REPUBLIC OF TÜRKİYE

ALTINBAŞ UNIVERSITY

Electrical and Computer Engineering

ENHANCING COVID-19 PATIENT PREDICTION WITH OVER-SAMPLING AND GENETIC FEATURE SELECTION

Yasir Ali Mohammede Ali AL-TAHHAN

Master's Thesis

Supervisor

Asst. Prof. Dr. Oğuz Ata

Istanbul, 2022

ENHANCING COVID-1 9 PATIENT PREDICTION WITH OVER-SAMPLING AND GENETIC FEATURE SELECTION

Yasir Ali Mohammede Ali AL-TAHHAN

Electrical and Computer Engineering

Master's Thesis

ALTINBAŞ UNIVERSITY

2022

The thesis titled Enhancing covid-19 patient prediction with over-sampling and genetic feature selection prepared by Yasir Al-Tahhan and submitted on 20/12/2022 has been **accepted unanimously** for the degree of Master in Electrical and Computer Engineering.

Asst. Prof. Dr.Oğuz ATA

Supervisor

Thesis Defense Committee Members

Asst. Prof. Dr. Oğuz ATA	Department of Software Engineering, Altinbas University	_____
Prof. Dr. Hasan Hüseyin BALIK	Department of Computer Engineering, Yildiz Technical University	_____
Asst. Prof. Dr. Aytuğ BOYACI	Department of Computer Engineering, National Defence University	_____

I hereby declare that this thesis meets all format and submission requirements of a …… (Master's) thesis.

Submission date of the thesis to the Graduate Education Institute: ___/___/___

I hereby declare that all information/data presented in this graduation project has been obtained in full accordance with academic rules and ethical conduct. I also declare all unoriginal materials and Conclusions have been cited in the text and all references mentioned in the Reference List have been cited in the text, and vice versa as required by the abovementioned rules and conduct.

Yasir Ali Mohammede Ali AL-TAHHAN

Signature

DEDICATION

I'd want to thank Asst. Prof. Dr.Oğuz Ata from the bottom of my heart. For all of his expertise and assistance during my Master's degree studies and effort to accomplish my thesis. I realized my ambition, and I'd want to thank everyone of my family members for their support during my studies.

ABSTRACT

ENHANCING COVID-19 PATIENT PREDICTION WITH OVER-SAMPLING AND GENETIC FEATURE SELECTION

Yasir, Al Tahhan

M.sc., Electrical and Computer Engineering, Altınbaş University

Supervisor: Asst. Prof. Dr. Oğuz Ata

Date: December / 2022

Pages: 46

SARS-CoV 2, the cause of COVID-19 (coronavirus disease) become a global pandemic. Because the number of patients is growing more and more every day, evaluating test data takes time, which leads to the emergence of drugs and the discovery of limitations. As a result of these limitations, a clinical policy system that includes predictive algorithms is required. By identifying diseases, predictive algorithms can relieve pressure on health care systems. In this study, a combination of over-sampling and genetic feature selection (GFs) with laboratory data is proposed to build machine-learning clinical prediction models that identify potentially infected patients with COVID-19. Firstly, the oversampling method is used to overcome the problem of imbalance by uniforming the distribution data for all classes. Second, the GFs is used to remove the noise and irrelevant features from the data. To compute and evaluate our performance of the proposed method, many metrics were used: accuracy, recall, F1-score, AUC, and precision scores. It is proven here that random forest (RF), decision tree (DT), and multilayer perceptron (MLP), can classify COVID-19 patients with more than 96% accuracy.

Keywords: COVID-19, SMOTE, Genetic Algorithm, Imbalanced Data, Random Forest.

TABLE OF CONTENTS

Pages

LIST OF ABBREVIATIONS

ML	MACHINE LEARNING
GA	GENETIC ALGORITHM
SMOTE	SYNTHETIC MINORITY OVERSAMPLING TECHNIQUE
GFs	GENETIC FEATURE SELECTION
RF	RANDOM FOREST
DT	DECISION TREE
MLP	MULTILAYER PERCEPTRON
AI	ARTIFICIAL INTELLIGENCE
COVID-19	CORONAVIRUS DISEASE

LIST OF TABLES

Pages

LIST OF FIGURES

Pages

CHAPTER ONE

1. INTRODUCTION

1.1 BACKGROUND

Despite these efforts, a fundamental problem is the lack of routine medical records that detail the feedback loop between human interaction and illness. While the use of the phrase "beyond distressing," like many others in the text, grabbed the reader's attention, it is not a grammatical mistake. Things have changed considerably in the last few years. The COVID-19 epidemic has overrun the arena, and one of the few assets remaining to combat the issue is non-drug interventions (NDI) [1]. The National Prevention Initiative (NPI) is a collection of top-down (government) and bottom-up (individual) actions aimed at breaking the infection cycle by changing important parts of human behavior. Travel limits, night classes, social distance, social event rules, face masks, improved shaving, remote employment, college closings, and locks are just a few examples. The geographical and temporal variety of these techniques, as well as their cost of the alternative, present unparalleled opportunities to conceive about, quantify, and model the link between human behavior and infectious illness[2]. Modern era and records applications, such as those developed by Google, Apple, and Facebook, as well as major telecom operators such as Vodafone, Telefonica, and Orange, and small businesses such as Cuebiq, SafeGraph, and Unacast, provide an unrivaled lens for achieving a satisfactory NPI understanding impact. In light of this, I'd want to outline some of the most significant observations, proof, methodology, and know-how gathered throughout the first year of the COVID-19 epidemic[3]. Due to the excellent work of the look at the network, it has become difficult to manually review all COVID-19 papers. A PubMed search for "COVID" yielded over 71,000 results, giving you an idea of the variation. On Google Scholar, we'll find over 135,000 matches. Only a few study papers are smaller than

most as part of this big undertaking. Regardless, the NPI has had an impact on nearly every aspect of human life.

The World Health Organization classified Coronavirus disease 2019 (COVID-19) a pandemic on February 11, 2020, and this virus is a category 2 coronavirus, according to the International Committee on Taxonomy of Viruses (ICTV) (hyperacute respiratory syndrome). SARS-CoV-2 is a sarcoma-causing virus. This sickness has been categorized as a dangerous and lethal condition by fitness firms, and the industry requires adequate treatment choices and other care services to handle it [4]. The pandemic has created a threat to worldwide health systems since it necessitates unnecessary surveillance and hospitalization to detect and treat infected people, particularly if the spread of the new virus is not controlled. The World Health Organization (WHO) [5] has declared the COVID-19 outbreak a pandemic, necessitating the rapid deployment of infrastructure and procedures to classify those most at vulnerable of sickness, and death. COVID-19 has several side effects in people. However, more than 80% of people with mild to severe illness improve without needing to be hospitalized [6]. The most popular signs and indication of this virus are: 'fever', 'dryness', and 'weariness' these signs and symptoms appear gradually in all infected persons. In addition, additional symptoms such as chest discomfort or exhaustion, as well as loss of capacity to talk or move, may arise in some patients [3]. It is unknown what effect this variation has on the production of defensive immune responses and the function of antibodies in disease progression. When treatment regimens for moderate and severe COVID-19 cases evolve, our understanding of the therapeutic effects of this research on the immune response to SARS-CoV-2 is limited.

As an Artificial intelligence AI application, Machine learning (ML) algorithms have the unique power of learning from experience and improving themselves without the need for particular programming [7]. The examples below will be used to teach and assist grasp ML. Assume the goal is to create an AI medical diagnostic tool that can predict if a patient will develop a specific form of cancer. The computer training data provided is made up of a broad group of samples. The first stage in training the computer is to have the doctor take down all of the symptoms that may be associated with the cancer type you're interested in [8]. The training data will then be utilized

to collect information from each medical report on these symptoms (i.e. the feature retrieval process). The studied data collection is then utilized to train the system using a sophisticated machine learning algorithm, which comprises information about recorded conditions collected from each medical record and linked outcomes (whether the patient has been diagnosed with cancer or not). During the training phase, the computer automatically identifies the relationship between indicators and outcomes; H. Cancer (malignant) or natural (benign). Following screening exams, a professional procedure based on medical history for previously reported symptoms may be utilized to diagnose potential patients. Machine learning algorithms are classified into two types: unobserved algorithms and managed algorithms [8]. According to the preceding example, the supervised algorithm requires a data collection comprising human-tagged training data [9].

The most common issue with most medical data sets is imbalanced data. Several open-source methods and packages [10] 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 multiple category record 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. Observers should use numerous indications [11] in their evaluation to avoid discrepancies and acquire a comprehensive report on the genuine effectiveness of the classifiers. In a multi-indicator measurement approach, the difference between the monitored indication and the lowest metric is chosen as the final score for model selection accuracy. This recommendation [12] provides an alternative viewpoint that is unaffected by imbalance register discrepancies. However, there will be drawbacks to employing this strategy. When choosing a standard, for example, one of the lowest type classifiers may have unique indicators, and no classifier may be displayed to govern all other indications on all tags

inside the class. For example, one classifier represents total consistency, while the other represents group scale precision.

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 [13]. Once the best features are identified, ML method will rapidly establish the relationship between the retrieved characteristics and the intended output. This machine learning approach is effective when the extracted properties, like in the example above, can be identified manually and the function (the list of diseases) can be selected by a expert system.

Metaheuristic algorithms have recently been applied to tackle real-world complicated issues in domains such as economics, biomedical, and engineering [14]. The fundamental components of a metaheuristic algorithm are intensification and diversity. A good balance of these aspects is essential to handle the real-life challenge effectively. Most metaheuristic algorithms are motivated by biological evolution and swarming behavior. These algorithms are classified into two types: meta population-based algorithms and single-solution algorithms as seen in Figure 1.1. A single-solution metaheuristic algorithm starts with a candidate solution and finishes it with a local search. On other hand, resulting solution of single-solution based on a meta-heuristic may become trapped in the local optima [15]. Population-based metaheuristic employs a variety of possible solutions throughout the search process. These metaheuristics protect population diversity and keep solutions from becoming locked in local optima. There are several well-known population-based metaheuristic algorithms such as genetic algorithm (GA) [15], ant colony optimization (ACO), and particle swarm optimization (PSO) [16].

