Journal of Positive School Psychology http://journalppw.com
2022, Vol. 6, No. 5, 24 – 32

A Machine Learning technique to analyze and detect Corona Virus

[1]Md. Ahsan Arif

[1]Associate Professor and HEAD, The International University of Scholars (IUS), Dhaka, Bangladesh, ahsan@ius.edu.bd

Abstract

COVID 19 has expanded repeatedly over the whole world, and the number of infected people has been increasing tremendously. COVID 19 has stormed the world in a blink resulting in millions of deaths with economic downfall around the globe. It has triggered a disastrous paradigm shift for the world. Given the unavoidable circumstances, testing for the virus on a rapid daily basis for million people yields the importance of partaking next steps in virus control. The supply chain of traditional Check-up and report time is exorbitant and has the avenue of exceeding the possibility of misreporting. As a result, we have presented Machine learning-based methods for COVID-19 identification. To improve the COVID 19 prediction algorithm, this study indicates the use of exhaustive profiling, SMOTE (Synthetic Minority Oversampling Technique), a classification model, and a deep learning model. This paper goals to provide Machine learning classifier algorithms and Neural Networks with selected attributes to obtain better accuracy and efficacy with a subsequent comparison with different algorithms.

Keywords: COVID-19, Data Mining, Machine Learning, Classification, Deep Learning, Exhaustive Profiling, One Hot Encoding.

I INTRODUCTION

Corona virus disease 2019 (Covid-19, earlier known as "2019 novel corona virus") is a transmissible disease caused by Severe Acute Respiratory Syndrome Corona virus 2 (SARS-COV-2) [1]. Covid-19 was discovered for the first time in December 2019 in Wuhan, China [2]. The World Health Organization (WHO) announced Covid-19 outbreaks an epidemic in March 2020 [3]. Covid-19 is reportedly suspected of being the cause of an enormous number of reported morbidity and mortality worldwide, owing to a lack of containment as well as disinformation [4]. As of April 12, 2021, there are more than 136.19 million global cases of Covid-19, including 2.93 million premature deaths confirmed across 192 countries [5]. More than 219 countries are challenged heavily through the ensuing pandemic and it is increasing rapidly on a daily basis. There has not been a suitable vaccine produced for the treatment of this disease until now. Corona virus is actually only found in the more mature individual people. Even an infant is not immune to it, despite the reality that it can happen at any age. Symptoms of Covid-19 will range from mild to serious at this time. Some people have no noticeable symptoms at all. Fever, cough, shortness of breath, headache, and sore throat are the most common Covid-19 symptoms. Many aspects of the epidemic are still posing problems for the global medical community, such as the medical equipment insufficiency and increasing hospital bed demand. In the advanced world, there has been a scarcity of the most effective Covid-19 diagnostic tool, which uses reverse transcriptase polymerization chain reaction (rt-

PCR) [6]. At this point in time, it is essential to accurately forecast Covid-19 in order to halt the rise in Covid-19 [7].Effective Covid-19 prediction will gradually reduce the enormous pressure on the medical healthcare system. Researchers in previous studies gathered data from sick patients, and some used databases from various countries' government websites. The Israeli Ministry of Health regularly updated their Covid-19 results. The Israeli Ministry of Health dataset holds exactly 27,42,596 official documents of the Covid-19 studied patients at this time. We were able to make good use of this data collection. The Israeli Ministry of Health officially released official data for each person who was accurately checked for sars-cov-2 via rt-PCR assay of a nasopharyngeal swab as part of the Covid-19 campaign [8]. The rt-PCR assay was used to validate any single negative and positive Covid-19 case in this dataset. To achieve the best result, we compared machine learning classification algorithms and artificial neural network models to forecast Covid-19 based on specific Eight Key features as well as primarily based on symptoms in this study. We have discovered that using the SMOTE sampling approach will potentially boost an unbalanced dataset and achieve the best possible results. As a result, it would be of tremendous help to the medical field.

