. The single-point crossover and the uniform crossover are two of the most well-known crossover techniques. The paternal chromosomes are separated at a single crossover site that is chosen at random in the single-point crossover procedure. Offspring are created by swapping the genetic material beyond this dividing point. The probability of employing a common crossover is confined to a specific range. In contrast, in the uniform crossover approach, 'n' crossover points are randomly selected. Subsequently, the parental chromosomes are partitioned at these chosen points and then reassembled by alternating between the parental sources.

The one-point crossover mechanism is one where two parent feature sets are recombined to produce the offspring. The crossover of information in the feature sets would thus help in searching for new amalgamations of the features, which, when put together, may work in a synergistic way to enhance the model's performance.

III. Mutation operators

When one group passes its genetic variety to the following generation, this is referred to as a mutation and is an event in the maintenance of genetic diversity. The three most prevalent types of mutations are translocation, simple inversion, and scramble mutation. The displacement mutation (DM) alters a substring inside a single solution. Both the legitimacy of the final answer and the unpredictability of the offset mutation are ensured by the random selection of the change's position within the chosen substring. Exchange mutations and insertion mutations are two separate subtypes of displacement mutation.

To replace or move a piece of a person's solution, exchange mutations, and insertion mutations factors are used. The SIM operator has an abbreviation for "simple inversion-mutation operator," which has been well represented in the given process, since the sub-chain is inverted through the reversal of the original string by the user in one solution between two specified positions. The SIM operator serves as a string inverter, reversing the series and placing it in a random position [23].

In this study, to assure the diversity of features and void pre-convergence, we employed a uniform mutation process. This would introduce the feature set to random alteration, granting an opportunity for the selection of new information from those well ignored by another selection method. The Algorithm has been suitably parameterized so as to get most predictive features for COVID-19 datasets. Mutation rate, crossover probability, and population size are some of the parameters fine-tuned through preliminary tests to make sure the algorithm converges towards the optimal feature subset. This assists in improving the predictive performance of the model without overfitting.

3.5. Feature selection

It is thus very important that a judicious selection of features be made for the machine learning task. In fact, feature selection and relevance are considered very

important since the datasets of today are high-dimensional and, therefore, most often contain lots of redundant and irrelevant features. This is the reason why the selection of features becomes very important. It works, therefore, to find the subset of features that give classification accuracy while reducing the computational burden of the learning model and at the same time filtering noise out of the input features. The relevance of the feature selection technique will be the ability for proper alignment of the chosen context of the problem and, in effect, the possibility of revealing underlying data patterns.

Genetic algorithms are one of the most commonly used-based evolutionary search for the most optimized set of features. Population creation begins with building subsets out of available features [24]. To evaluate these subgroups within the population, a predictive model tailored to the specific task comes into play. Following a comprehensive assessment of every individual in the population, a competition ensues to determine which subgroups will persist and advance to the next generation. Through a combination of crossover, which involves updating the winning groups, and mutation, the tournament victor transitions into the next generation while also undergoing a process of random feature removal.

3.6. Machine learning

Machine learning offers a large area of application in not only drawing complex models but also mining quite valuable medical knowledge in order to provide a different angle towards clinicians and medical specialists [25]. These models will be able to predict and play a pivotal role in improving the decision-making process for inpatient care in clinical practice. On the other hand, in the case of a fairly good adherence level to clinical guidelines, they can work independently and offer a diagnosis for a wide variety of diseases [26].

I. Random forest (RF)

One noteworthy machine learning technique in this context is the Random Forest (RF), which represents a supervised learning method known for its efficiency and simplicity. Multiple decision trees are created from randomized data samples, the prediction data is extracted from each tree, and the results are averaged. The Random Forest technique outperforms the performance of a single decision tree [27]. This is done while minimizing overfitting.

II. MLP

The multilayer perceptron (MLP) is a category of feedforward artificial neural network. The basic perceptron acts as a binary classifier, while the MLP comes in as a complex classification under the objective of very sophisticated tasks [28]. Research on the XOR issue served as the inspiration for MLP, which was later developed. The input layer, the hidden layer, and the output layer are the minimal number of layers that make up the MLP algorithm's structure. In this system, sensors give each input signal a different weight. As a result, each link connecting a sensor in one layer to the next layer produces a unique output.