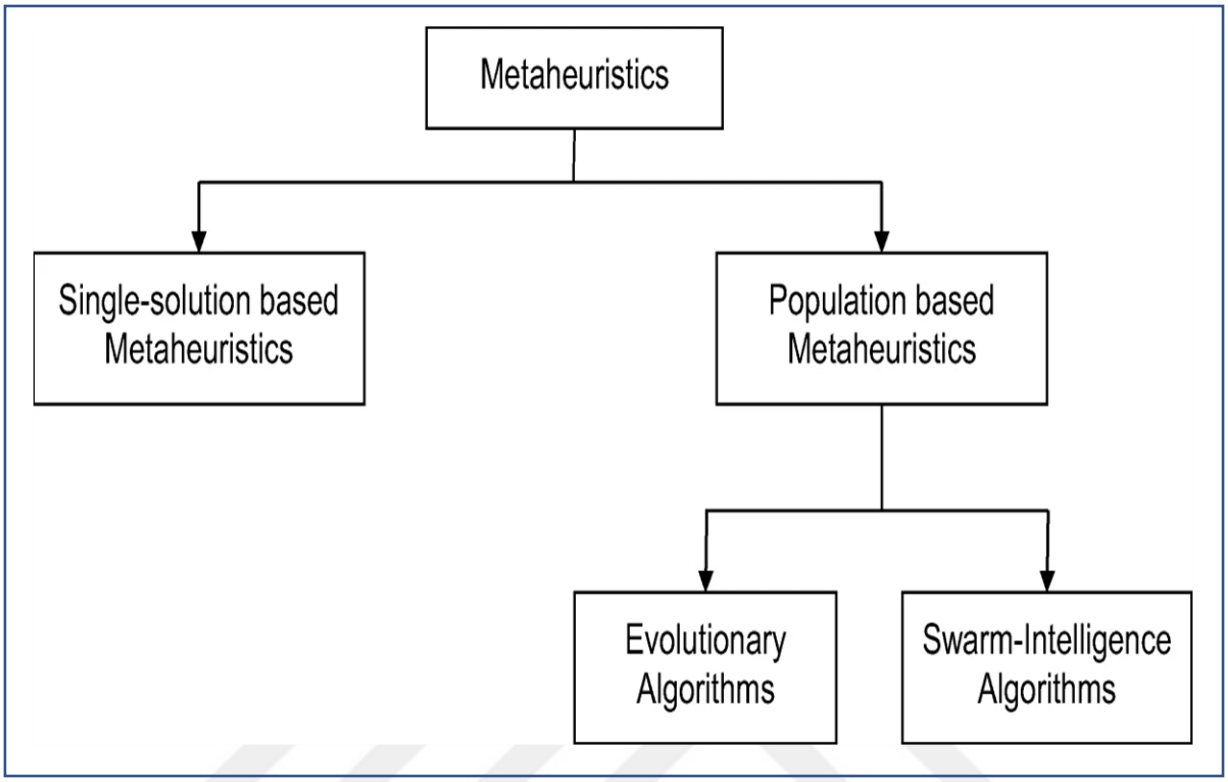

Figure 1. 1 Metaheuristic Algorithm Classification.

The Genetic Algorithm (GA) is a meta-heuristic algorithm inspired by the process of biological evolution. GA mimics Darwin's survival of fittest concept in nature. GA was suggested in 1992 by J.H. Holland. Chromosome mapping, fitness selection, and biologically inspired operators are three crucial parts of the GA procedure. Holland also incorporated a feature known as inversion, which is commonly used in GA implementations [17]. Genes are binary codes (0,1) that are used to represent chromosomes. Genetic variables and regular population interchange are used in their treatment. Each chromosome in a population is assigned a different value using the fitness function [15]. The genetic algorithm approach is illustrated in Figure 1.2, and the three main processes in GA are Selection; mutation; and crossover, all of which are biologically inspired aspects. Based on their suitability for selection, chromosomes are chosen for next GA the processing. To produce offspring, the crossover operator selects a random location and changes the order of the chromosomes. Some chromosome bits are flipped arbitrarily dependent on chance in mutation [18].

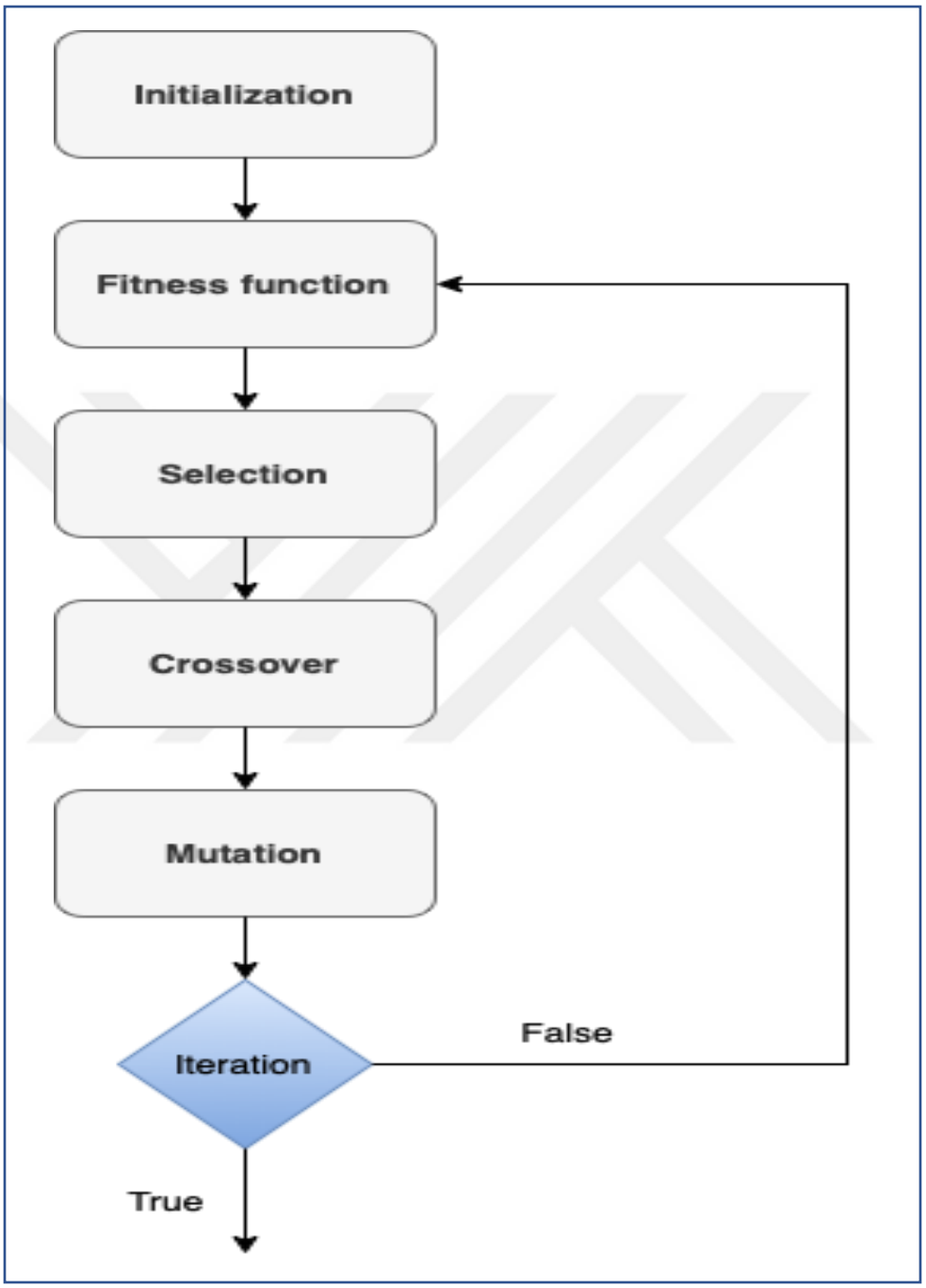

Figure 1. 2 Genetic algorithm procedure diagram.

1.2 PROBLEM STATEMENT

The problem of imbalance in clinical data sets is not addressed, and this will bias the models to learn from the larger class rather than learn from all classes. Selecting a subset of high important features information training is called feature selection. There are important factors to consider at the same time as deciding on a feature. First, the feature has to limit redundant capability and clear out the noise, which can bring about a sizeable lack of detecting accuracy. The results of this research will be utilized to see if the machine learning model can be improved and how accurate it is as a system for detecting imbalanced data.

1.3 THE AIM OF THESIS

One of the problems of ML is the unbalanced dataset, where the range of samples in one class is too low for the sample set in another class, a class containing many samples is referred to as a major class, while a class with few numbers of samples is called a minor class. Because the proportion of infected human increases dramatically and without preceding reviews, approximately this disease, using ML strategies might be beneficial to diagnose this virus [19]. Therefore, Synthetic Minority Over-Sampling Technology (SMOTE) comes as one of the powerful machine learning approach for overcoming unbalanced data issues by generating new data in minority classes in the dataset by using k-nearest neighbours to find nearly equal classes. On the other hand, the low accuracy of the classification could be attributed to the fact that the negatively significant features were not carefully extracted from the original data set. In this study, we used a Hyper-ML model to classify COVID19 disease by applying SMOTE and a genetic feature selection approach.

1.4 OBJECTIVES

The main study objectives can be summarized bellow

i) Combine oversampling with GA feature selection to enhance the prediction with unbalanced COVID-19 data
ii) Apply SMOTE oversampling to deal with high imbalance problems in COVID-19 datasets.

iii) Remove inconsistency from features in the data set and reduce processing time using Genetic Features Selection.
iv) Enhancing the overall performance prediction of COVID-19, reducing the error rate and comparing with the previous studies.
v) To conduct a study to predict numeric medical datasets through using optimization of deep learning model tools with laboratory results instead of X-rays.

1.5 THESIS OUTLINE

The following sections are a description of the dissertation details:

Chapter 2: includes a literature review as well as background descriptions of the COVID-19 data sets and the imbalanced dataset.

Chapter 3: Introduces the design and implementation technique, as well as the recommended methodologies for oversampling and the Genetic feature selection model for medical data.

Chapter 4: The proposed methodology results are provided, beginning with the ML models experimental results and finishing with the suggested methodology experimental results.

Chapter 5: The study is concluded, and recommendations for further work are given.

CHAPTER TWO

2. LITERATURE REVIEW

2.1 INTRODUCTION

The Coronavirus has recently gained the attention of scientific researchers due to health organizations' incapacity to detect the causes of this disease due to its genetic structure[20], ubiquitous nature, and potential hazards. Artificial learning algorithms have been employed in a range of fields, including supporting health institutions and improving medical field efficiency in terms of accurate diagnosis and minimizing the dangerous epidemic [21]. Researchers applied a range of algorithms and strategies to improve the categorization efficiency of Covid-19 patients to acquire trustworthy results and learn more about the disease's origins. However, predicting a patient's clinical result is difficult since various confounding variables, including imbalanced data, uninteresting features, and training time might influence a patient's classification. By minimizing these issues, health organizations will be able to better identify and characterize medical outcomes.