2 Related Work

Albeit significant research work has been carried out for battling COVID as well as COVID related additional problems [7], most of the research works had a backlog in gathering homogeneous and balanced data, which resulted in the skewed dataset that eventually led to a lesser metric of values in terms of accuracy and efficacy of the research works [9]. Kolla Bhanu Prakash and his colleagues used Machine learning algorithms to build various prediction models. The model's efficiency is computed and analyzed. Random Forest Regression and Random Forest Classifier outperformed SVM, KNN+NCA, Decision Tree Classifier, Gaussian Naive Bayesian Classifier, Multi linear Regression, Logistic Regression, and XGBoost Classification in their research [10].Amir Ahmad and his colleagues have made several recommendations to Machine learning practitioners in order to enhance the efficiency of machine learning approaches in the prediction of verified cases of COVID [11]. The research of Furqan Rustam and his team included the creation of Supervised Machine Learning Models for COVID forecasting [12]. Shreshth Tuli and their collaborators used cloud computing to forecast the COVID pandemic's development and pattern and to create a Machine learning model [13]. In R Sujatha's and the team's study, they are developing a machine learning model to forecast COVID in India [14]. According to Peipei Wang et al., they used a logistic model and Machine learning techniques to forecast COVID disease patterns [15]. Dan Assaf and the rest of the team wanted to see whether Machine-learning algorithms could forecast the outcome of non-critical COVID patients based on clinical criteria at the time of admission [16]. Lamiaa A. Amar, Ashraf A. Taha, and others used Egypt's COVID Dataset to forecast the number of patients that would be infected with COVID and to estimate the scale of the final epidemic [17]. For clinical text results, Akib Mohi Ud Din Khanday et al. used Machine Learning methods to detect the COVID [18]. Adam L. Booth and his colleagues focused on the records of 398 patients, 43 of whom were dead and 355 of whom were living. They identified five serum parameters, including c-reactive protein, blood urea nitrogen, serum calcium, serum albumin, and lactic acid, using the data and created a Machine learning algorithm that can predict death up to 48 hours before the patient passes away [19][27].

3 Methodology

3.1 Source of Dataset

We used the data from the paper referred to, "Machine learning-based prediction of COVID-19 diagnosis based on symptoms" [6]. They took the Hebrew dataset from the Israeli

Ministry of Health's website [8] [20] and converted it to English. It was last revised on November 15, 2020, with 27,42,596 people checked (including 2,20,975 COVID 19 affected people and 24,80,403 COVID-19 non-affected people).

3.2 Materials

We have made use of Google Research's Colaboratory [21], also known as "Colab," is a commodity. It enables anyone to write and run any random Python code using a browser. It's ideal for computer learning, data mining, and educational purposes. Colab is a hosted Jupyter notebook program that doesn't need any configuration and gives you free access to computing services, including GPUs. Colab is completely free to use.

3.3 Data Pre-processing

Data Pre-processing, such as Data Transformation, Data Cleaning/Cleansing, Data Integration, Data Reduction, and Data Discretization, is often needed before applying the Machine Learning algorithm. It also entails locating relevant/irrelevant or incomplete data, eliminating noise or outliers, and gathering the appropriate data from the model or account for noise removal. Typically, raw data comprises a combination of important and unrelated characteristics. Irrelevant attributes limit the precision of machine learning models, and the model's prediction cannot be accurately predicted. Our dataset contains a total of 27,42,596 records, including 2,20,975 COVID-19 affected people and 24,80,403 COVID-19 non-affected people. For the output of the corona effect, the dataset contains unbalanced results. As a result, data pre-processing is needed in the database to balance the distorted data. To find the best solution, we tried 18 different variations, also known as exhaustive profiling, such as removing null values, assuming the null is positive or negative, and so on.

Furthermore, we have used Synthetic Minority Over-sampling Technique (SMOTE) [22] [23] [26] [27] for balancing the skewed data, k-fold fold cross-validation to fit the model. The Synthetic Minority Oversampling Technique (SMOTE) [22] [23] produces synthetic samples at random by taking each minority class sample and joining any/all of the k-nearest neighbors along with line segments in a less application-specific manner, working in "function space" rather than "data space." Based on the number of over-sampling needed, randomly generated synthetic samples are arbitrarily chosen from the k-nearest neighbors. The synthetic samples are created at random in the following ways: Calculate the disparity between the samples and the samples closest to them. Then multiply the calculated difference with an arbitrary number between 0 and 1, and append it to the samples. SMOTE effectively broadens the decision-making area of the minority class. An example of SMOTE sampling is seen in the diagram 1 below.