III. Decision tree (DT)

This algorithm belongs to the category of supervised learning methods and is structurally similar to decision trees. It uses a structure resembling a tree to express

groups of decisions that have an impact on certain ratings, linking them with reflecting ratings or evaluations for both the tree's leaves and branches [29]. The root node, which acts as the highest node, is located at the decision tree's tip. This node oversees selecting the appropriate split depending on the values of the characteristics. The building of a decision tree happens when an input value is sent through the tree starting at the root node and ending at a leaf node.

IV. XGB

An intricate supervised machine learning technique built on the decision tree approach is called extreme gradient boosting (XGB), and it was first described by Pathy et al. [30]. The loss function in the XGB is regularized in the objective function as opposed to the gradient boosting framework, enabling the smoothness of the final learning weights, and lowering the risk of overfitting. The outcome is predicted by this decision tree-based technique using a tree ensemble model, which is based on the K regression trees. The total of each tree's projections determines the final degeneration of a particular sample.

In this study, hyperparameters were used to tune 4 ML models using a trial-and-error method as shown in Table 4. In the first model, the maximum depth of trees (max depth) was used to regulate the maximum depth of each tree in the forest, while the RF utilized the number of trees (n estimators) to change the number of trees. The second model employs multiple linear programming (MLP) to determine the ideal arrangement of hidden layers and neurons. To manage the maximum depth of the tree, the third model DT employs maximum depth (max depth), much as RF. The fourth model, XGBoost, used the learning rate to control the contribution of each tree to the final prediction and the maximum depth (max_depth) such as DT and RF.

Random Forest, MLP, Decision Tree, JSON, and XGB are selected based on their known performance over this healthcare dataset with characteristic non-linear and complex data patterns. The reason the Random Forest model was chosen is that it is robust to overfitting and can be used on large datasets with high dimensionality. The reason for the selection of MLP was to use layers and neurons, giving the capability of modelled complex relationships. Finally, the decision tree was included for the interpretability of the model a very important criterion in clinical decision support. Finally, XGB was undertaken to consider its performance in the task of classification and to be scalable.

Table 4. Hyperparameters for COVID-19 datasets.

Dataset	Model	Hyperparameters
Dataset I	RF	n_estimators = 100, max_depth = 9
	MLP	hidden_layer_sizes= 100, Learning Rate= 0.01
	DT	max_depth= 8
	RF	n_estimators = 120, max_depth = 9
Dataset II	RF	n_estimators = 120, max_depth = 9
	MLP	hidden_layer_sizes= 100, Learning Rate= 0.01
	DT	max_depth= 10
	XGB	Learning Rate= 0.01, max_depth= 15

3.7. Normalization

The dataset for COVID-19 was normalized with data normalization to standardize for proper organization and, to a certain extent, a uniform format and content of information in all fields throughout the dataset, enhancing the relevancy of the data [31]. The dataset was adjusted to achieve a mean of 0 and a variance of 1. Normalization is calculated as follows:

$$Z = \frac{x - \mu}{\sigma}$$

where x denotes the values associated with each attribute, μ signifies the means of individual attributes, σ represents the standard deviation of the dataset, and Z indicates the attributes in a standardized representation.

3.8. limitations and assumptions

Data preprocessing assumes missing data is completely random (MCAR), but this may not be the case in the data that originates from the healthcare industry. ADASYN over-sampling, while mitigating the class imbalance to some extent, has a disadvantage: it operates on the assumption that the generated synthetic samples do not effectively approximate the true distribution of the minority class. We warn that this may not be true, especially in multi-modal situations. These genetic features were selected with the underlying assumption that, through an evolutionary search process, a globally optimal set of features could be found. However, because genetic algorithms are heuristic in nature, we assume there is a possibility of converging to local optima and being sensitive to initial conditions.