In another study [22],the author proposed a new COVID-19 prediction model, in which a high feature rating was selected to remove irrelevant features from the Hospital Albert Einstein data set. The COVID data set was split into training data and testing data to evaluate several ML models, and the best evaluation models were Multilayer Perception (MLP), and Logistic Regression (LR) with an accuracy of 93%, 92%, respectively.

To analyze the efficacy of the proposed approaches, the author offered five different machine learning algorithms to detect patients at risk of positive COVID-19 infection [23], The data set was split into 80 percent train and 30 percent test, with several evaluation measures like as

accuracy, sensitivity, specificity, and AUC utilized to evaluate the prediction. With an AUC of 85 percent, support vector machines produced the best predictions.

The author used ML classifications in the study [24] to diagnose the virus. As they analyzed several algorithms available to process data from afflicted patients and identified AdaBoost-RF as the best approach, they used grid search optimization to change the hyperparameters of the AdaBoost with Random Forest (AdaBoost-RF). The model predicts the severity of COVID-19 patients based on regional and demographic data.

The expert model was created using an Artificial neural network(ANN), with a deep extreme learning machine (DELM). The model has a lot of promise for identifying coronavirus outbreaks in remote locations, and it has even been used to do rudimentary motions. Many options and activation characteristics have been defined for the optimal binding of the different DELM parameters [25], and many options and activation features have been employed for the optimal binding of the different DELM parameters to attain the optimal form.

The author [26] proposed four deep learning models(DLM) with two hyper-deep learning models to evaluate clinical data for COVID-19. DLM was developed to evaluate the classification performance with 18 clinical features and 600 patients. Dataset was trained ,and tested with different approaches. To increase the overall performance of the models, several of the hyper-parameters are modified by trial and error. The CNN LSTM model had the greatest accuracy of 92 % and recall of 94 %.

The authors [27] used deep learning to create an artificial intelligence computer that could recognize typical lung patterns for COVID-19 and validate illness severity as well as developments in the usage of thick chest CT scans. Bioimaging manufacturers have been able to review several medical radiology reports with the use of artificial intelligence, demonstrating that the primary current deep learning technology can assist radiologists in diagnosing and complying with the ongoing treatment of COVID-19 patients on CT scans.

The authors [28] blood tests provided by the Covid-19 dataset are used to predict a positive prognosis for patients by applying four ML models. Resampling method applied to reduce the bias distribution in dataset. Shapley Additive Explanations (SHAP) approach is used to calculate the attractiveness of each feature in dataset. Four blood criteria have been identified as the most important for diagnosing COVID-19. Optimal hyperparameters has been used to improve the performance of all ML models. The best results achieved are 91% of accuracy and 92% of AUC.

Previous research with this dataset has not solved the problem of most classes in the data, resulting in a bias in the models in the training process towards the classes with the most samples. Finally, using this study, we concentrated on evaluating and resolving the problem that we discussed by using our system to achieve higher prediction accuracy than earlier methods.

2.2 HISTORY OF IMBALANCED DATA

Classification of unbalanced data sets is a relatively new area of research in the context of a broader machine learning science that seeks to take advantage of unbalanced data distributions. In a data set containing two classes and several classes, the data set is unbalanced if the sample from one class contains more classes than the sample from the other classes. The majority of common machine learning algorithms fail with this data set because they favor the majority portion of the dataset, resulting in inferior predicted performance when compared to the minority class. As a consequence, analyzing uncommon but noteworthy occurrences becomes challenging. To reduce the total error rate, they assume the same misclassification cost for all samples [29].

This record's misclassification costs are the same for each class, which is wrong. When anticipating a software fault, for example, if a disturbed component is referred to as a positive class, and a broken module is referred to as a negative class, the absence of a flaw (false negative) during the software development phase is considerably more costly than a false positive error [30].

Unbalanced data is handled by modifying machine learning methods at the algorithm level. Adjustments should be made to account for variable misclassification costs for each class, which is considered a cost-responsive approach that focuses on lowering cost mistakes rather than improving accuracy [29]. Another enhancement is to select an accurate inductive bias. Changing the average probability on a worksheet, or taking into consideration any minimal support for different groups in the assignment rule, such as whether a learner uses a decision tree. Several more techniques, such as in-depth ensemble and algorithmic approaches, are advised in the training algorithm to improve it for a minority class study. The ensemble form, which we will focus on in this analysis, is one of the most well-known methods in this category.

By combining a simple group of learners for the classification method, an ensemble classifier, also known as a multi-classifier technique, enhances training efficiency. The results of each primary classifier are gathered and applied to the present study's classification decisions. A stable aggregation of base classifiers is formed on a weighted version of the training results, with an emphasis on incorrect classifier samples at each stage of evolution [30].

An unbalanced data set is exacerbated by an unbalanced classification task of several classes, in which both minority and majority classes can exist, resulting in data distribution bias. In this situation, one class may be a minority in comparison to another, yet it may be the majority in others. As a result, various new challenges emerge that did not exist in the two-tier scenario [31].

Emerging technologies such as the Internet of Things (IoT) and an online social network (OSN) have enabled the generation of massive volumes of data, known as big data, which presents certain learning issues due to the algorithms it contains. It is due to their distinguishing properties, such as length, speed, variety, and consistency [32]. Furthermore, imbalanced data is common in this type of data set, and when looking at properties with several classes, learning from a vast, unbalanced data set with multiple classes becomes a tough problem that, despite the data, has not been thoroughly studied. Frequently appear in real-world categorization problems [33].

CHAPTER THREE

3. IMPLEMENTATION AND DESIGN MODELS

3.1 INTRODUCTION

The primary goal of this project is to identify the best collection of functions and handle imbalanced data in order to improve the accuracy of the medical data categorization system. This chapter dives deep into the proposed system's design and implementation. To solve the issue of unbalanced data, the design employs different ML models.

3.2 THE PROPOSED METHODOLOGY

We present our methods in the method section, which we utilized to address the difficulties in the medical dataset. The suggested methodology's main goals are to create a highly efficient machine learning system that can reinforce imbalanced and other data sets.

The six steps of our suggested technique are as follows: (1) Preprocessing a COVID-19 data set, (2) Oversampling binary data, (3) Genetic feature selection, (4) Splitting with train and test sets, (5) Theoretical backgrounds of different ML algorithms, and (6) Evaluation of ML efficacy on various metrics The methodological stages are outlined in Figure 3.1 and the subsections that follow.

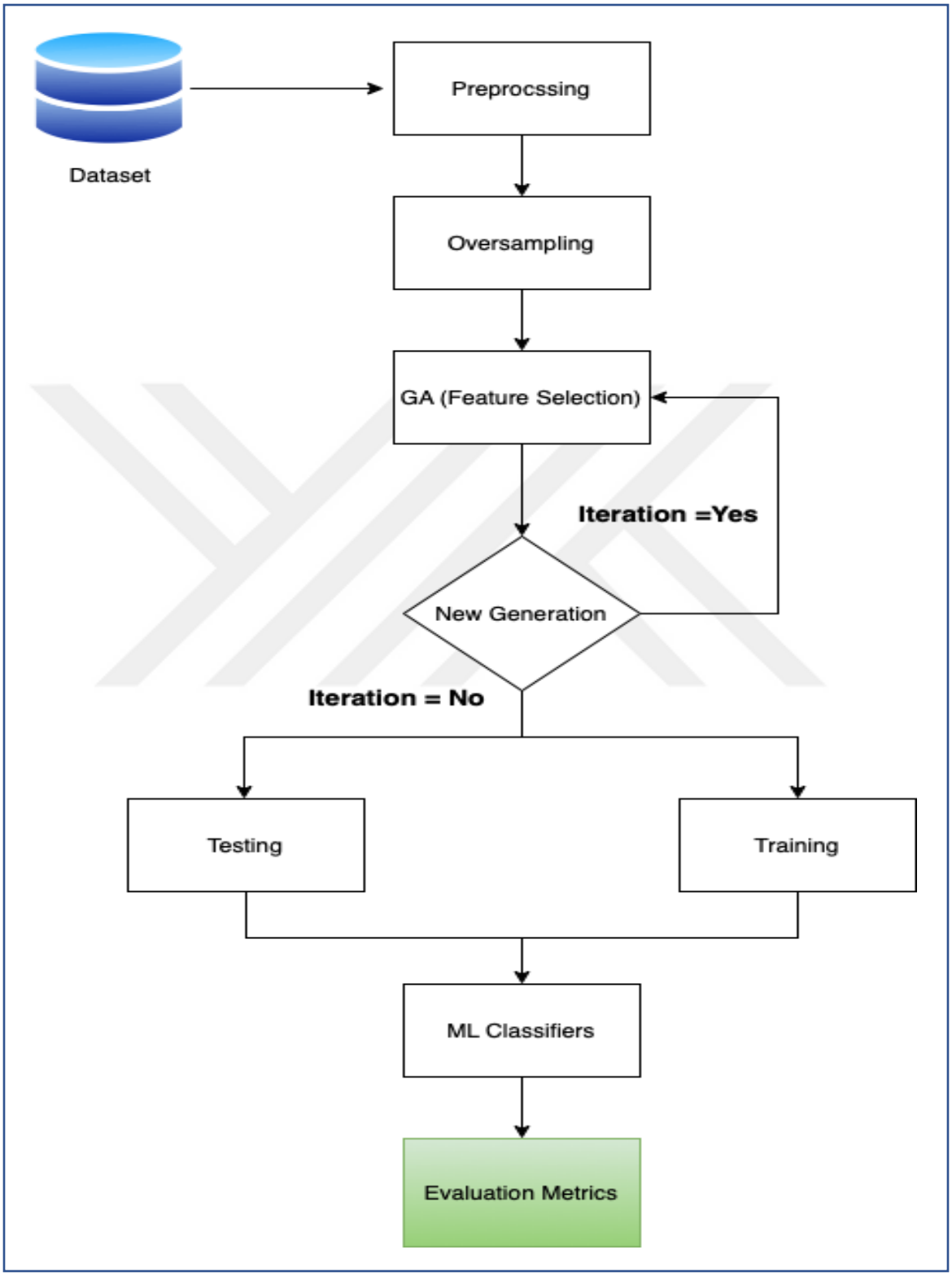

Figure 3. 1 Basic learning process for developing predictive models.

