IIUC Studies, Vol.-20, Issue-1, June 2023
DOI: https://doi.org/10.3329/iiucs.v20i1.69055

A comparative study on machine learning algorithms for improved prediction measures for COVID-19

Md. Ziaur Rahman
Department of Computer Science and Engineering
International Islamic University Chittagong (IIUC), Bangladesh

Abstract

The Corona-virus (COVID-19) is an emerging disease responsible for infecting millions of people since the first notification until nowadays. Corona virus causes respiratory ailment like influenza with symptoms for example, cold, coughs, fatigue, fever and gradually increases the breathing problem. The disease and symptoms are changing frequently thus due to time constraints it is literally impossible to test. Analysis of Covid-19 data using machine learning paradigm is becoming a major interest of the researchers in this situation. The aim of this study is to develop a better predicting model for Covid-19 patients. Patients feature can be assessed statistically and traditionally. But with this day and age of advanced machine learning approaches Covid-19 can be predicted using machine learning techniques with better accuracy. In this work four well known machine learning approaches was used for better prediction in Covid-19. However, this study focuses on optimizing machine learning approaches. Two optimization approaches employed for Grid Search and Random Search are used for fine tune in prediction.

Keywords COVID-19, Machine Learning, Optimization, Prediction.

Paper type Research paper

1. Introduction

In recent years, as a general introduction of Covid-19, the highly contagious viral illness over the worldwide caused by severe acute respiratory syndrome corona virus 2 (SARS-CoV-2) (Ahmed, Rahman, & Hoque, 2020). This Covid-19 outbreak was declared by the World Health Organization as a public health emergency. In order to identify and isolate the contagious elements, diagnosis of COVID-19 is important (Ahmed, Rahman, & Hoque, 2020). The COVID-19 virus primarily spreads through the air when an infected person releases droplets from their mouth or nose while speaking, coughing, or sneezing. The transmission occurs when a healthy individual is close to the contaminated droplets (Ahmed, Rahman, & Hoque, 2020). The respiratory spreads vary

IIUC Studies
Vol.-20, Issue-1, June 2023
pp. 167-186
© IIUC
ISSN 2408-8544

vastly with individual factors age, gender, observations of fever, and travel history. Like normal fever, COVID-19 people also become affected with fever, fatigue and dry cough. Some patients may have sore throat, diarrhea, headache and nasal congestion myalgia, gastrointestinal symptoms, and Ansonia (Amoiralis, Tsili, Kladas, & Souflaris, 2012). The affected person shows early-stage symptoms and affected by the major symptoms within 2-14 days (Ahmed, Rahman, & Hoque, 2020; Amoiralis, Tsili, Kladas, & Souflaris, 2012). The increasing numbers of managing COVID-19 cases is a tremendous challenge for health care facilities in worldwide; though, there is not sufficient information about the virus. The severity of cases in this outbreak, and the survival rate from the infection, are putting great pressure on physicians and medical services, leading to a shortage of intensive care resources and electronic health facilities. The sudden pervasiveness of severe acute respiratory syndrome has been leading each country into a prominent crisis worldwide (Hakim, Uddin, & Hoque, 2020). People have been infected by the predominant virus vastly, resulting in various measures being enforced including country lockdowns, curfews and travel restrictions has been given by the Govt. for the people safety. Only 7 types of corona viruses can infect humans. The fast among them was discovered by scientists in 1965 (Hakim, Uddin, & Hoque, 2020). Although, almost 30% of this virus was only limited to the Middle East it was much deadlier than the SARS virus with a mortality rate (Hoque, Ahmed, Uddin, & Faisal, 2020). However, the comparison between SARS and MERS, the SARS-COV-2 virus is responsible for Covid-19 is a different beast altogether. In Bangladesh, the first corona virus case was reported on 8th March, 2020 (Hoque, Kabir, & Hossain, 2018). The COVID-19 virus has had a significant impact globally, infecting many individuals. The government is aware of the critical factors involved in the transmission of the virus, which include respiratory droplets from coughing or sneezing. The growing importance of healthcare epidemiology and demographic analysis has led to an expansion of electronic health data (Hoque, Ahmed, Uddin, & Faisal, 2020). The availability of electronic health system and data is increasing day by day. Corresponding weight has given by the measurement to the feature and the same input data can be used for training machine learning algorithms to improve its decision-making and reliability in terms of predicting Covid-19 symptoms at early stage (Josue, Arifianto, Saers, Rosenlind, & Hilber, 2020).

In data measurement, specific indicators that reflect symptoms are used to determine the progress of the Covid-19 situation. Predictive modelling may be applied, but the accuracy of the data depends on the relationship between previously collected data and various mathematical formulas. If this statistical formula is not reliable for prediction, in this matter machine learning comes to the rescue. Machine learning sector is an advanced modern branch of computer science that enables the machine to think like a human brain and it can also perform the task like human being. Instead of doing mathematical calculations with some formulas, the machine learning approach take into account the previous values and use them in predicting some sort of patterns. In addition, machine learning is preferable to statistical approaches in the application of predictions because machine learning approaches are more powerful in predicting approaches than statistical approaches. Statistical Approaches is expensive, it takes more time for execution. When evaluating data, only certain collected variables are considered—specifically, those that pertain to the formula. In contrast, machine learning considers all the available variables and creates patterns based on their relationships. As a result, predictions are more accurate. Because of the considerable benefits of machine learning over statistical measurement, machine learning methods are increasingly being used in the medical sector as a way to predict positive cases, predict important symptoms, performance evaluation etc. Depending on the value of certain parameters and their works, choosing the appropriate parameter can lead to better results from these methods. Therefore, the Tuning approach of number parameters has also been used like Grid Search (GS) and Random Search (RS). In addition, a number of feature selection techniques have also been introduced for the selection of important features.