We tuned very carefully the hyperparameters of our machine learning models, but we have to admit that there is the danger of overfitting the data due to the very intensive optimization of parameters. We used the cross-validation as a tool to escape from this danger, but surely, the problem of generalization of the model is one of the most relevant ones to solve in the domain. Last, not the least, where the selection of the model fell on Random Forest, MLP, Decision Tree, and XGB, these last, due to the robustness of its performance in datasets similar to the one in the present study. We also recognize the fact that most of these models come with intrinsic biases and assumptions that are not valid and therefore hamper the performance in new data scenarios.

4. Experimental results

This introduced a challenge, particularly in counterbalancing the challenge introduced by the imbalanced nature of COVID-19 and other medical datasets; therefore, we introduced a novel approach in training our machine-learning classification model. The first step was solving the problem of missing data - a vital process in increasing the statistical strength of our models. The ADASYN oversampling approach was used to compensate for the challenge of imbalance, more so in a situation where there is skewness in the distribution of classes in the dataset.

We also applied the Genetic Algorithm (GA) feature selection method, which allowed reducing, in our case, the training time of our model by way of reducing the features to be used in it while eliminating less important ones in a dataset that had minimal effect. This step led to increased performance for our model.

Recognizing the need for consistency in data pre-processing, we standardized the data across a uniform distribution range. This standardization was particularly important because the laboratory findings (features) in the dataset lacked well-defined limits. In the learning phase, our machine learning models were trained using 80% of the dataset, reserving the remaining 20% for testing the classifiers. This methodology was consistently applied to both datasets I and II, ensuring a robust evaluation of model performance. The dataset (I) has 5,644 patients, and the number was reduced to 600 using threshold removal of missing values for features with more than 90% missing values. Following preprocessing, the 4 ML models each attained the highest accuracy rates. As a result, the suggested method considerably increased the three classifiers' accuracy rates.

Different experimental settings were employed in a trial-and-error manner to determine the best GA parameters with RF model for both suggested datasets. Several factors are adjusted in GA, including population size (I), crossover (II), mutation (III), and iteration (IV). The best crossover numbers (0.4, 0.6), the best mutation set at (0.2, 0.4), the population feature size set at three alternative sizes (60, 80), and the fourth GA parameter (Maximum Iteration) set at 10. The GA parameters that we employed in this investigation are listed in detail in Table 5.

Table 5. Genetic algorithm parameters with optimal feature selection.

Dataset	MI	Populations	Crossover	Mutation	Selected	ACC
Dataset I	10	60	0.4	0.2	29	96
	10	60	0.6	0.2	25	97
	10	60	0.6	0.4	22	97
	10	80	0.4	0.2	30	96
	10	80	0.6	0.2	20	97
	10	80	0.6	0.4	18	97
Dataset II	10	60	0.4	0.2	16	94
	10	60	0.6	0.2	14	96
	10	60	0.6	0.4	13	96
	10	80	0.4	0.2	11	97
	10	80	0.6	0.2	10	97
	10	80	0.6	0.4	7	97

The overall performance results demonstrated that the suggested strategy is effective when used with various classification models. The anticipated accuracy metrics generated by each COVID-19 model for dataset I as shown in Table 6. In this experiment, oversampling and GA feature selection with an RF classifier were applied to the dataset I as part of the preprocessing step. All models' classification performance results were also carefully tested.

Table 6. Results of ML models for COVID-19 Dataset I.

Method	Accuracy	Precision	Recall	F1-score
RF	99%	99%	99%	99%
DT	92%	92%	92%	92%
MLP	96%	96%	96%	96%
XGB	97%	97%	97%	97%

In Table 7, COVID-19 dataset II has been used with 4 ML models. After preprocessing, the features decreased from 17 to 7 optimal features using GA and RF model. The RF model reached highest accuracy of 99% with using the optimal features from GA and tunning hyperparameters.

Table 7. Results of ML models for COVID-19 Dataset II.

Method	Accuracy	Precision	Recall	F1-score
RF	99%	99%	99%	99%
DT	95%	95%	95%	95%
MLP	96%	96%	96%	96%
XGB	98%	98%	98%	98%

In binary classification problems, the AUC rating is most employed to identify accurate class classifying models. An AUC score is obtained by dividing the total number of true positives by the total number of false positives. It implies that a greater rate yields better class prediction. It is good to have a ratio of 0.8 to 0.9, considering one is advised not to have any value exceeding 0.9. On the other hand, the AUC is of critical importance in outcome prediction in COVID-19, as it shows good predictive accuracy in determining the precision of the model in positive case findings, making it a very important model output for management decision support in treatment and allocation of healthcare resources.