3.3 DATA DESCRIPTION

The COVID-19 data used in this study were collected from (5,644) patients at Albert Einstein Hospital in Brazil. Best practices and standards have been used to completely anonymize data [34]. Clinical data have previously been normalized to obtain a uniform distribution. The data contains blood test results for all patients tested for Sars-CoV-2: positive and negative. This publicly available data contains 111 properties from 5,644 people, including numerous medical tests. However, the data set is highly imbalanced, with relatively few positive patients compared to a large number of negative patients. As shown in Table 3.1, the missing value was processed, and then the remaining data was processed by over-sampling method to balance the data

Table 3. 1 Comparative between non-patients and patients in COVID-19 dataset.

Dataset	Positive	Negative	Total
Original	99	501	600
SMOTE	501	501	1002

3.4 DATASET PREPROCESSING

Data preprocessing is a collection of sub-operations in the form of several processes that may be performed on datasets for analysis and formation [35]. Data pre-processing in this research involves cleaning data, oversampling, and Feature selection. Cleaning data used to reduces mistakes and distortion from the dataset. In addition, the oversampling: The SMOTE approach is used to overcome the problem of unbalanced by randomly generating new-samples from minority data-samples, and their neighbors. Another ML strategy is feature selection, which decreases the diversity of inputs (attributes) to choose features that may be most significant in building the model. As a consequence, the normalization approach is used to convert all the values into a certain distribution in order to identify a stronger relationship within data.

The COVID dataset has a large number of missing data, If given as input all at once, this might cause an error. To do this, we have omitted any features with missing values greater than 90% to ensure we have minimum metrics for all features. Then, as in previous studies on this dataset, we eliminated the records with high missing data [26], reducing the patient number from 5,644 patients to 600 patients. Text data was another issue in this dataset since machine learning employs mathematical equations, which require all data provided to be numeric inputs, thus we labeled the categorical variables with a unique number for each record in the features.

3.5 OVERSAMPLING TECHNIQUE

Over-sampling method, quite different from the under-sampling method, in which majority class samples are not evaluated but instead minority samples are optimized to improve classification performance. Synthetic-minority oversampling technique (SMOTE) [36] creates a new minority sample by interpolating a homogeneous, randomly selected surrounding sample, which increases the minority sample's discovery rate. As a result of the criteria used to produce additional samples, overlapping noise data, boundary data, and other difficulties may occur. As a remedy to the challenges outlined above, cross-validation utilizing the SMOTE approach has been proposed.

Combining K-means and SMOTE and combining K-means and SMOTE was proposed by the author [37]. To combine samples from minority groups, K-means clustering was applied. Following that, very safe samples are sorted using SMOTE's linear combination approach, which properly handles sample noise.

Fuzzy-Firefly SMOTE was proposed in [38] utilizing fuzzy groups to aggregate minority classes and then using the Firefly technique to create minority samples. Statistical tests and assessments have revealed that it is more efficient than other over-sampling strategies. Hyper SMOTE-RF also presents a novel technique. The authors of suggest a set of redundancy sampling procedures that are primarily based on the feature area and extract the remaining community functions using

sliding and oversampling windows in the characteristic region. Finally, the biological data is categorized using the learned RF model.

The most common reason for imbalanced data in health records is that there are fewer non-patients than there are patients. This issue can prevent machine learning from performing proper calculations. As in our training data, the non-patients data make up a significant majority of classes, resulting in a bias in classification accuracy that influences non-patients. To solve this problem oversampling with SOMTE is used. SOMTE is a complex machine learning strategy that solves this problem by randomly producing additional samples between minority group samples and their neighbors, increasing the number of minority samples class to be balanced [39].

3.6 GENETIC ALGORITHM

Genetic Algorithm (GA): An optimization technique inspired by natural selection. This algorithm is based on population search and applies the idea of survival of the fittest [15]. New-Populations are formed by continually applying genetic factors to members of an existing population.

GA process begins with the initialization of the population, selecting a certain number of chromosomes at random. Then, fitness functions are used to calculate each chromosome produced by the population. In the selection method, two chromosomes are selected based on score fitness [15]. Crossover is the most important step in GA, the two chromosomes (parents) that were produced from the selection method are mated by their genes to create new offspring. Then, the mutation will be used randomly for some genes in the new offspring to maintain population diversity. The new offspring acquired by mutation is represented as a new population. Following are the GA procedures:

3.6.1 Selection Techniques

Selection process: It is a critical stage in GA that determines whether or not certain chromosomes will participate in the reproductive process based on their high fitness. The selection process is

also called the reproduction process [40]. There are many well-known strategies used in the selection process:

The wheel is then randomly turned to identify particular solutions which will be employed in the next generation's building [40]. However, it has several drawbacks, including inaccuracies caused by the randomness of the algorithm. The roulette wheel selection technique was developed by (Brindle and De Jong) by including the idea 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 [40].

Tournament selection: introduced by Brindle In 1983 for the first time. Individuals are chosen in pairs using a stochastic roulette wheel based on their fitness scores. The individuals with the highest fitness value are added to the next generation group after selection. If an individual succeeds in reaching the final set of solutions, it is compared with all other individuals [40].

The SUS (stochastic universal sampling) approach is a replacement for the current roulette-wheel selection mechanism. At equally spaced intervals, it chooses a new individual from a list of people from the same generation [41]. It assures that everyone has an equal opportunity of being chosen to participate in the generational crossover. Although SUS outperforms traditional Roulette wheel selection in the Traveling Salesman Issue [42], it outperforms SUS when the problem size grows.

3.6.2 Crossover Operators

Crossover operators combine the specific genes from one or more parents to produce offspring. A single point crossover and a uniform crossover are two well-known crossover operators. Single point: [43] randomly detects a single crossover point before dividing the parents at this crossing

point, producing offspring by exchanging the tail. The chance of a common crossover is within a given range. The n crossover points are chosen at random, after which they separate along those lines and reconvene, rotating between parents. Figure 3.2 shows the genetic information following crossover.

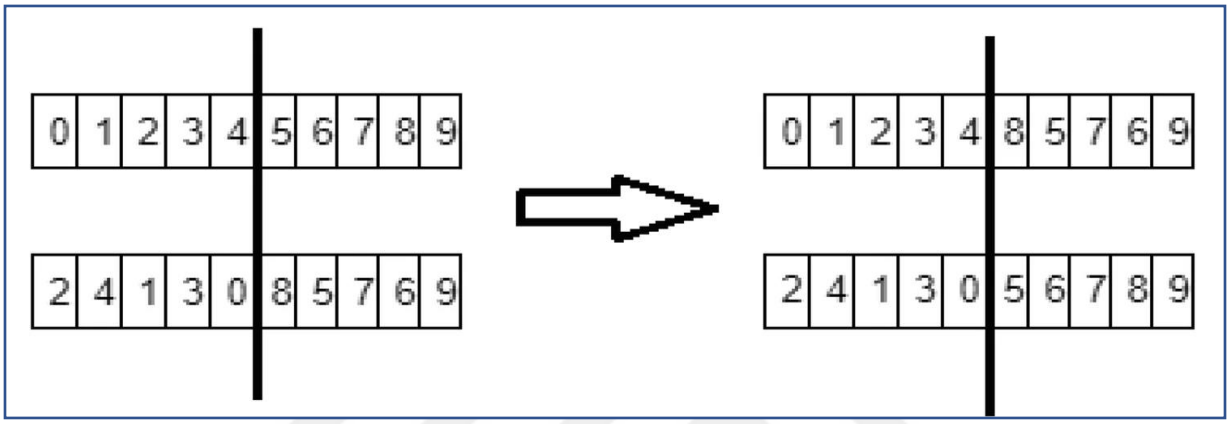

Figure 3. 2 Swapping before and after uses crossover points [15].

Uniform crossover: assigns (Heads) to one parent and (Tails) to the other. Where, for each gene in the first child, a coin is flipped, and an inverted copy of the gene is made for the second child. Inheritance is not dependent on rank. Figure 3.3 shows the swapping of the uniform crossover procedure.

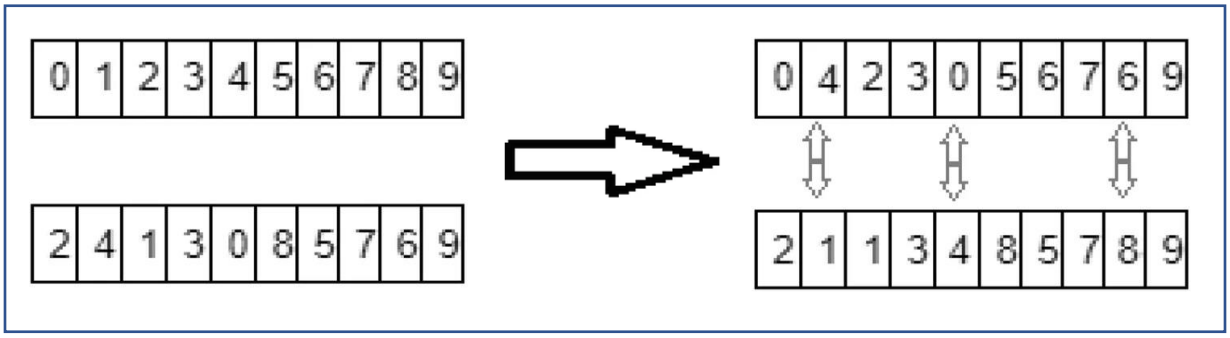

Figure 3. 3 Swapping genes information in Uniform crossover [15].

3.6.3 Mutation Operators

A mutation is a genetic diversity-maintaining operator that happens when one population passes its genetic diversity to the next. Translocation, simple inversion, and scramble mutation are three well-known mutation causes. Within a single solution, the displacement mutation (DM) changes a substring. The offset location is picked at random from the substring given, guaranteeing that both the final solution and the random offset mutation are legitimate. The exchange mutation and the insertion mutation are two types of DM alterations. Using exchange mutation and insertion mutation factors [40], a part of the individual solution is replaced by another segment or moved. The SIM operator (Simple inversion-mutation operator) in one solution reverses the sub-chain between any two defined locations. SIM is a string inverting agent that reverses a string and positions it at random [40].

3.7 FEATURE SELECTION

A fundamental part of any machine learning process is feature selection. However, there is a large percentage of data available these days. Several redundant and useless features often affect the classification performance of high-dimensional datasets [44]. In this situation, feature selection becomes critical. Feature selection aims to find a feature subspace that maintains classification accuracy while lowering the learning model's high computing cost and removing noise. The capacity of a feature selection technique to match the issue framework and find the fundamental patterns within the data is critical to its applicability.

Genetic algorithms employ an evolutionary technique to find the best set. The first stage in feature selection is to create a population from subsets of the available features [45]. A prediction model for the target task is used to evaluate the subgroups from this population. After each member of the population has been taken into account, a tournament is held to choose which subgroups will survive until the following generation. The tournament winners make form the next generation, with some cross-over (winning groups update) and mutation (randomly remove some features).

