Ilkogretim Online - Elementary Education Online, 2021; Vol 20 (Issue 3): pp. 4156-4162
http://ilkogretim-online.org
doi: 10.17051/ilkonline.2021.03.429

Ensemble Genetic Algorithms For Blood Covid-19 Diagnosis

Devvret Verma Department of Biotechnology, Graphic Era Deemed to be University, Dehradun, Uttarakhand India, 248002 devvret@geu.ac.in

Kumud Pant Department of Biotechnology, Graphic Era Deemed to be University, Dehradun, Uttarakhand India, 248002 kumud.pant@geu.ac.in

Poonam Verma Department of Computer Application Graphic Era Hill University, Dehradun, Uttarakhand India, 248002 pverma@gehu.ac.in

ABSTRACT

The state of the world is urgent at this time. The recently discovered corona virus causes Covid-19, a devastating pandemic disease. To date, there are 3.23 million ongoing cases in India, with 59,449 fatalities. There is a lot of effort being put in by scientists and doctors to find a cure and vaccination for this. Medical advancements are being aided by research in the areas of machine learning and artificial intelligence, which are being used to forecast the development of disease and to detect the existence of the virus in the human body. The researchers here hope to learn more about COVID-19 by examining the virus in human blood samples. The study's blood samples have more than a hundred different characteristics. Therefore, the genetic algorithm has been used for feature reduction in high-dimensional data processing. This research will use a genetic algorithm to predict the presence of COVID-19 in a blood sample. About 5,644 patients' records with 111 different characteristics are included in the sample. The dimensionality reduction technique will be based on a genetic algorithm, similar to that utilised in the optimization algorithm for ant colonies for disaster relief. The programming language python is used throughout this study, and the metrics sensitivity, specificity, accuracy, and area under the curve (AUC) are used to assess the model's efficacy. The applied model has a 92% AUC, a sensitivity of 96.76 percent, a specificity of 98.80 percent, and an accuracy of 98.7 percent. The results showed that the custom algorithm was superior to the best existing solutions.

I. INTRODUCTION

To this day, COVID-19 remains one of humanity's greatest dangers. Scientists from all over the world are desperately trying to find a way to stop the microscopic undetectable danger they pose. The field of computer science is making significant contributions to COVID-19

research alongside the field of medicine by speeding the analysis of crucial data. Disease early detection utilising machine learning-based classifiers and prediction models is also feasible. Here, we offer a machine learning algorithm for COVID-19 disease categorization using Logistic regression. The model's 92% accuracy was achieved by using data from Kaggle for training. COVID-19: Any virus in the coronavirus family has the potential to cause disease in humans. Some mild to severe respiratory symptoms may be experienced by infected people. An array of respiratory symptoms, such as fever, coughing, breathlessness, and difficulty breathing, are associated with COVID-19. It has been reported that this sickness can cause pneumonia, severe acute syndrome, renal failure, and even mortality. Comorbid illnesses, such as hypertension, cardiovascular disease, kidney disease, a blood infection, anxiety, and depression, have also been shown to impair a patients ' lives to the point where they require immediate medical attention. WHO predicts that by August 23, 2020, there will be approximately 1.7 million more instances of COVID-19 more than 39K more deaths over the world.

In comparison to time-consuming and expensive diagnostic methods like imaging and RT-PCR, the COVID-19 early-detection blood tests ML system will be quick, simple, inexpensive, and accessible (Tripathi and Agrawal, 2020). Large effects are expected to be felt in low- and middle-income countries where PCR test kits, laboratory supplies, and diagnostic facilities are few. This rapid, low-cost procedure for potentially infected people can help slow the spread of the pandemic, which is a major benefit. The following are some of the work's contributions:

- ✓ Developed an ensemble method for predicting
- ✓ Infection with the COVID-19 virus utilizing a genetic algorithm and a machine learning classifier.
- ✓ Area under the curve [AUC], Accuracy, sensitivity and specificity are used to evaluate the suggested model.
- ✓ Finally, the model's performance is assessed by contrasting it with that of previously established ML methods.