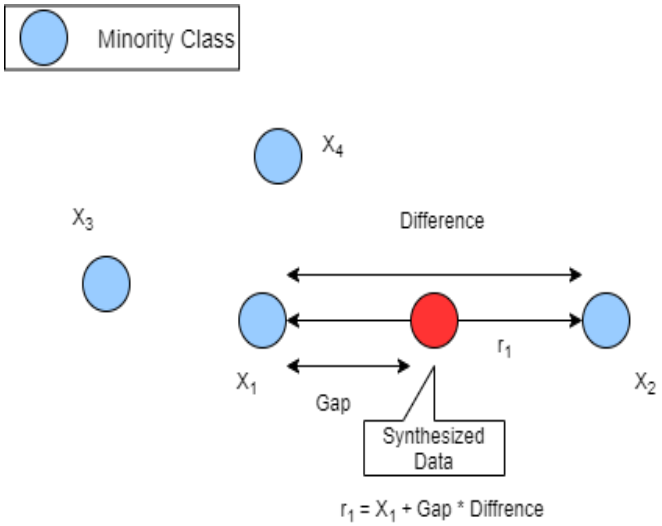

Figure 1: *An example of SMOTE sampling*

The dataset contains nine columns (cough, fever, sore-throat, shortness-of-breath, headache, corona-result, age-60-and-above, gender, test-indication) where seven columns (cough, fever, sore-throat, shortness-of-breath, headache, age-60-and-above, gender) have binary values, and two of them (corona-result and test-indication) contain non-binary values, so we have used One hot encoding on 'test-indication' column to get binary value. In the corona-result column, we have non-binary values (Positive, Negative, and Other). We have performed the Exhaustive profiling methodology on three columns (gender, age-60-and-above, and corona-result). The columns

'gender', and 'age-60-and-above' contain null values, and corona-result contains an irrelevant value 'other.' Hence, we have decided to perform the Exhaustive profiling on 'null' values and 'other', like gender null=drop/male/female, age-60-and-above null=drop/yes/no, and corona-result 'other'=drop/negative. We have pre-processed the data through the Exhaustive-profiling. It has provided the most efficient solution by dropping 'null' values of gender, age, and the 'other' value of the corona result column. Table 1 shows the best results from 18 combinations by the Exhaustive profiling. And, Fig. 2 is explaining our process for data processing.

Table 1: *Data pre-processing with exhaustive profiling*

Exhaustive Profiling Step	Result Analysis		Decision Tree	Random Forest	Logistic Regression	Naïve Bayes	Neural Network
Drop Null and Irrelevant Data	Accuracy		80.08%	80.08%	79.79%	79.93%	80.08%
	Precision	Positive	93%	93%	93%	92%	93%
		Negative	73%	73%	73%	73%	73%
	Recall	Positive	65%	65%	65%	65%	65%
		Negative	95%	95%	95%	95%	95%
	F1_Score	Positive	77%	77%	77%	77%	77%
		Negative	83%	83%	83%	82%	83%

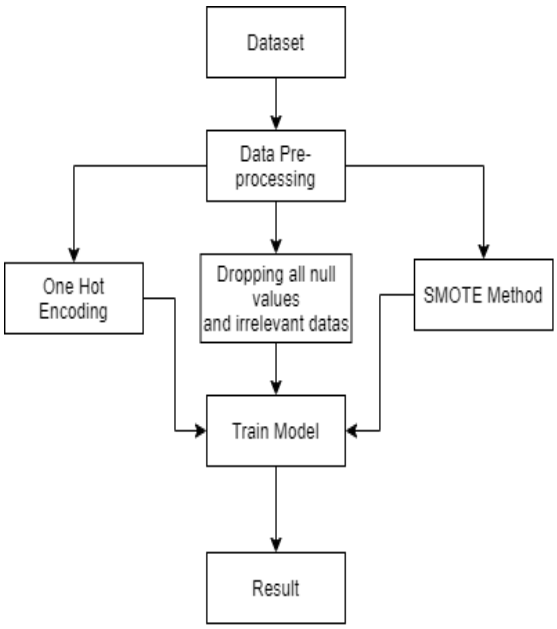

Figure 2: *Data processing steps*

3.4 Research Design

To accurately predict the COVID-19 on the Israeli dataset [6], we used Machine Learning algorithms such as Decision Tree, Random Forest, Logistic Regression, Naive Bayes, and Deep Learning algorithms such as Neural Network.