If the full data set, which contains a large number of characteristics, is utilized in this study, the calculations will be difficult and time-consuming. As a result, a future selection strategy is employed as an essential procedure for pre-processing. As a result, the amount of features must be reduced in order to obtain the most important features that may be employed as a first step in the model's learning process. Therefore, the results will be easier to understand, the overall efficiency of the task is improved, and the classification accuracy is improved.

3.8 MACHINE LEARNING

Machine learning is used to create complicated models and extract medical information, providing new insights for clinicians and specialists [45, 46]. In clinical practice, predictive machine learning models can highlight improved rules in inpatient care decision-making. They are also able to independently diagnose various diseases, provided clinical guidelines are followed [48]. According to [49], the use of these models in drug prescribing could save physicians time and open up new medical possibilities in the detection of pathologies. In this study, three ML models were used to classify the COVID-19 dataset:

Random Forest (RF) is a supervised machine learning technique, it is also an efficient and simple method, as illustrated in Figure 3.4. The random forest approach generates decision trees from randomized chosen data samples, extracts predictions data from each tree, and decides on the best solution, by eliminating overfitting through result averaging, it outperforms a single decision tree. [50]. The random forest technique produces great results because trees defend one other from individual faults. While individual trees may provide wrong responses, several others can lead the trees to the correct conclusion as a collective [51].

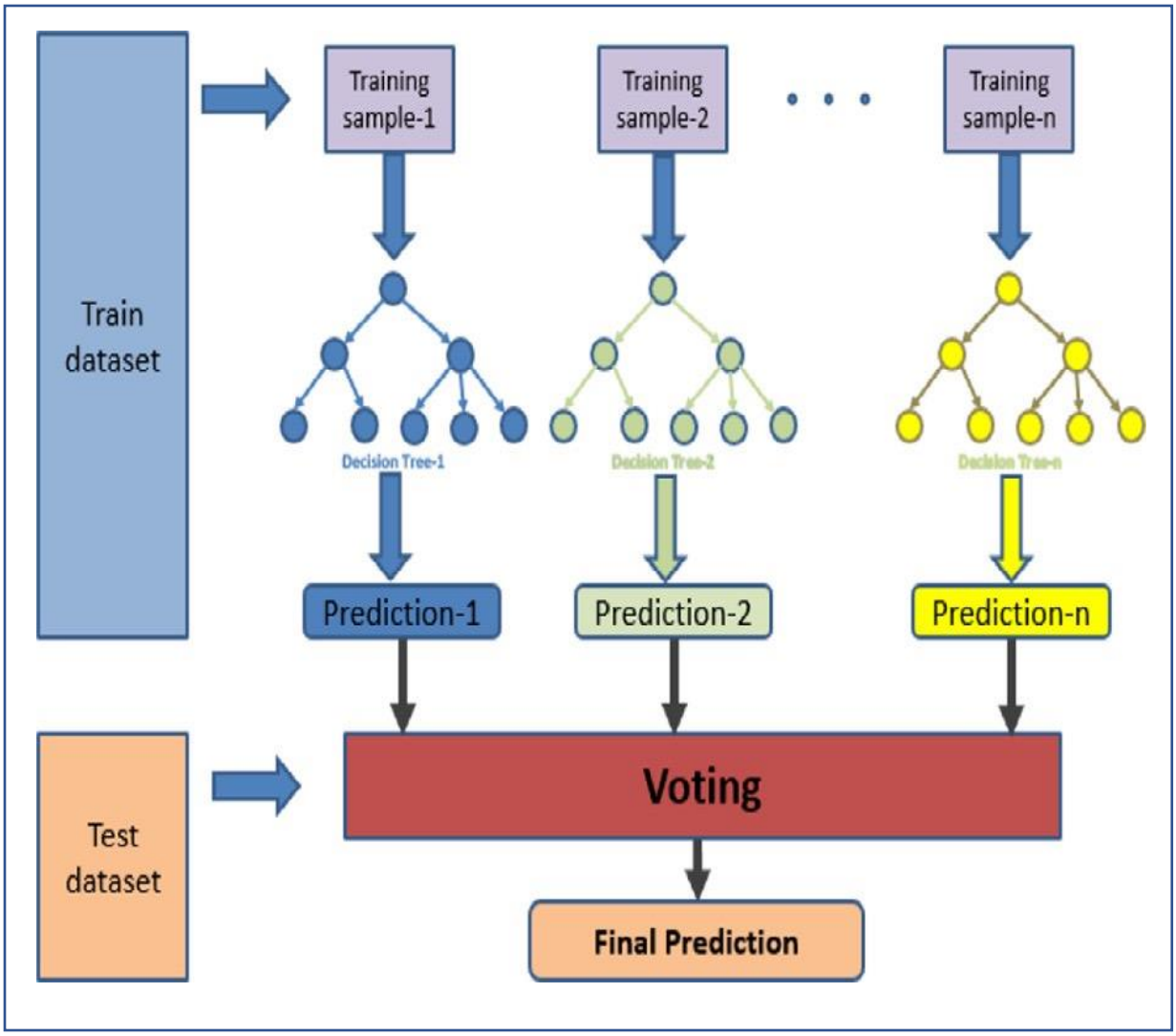

Figure 3. 4 Random Forest Method [52].

Multilayer perceptron (MLP): is a kind of artificial neural network that is fed forward. A perceptron is a binary classifier method in its simplest form. MLP is a multi-perceptron classification method that can answer more difficult tasks [53]. MLP arose as a consequence of research into the XOR problem. At least three layers make up the MLP algorithm: the input layer, the hidden layer, and the output layer. Sensors utilize various weights for each input signal. A different output is shown by each connection linking a sensor in one-layer to the next-layer.

Decision Tree (DT): is a member of the family of supervised learning algorithms. It is a structural method similar to a decision tree, which depicts groups of options that contribute to those ratings, with reflective ratings or judgments for leaves ANd branches [54] , as shown in Figure 3.5. The highest node of the choice tree is that the root node. He learns the way to divide by the worth of an attribute. A tree is formed by making an input value through a tree that starts at the base node and ends at the leaf.

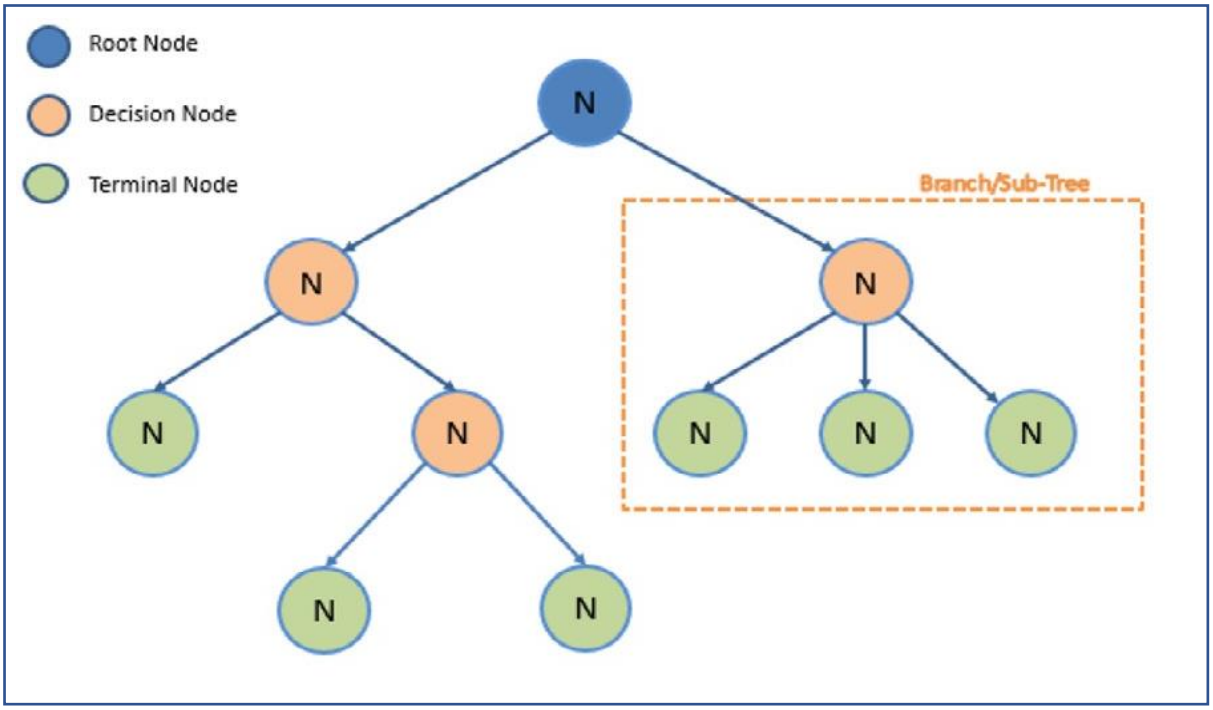

Figure 3. 5 **Decision Tree Method [52].**

3.9 NORMALIZING

The different distribution of the feature values in the COVID-19 dataset causes noise in the classification performance. As a result, the dataset [19] is normalized to place it in a homogeneous range between [0,1]. The following equation is used to compute normalization:

$$f_i^{new} = \frac{f_i^{old} - \min(F)}{\max(F) - \min(F)} \quad (1)$$

The function's current value is fi and the characteristic curves "minimum and maximum" values are min (F) and max (F).

3.10 EVALUATION METRICS

In this work, we employed a variety of measures to assess the efficacy of each built predictive model based on several supervised machine learning algorithms [55]. The evaluation metrics employ a confusion matrix to predict accurately and inaccurately evaluated outcomes by employing several classification metrics:

i) True Positive (TP): The total number of positive patients that are accurately classified as positive.
ii) False Positive (FP): The total number of positive patients that are misclassified as negative.
iii) True Negative (TN): The total number of negative patients classified as negative.
iv) False Negative (FN): The total number of patients misclassified as negative.