For the dataset I the investigation showed that the RF approach was the most effective; the outcome was 99.7%. All the other approaches had outstanding grades, scoring higher than 90%, which is another plus. Based on AUC values, COVID-19 may be predicted using machine learning methods. Because it demonstrates how to discern between sick and healthy people, the AUC degree is crucial in the field of medical study. AUC analysis for ML algorithms is shown in Fig. 2 as a graph.

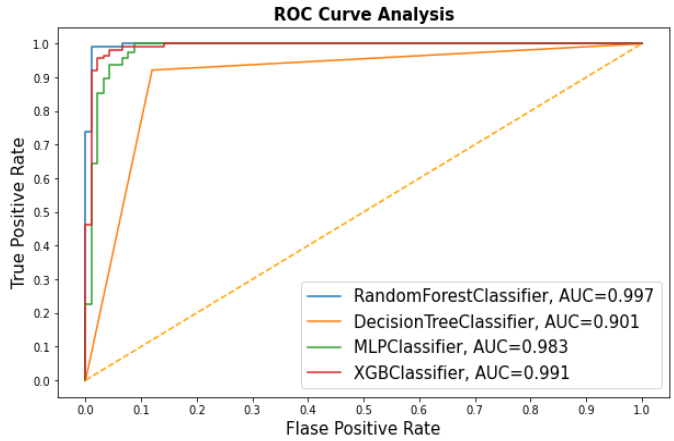

Fig. 2. AUC analysis for all ML models for Dataset I.

The investigation of Dataset II showed that the XGB model reached a 99.6% AUC, whereas the other models reached more than 94%, as shown in Fig. 3. The

more closely the curve traces the upper and left borders of the ROC space, respectively, the more precise the model. A curve that is close to the diagonal line (a 45-degree angle) suggests that random guessing is a more accurate method of prediction. Overall performance is measured by the area under these curves, or AUC.

The shape and area under these curves indicate the discriminative power of the various COVID-19 outcomes among the models. The larger area under the curve represents that the model will correctly predict, in a higher proportion of cases, a randomly picked positive instance to be more likely positive than a randomly picked negative instance.

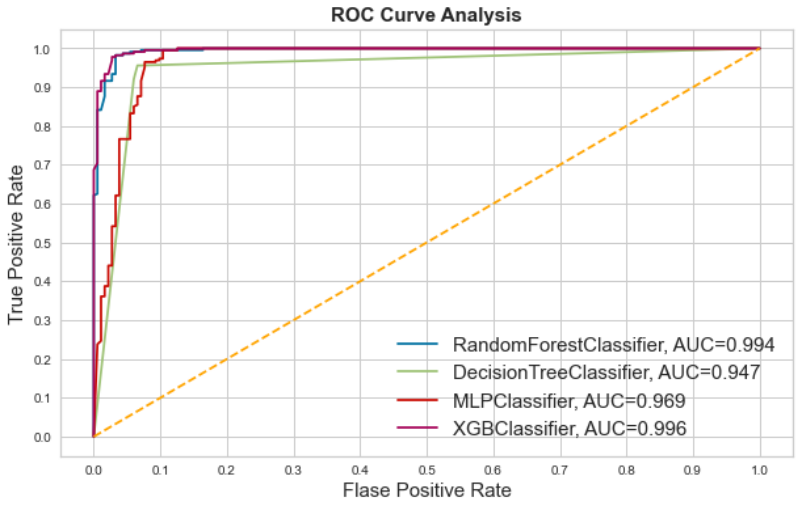

Fig. 3. AUC analysis for all ML models for Dataset II.