In this study I tried to review the performance of machine learning classifier named SVM, RF, DT and KNN. The prediction of the learning approach employs from a number of Covid_19 symptoms data. Performance appraisal was not primarily focused on machine learning method, this study actually focused on comparing better SVM, RF, DT and KNN also using two optimization approaches called GS and RS. In addition, I used feature selection techniques called ExtraTressClassifier and Correlation approaches to find out which feature is the best in the models.

The model for predicting Covid-19's progression is a resource for patients and medical professionals to make informed decisions. Analysing with statistical methods, incorporating machine learning results is more accurate for predictions. Considers dependent and independent variables to make predictions. Forecasting independent variables in the case of Covid-19 positive patient indicates a number of symptom ratios that can be considered indicators for each Covid-19 positive patients. Each of these indicators has the same level of importance in predicting performance. To solve this problem, two feature selection techniques have been used and the

Correlation matrix is a very popular technique between them. Furthermore, to improve learning performance techniques related to parameters need to be fine-tuned. Therefore, a number of parameters tuning approaches have also been used. The parameter tuning approach further needs to be tested in predicting the Covid-19 patients. Therefore, the researcher tried to promote the performance in prediction of Covid-19 positive cases with the help of machine learning classifiers and comparison it with other models. In addition, both GS and RS are tested for the better result. And finally, when predicting theCovid-19 patients, the researcher wanted to show between SVM, RF, DT and KNN in which classification algorithm works best when GS and RS are used.

The remainder of this paper is organized as follows – Section 1 includes the introduction. Related Works has been discussed in section 2. Materials and methods incorporating system descriptions of the research are discussed in Section 3. Sections 4 shows result from the experiment. The expected objectives are full filled in this section. Section 5 contains the conclusion. References have been added at the end of the paper.

2. Literature review

In this section, the key points of the reviewed literature have been discussed. In the study by Ahmed, Rahman, and Hoque (2020), the models' efficiency and quality were assessed by evaluating their sensitivity and specificity as performance metrics.. The SVM and Naïve Bayes models yielded the best results in various clinical prediction tasks within the epidemiology dataset, which examined the active and cured COVID-19 cases. It can also reduce healthcare system and tools of rapid diagnosis.

A model is described that uses cross-validation of blood tests from COVID-19 positive patients to identify various bacterial and viral infections. The model was evaluated using the ROC and AUC scores, which showed a result of 81.9% and 0.97% respectively. The XGB Algorithm was used to determine the most important features (Ahmed, Rahman, & Hoque, 2020). The early-stage analysis conducted in the study involved using five machine learning algorithms based on the cumulative patient cases of COVID-19 infections. The study used a new dataset from mainland China and focused on different clinical features of the infected patients. The researchers used various classifiers to evaluate the information criterion and assess their performance. Out of all the classifiers, SVM showed the best accuracy (Ahmed, Mortuza, Uddin, Kabir, Mahiuddin, & Hoque, 2018).

The study analyzed and compared seven performance metrics (accuracy, sensitivity, specificity, precision, recall, F-measure, ROC, and AUC). The results showed that the AUC of the Random Forest (RF) classifier outperformed the other models, displaying a higher level of performance. Additionally, kappa statistics was applied to enhance the ability to predict a patient's survival outcome (Amoiralis, Tsili, Kladas, & Souflaris, 2012). The study compared different machine learning models to predict COVID-19's spread in Bangladesh. The goal was to enhance these models for more accurate forecasts of the pandemic's impact. The Facebook Prophet method outperformed other models. It was effective because it used health data to predict the disease. The models worked better with various measures of effectiveness (Hoque, Ahmed, & Hannan, 2020). The SVM model also showed promising results. Compared to the Prophet method, it needed more complex data but improved predictions for COVID-19 using a special mathematical method, which altered the forecast for India's population (Kabir, Rashid, Gafur, Islam, & Hoque, 2019). Robust Weibull model based on iterative weighting, showed that the model was able to make statistically better predictions than the baseline Gaussian model (Kolyanga, Kajuba, & Okou, 2014). The Gaussian model presents an over-optimistic view of the COVID-19 scenario. It predicts the severity of the spread of SARS-CoV-2 across countries worldwide in real-time applications. However, if the model demonstrates poor fitting, this approach could lead to suboptimal decision-making, potentially worsening the public health situation (Kabir, Rashid, Gafur, Islam, & Hoque, 2019).

Researchers analyzed three indicators - LDH, hs-CRP, and lymphocyte levels - along with other clinical data to predict the course of COVID-19 in patients (Rosas et al., 2005). This approach aimed to identify high-risk patients early on, before their condition worsened significantly. Additionally, it could help track overall mortality trends in COVID-19 patients. The researchers developed a model that is efficient, clinically applicable, and easy to interpret for healthcare professionals. They compared its performance to other common methods like random forest and logistic regression.