The aforementioned four evaluation metrics are used to evaluate the classification results:

Accuracy **:** An accuracy ratio scale is a useful tool for assessing algorithm execution activity for approaches employed before and after processing. The accuracy ratio scale is one of the most essential ways to assess algorithm performance before and after processing. (TP + TN) is the scale of correctness and error ratios. Using the following equation, we can calculate the accuracy:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} .100 \quad (2)$$

Precision**:** It is a scale used to measure the percentage of correct samples (TP) compared to false samples (FP). This scale is affected by the amount of unbalanced data, which is reflected in the results of the following equation:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} . 100 \quad (3)$$

Recall**:** A crucial measure for determining the number of right data predictions and the impact of information imbalance on the outcomes. The recall is calculated using Equation (3):

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (4)$$

F1 score: It is an important metric for determining the model's usefulness and efficiency, as well as the correctness and validity of the outputs. It is also said to be one of the most important indicators for ensuring data balance and maximizing efficiency. The F1 score is computed using Equation (4):

$$\text{Score} \frac{\text{Precision} . \text{Recall}}{\text{Precision} + \text{Recall}} . 2 \quad (5)$$

All of these experimental metrics are insufficient for assessing students in imbalanced data sets. Accuracy is also a deceptive evaluation criterion that biases the majority class and predicts who were in the minority. We want a broader assessment taking into consideration many characteristics for classification issues involving an imbalanced distribution of classes, such as assessing a classifier’s ability to balance between two classes and perceiving the two classes in the same way [19]. The "Area under the curve (AUC)" is utilized in our experiment for the first time since it displays stability in the face of imbalanced distribution of data.

In binary classification, the AUC score is used to choose the optimal model for class predictions. The AUC score is calculated by dividing the amount of the TP rate by the amount of the FP rate.

CHAPTER FOUR

4. RESULTS AND DISCUSSION

4.1 EXPERIMENTAL RESULTS FOR COVID-19

Covid-19, like other medical datasets, has an unbalanced dataset problem, which causes classification and results to be skewed, thus we use a new approach to train ML classification model. To boost the statistical power of models, we deleted the high missing data. To get around most classes in the dataset, use SMOTE oversampling. In addition, the GA feature selection approach was used, which removes features that have the least influence on the data and minimizes training cost while finding the most distinct features.

In this study, Decision Tree, MLP, and RF Classifiers were utilized to determine whether or not patients have disease. We investigated the evaluation of all ML models by dividing the data into 80-20 with training test split technique, which resulted in remarkable overall prediction performance. We modified the GA using a trial-and-error method to identify the ideal parameters and features using a Decision Tree Classifier.

Table 4.1 displays the overall performance of the original dataset. Our machine learning models employed 80% of the data set to train and 20% for evaluating the classifiers throughout the learning process. The remove missing values method used on the original dataset, resulting in a reduction of 5,644 patients to 600. DT and RF offered the best evaluation results, with 88% accuracy, respectively.

Table 4. 1 The outcomes of testing the classifier using the original data set

Method	Accuracy	Precision	Recall	F1-score
DT	88%	90%	88%	89%
RF	86%	87%	86%	86%
MLP	88%	86%	88%	86%

Through a trial-and-error approach, different settings of the experiments were used to find the optimal parameters for GA. In GA different parameters are tuned such as (I) population size, (II) crossover, (III) mutation, and (IV) Iteration. Population feature size was set at three different sizes (40, 60, 80), the best crossover numbers (0.4, 0.6), the best mutation set at (0.2, 0.4), and also the fourth GA parameter (Iteration) has been set to 10. In Table 4.2 the GA parameters that we used in this study are detailed.

Table 4. 2 Genetic algorithm parameters with optimal feature selection.

Max Iteration	Populations	Crossover	Mutation	Selected	Accuracy
10	40	0.4	0.2	35	96
10	40	0.6	0.2	26	96
10	40	0.6	0.4	25	97
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

After preprocessing, the best accuracy rates were achieved by the three classifiers, respectively. As a consequence, the suggested technique significantly improved the accuracy rate of the three classifiers. The total performance results showed that the proposed method works well with different classification models. Table 4.3 compares the predicted accuracy metrics produced from each COVID-19 model. As part of preprocessing procedure in this investigation, oversampling and GA feature selection were done to the dataset. Furthermore, the classification performance outcomes of all models were thoroughly reviewed. As a result, the proposed COVID-19 diagnostic method can assist medical practitioners in accurately identifying COVID-19 patients.

Table 4. 3 Results of performance evaluations for all ML models.

Method	Accuracy	Precision	Recall	F1-score
DT	97%	97%	97%	97%
RF	98%	98%	98%	98%
MLP	96%	96%	96%	96%

4.2 COVID-19 ANALYSIS RESULTS

Data are growing very faster every day and it is not possible to process the data manually. Data analysis programs allow understanding the information in deep. Data analysis is used in machine learning to determine the performance of the model. The result of the model can decide whether the model reaches a good performance or not. The confusion matrix and AUC are employed in this work to evaluate the effectiveness of our suggested technique.

A Confusion-matrix is a table that summarizes the effectiveness of a rating scheme. The number of rights and unsuccessful predictions is divided into groups and summarized by a numerical value. As shown in Figure 4.1.

The AUC rating is most commonly employed in binary classification tasks to find a good model for class classification. The AUC rating is calculated by dividing the total of true positives by the false - positive results. It implies that a higher rate is more accurate in class prediction. As a consequence, a good ratio is between (From 0.8 and 0.9), with a value greater than 0.9 is preferable. The RF method was proven to be the most successful in this study, and the result was 100%. Furthermore, the scores for the other methods were excellent, with all of them scoring higher than 97 %. Machine learning algorithms may be used to predict COVID-19 based on AUC values. As a result, the AUC degree is critical in medical science since it shows how to distinguish between sick and healthy people. Figure 4.2 displays an AUC analysis graph for ML algorithms.

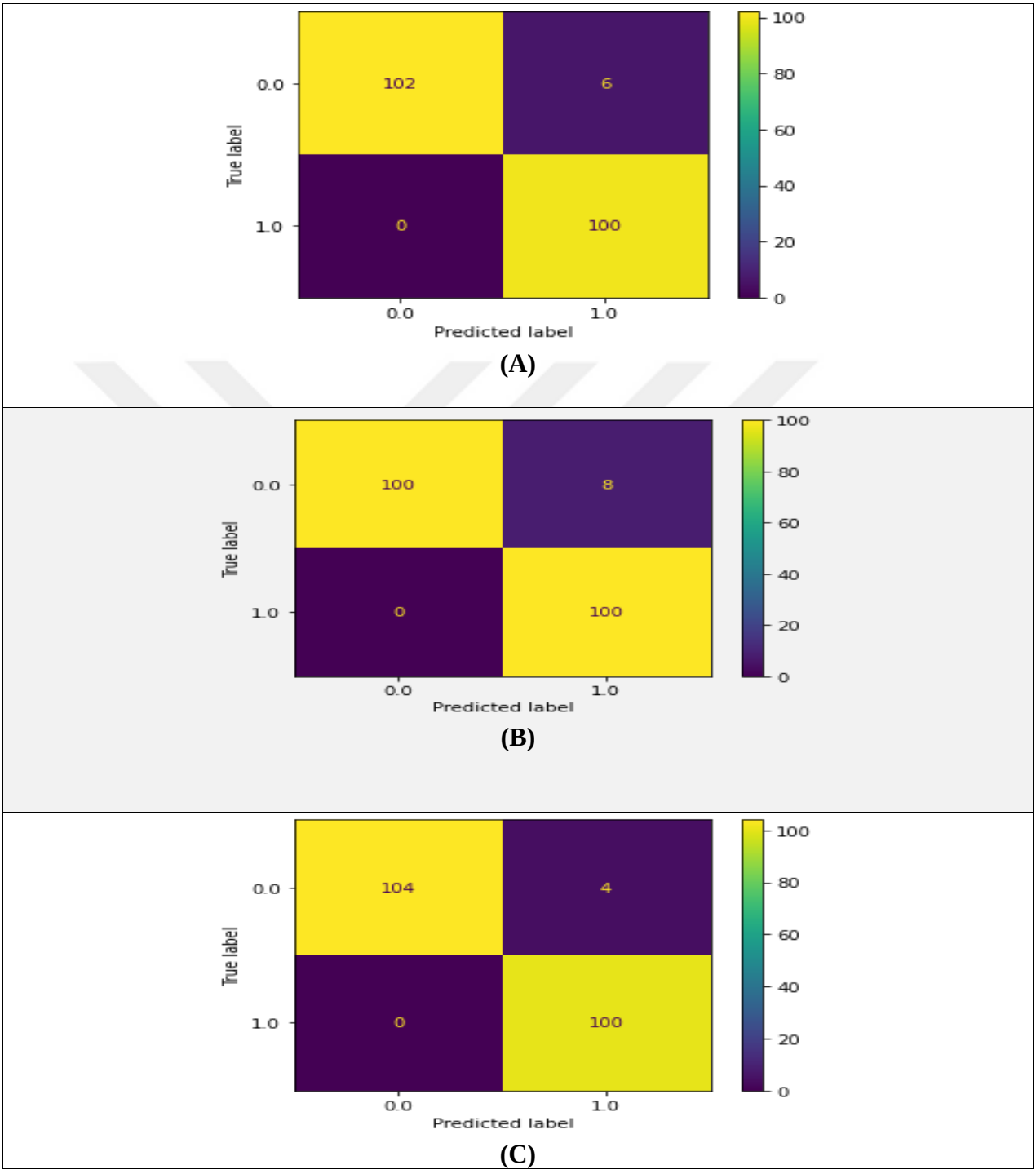

Figure 4. 1 Confusion matrix. Three machine learning are used confusion matrix (A) Decision tree method (B) MLP method, (C) RF method. The diagonal matrix for each method's outcome is represented by letters A through C, whereas the RF method produces the best results.