3.4.1 Decision Tree

A Decision Tree is a supervised Machine Learning method for solving classification and regression problems by persistently separating data dependent on an individual parameter. The choices are made in the leaves of a Decision Tree algorithm, and the data is divided into nodes. In the Classification Tree, the decision variable is stated explicitly (outcomes in either Yes or No). Both classification and regression problems can be solved with it. It is simple to understand and manipulate categorical and quantitative data. Furthermore, the tree traversal algorithm's efficiency allows it to fill missed values in attributes along with the best possible value, resulting in significant results. Over-fitting can also be solved using Decision Tree [24][26]. Decision trees are simple to build and comprehend. When the dependent variables are categorical [25], there are two distinct ways for creating a decision tree: one is the combination of information gain (1), and the other is Gini impurity (2).

Information gain = - $\sum^{m} P_i \log_2(P_i)$ (1)

Gini Impurity = 1 - $\sum_{i=1}^{n} (P_i)^2$ (2)

3.4.2 Random Forest

Random forest is a machine learning technique that overcomes these obstacles by constructing several decision trees, each of which collaborates on each data set. A random forest is essentially a bagging (bootstrap aggregating) algorithm. It integrates the output from multiple decision trees to provide the decision/prediction [25]. To solve the classification problem, we can use the Gini Index (3) or Entropy (4) to determine how nodes in a Random Forest decision tree should be related.

$$\text{Entropy} = \sum_{i=1}^{c} - P_i * \log_2(P_i) \quad (3)$$

$$\text{Gini} = 1 - \sum_{i=1}^{c} (P_i)^2 \quad (4)$$

3.4.3 Logistic Regression

To deal with a classification problem, we use Logistic Regression. It has used to forecast conditional (categorical) cases. As a result, the categorical target variable is predicted using Logistic Regression. The linear equation is passed into a sigmoid activation function (5) in Logistic Regression, which is an extension of Linear Regression. Logistic Regression represents the simplicity of implementation, computational capability, and capability from a training perspective [24] [25]. Mathematically the Logistic Regression has presented below:

$$Y = \frac{1}{(1+e^{-(a+b*x)})} \quad (5)$$

3.4.4 Naive Bayes

The Naive Bayes algorithm relies on conditional probability and is simple to use. The initial assumption is provisional independence. For this reason, it is called “naive”. The assumption that all input attributes are independent of another one. Naive Bayes (NB) Algorithm is easy to implement and provides good performance. This algorithm works with fewer training data and the number of predictors measure in a straight line. Moreover, it can handle binary classification problems and build probabilistic predictions. Besides, it can deal with consecutive and discrete data as well [24]. The Naive Bayes (NB) Algorithm depends on the Bayes theorem of probability to determine the class of unfamiliar data sets [26]. The mathematical term of Naive Bayes has presented below:

$$P(A|B) = \frac{P(B|A) * P(A)}{P(B)} \quad (6)$$

3.4.5 Neural Network

A Neural Network is a supervised learning algorithm that uses a combination of hyperparameters to estimate the likely complex relationships between input and output. Neural networks were created because linear/logistic regression can only estimate certain shapes within data, and complex functions can only approximate them. The more complex the function (some ways to avoid overfitting), the better the precision can be predicted [25].

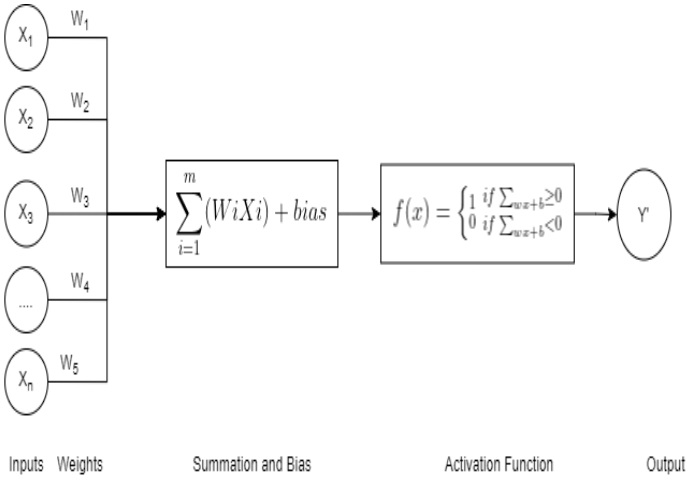

Figure 3: *A Representation of Neural Network*

4 Results, Analysis and Discussions

In this article, we used 18 separate Exhaustive-Profiling techniques to figure out which attributes are bad and which are good for our predictive models. We have performed Exhaustive-Profiling on three columns (gender, age-60-and-above, and corona-result). Because gender and age-60-and-above have many null values, and corona-result contains an irrelevant value 'other'. Our Exhaustive profiling was like gender null=drop/male/female, age-60-and-