Researchers compared the performance of three methods for predicting COVID-19 outcomes: MLP (Multilayer Perceptron), VAR (Vector Autoregression), and a method based on maximum autoregression with information criterion (Suechoey et al., 2005). Their analysis focused on data from several hospitals and aimed to understand the immune system response of infected individuals. The MLP method showed the best results in terms of accuracy and efficiency based on various performance metrics.

This study analyzed the risk of positive RT-PCR tests in patients after they were discharged from the hospital (Suechoey et al., 2005). The researchers found that including more positive RT-PCR samples would improve the accuracy of their analysis. Additionally, they proposed using a nucleic acid test to easily identify patient characteristics. Their main goal was to determine the true positive rate of RT-PCR tests in recovered patients and develop a reliable model using a random forest approach to predict who might test positive again after discharge.

Researchers built a Random Forest (RF) algorithm to predict the outcomes (prognoses) of COVID-19 patients at an early stage of the disease (Zhan et al., 2014). This model relied on the most relevant clinical information available during hospitalization. The test results were very promising, with an Area under the ROC Curve (AUC) of 100%. AUC is a metric used to assess the performance of classification models.

This study explored using XGBoost, a machine learning model, to predict the risk of death in COVID-19 patients (Zhan et al., 2014). While the model showed promising results, there are limitations to consider. Firstly, the study lacked a large and diverse dataset, which can affect the model's generalizability. Secondly, it was a retrospective study conducted at a single center, limiting its overall scope. Despite these limitations, the research provides valuable initial insights into the clinical course and outcomes of severe COVID-19 patients. The authors suggest that XGBoost, although a "black box" model (meaning its internal workings are not easily interpretable), has the potential to be a powerful tool for risk assessment in COVID-positive patients. Further research with larger and more diverse datasets is needed to validate and improve this model.

A comparative experimental model between artificial neural networks44, extra trees, random forests, catboost, and extreme gradient boosting capable of predict negative prognostic outcomes with high overall performance for COVID-19 (Zhan, Goulart, Falahi, & Rondla, 2014). The outcomes show overlapping, which may have calculated the performance of the predictive models, even the majority cases of the outcomes were independent in the technique.

A review of existing literature suggests that there's a gap in using machine learning models to extract clinical features and build accurate prediction systems for COVID-19. This research aims to address this gap by implementing machine learning models for improved COVID-19 prediction. Additionally, the use of Gradient Boosting (GS) and Random Forests (RS) models for predicting positive cases based on symptoms in the medical field appears limited. This study also emphasizes the need for further research to bridge these identified knowledge gaps in COVID-19 prediction within the medical sector.

3. Materials and methods

This section will describe research methodology. Firstly, a discussion will be introduced about the process of collecting data and explaining the data. Then the preprocessing technique will be described simultaneously. After that the classifiers and their working principle will be discussed. Finally, influencing factors and features of determining the outcome of CCOVID-19 will be figured out and analyzed.

3.1 Proposed model

This research focuses on comparing three machine learning techniques (SVM, RF, and KNN) for their ability to predict and detect COVID-19 positive patients based on symptoms. Additionally, the performance of feature selection techniques employing correlation matrices and ExtraTreesClassifier was evaluated. Furthermore, the optimization process using GridSearchCV and RandomizedSearchCV for tuning hyperparameters in both Gradient Boosting (GS) and Random Forests (RS) was reviewed.

To achieve these goals, the first step involved collecting a COVID-19 dataset. Following data collection, pre-processing steps were applied, including null value removal, min-max normalization, and dataset balancing. Next, the focus shifted to performance evaluation. Here, I aimed to improve the performance of various models including SVM, RF, Decision Trees (DT), and KNN by optimizing their hyperparameters using both GS and RS. Specifically, for SVM, I optimized the "C" and "gamma" parameters, while for RF, I focused on "max_features" and "max_depth". Similarly, hyperparameter optimization was conducted for DT ("max_depth" and "min_samples_split") and KNN ("n_neighbors" and "leaf_size"). Subsequently, the performance of GS and RS-optimized models (SVM, DT, KNN, and RF) was compared.

To identify key features, both correlation matrix and ExtraTreesClassifier approaches were used. Following feature selection with these methods, the performance comparison between the optimized models (SVM, RF, DT, and KNN) was revisited. Finally, a 5-fold cross-validation approach was employed while training the classifiers to ensure robust evaluation. The techniques used in my research include:

➡ GS-SVM
➡ RS-SVM
➡ GS-RF
➡ RS-RF
➡ GS-DT
➡ RS-DT

- GS-KNN
- RS- KNN

This research was conducted using the block diagram shown in Figure 1

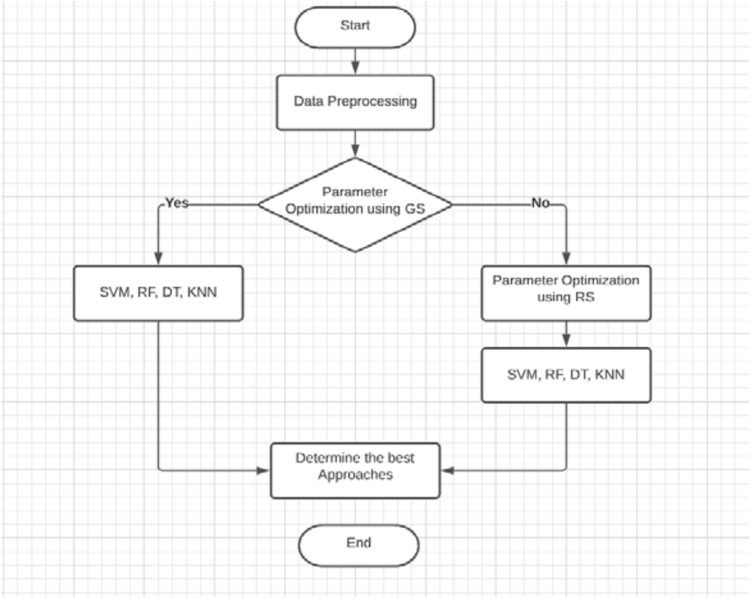
Start
Data Preprocessing
Parameter Optimization using GS
Yes
No
SVM, RF, DT, KNN
Parameter Optimization using RS
SVM, RF, DT, KNN
Determine the best Approaches
End

Figure 1
Proposed model diagram

3.2 Data description

Since my research aimed to predict COVID-19 positivity based on symptom data, the researcher required a dataset containing information on various patient indicators. To find a suitable dataset, the researcher extensively searched for reliable resources. Fortunately, the researcher found a dataset on Figshare (https://figshare.com/articles/dataset/S1_Data_-/13355374) that collects data on different symptoms experienced by COVID-19 patients.

The downloaded dataset included information on various patient symptoms categorized by age and gender. However, the data required some cleaning and pre-processing to fit the required needs analysis. This involved arranging the data into a desired format and removing irrelevant features. After this process, the final dataset contained 43 indicators representing different symptoms and their percentages, along with 4,000 data instances. It is important to note that the original dataset contained more instances. So, some data, which appeared irrelevant or potentially misleading for the analysis, were eliminated.

Below is a list of the 43 indicators included in the prepared dataset:
-- Body Temperature
-- Pulse Rate
-- SPO2
-- Respiratory Rate
-- Systolic BP
-- Diastolic BP
-- Age
-- Overseas Heal Facilities
-- Australian Health Facilities
-- COVID Contact Status
-- Overseas Travel
-- Fever > 38 C
-- Fever Subjective
-- Sore Throat
-- Shortness of breath
-- Cough
-- Anosmia
-- Coryza
-- Diarrhoea
-- Malaise
-- Ageusia
-- Asymptomatic
-- Headache
-- Sinusitis
-- Other GI symptoms
-- Chest pain
-- Cardiovascular disease
--Diabetes
--Hypertension
-- ACEI/ARB treatment
--Smoking
--Chronic renal
--Immunosuppressed
--CRD
--Pregnancy
--COVID discharge disposition
--Swab testing
--O/s travel
--Any fever

--Loss of taste smell
--Number of symptoms
--Gender
--Covid Detected

3.3 Data preprocessing

After data collection, the next crucial step involves pre-processing the information. Real-world data often contains various inconsistencies: missing values (null values), inconsistencies (incomplete or noisy data), outliers (overvalued values), irrelevant features, and more. These issues can hinder the effectiveness of machine learning models. Data pre-processing addresses these challenges to ensure that the data is well-structured and suitable for machine learning algorithms. It involves "cleaning" the data by removing irrelevant information (inessentiality) and aligning it into a format that facilitates efficient learning. In this study, the researcher applied the following pre-processing techniques to address potential problems and improve the quality of the data:

- Removal of Null Values
- Min-max Normalization
- Class Balancing

3.4 Removal of Null values

A particular field is not showing a specific value or empty attributes in a dataset or table. The significance of a particular piece of information can sometimes be underestimated by researchers. While a consistent deviation in the measured value represents a systematic error, the presence of an unexpected value in the dataset is a clear indication of an anomaly.

3.5 Min-Max normalization

Min - Max normalization is a process which measure data in the range of 0 and 1 by default. While min-max normalization scales data to a specific range (often 0 to 1), standardization offers an alternative approach. Standardization transforms the data to have a zero mean and a unit standard deviation. This technique can be particularly useful if anyone anticipates the need to scale the data to a different range in the future. By applying normalization to the dataset, features with different scales or magnitudes can be brought to a common scale, which facilitates learning algorithms in analysing the dataset. Since the collected dataset includes various attributes with different units, min-max standardization can be employed to scale them from 0 to 1.

3.6 Class balancing

Class imbalance algorithm contains an equal proportion of ideas in each class. Class anomalies are present in a wide range of areas, including medical diagnostics, spam filtering, and fraud detection, which can lead to errors in machine learning algorithms. To address the issue of class imbalances, oversampling techniques may be employed. These techniques involve creating additional copies of the minority class to balance the number of instances with the majority class.

In the picture, it shows the imbalance in our dataset which does not have the same number of instances in each class presented in figure 2.