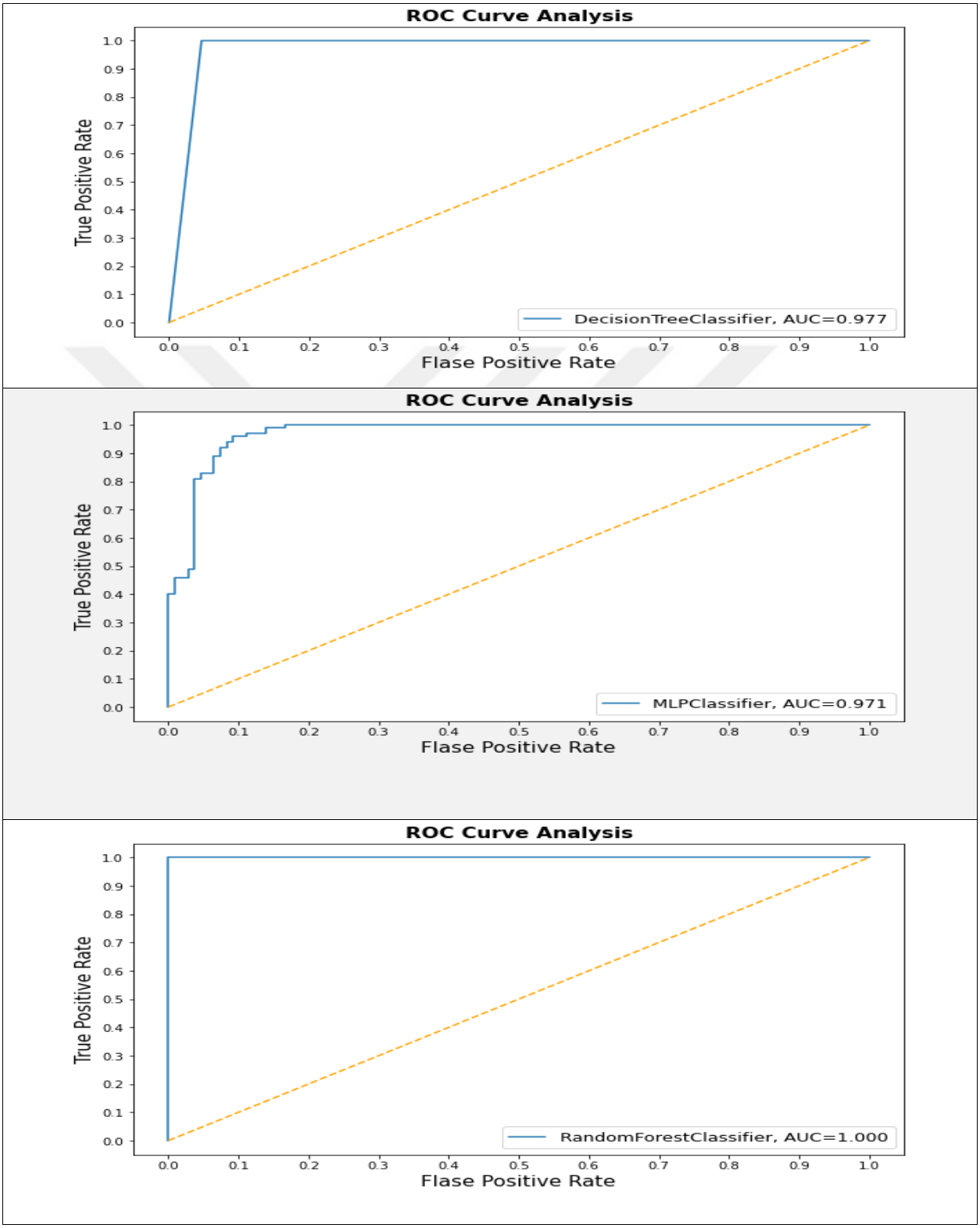

Figure 4. 2 AUC analysis for all ML models.

Table 4. 4 Comparison of evaluation results.

Study	Method	Accuracy	F1-Score	AUC	Publication Year
Proposed	**RF**	**98.2%**	**98.1%**	**99.9%**	**-**
Mondal et al. [22]	MLP	93%	-	-	Apr/ 2020
Batista et al. [23]	SVM	80%	-	85%	Jul/ 2020
Alakus and Turkoglu [26]	CNN LSTM	92%	93%	90%	July/ 2020
Prabhu et al. [28]	ML	91%	-	91%	Feb/ 2022

The outcomes of a comparison of this research with previous researchers are summarized in Table 4.4. In research [22] ,[23] and [28] the authors proposed different machine learning methods to predict and analyze the outcome of COVID-19 Patients. In another research [26] the author proposed several deep learning models to enhance the prediction with COVID-19 Patients best result was obtained with the hyper CNN-LSTM model.

CHAPTER FIVE

5. CONCLUSION AND FUTURE WORK

5.1 CONCLUSION

COVID-19 was proclaimed a worldwide pandemic after its first confirmed appearance in China. The disease had spread to more than 200 countries and regions, with the United States having the most cases reported worldwide. China's condition has improved since early March 2020 as a result of a variety of efforts, including rigorous quarantine laws and travel restrictions. COVID-19 has no effective therapy at the moment; existing medications just address the symptoms. ML is used to accurately detect patients with COVID-19. In this study, different ML methods are applied, first oversampling applied to overcome the imbalanced dataset, where the data set has a fivefold large distribution with a negative class (non-patients), and this results in an inaccurate classification metrics of the positive class (patients). For Feature selection GA applied to remove irrelevant features and reduce processing training-time. A trial-and-error approach was used to adjust different parameters of GA to reach the optimal feature for the COVID-19 dataset. Since ML is based on the mathematical equation the text data are converted to the numeric by using the label encoding method. Also, data are normalized in a certain range to improve performance. To validate our overall model performance, the data was divided into (80%) training and (20%) test sets, with different evolution metrics such as precision,recall , F1 score and accuracy. A rating model prediction tool with the RF method was selected as the best accuracy metric with 98%. While with DT we got 97% accuracy as a second-best model, also, MLP has been selected as the third model with accuracy reached to 96% and 97% AUC for both DT and MLP. The main limitation of this study is the limited number of patients included in the data, as well as some gaps in laboratory findings. Furthermore, the information got imbalanced, so we balanced it using the oversampling method.

5.2 FUTURE WORKS

Future works will focus on the following areas:

a) It might create a new generation of oversampling techniques for working with different dimensions of data.
b) A newer COVID-19 dataset can be used instead of an out-of-date dataset obtained for testing reasons.
c) Extending the suggested approach to other classification systems, such as disease diagnosis, with a corresponding modification in the dataset employed.
d) Using other evolutionary or metaheuristic approaches, such as PSO or ABC, and ACO.

REFERENCES

[1] N. Perra, "Non-pharmaceutical interventions during the COVID-19 pandemic: A review," *Physics Reports*, vol. 913. Elsevier B.V., pp. 1–52, May 23, 2021. doi: 10.1016/j.physrep.2021.02.001.

[2] C. Sohrabi *et al.*, "World Health Organization declares global emergency: A review of the 2019 novel coronavirus (COVID-19)," *International Journal of Surgery*, vol. 76. Elsevier Ltd, pp. 71–76, Apr. 01, 2020. doi: 10.1016/j.ijsu.2020.02.034.

[3] C. Jiehao *et al.*, "A Case Series of Children With 2019 Novel Coronavirus Infection: Clinical and Epidemiological Features," *Clinical Infectious Diseases*, vol. 71, no. 6, pp. 1547–1551, Sep. 2020, doi: 10.1093/CID/CIAA198.

[4] M. I. Arshad, H. A. Khan, B. Aslam, and J. A. Khan, "Appraisal of One Health approach amid COVID-19 and zoonotic pandemics: insights for policy decision," *Tropical Animal Health and Production*, vol. 53, no. 1. Springer Science and Business Media B.V., Dec. 01, 2021. doi: 10.1007/s11250-020-02479-0.

[5] "Coronavirus disease (COVID-19)." https://www.who.int/emergencies/diseases/novel-coronavirus-2019 (accessed Dec. 13, 2020).

[6] "Field Briefing: Diamond Princess COVID-19 Cases." https://www.niid.go.jp/niid/en/2019-ncov-e/9407-covid-dp-fe-01.html (accessed May 17, 2021).

[7] T. G. Dietterich, "Ensemble methods in machine learning," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2000, vol. 1857 LNCS, pp. 1–15. doi: 10.1007/3-540-45014-9_1.

[8] O. Avci, O. Abdeljaber, S. Kiranyaz, M. Hussein, M. Gabbouj, and D. J. Inman, "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*, vol. 147. Academic Press, p. 107077, Jan. 15, 2021. doi: 10.1016/j.ymssp.2020.107077.

[9] G. Li and E. De Clercq, "Therapeutic options for the 2019 novel coronavirus (2019-nCoV)," *Nature Reviews Drug Discovery 2021 19:3*, vol. 19, no. 3, pp. 149–150, Feb. 2020, doi: 10.1038/d41573-020-00016-0.

[10] C. Zhang *et al.*, "Multi-Imbalance: An open-source software for multi-class imbalance learning," *Knowl Based Syst*, vol. 174, pp. 137–143, Jun. 2019, doi: 10.1016/J.KNOSYS.2019.03.001.

[11] M. Sokolova and G. Lapalme, "A systematic analysis of performance measures for classification tasks," *Inf Process Manag*, vol. 45, no. 4, pp. 427–437, Jul. 2009, doi: 10.1016/J.IPM.2009.03.002.

[12] J. Shreve, H. Schneider, and O. Soysal, "A methodology for comparing classification methods through the assessment of model stability and validity in variable selection," *Decis Support Syst*, vol. 52, no. 1, pp. 247–257, Dec. 2011, doi: 10.1016/J.DSS.2011.08.001.

[13] M. Elhoseny *et al.*, "A new multi-agent feature wrapper machine learning approach for heart disease diagnosis," *Computers, Materials and Continua*, vol. 67, no. 1, pp. 51–71, 2021, doi: 10.32604/cmc.2021.012632.

[14] V. Kumar, J. K. Chhabra, and D. Kumar, "Parameter adaptive harmony search algorithm for unimodal and multimodal optimization problems," *J Comput Sci*, vol. 5, no. 2, pp. 144–155, Mar. 2014, doi: 10.1016/J.JOCS.2013.12.001.

[15] S. Katoch, S. S. Chauhan, and V. Kumar, "A review on genetic algorithm: past, present, and future," *Multimed Tools Appl*, vol. 80, no. 5, pp. 8091–8126, Feb. 2021, doi: 10.1007/S11042-020-10139-6/FIGURES/8.

[16] A. Chakraborty and A. K. Kar, "Swarm Intelligence: A Review of Algorithms," *Modeling and Optimization in Science and Technologies*, vol. 10, pp. 475–494, 2017, doi: 10.1007/978-3-319-50920-4_19.

[17] N. B. Bahadure, A. K. Ray, and H. P. Thethi, "Comparative Approach of MRI-Based Brain Tumor Segmentation and Classification Using Genetic Algorithm," *Journal of Digital Imaging 2018 31:4*, vol. 31, no. 4, pp. 477–489, Jan. 2018, doi: 10.1007/S10278-018-0050-6.

[18] A. Sohail, "Genetic Algorithms in the Fields of Artificial Intelligence and Data Sciences," *Annals of Data Science 2021*, pp. 1–12, Aug. 2021, doi: 10.1007/S40745-021-00354-9.

[19] H. Faris *et al.*, "Improving financial bankruptcy prediction in a highly imbalanced class distribution using oversampling and ensemble learning: a case from the Spanish market," *Progress in Artificial Intelligence*, vol. 9, no. 1, pp. 31–53, 2020, doi: 10.1007/s13748-019-00197-9.