The following is the outline of the paper's structure: Work pertinent to the discussion is included in Section 2, while Sections 3 and 4 discuss the work to be done and how the results should be interpreted, respectively. Section 5 then discusses the conclusion and references.

I. LITERATURE SURVEY

The finding that researchers could classify new cases of COVID-19 only using 25 ml of serum from samples taken is a milestone that could aid the global effort to prevent the development of the COVID-19 population via rigorous interaction surveillance (Garg et al., 2020). The 2019 model coronavirus disease outbreak, caused by the severe acute

respiratory syndrome caused by coronavirus 2 (SARS-CoV-2), is fast spreading across the globe and has already resulted in a substantial number of deaths (Kumar et al., 2020). More over 2.3 million reported cases had been documented in 216 countries and territories as of July 11, 2020, and the epidemic had claimed the lives of more than 1.5 million people by that date (WHO, 2020). The global spread of the COVID-19 epidemic affected nearly every area of human existence. Since then, a lot of work has been done in every corner of the globe to pinpoint the source of the virus and find a cure as soon as possible. At the moment, the most reliable method for diagnosing COVID-19 is true polymerase chain reactions (RT-PCR) followed by DNA sequencing and identification. However, it takes time, money, and specialised equipment, and there is an about 20% possibility of getting a false null hypothesis (Kermali et al., 2020). Furthermore, RT-PCR test kits are not easily available in any part of the world.

Recent clinical studies have shown that COVID-19 has a profound impact on blood properties (Bao et al., 2020), highlighting the need of first COVID-19 testing in the diagnosis and management of this disorder (Arnold et al., 2020). A disease's presence or absence can be confirmed by a diagnostic test, whereas early screening can offer a probabilistic early prediction. This research (Guncar et al., 2018) found that it is not easy to acquire all the details from a person's normal blood testing. This means that routine blood test data may be used to train and test a variety of machine learning algorithms. For this reason, researchers have started working on methods to detect COVID-19 in routine blood samples (Brinati et al., 2020).

II. PROPOSED METHODOLOGY

Although various means of identification have been developed, this study is the first to focus on doing so using a blood sample from a patient with COVID-19 . To improve disease forecasting, we integrate ML with genetic analysis in this research. There are two stages to the analysis of the data. Following the use of evolutionary methods to reduce the feature dimensions, a random forests classifier is then utilised for classification. Dimensionality reduction is accomplished with the aid of genetic algorithms by selecting features. In order to correctly categorise cases of COVID-19, this technique incorporates the algorithm for optimizing ant colonies and a relief algorithm.

✓ Ant colony optimization method

Wrapper-based chosen techniques like the ACO algorithm are used to solve computational problems. Probabilistic methods are frequently used to find the optimal function subset because they reduce the time spent searching for the best possible path in a graph. Because of its great tensile strength and accuracy, the ACO algorithm has emerged as the state-of-the-art method for handling the challenging problem of optimization, in this case the optimization of feature selection. To define edge characteristics between all the other

networks and the accurate feature, networks require a problem to solve. To find the best possible subset of features, you can use a crossover halt criteria to generate an anti-graph path that visits the fewest number of nodes.

- ✓ **Relief method**

Effective relief algorithms are one of the most popular kinds of filtration functional selection models. This method works well with both nominal and continuous characteristics, as well as missing data and noisy tolerances. Theoretically, higher quality data can be achieved by increased categorization collaboration. However, there is not a single difference between the samples.