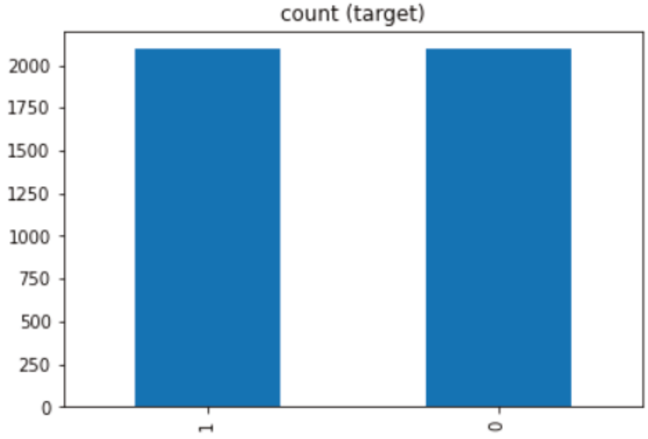

Figure 2
After class balancing

3.7 Feature selection and parameter optimization

In machine learning, feature selection is a central idea in which greatly influences the outcomes of the learning model. Irrelevant or partially related features cannot give positive affect model execution. Feature selection and data cleaning should be the first and foremost priority. In designing one's model feature selection, the first and foremost concept in machine learning is selecting the features naturally or physically. It is also playing the most important role in predicting and forecasting dependent variables or yields. If dataset have prominent features, it can reduce the performance of the models. Completion of feature selection can reduce over correlation in shorten training time, and improve accuracy and performance.

Machine learning has seen significant advancements with the development of modern models that operate using numerous parameter keys. These keys enable the models to function efficiently, and their default values can be adjusted as necessary. It is important to note that real-world datasets differ from virtual ones and may require unexpected removal. Applying the same parameters to different datasets in similar models does not guarantee consistently good results. Therefore, it is crucial to balance these parameters to achieve high and accurate predictive performance for the required dataset. The process of controlling a model's performance by adjusting its learning parameters is known as parameter tuning. Examples of hyperparameters influence the performance of the SVM algorithm include C and Gamma, but in deep learning parameters, such as the value of K, max depth, weight, and leaf size, are significant. In the study, the techniques used for parameter optimization are outlined below:

- Grid Search (GS)
- Random Search (RS)

4. Results and analysis

In my research, the researcher provides a comprehensive analysis of feature selection using the ExtraTreesClassifier. Following this, the researcher discussed the performance of SVM, RF, and KNN algorithms, which were optimized through Grid Search and Random Search methods. Finally, a comparative discussion on the performance of different perspectives of Covid-19 symptoms in positive patients has been presented.

4.1. Feature Selection using ExtraTressclassifier

As previously stated, their study has used two feature selection methods: the ExtraTreesClassifier and the Correlation Approach. Both methods aim to minimize the number of features while ensuring the selection of optimal ones. This research utilizes feature importance to identify the factors that impact the target variable. Feature importance is quantified by assigning scores to each feature, with higher scores indicating greater significance. Figure 3 illustrates the top features identified by the ExtraTreesClassifier, highlighting that the number of symptoms is the most significant feature.