As shown in Figs. 4 and 5, the confusion matrix was used for both COVID-19 datasets. In confusion matrix, a high count of TP and TN indicates effective model performance in correctly identifying COVID-19 cases and non-cases, which is vital in clinical decision-making. This shows in the TP and TN analyses of the matrices the competency of this model in accurate diagnosis, an exemplary strength in the context of pandemic response. This has seemed to weigh heavily in healthcare settings. False positives can lead to unnecessary medical interventions; false negatives are even worse because they represent missed or delayed diagnoses. Our discussion now delves into the implications of these outcomes, assessing the balance each model strikes between sensitivity (reducing FNs) and specificity (reducing FPs). Table 8 shows the results of a comparison of the current study with previous studies.

Mondal et al. [11] Of all these, different ML models were applied to improve the prediction for the COVID-19 dataset 1, and the improved prediction has shown the highest accuracy produced by MLP at 93%. Similarly, in another study, Alakus and Turkoglu [15] yielded the best performance in predicting COVID-19 using different DL models applied to COVID-19 dataset 1. In this dataset, CNN-LSTM had an accuracy of 92.30%.

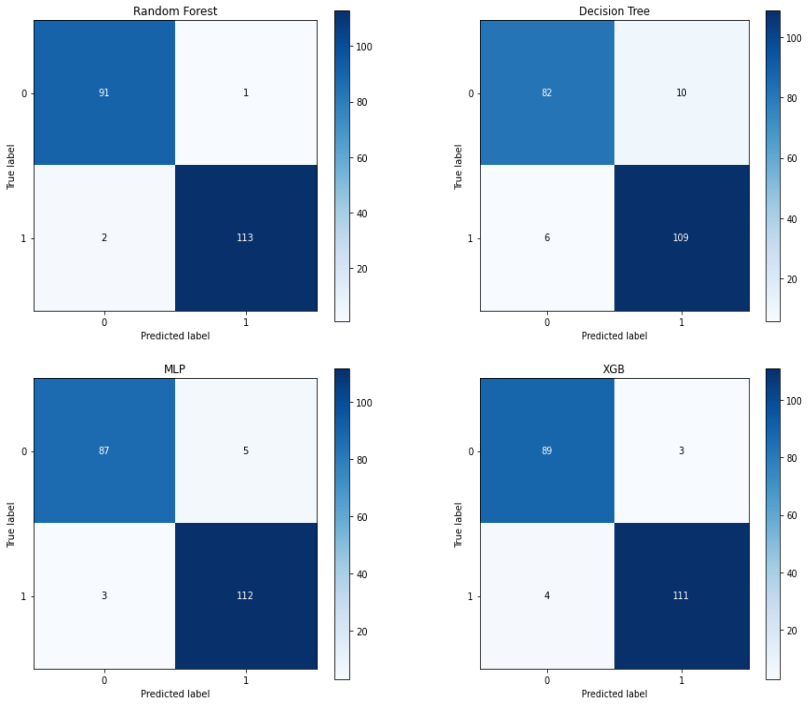

Fig. 4. Confusion matrix for dataset I.

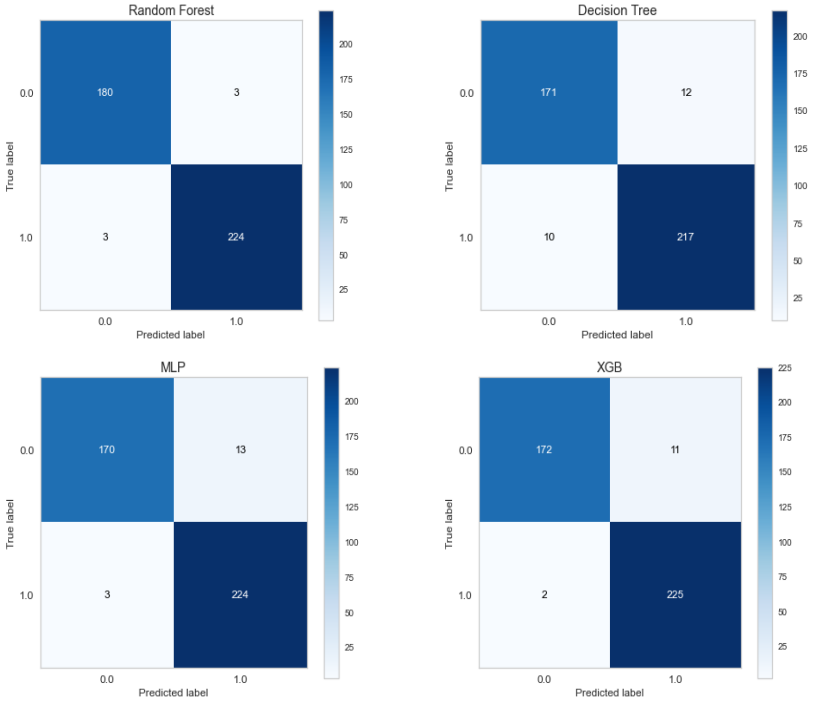

Fig. 5. The confusion matrix for dataset II.