These ideas can be represented in terms of discrete operations within the relief algorithm: The training samples are sorted as follows: first, a sample xi is selected; then, the k nearest neighbours of xi are selected and recorded as H; and finally, the k not-similar closest neighbours of xi are selected and written as M c. The function's weight can be determined by modifying the weight value of the exceptionally attractive based on the distances in between present sample and its neighboring pixels. With enough time, this method can be employed to accurately duplicate the relative importance of any trait**.**

- ✓ **Combining relief and ant colony optimization reduces dimensionality**

To address the problem of high dimensionality, the authors of this study offer a hybrid filter-wrapper technique for selecting features from blood samples. Features from the blood data set are selected at random, with the average value being x plus k the values of the nearest neighbours in the same class. According to the guidelines in Note 1, the sample mean should be used instead of randomly selected instances to ensure that all categories are adequately represented. Because this criterion is indistinguishable from the actual data set, it rules out the use of a random number generator to select samples. When you take this extra precaution, you can rest assured that your measurements are precise and that your weighted random sample is as small as possible. Let's look into the algorithm used to pick out features from blood samples.

Figure 1 depicts the process of de-identifying blood sample data. There are correlation weights included in the improved algorithm's stress metrics. Reversing the order of search results can help filter out noise and unearth hidden associations between categories. Secondly, the enhanced ACO method assigns a smaller collection of features from which to make a prediction, and organises them according to a relief-enhanced weight. Weighing relevant features for a second search based on the results of the first. The optimal solution is to use the criterion with the highest classification accuracy. You can see the reduced number of characteristics identified by the random forest classifier and the prediction models in Figure 1 of the aforementioned work's flowchart**.**

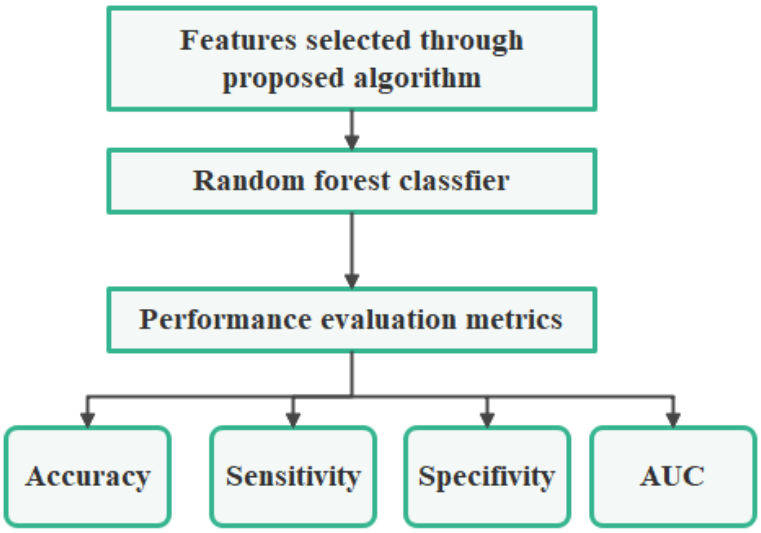

Fig 1: Flowchart of developed model

Segmentation is the last step of a model implementation, and it involves applying a classifier to the remaining attributes. A random forest classifier is used in this investigation. Genetic algorithms pair well with random forest, even though some publications use other classifiers. Every common classifier was beaten by random forests, including the popular gradient boost trees. The random forest algorithm uses bootstrapping and feature-selection at the node split level to diversify its model.

The bootstrap sample is a type of classification known as "bagging." To use, random forests need two parameters: the number of trees and the number of features evaluated at each node.

III. RESULTS AND ANALYSIS

This investigation is implemented in the programming language Python, and the model's efficacy is assessed using several parameters, such as its accuracy, sensitivity, specificity, and area under the curve (AUC). The results have proven that the proposed method is better than existing algorithms. In order to classify the blood samples, we will utilise a random forest as the classifier and feed it the cleansed data set.