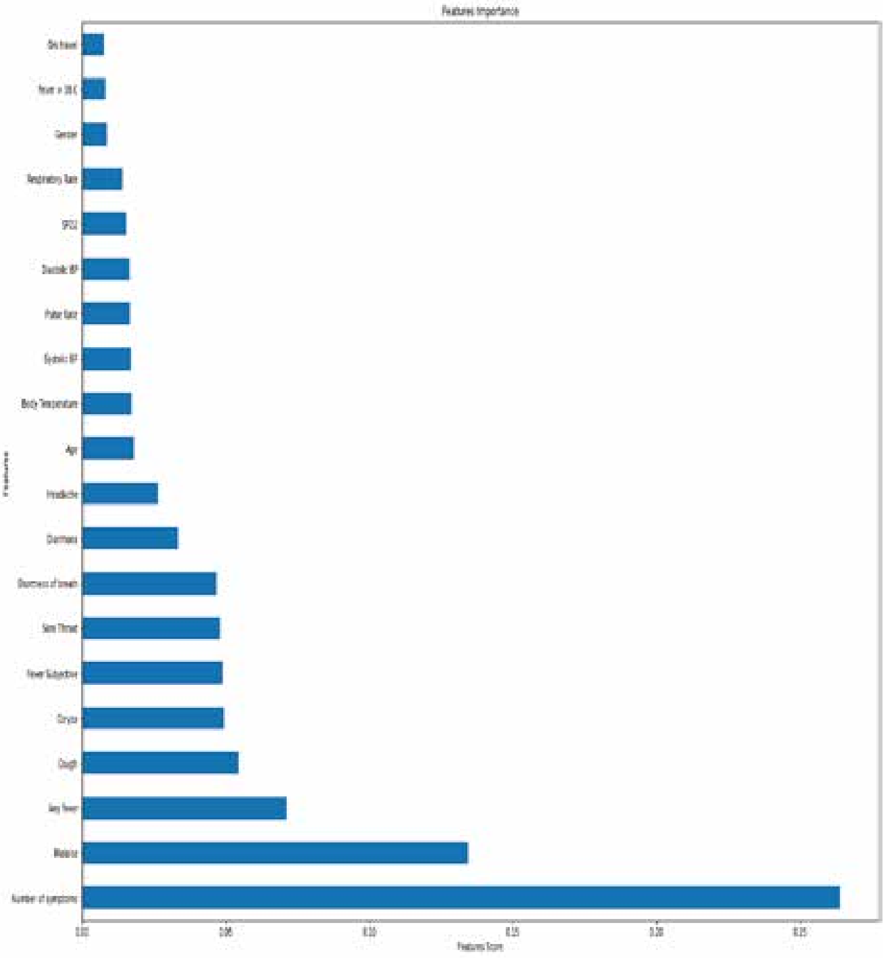

Figure 3
Features importance

4.2 Correlation based feature selection

When the researcher selected features using a correlation approach, a heat map was developed that showed the correlation between dependent variables and independent variables, and key features were selected based on the correlation value. Based on the correlation, the following five characteristics were chosen for both SVM and RF. The chosen features were the same because the selection of correlation-based features is a statistical measurement and does not depend on the classifiers, while GA selects features based on the accuracy given by the classifiers.

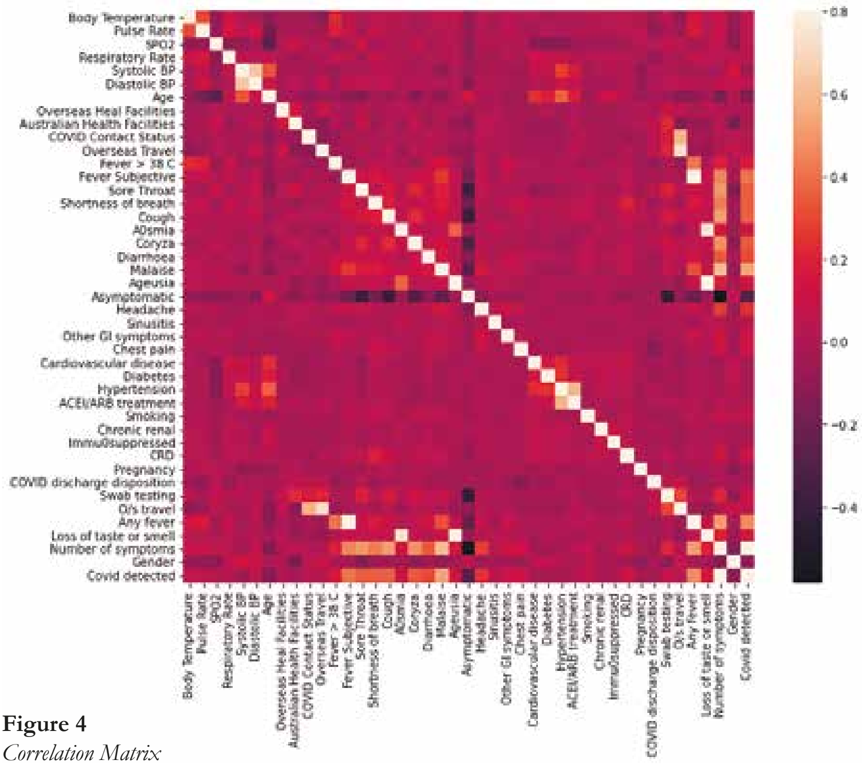

Figure 4
Correlation Matrix

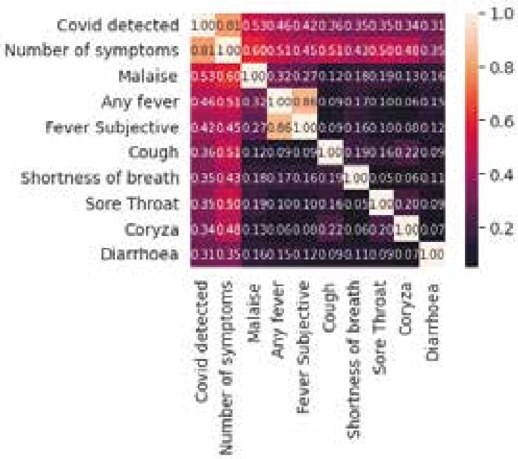

Figure 5
Correlation matrix heat map

4.3 Result of parameter optimization

The parameter optimization technique was assessed using Grid Search (GS) and Random Search (RS) for Support Vector Machine (SVM), Random Forest (RF), and K-Nearest Neighbors (KNN). For the SVM, 'C' and 'Gamma' parameters were considered, while 'Max_depth' and 'max_features' were the parameters for RF. For KNN, 'n_neighbors' and 'leaf_size' were the two parameters taken into account. The optimal values of the parameters for SVM, RF, and KNN were determined through parameter optimization. Subsequently, a 5-fold cross-validation technique was employed to train the classifiers as part of the parameter optimization process. Each model yielded different values for the selected parameters. Below is a selection of optimized parameter values for SVM and RF, along with the features:

Table 2

Comparison of different parameters using GS

Cross Validation	SVM with GS		RF with GS		DT with GS		KNN with GS	
No of folds	C	gamma	Max_depth	n-estimators	param_max_depth	param_max_features	param_n_neighbors	param_leaf_size
1	64.390	4.364	149	10	580	8	45	6
2	5.614	2.164	85	9	299	9	22	7
3	9.160	1.665	38	8	264	8	26	5
4	68.356	4.198	266	11	603	8	4	6
5	35.468	4.265	115	9	639	9	40	4