above null=drop/yes/no, and corona-result 'other'=drop/negative. It was the combination of values among these three columns (gender, age-60-and-above, and corona-result). After executing the Exhaustive-Profiling and the comparative study, we have discovered that dropping the null values for gender, age-60-and-above, and 'other', an irrelevant value in corona-result, provides the highest and balanced Precision, Recall, and accuracy for the output class. In the below, Table 2, and Table 3 represent the comparative study of the algorithms for the COVID-19 prediction model before and after using SMOTE sampling. Table 2, and Table 3 have the Accuracy for each algorithm and precision, recall, f1-score for each output class (Corona Positive and Corona Negative).

Table 2: *Accuracy, Precision, and Recall before using SMOTE*

		DecisionTree	Random Forest	Logistic Regression	**Naïve** Bayes	NeuralNetwork
Accuracy		93.00%	93.00%	92.03%	91.71%	92.99%
Precision	Positive	65%	65%	71%	56%	56%
	Negative	96%	96%	93%	96%	96%
Recall	Positive	59%	59%	29%	65%	59%
	Negative	97%	97%	99%	95%	97%
F1_Score	Positive	62%	62%	42%	60%	62%
	Negative	96%	96%	96%	95%	96%

Table 3: *Accuracy, Precision, and Recall after using SMOTE*

		DecisionTree	Random Forest	Logistic Regression	**Naïve** Bayes	NeuralNetwork
Accuracy		80.08%	80.08%	79.79%	79.93%	80.08%
Precision	Positive	93%	93%	93%	92%	93%
	Negative	73%	73%	73%	73%	73%
Recall	Positive	65%	65%	65%	65%	65%
	Negative	95%	95%	95%	95%	95%
F1_Score	Positive	77%	77%	77%	77%	77%
	Negative	83%	83%	83%	82%	83%

So, by comparing Table 2 and Table 3, SMOTE over-sampling method performs well to make our result well-balanced. SMOTE over-sampling method increases and downgrade the data to prepare the dataset balanced. Our dataset performed unconventionally previously. Therefore by using SMOTE over-sampling, we obtained a more impressive result than before.

The ROC curve (Receiver Operating Characteristic curve) with TPR (True Positive Rate, also known as Recall) against FPR (False Positive Rate, the frequency of a false alarm) is plotted in Fig 4 and Fig 5, before and after using SMOTE sampling for the different algorithms, where TPR is on the y-axis and FPR is on the x-axis. The AUC (Area Under the ROC Curve) plots the criterion of separability, while the ROC plots a probability curve. It illustrates how well the model can distinguish between groups. The higher the AUC, the easier it is to distinguish between positive and negative classes. For some cut-off

values of the predictive measure, each point of the ROC curve reflects a True-Positive/False-Positive data pair. The point (0, 1) reflects a precise finding of 0% False-Positives and 100% True-Positives.