Table 2 shows that the suggested method, which utilises a genetic algorithm ensemble in conjunction with a machine learning algorithm, has a success rate of 99.**3%.**

Table 2: A comparison of the completed work to previously developed models

Model	Accuracy (%)	Sensitivity(%)	Specificity(%)	AUC(%)
ANN with SMOTE	88	45	92	84
Logistic Regression	85	87	81	84
Bayes net	96	97	94	0.1
SVM	0.1	69	86	89
Proposed with random forest	99	97	99	93

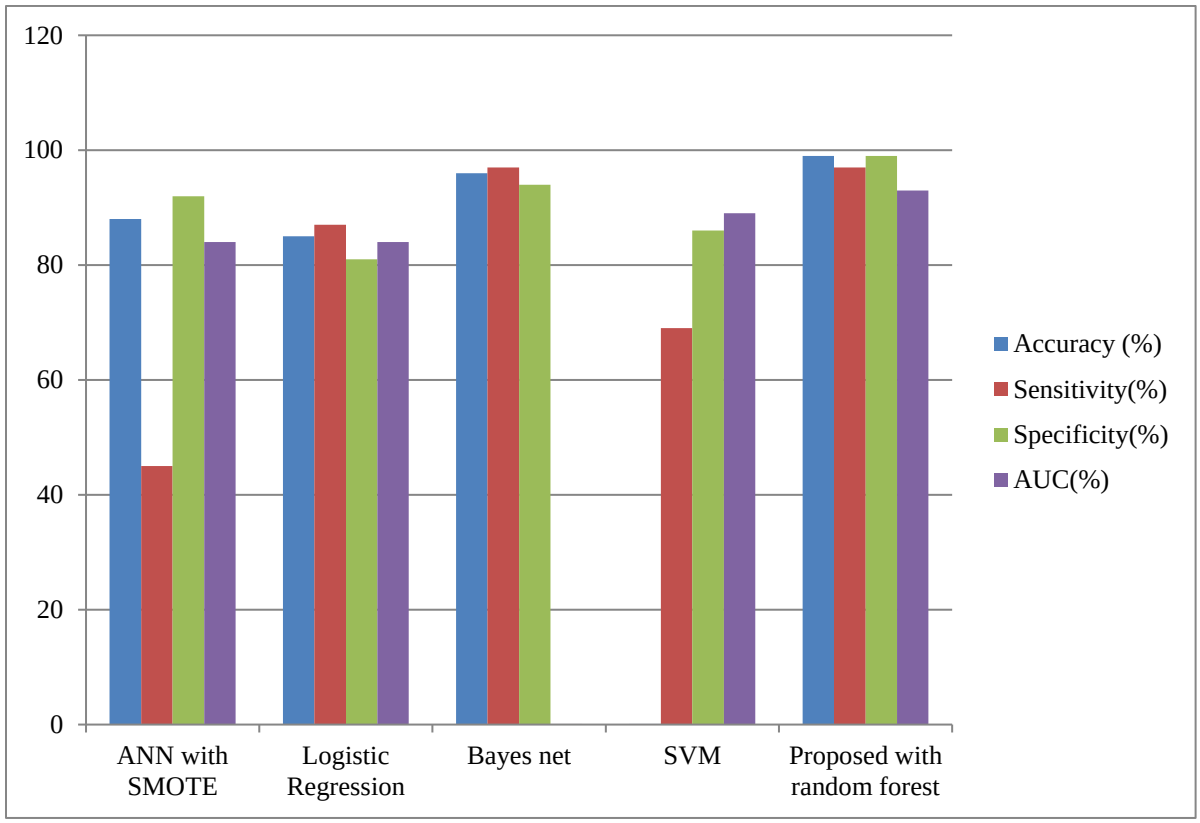

Fig 2 : graph comparison of proposed with existing works

IV. CONCLUSION

In this study, blood samples were used to analyse the virus known as COVID-19. The blood samples analysed thus far have revealed over a hundred unique traits. Therefore, the evolutionary algorithm was used to perform feature reduction on high-dimensional data. The relief approach was used in tandem with aco optimization to choose features from the top data in this study. Following these steps, the blood test data set is reduced in size by feature selection and then classified using a random forest classifier. The used model

achieved a 98.7% accuracy rate, a sensitivity of 96.76%, a specificity of 98.80%, and an area under the curve (AUC) of 92. The results showed that the developed algorithm outperformed the state-of-the-art algorithms by a significant margin. Research into gene expression will allow for the identification of COVID-19 disease in the future through the use of gene expression classification.