Table 8. Comparison of evaluation results for Dataset I.

Study	Dataset	Method	ACC
Proposed	Albert Einstein Hospital	**RF**	**99%**
Mondal et al. [11]	Albert Einstein Hospital	**ML**	93%
Alakus and Turkoglu [15]	Albert Einstein Hospital	**DL**	92%

Table 9 shows the results of a comparison between our study and previous studies. Mohammedqasem et al. [32] applied different models to improve the prediction for COVID-19 dataset II, RNN showed the highest accuracy of 98%. In another study, A. Mary et al. [18] applied different DL models to improve the prediction for COVID-19 dataset II, Stacked RNN showed the highest accuracy of 97%.

Table 9. Comparison of evaluation results for Dataset II.

Study	Dataset	Method	ACC
Mohammedqasem et al. [32]	Novel Coronavirus 2019	RNN	98%
Sowjanya et al. [17]	Novel Coronavirus 2019	Stacked RNN	97%
Proposed	Novel Coronavirus 2019	RF	99%

5. Conclusion

This paper applies machine learning techniques to two representative datasets of COVID-19, which are inherently imbalanced. Among the core limitations, according to our views, would be the relatively small sample of patients, possibly undermining the statistical power and generalization of the findings. More so, the testing based on the proposed system will have data imbalances and, however, may introduce great bias. Hence, it might affect our predictions' validity. These limitations underscore the need for concerted efforts to improve data collection across the board in healthcare research.

This was an intended reason for the choice of ADASYN for oversampling, with an understanding that ADASYN is able to produce synthetic samples that would be adaptive and at the same time representative of the minority class, hence the class imbalance experienced in our datasets. So, it evidently shows how effective this method was, as both F1 and Ajson's scores rose in their utilization post-oversampling. Data standardization was also necessary, considering the range of features and their totally different scales. It also standardizes them so that all features contribute equally to the predictions of the model and no single feature dominates due to scale differences.

Our selection of Random Forest (RF) as the best-performing model was based on its superior accuracy and robustness in handling imbalanced datasets, as evidenced by its high-performance metrics. RF's ability to handle large feature sets and provide important insights into feature importance drove its selection over other models. The parameters of the Genetic Algorithm (GA) were fine-tuned with trial and error to have a very optimized feature selection. The parameters are population size, mutation rate, and crossover rate, all adjusted based on the sensitivity of the model in relation to its performance in achieving the highest prediction accuracy possible without overfitting.

The choice of the evaluation metrics precision, recall, F1-score, and accuracy were important in the study. These metrics gave a representative sample of the performance of the models in relation to accuracy and the balance between

sensitivity and specificity, which is important in sound medical diagnostics. Finally, our study bears wider significance in that it contributes to the field of imbalanced datasets in health care. By balancing the data and applying the model parameter optimization, this proposed machine learning model can provide promissory results as an example for diagnosis in COVID-19.

Abbreviations

ACC	Accuracy
ACO	Ant colony optimization
AdaBoost-RF	AdaBoost with Random Forest
ADASN	Adaptive Synthetic Sampling Approach for Imbalanced Learning
ANN	artificial neural network
DL	Deep Learning
DM	displacement mutation
DT	The Decision Tree
GA	genetic algorithm
LR	Logistic Regression
MI	Max Iteration
ML	Machine Learning
MLP	Multilayer Perception
NDI	The National Prevention Initiative
PSO	particle swarm optimization
RF	Random Forest
SIM	simple inversion-mutation operator
XGB	extreme gradient boosting