[20] L. J. Muhammad, E. A. Algehyne, S. S. Usman, A. Ahmad, C. Chakraborty, and I. A. Mohammed, "Supervised Machine Learning Models for Prediction of COVID-19 Infection using Epidemiology Dataset," *SN Comput Sci*, vol. 2, no. 1, pp. 1–13, Feb. 2021, doi: 10.1007/S42979-020-00394-7/FIGURES/11.

[21] X. Jiang *et al.*, "Towards an artificial intelligence framework for data-driven prediction of coronavirus clinical severity," *Computers, Materials and Continua*, vol. 63, no. 1, pp. 537–551, Mar. 2020, doi: 10.32604/cmc.2020.010691.

[22] M. R. H. Mondal, S. Bharati, P. Podder, and P. Podder, "Data analytics for novel coronavirus disease," *Inform Med Unlocked*, vol. 20, p. 100374, Jan. 2020, doi: 10.1016/J.IMU.2020.100374.

[23] B. Afm *et al.*, "COVID-19 diagnosis prediction in emergency care patients: a machine learning approach," *medRxiv*, p. 2020.04.04.20052092, Apr. 2020, doi: 10.1101/2020.04.04.20052092.

[24] C. Iwendi *et al.*, "COVID-19 patient health prediction using boosted random forest algorithm," *Front Public Health*, vol. 8, no. July, pp. 1–9, 2020, doi: 10.3389/fpubh.2020.00357.

[25] M. A. Khan, S. Abbas, K. M. Khan, M. A. A. Ghamdi, and A. Rehman, "Intelligent forecasting model of covid-19 novel coronavirus outbreak empowered with deep extreme learning machine," *Computers, Materials and Continua*, vol. 64, no. 3, pp. 1329–1342, 2020, doi: 10.32604/cmc.2020.011155.

[26] T. B. Alakus and I. Turkoglu, "Comparison of deep learning approaches to predict COVID-19 infection," *Chaos Solitons Fractals*, vol. 140, p. 110120, Nov. 2020, doi: 10.1016/J.CHAOS.2020.110120.

[27] Z. Li *et al.*, "From community-acquired pneumonia to COVID-19: a deep learning–based method for quantitative analysis of COVID-19 on thick-section CT scans," *Eur Radiol*, vol. 30, no. 12, pp. 6828–6837, Dec. 2020, doi: 10.1007/s00330-020-07042-x.

[28] K. Chadaga, C. Chakraborty, S. Prabhu, S. Umakanth, V. Bhat, and N. Sampathila, "Clinical and Laboratory Approach to Diagnose COVID-19 Using Machine Learning," *Interdiscip Sci*, vol. 14, no. 2, pp. 452–470, Jun. 2022, doi: 10.1007/S12539-021-00499-4/FIGURES/9.

[29] J. Tanha, Y. Abdi, N. Samadi, N. Razzaghi, and M. Asadpour, "Boosting methods for multi-class imbalanced data classification: an experimental review," *J Big Data*, vol. 7, no. 1, pp. 1–47, Dec. 2020, doi: 10.1186/S40537-020-00349-Y/FIGURES/5.

[30] Y. Abdi, S. Parsa, and Y. Seyfari, “A hybrid one-class rule learning approach based on swarm intelligence for software fault prediction,” *Innovations in Systems and Software Engineering 2015 11:4*, vol. 11, no. 4, pp. 289–301, Sep. 2015, doi: 10.1007/S11334-015-0258-2.

[31] J. Bi and C. Zhang, “An empirical comparison on state-of-the-art multi-class imbalance learning algorithms and a new diversified ensemble learning scheme,” *Knowl Based Syst*, vol. 158, pp. 81–93, Oct. 2018, doi: 10.1016/J.KNOSYS.2018.05.037.

[32] J. L. Leevy, T. M. Khoshgoftaar, R. A. Bauder, and N. Seliya, “A survey on addressing high-class imbalance in big data,” *J Big Data*, vol. 5, no. 1, pp. 1–30, Dec. 2018, doi: 10.1186/S40537-018-0151-6/TABLES/5.

[33] B. Abu-Salih *et al.*, “Time-aware domain-based social influence prediction,” *J Big Data*, vol. 7, no. 1, pp. 1–37, Dec. 2020, doi: 10.1186/S40537-020-0283-3/TABLES/11.

[34] “Diagnosis of COVID-19 and its clinical spectrum | Kaggle.” https://www.kaggle.com/einsteindata4u/covid19 (accessed Nov. 19, 2021).

[35] R. Geetha, S. Sivasubramanian, M. Kaliappan, S. Vimal, and S. Annamalai, “Cervical Cancer Identification with Synthetic Minority Oversampling Technique and PCA Analysis using Random Forest Classifier,” *Journal of Medical Systems 2019 43:9*, vol. 43, no. 9, pp. 1–19, Jul. 2019, doi: 10.1007/S10916-019-1402-6.

[36] D. Elreedy and A. F. Atiya, “A Comprehensive Analysis of Synthetic Minority Oversampling Technique (SMOTE) for handling class imbalance,” *Inf Sci (N Y)*, vol. 505, pp. 32–64, Dec. 2019, doi: 10.1016/J.INS.2019.07.070.

[37] G. Douzas, F. Bacao, and F. Last, “Improving imbalanced learning through a heuristic oversampling method based on k-means and SMOTE,” *Inf Sci (N Y)*, vol. 465, pp. 1–20, Oct. 2018, doi: 10.1016/J.INS.2018.06.056.

[38] H. Lee, J. Kim, and S. Kim, “Gaussian-Based SMOTE Algorithm for Solving Skewed Class Distributions,” *International Journal of Fuzzy Logic and Intelligent Systems*, vol. 17, no. 4, pp. 229–234, Dec. 2017, doi: 10.5391/IJFIS.2017.17.4.229.

[39] B. S. Raghuwanshi and S. Shukla, “SMOTE based class-specific extreme learning machine for imbalanced learning,” *Knowl Based Syst*, vol. 187, p. 104814, Jan. 2020, doi: 10.1016/J.KNOSYS.2019.06.022.

[40] K. Jebari and M. Madiafi, “Selection Methods for Genetic Algorithms Smart cities View project fuzzy clustering techniques View project Selection Methods for Genetic Algorithms,” *Int. J. Emerg. Sci*, vol. 3, no. 4, pp. 333–344, 2013.

[41] W. Abdulal and S. Ramachandram, “Reliability-aware genetic scheduling algorithm in grid environment,” *Proceedings - 2011 International Conference on Communication Systems and Network Technologies, CSNT 2011*, pp. 673–677, 2011, doi: 10.1109/CSNT.2011.145.

[42] S. Sharma and K. Gupta, “Solving the traveling salesmen problem through genetic algorithm with new variation order crossover,” pp. 274–276, Aug. 2012, doi: 10.1109/ETNCC.2011.6255903.

[43] K. G. Dhal, S. Ray, A. Das, and S. Das, “A Survey on Nature-Inspired Optimization Algorithms and Their Application in Image Enhancement Domain,” *Archives of Computational Methods in Engineering 2018 26:5*, vol. 26, no. 5, pp. 1607–1638, Sep. 2018, doi: 10.1007/S11831-018-9289-9.

[44] E. Alhenawi, R. Al-Sayyed, A. Hudaib, and S. Mirjalili, “Feature selection methods on gene expression microarray data for cancer classification: A systematic review,” *Comput Biol Med*, vol. 140, p. 105051, Jan. 2022, doi: 10.1016/J.COMPBIOMED.2021.105051.

[45] N. Maleki, Y. Zeinali, and S. T. A. Niaki, “A k-NN method for lung cancer prognosis with the use of a genetic algorithm for feature selection,” *Expert Syst Appl*, vol. 164, p. 113981, Feb. 2021, doi: 10.1016/J.ESWA.2020.113981.

[46] O. S. Tătaru *et al.*, “Artificial Intelligence and Machine Learning in Prostate Cancer Patient Management—Current Trends and Future Perspectives,” *Diagnostics 2021, Vol. 11, Page 354*, vol. 11, no. 2, p. 354, Feb. 2021, doi: 10.3390/DIAGNOSTICS11020354.

[47] T. Nakamura and T. Sasano, “Artificial intelligence and cardiology: Current status and perspective,” *J Cardiol*, vol. 79, no. 3, pp. 326–333, Mar. 2022, doi: 10.1016/J.JJCC.2021.11.017.

[48] P. Ghosh, S. Azam, A. Karim, M. Hassan, K. Roy, and M. Jonkman, “A Comparative Study of Different Machine Learning Tools in Detecting Diabetes,” *Procedia Comput Sci*, vol. 192, pp. 467–477, Jan. 2021, doi: 10.1016/J.PROCS.2021.08.048.

[49] Y. Bian and X. Q. Xie, "Generative chemistry: drug discovery with deep learning generative models," *Journal of Molecular Modeling 2021 27:3*, vol. 27, no. 3, pp. 1–18, Feb. 2021, doi: 10.1007/S00894-021-04674-8.

[50] M. Belgiu and L. Drăgu, "Random forest in remote sensing: A review of applications and future directions," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 114, pp. 24–31, Apr. 2016, doi: 10.1016/J.ISPRSJPRS.2016.01.011.

[51] Z. Masetic and A. Subasi, "Congestive heart failure detection using random forest classifier," *Comput Methods Programs Biomed*, vol. 130, pp. 54–64, Jul. 2016, doi: 10.1016/J.CMPB.2016.03.020.

[52] S. Al and M. Dener, "STL-HDL: A new hybrid network intrusion detection system for imbalanced dataset on big data environment," *Comput Secur*, vol. 110, p. 102435, Nov. 2021, doi: 10.1016/J.COSE.2021.102435.

[53] M. Desai and M. Shah, "An anatomization on breast cancer detection and diagnosis employing multi-layer perceptron neural network (MLP) and Convolutional neural network (CNN)," *Clinical eHealth*, vol. 4, pp. 1–11, Jan. 2021, doi: 10.1016/J.CEH.2020.11.002.

[54] M. M. Ghiasi and S. Zendehboudi, "Application of decision tree-based ensemble learning in the classification of breast cancer," *Comput Biol Med*, vol. 128, p. 104089, Jan. 2021, doi: 10.1016/J.COMPBIOMED.2020.104089.

[55] J. Liu *et al.*, "Overlapping and discrete aspects of the pathology and pathogenesis of the emerging human pathogenic coronaviruses SARS-CoV, MERS-CoV, and 2019-nCoV," *J Med Virol*, vol. 92, no. 5, pp. 491–494, May 2020, doi: 10.1002/JMV.25709.