Table 3

Comparison of different parameters using RS

Cross Validation	SVM with GS		RF with GS		DT with GS		KNN with GS	
No of folds	C	gamma	Max_depth	n-estimators	param_max_depth	param_max_features	param_n_neighbors	param_leaf_size
1	15.362	8.126	87	17	9	2	9	6
2	4.823	9.626	37	52	5	20	4	7
3	73.451	9.422	70	37	2	10	5	5
4	47.104	4.529	55	25	1	20	8	6
5	38.151	2.219	87	20	6	10	7	4

Table 4

Accuracy comparison between SVM and RF with GS

Cross Validation			SVM with GS				RF with GS	
No of folds	C	gamma	Train Accuracy	Test Accuracy	Max_depth	n-estimators	Train Accuracy	Test Accuracy
1	64.390	4.364	99.264	98.197	149	10	100.0	99.836
2	5.614	2.164	99.057	97.709	85	9	100.0	99.502
3	9.160	1.665	98.532	98.685	38	8	100.0	99.671
4	68.356	4.198	99.161	98.526	266	11	100.0	100.0
5	35.468	4.265	98.639	96.295	115	9	100.0	99.343
Average			98.912	97.865			100.0	99.671

Table 5

Accuracy comparison between SVM and RF with RS

Cross Validation			SVM with GS				RF with GS	
No of folds	C	gamma	Train Accuracy	Test Accuracy	Max_depth	n-estimators	Train Accuracy	Test Accuracy
1	15.362	8.126	100.0	99.836	87	17	100.0	99.836
2	4.823	9.626	100.0	99.343	37	52	100.0	99.672
3	73.451	9.422	100.0	99.016	70	37	100.0	100.0
4	47.104	4.529	100.0	99.508	55	25	100.0	99.671
5	38.151	2.219	100.0	100.0	87	20	100.0	99.502
Average			100.0	99.540			100.0	99.737

Table 6

Accuracy Comparison between DT and KNN with GS

Cross Validation	DT with GS				KNN with GS			
No of folds	param_ max_depth	param_ max_features	Train Accuracy	Test Accuracy	param_n _neighbors	param _leaf_size	Train Accuracy	Test Accuracy
1	580	8	100.0	99.002	45	6	99.361	98.753
2	299	9	100.0	99.162	22	7	99.601	99.073
3	264	8	100.0	99.027	26	5	100.0	99.861
4	603	8	100.0	99.276	4	6	99.861	99.274
5	639	9	100.0	100.0	40	4	99.402	98.642
Average			100.0	99.293			99.645	99.246

4.4 Performance comparison

At this work, the important features have been selected using a ExtraTressclassifier and correlation approach. In addition, parameter optimization has been conducted for SVM, RF, and KNN classifiers. The next step is to compare the accuracy among these classifiers. The accuracies of SVM, RF, and KNN have been compared, taking into account the features selected by the optimized parameters through both the correlation approach and random search.

Table 7

Accuracy comparison between DT and KNN with RS

Cross Validation	DT with GS				KNN with GS			
No of folds	param_ max_depth	param_ max_features	Train Accuracy	Test Accuracy	param_n _neighbors	param _leaf_size	Train Accuracy	Test Accuracy
1	9	2	100.0	99.895	9	6	100.0	99.034
2	5	20	100.0	99.876	4	7	100.0	99.267
3	2	10	100.0	99.857	5	5	100.0	99.564
4	1	20	100.0	100.0	8	6	100.0	99.058
5	6	10	100.0	100.0	7	4	100.0	99.546
Average			100.0	99.925			100.0	99.38

Table 8

Accuracy comparison

	Train Accuracy	Test Accuracy
SVM - GS	98.912	97.865
SVM -RS	100.0	99.540
RF-GS	100.0	99.671
RF-RS	100.0	99.737
DT-GS	100.0	99.293
DT-RS	100.0	99.925
KNN-GS	99.645	99.246
KNN-RS	100.0	99.38

By optimizing the parameters, one can achieve the optimal selection using both Grid Search and Random Search. The parameters of SVM, RF, and KNN within the training set were subject to change with each variation in the parameters and folds. When applying 5-fold cross-validation, the training set comprised 75% of the entire dataset, and the test section accounted for 25%. It is to be noted that the total percentage for the training and test sets should add up to 100%.

In the above performance Table 8, the classifiers were compared each other and used with different parameter optimization and feature selection technique. Table 8 shows that the result of the classifiers were trained with parameter optimized by GS, train accuracy for SVM was 98.91% and test accuracy was 97.86% while for RS the train accuracy was 100% and test accuracy was 99.54%. Similarly it also shows that the result with parameters optimization by RS produces better accuracy than GS. The comparison in the Table 8, DT-RS shows the better result other than the classifiers. However, if we compare the entire Table 8, we can clearly observe that DT-RS generated the highest performance with test accuracy of 99.92%. In a nutshell, it can also be notated that SVM and KNN performs less better in prediction of Covid-19 positive case with parameter optimized by GS and RS.