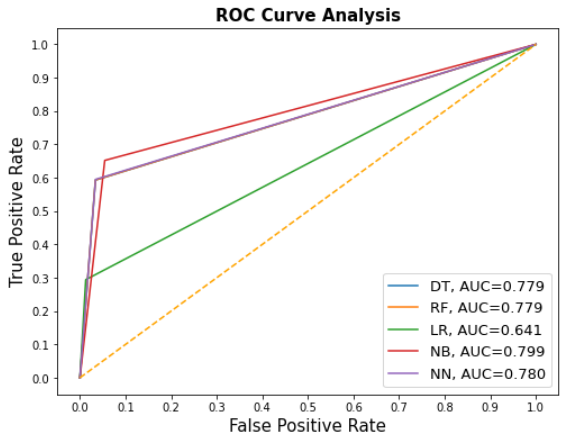

Figure 4: *ROC Curve Before using SMOTE*

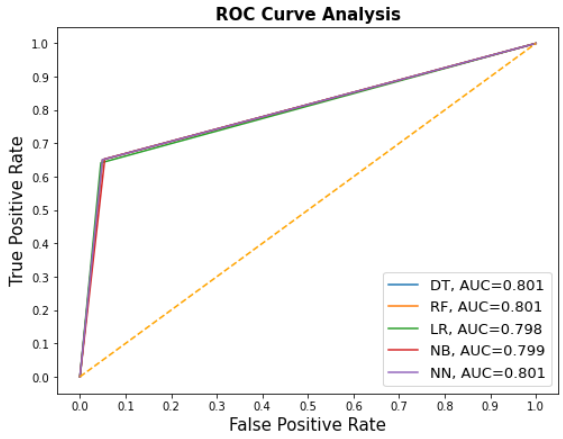

Figure 5: *ROC Curve After using SMOTE*

Fig 4 and Fig 5 are demonstrating the trade-off between sensitivity and specificity. Conspicuously, sensitivity and specificity are oppositely symmetrical. When sensitivity increases, the specificity decreases, and when sensitivity decreases, the increases. The general rules for interpreting AUC value is Table 4 [27].

Table 4: *Rules of AUC*

0<AUC≤0.5	No Distinction
0.5<AUC≤0.6	Poor Distinction
0.6<AUC≤0.7	Acceptable Distinction
0.7<AUC≤0.8	Excellent Distinction
0.9≤AUC	Outstanding Distinction

In Fig 4, the sensitivities for the different algorithms (Decision Tree, Random Forest, Linear Regression, Naive Bayes, Neural Network) before using SMOTE sampling are 59%, 59%, 29%, 65%, and 59% respectively, and AUC's are 77.9%, 77.9%, 64.1% 79.9%, and 78.0% where Linear Regression classifier provides the lowest sensitivity and AUC. After using SMOTE sampling, in Fig 5, the sensitivities for the different algorithms (Decision Tree, Linear Regression, Random Forest, Naive Bayes, Neural Network) are 65\% respectively to all, and AUC's are 80.1%, 80.1%, 79.8%, 79.9%, and 80.1% where Linear Regression classifier's sensitivity and AUC have increased significantly along with other algorithms. Hence, the sensitivity and AUC have increased significantly after using the SMOTE sampling. According to the Table 4 [27], it is an excellent distinction.

5 Conclusion

Covid-19 affective cases have been broadening continuously with exceedingly, but many medical systems provide insufficient equipment and may occur human error in the prediction of corona virus. Moreover, this situation may get challenging to prevent. Many countries are trying to invent a cure against it, but they couldn't succeed. Therefore, in this situation, the covid-19 prediction becomes more important along with other diseases. The whole world is battling against it to prevent it as soon as possible. Our study has highlighted the prediction of corona virus by the classifier algorithms and deep learning algorithm based on COVID-19's symptoms. Moreover, we have used Exhaustive profiling and SMOTE sampling to handle the imbalanced data, along with Classification algorithms and Deep learning algorithm to predict corona virus. In our research, we have learned all the models are performed better after using SMOTE sampling for the imbalanced dataset and explained each algorithm's accuracy, recall, and precision. We need to work on recall in the future to get better results for COVID-19 as it becomes a pandemic.

Code Availability

The model analytical code for reproducing the predictions and results are available at:https://github.com/mdaminulislamtushar/Thesis

Reference

[1] World Health Organization (2020) Naming the corona virus disease (COVID-19) and the virus that causes it.

[2] World Health Organization et al. (2020) Novel corona virus—China, disease outbreak news: update.

[3] World Health Organization and others (2020) WHO Director-General's opening remarks at the media briefing on COVID-19-11 March 2020, Geneva, Switzerland.

[4] The Center for Systems Science and Engineering (CSSE) at Johns Hopkins University (JHU) (2020) Covid-19 dashboard by the center for systems science and engineering (csse) at Johns Hopkins University, USA.

[5] Yazeed Zoabi, Shira Deri-Rozov and Noam Shomron (2021) Machine learning-based prediction of COVID-19 diagnosis based on symptoms. npj Digital Medicine, Nature Publishing Group,4, 1–5.

[6] Israeli Ministry of Health (2020) COVID-19 Government Data from Israeli Ministry of Health website. https://data.gov.il/dataset/covid-19

[7] Ehsan Allah Kalteh and Abdolhalim Rajabi (2020) COVID-19 and digital epidemiology.Journal of Public Health, Springer, 1–3.

[8] Kolla Bhanu Prakash, S. Sagar Imambi, Mohammed Ismail and others (2020) Analysis, prediction and evaluation of covid-19 datasets using machine learning algorithms.International Journal,8(5).