References

1. Perra, N. (2021). Non-pharmaceutical interventions during the COVID-19 pandemic: A review. *Physics Reports,* 913, 1-52.
2. Dietterich, T.G. (2000). Ensemble methods in machine learning. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 1857, Springer, Berlin, Heidelberg, 1-15.
3. Avci, O.; Abdeljaber, O.; Kiranyaz, S.; Hussein, M.; Gabbouj, M.; and Inman, D.J. (2021). A review of vibration-based damage detection in civil structures: From traditional methods to Machine Learning and Deep Learning applications. *Mechanical Systems and Signal Processing*, 147, 107077.
4. Elhoseny, M.; Mohammed, M.A.; Mostafa, S.A.; Abdulkareem, K.H. Maashi, M.S.; Garcia-Zapirain, B.; Mutlag, A.A.; and Maashi, M.S. (2021). A new multi-agent feature wrapper machine learning approach for heart disease diagnosis. *Computers, Materials & Continua*, 67(1), 51-71.
5. Zhang, C.; Bi, J.; Xu, S.; Ramentol, S.; Fan, G.; Qiao, B.; and Fujita, H. (2019). Multi-Imbalance: An open-source software for multi-class imbalance learning. *Knowledge-Based Systems*, 174, 137-143.
6. Chen, H.; Li, T.; Fan, X.; and Luo, C. (2019) Feature selection for imbalanced data based on neighborhood rough sets. *Information Sciences*, 483, 1-20.

7. Sharifai, G.A.; and Zainol, Z. (2020). Feature selection for high-dimensional and imbalanced biomedical data based on robust correlation based redundancy and binary grasshopper optimization algorithm. *Genes*, 11(7), 717, 1–26.
8. Kumar, V.; Chhabra, J.K.; and Kumar, D. (2014). Parameter adaptive harmony search algorithm for unimodal and multimodal optimization problems. *Journal of Computational Science*, 5(2), 144-155.
9. Katoch, S.; Chauhan, S.S. and Kumar, V. (2021). A review on genetic algorithm: past, present, and future. *Multimedia Tools and Applications*, 80(5), 8091-8126.
10. Chakraborty, A.; and Kar, A.K. (2017). *Swarm intelligence: A review of algorithms*. In Patnaik, S.; Yang, X.S.; and Nakamatsu, K. (Eds.). *Nature-Inspired Computing and Optimization.* Springer International Publishing, 10, 475-494.
11. Mondal, M.R.H.; Bharati, S.; Podder, P. and Podder, P. (2020). Data analytics for novel coronavirus disease. *Informatics in Medicine Unlocked*, 20, 100374.
12. Samuel, J.; Ali, G.G.M.N.; Rahman, M.K.; Esawi, E.; and Samuel, Y. (2020). COVID-19 public sentiment insights and machine learning for Tweets classification. *Information*, 11(6), 314.
13. Iwendi, C.; Bashir, A.K.; Peshkar, A.; Sujatha, R.; Chatterjee, J.M.; Pasupuleti, S.; Mishra, R.; Pillai, S.; and Jo, O. (2020). COVID-19 patient health prediction using boosted random forest algorithm. *Frontiers in Public Health*, 8, 562169.
14. Khan, M.A.; Abbas, S.; Khan, K.M.; Ghamdi, M.A.A.; and Sakhawat, A.R. (2020). Intelligent forecasting model of COVID-19 novel coronavirus outbreak empowered with deep extreme learning machine. *Computers. Materials and Continua*, 64(3), 1329-1342.
15. Alakus, T.B.; and Turkoglu, I. (2020). Comparison of deep learning approaches to predict COVID-19 infection. *Chaos Solitons Fractals*, 140, 110120.
16. Li, Z; Zhong, Z.; Li, Y.; Zhang, T.; Gao, L.; Jin, D.; Sun, Y.; Ye, X.; Yu, L.; Hu, Z.; Xiao, J.; Huang, L.; and Tang, Y. (2020). From community-acquired pneumonia to COVID-19: a deep learning-based method for quantitative analysis of COVID-19 on thick-section CT scans. *European Radiology*, 30(12), 6828-6837.
17. Sowjanya, A. M.; and Mrudula, O. (2023). Effective treatment of imbalanced datasets in health care using modified SMOTE coupled with stacked deep learning algorithms. *Applied Nanoscience*, 13(3), 1829-1840.
18. Kaggle (2020). Diagnosis of COVID-19 and its clinical spectrum. Retrieved August 12, 2023, from https://www.kaggle.com/datasets/einsteindata4u/covid19
19. Geetha, R.; Sivasubramanian, S.; Kaliappan, M.; Vimal, S.; and Annamalai, S. (2019). Cervical cancer identification with synthetic minority oversampling technique and PCA analysis using random forest classifier. *Journal of Medical Systems*, 43(9), 286.
20. De la Cruz-Ruiz, F.; Canul-Reich, J.; Rivera-López, R.; and de la Cruz-Hernández, E. (2023). Impact of data balancing a multiclass dataset before the creation of association rules to study bacterial vaginosis. *Intelligent Medicine*. (in press).