5. Conclusion

To precise the prediction in Covid-19, we have to determine which optimizer and feature selection approach works better with SVM, RF, DT and KNN. For this purpose, data collected on symptoms of Covid-19. Then the data has been preprocessed. After the Covid-19 symptoms data was developed for machine learning, GS and RS were employed to optimize the parameters of SVM, RF, DT and KNN; and ExtraTressClassifier and correlation matrix were used for feature selection. It was found that the features selected by the RS gave the better result than the GS. Therefore, DT-RS can be a good feature selector and optimizer for predicting the Covid-19. Studying the previous works and researches, the researcher tried to work without any errors, but there are still shortcomings in this research. For example, the dataset of this study contains only 4000 instances which indicate that the dataset is small. However, the lack of sufficient performance data for Covid-19 with similar characteristics is the reason behind this. Predictive modeling in healthcare is a crucial aspect for the future.

References

Ahmed, M. R., Rahman, M. O., & Hoque, M. J. (2020). Smart home: an empirical analysis of communication technological challenges. *European Journal of Engineering and Technology Research, 5*(5), 571-575. https://doi.org/10.24018/ejeng.2020.5.5.1905

Ahmed, Z. U., Mortuza, M. G., Uddin, M. J., Kabir, M. H., Mahiuddin, M. & Hoque, M. J. (2018). *Internet of things based patient health monitoring system using wearable biomedical device.* Paper presented at the 2018 International Conference on Innovation in Engineering and Technology (ICIET), Dhaka, Bangladesh, (pp. 1-5). doi: 10.1109/CIET.2018.8660846.

Amoiralis, E. I., Tsili, M. A., Kladas, A. G. & Souflaris, A. T. (2012). Distribution transformer cooling system improvement by innovative tank panel geometries. *IEEE Transactions on Dielectrics and Electrical Insulation, 19*(3), 1021-1028. doi: 10.1109/TDEI.2012.6215108.

Hakim, M. L., Uddin, M. J. & Hoque, M. J. (2020). *28/38 GHz Dual-Band Microstrip Patch Antenna with DGS and Stub-Slot Configurations and Its 2×2 MIMO Antenna Design for 5G Wireless Communication.* Paper presented at the 2020 IEEE Region 10 Symposium (TENSYMP), Dhaka, Bangladesh, (pp. 56-59). doi: 10.1109/TENSYMP50017.2020.9230601.

Hoque, M. J., Ahmed, M. R., Uddin, M. J., & Faisal, M. A. (2020). Automation of traditional exam invigilation using CCTV and Bio-Metric. *International Journal of Advanced Computer Science and Applications, 11*(6), 392-399. http://dx.doi.org/10.14569/IJACSA.2020.0110651

Hoque, M., Ahmed, M., & Hannan, S. (2020). An automated greenhouse monitoring and controlling system using sensors and solar power. *European Journal of Engineering Research and Science, 5*(4), 510-515. https://doi.org/10.24018/ejers.2020.5.4.1887

Hoque, M., Kabir, S., & Hossain, M. K. (2018). Electricity crisis of Bangladesh and a new low-cost electricity production system to overcome this crisis. *International Journal of Scientific and Research Publications, 8*(7), 201-206. http://dx.doi.org/10.29322/IJSRP.8.7.2018.p7933

Josue, F., Arifianto, I., Saers, R., Rosenlind, J., & Hilber, P. (2020). *Transformer hot-spot temperature estimation for short-time dynamic loading.* Paper presented at the IEEE International Conference on Condition Monitoring and Diagnosis, Indonesia, (pp. 217-220). doi: 10.1109/CMD.2020.6416414

Kabir, M. H., Rashid, S. Z., Gafur, A., Islam, M. N., & Hoque, M. J. (2019). *Maximum energy efficiency of three precoding methods for massive MIMO technique in wireless communication system.* Paper presented at the IEEE International Conference on Electrical, Computer and Communication Engineering (ECCE), Bangladesh, (pp. 1-5). doi: 10.1109/ECACE.2019.8679238

Kolyanga, E., Kajuba, E. S., & Okou, R. (2014). *Design and implementation of a low-cost distribution transformer monitoring system for remote electric power grids.* Paper presented at IEEE International Conference on the Eleventh industrial and Commercial Use of Energy, South Africa, (pp. 1-7). doi: 10.1109/ICUE.2014.6904200

Rosas, C., Moraga, N., Bubnovich, V., & Fischer, R. (2005). Improvement of the cooling process of oil-immersed electrical transformers using heat pipes. *IEEE Transactions on Power Delivery, 20*(3), 1955-1961

Suechoey, B., Tadsuan, S., Thammarat, C., & Leelajindakrairerk, M. (2005, November). *An analysis of temperature and pressure on loading oil-immersed distribution transformer*. Paper presented at the 2005 IEEE International Power Engineering Conference, (pp. 634-638).

Zhan, W., Goulart, A. E., Falahi, M., & Rondla, P. (2014). Development of a low-cost self-diagnostic module for oil-immerse forced-air cooling transformers. *IEEE Transactions on Power Delivery, 30*(1), 129-137. doi: 10.1109/TPWRD.2014.2341454

Corresponding author

Md. Ziaur Rahman can be contacted at: ziaur.rahman@iiuc.ac.bd