[9] Amir Ahmad, Sunita Garhwal, Santosh Kumar Ray and others (2020) The number of confirmed cases of covid-19 by using machine learning: Methods and challenges.Archives of Computational Methods in Engineering, Springer, 1–9.

[10] Furqan Rustam, Aijaz Ahmad Reshi, Arif Mehmood and others (2020) COVID-19 future forecasting using supervised machine learning models.IEEE access, IEEE,8, 101489–101499.

[11] Shreshth Tuli, ShikharTuli, Rakesh Tuli and Sukhpal SinghGilld (2020) Predicting the growth and trend of COVID-19 pandemic using machine learning and cloud computing.Internet of Things, Elsevier,11, 100222.

[12] R. Sujath, Jyotir Moy Chatterjee and Aboul Ella Hassanien (2020) A machine learning forecasting model for COVID-19 pandemic in India.Stochastic Environmental Research and Risk Assessment, Springer,34, 959–972.

[13] Peipei Wang, Xinqi Zheng, Jiayang Li and Bangren Zhu (2020) Prediction of epidemic trends in COVID-19 with logistic model and machine learning technics. Chaos, Solitons & Fractals, Elsevier,139, 110058.

[14] Dan Assaf, Ya'ara Gutman, Yair Neuman and others (2020) Utilization of machine-learning models to accurately predict the risk for critical COVID-19. Internal and emergency medicine, Springer,15(8), 1435–1443.

[15] Lamiaa A.Amar, Ashraf A.Taha and Marwa Y.Mohamed (2020) Prediction of the final size for COVID-19 epidemic using machine learning: a case study of Egypt.Infectious Disease Modelling, Elsevier,5,622–634.

[16] Akib Mohi Ud Din Khanday, Syed Tanzeel Rabani, Qamar Rayees Khan and others (2020) Machine learning based approaches for detecting COVID-19 using clinical text data.International Journal of Information Technology, Springer,12(3), 731–739.

[17] Adam L. Booth, Elizabeth Abels and Peter McCaffrey (2021) Development of a prognostic model for mortality in COVID-19 infection using machine learning.Modern Pathology, Nature Publishing Group,34(3), 522–531.

[18] Israeli Ministry of Health (2020) COVID-19 Government Data Information from Israeli Ministry of Health website.https://data.gov.il/dataset/covid-19/resource/3f5c975e-7196-454b-8c5b-ef85881f78db/download/-readme.pdf

[19] Ekaba Bisong (2019) Google colaboratory: Building Machine Learning and Deep Learning Models on Google Cloud Platform, Springer,59–64.

[20] Nitesh V. Chawla, Kevin W. Bowyer, Lawrence Oand Hall and W. Philip

Kegelmeyer (2002) SMOTE: synthetic minority over-sampling technique.Journal of artificial intelligence research,16, 321–357.

[21] Tingting Pan, Junhong Zhao, Wei Wu and Jie Yang (2020) Learning imbalanced datasets based on SMOTE and Gaussian distribution.Information Sciences, Elsevier,512, 1214–1233.

[22] Susmita Ray (2019) A quick review of machine learning algorithms,2019 International conference on machine learning.Big data, cloud and parallel computing (COMITCon), IEEE, 35–39.

[23] Ayyadevara and V. Kishore (2018) Pro machine learning algorithms, Apress: Berkeley, CA, USA, Springer.

[24] Mokhairi Makhtar, Hasnah Nawang and Syadiah Nor Wan Shamsuddin (2017) Analysis On Students Performance Using Naive Bayes Classifier. Journal of Theoretical& Applied Information Technology, 95(16).

[25] Shengping Yang and Gilbert Berdine (2017) The receiver operating characteristic (ROC) curve.The Southwest Respiratory and Critical Care Chronicles,5(19), 34–36.

[26] Md. Ahsan Arif, Mausumi Islam Mau, Asma Jahan y Razia Tummarzia (2021) An Improved Prediction System of Students' Performance Using Classification model and Feature Selection Algorithm. En: International Journal of Advances in Soft Computing and its Applications 13.1 (march 2021), pp. 162-177.

[27] Md. Ahsan Arif (2011) Problems and Prospects: Universal Networking Language on Bangla Sentence Structure Perspective. En: International Journal of Engineering & Technology IJET-IJENS, Vol: 11, No: 04, 119-126.