21. Jebari, K.; and Madiafi, M. (2013). Selection methods for genetic algorithms. *International Journal of Emerging Science*, 3(4), 333-344.
22. Dhal, K.G.; Ray, S.; Das, A.; and Das, S. (2018). A survey on nature-inspired optimization algorithms and their application in image enhancement domain. *Archives of Computational Methods in Engineering*, 26(5), 1607-1638.
23. Alhenawi, E.; Al-Sayyed, R.; Hudaib, A.; and Mirjalili, S. (2022). Feature selection methods on gene expression microarray data for cancer classification: A systematic review. *Computers in Biology and Medicine*, 140, 105051.
24. Maleki, N.; Zeinali, Y.; and Niaki, S.T.A. (2020). A k-NN method for lung cancer prognosis with the use of a genetic algorithm for feature selection. *Expert Systems with Applications*, 164, 113981.
25. Tătaru, O.S.; Vartolomei, M.D.; Rassweiler, J.; Virgil, O.; Lucarelli, G.; Porpiglia, F.; Amparore, D.; Manfredi, M.; Carrieri, G.; Falagario, U.; Terracciano, D.; de Cobelli, O.; Busetto, G.M.; Giudice, F.D.; and Ferro, M. (2021). Artificial intelligence and machine learning in prostate cancer patient management current trends and future perspectives. *Diagnostics*, 11(2), 354.
26. Ghosh, P.; Azam, S.; Karim, A.; Hassan, M.; Roy, K.; and Jonkman, M. (2021). A comparative study of different machine learning tools in detecting diabetes. *Procedia Computer Science*, 192, 467-477.
27. Sheykhmousa, M.; Mahdianpari, M.; Ghanbari, H.; Mohammadimanesh, F.; Ghamisi, P.; and Homayouni, S. (2020). Support vector machine versus random forest for remote sensing image classification: A meta-analysis and systematic review. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 13, 6308-6325.
28. Desai, M.; and Shah, M. (2021). An anatomization on breast cancer detection and diagnosis employing multi-layer perceptron neural network (MLP) and Convolutional neural network (CNN). *Clinical eHealth*, 4, 1-11.
29. Ghiasi, M.M.; and Zendehboudi, S. (2021). Application of decision tree-based ensemble learning in the classification of breast cancer. *Computers in Biology and Medicine*, 128, 104089.
30. Pathy, A.; Meher, S.; and Balasubramanian, B. (2020). Predicting algal biochar yield using eXtreme Gradient Boosting (XGB) algorithm of machine learning methods. *Algal Research*, 50, 102006.
31. Senan, E.M.; Abunadi, I.; Jadhav, M.E.; and Fati, S.M. (2021). Score and correlation coefficient-based feature selection for predicting heart failure diagnosis by using machine learning algorithms. *Computational and Mathematical Methods in Medicine*, Volume 2021, Article ID 8500314.
32. Mohammedqasem, R.; Mohammedqasim, H.; Biabani, S.A.A.; Ata, O; Alomary, M.N.; Almehmadi, M.; Alsairi, A.A.; and Ansari, M.A. (2023). Multi-objective deep learning framework for COVID-19 dataset problems. *Journal of King Saud University - Science*, 35(3), 102527.