Machine Learning Classifier Model for Prediction of COVID-19

Jhimli Adhikari
Narayan Zantye College, Goa, India

COVID-19 pandemic has become a major threat to the world. In this study a model is designed which can predict the likelihood of COVID-19 patients with maximum accuracy. Therefore three machine learning classification algorithms namely Decision Tree, Naive Bayes and Logistic Regression classifier are used in this experiment to detect Covid-19 disease at an early stage. The models are trained with 75% of the samples and tested with 25% of data. Since the dataset is imbalanced, the performances of all the three algorithms are evaluated on various measures like F-Measure, Accuracy and Matthews Correlation Coefficient. Accuracy is measured over correctly and incorrectly classified instances. All the analyses were performed with the use of Python, version 3.8.2. Receiver Operating Characteristic (ROC) curves are used to verify the result in a proper and systematic manner. This framework can be used, among other considerations, to prioritize testing for COVID-19 when testing resources are limited.

Keywords: Covid-19, Classification, Machine learning, Outbreak, Prediction.

1. INTRODUCTION

The novel coronavirus disease 2019 (COVID-19) presents most shocking and urgent threat to global health. It has already spread over 185 countries and territories around the world and nearly seven lakhs thirty eight thousands of people [1] already died by this infectious disease. According to WHO (World Health Organization) reports [2] COVID-19 is a severe acute respiratory syndrome which is transmitted through respiratory droplets and contact routes. Thus, analysis of this disease requires major attention by the Government and several autonomous bodies to take necessary steps in reducing the effect of this global pandemic. It is extremely critical to detect the positive cases specially asymptomatic as early as possible so as to prevent the further spread of this epidemic and to quickly treat affected patients. In this context, Machine Learning based classifier modeling J. D. Kelleher [2015] is emerging as a powerful weapon in humanity's fight against disease outbreaks. There exist a large number of research works where machine learning algorithms have been applied to give efficient predictions in healthcare M. A. Ahmad [2018], A. Callahan [2017], P. Chowriappa [2013].

When a new pandemic hits, diagnosing individuals is challenging. Moreover, there are many countries where testing on a large scale is difficult and tests are likely to be expensive, especially in the beginning. Anyone who has any symptoms of COVID-19 is likely to be very concerned that they have contracted the disease, even if the same symptoms are indicative of many other, potentially milder diseases too. Some people have extremely mild cases while others find themselves fighting for their lives. The symptoms of the disease resemble a typical viral, respiratory infection with an incubation time of 14 days. In this article we have shown how classification algorithm such as Decision Tree, Naive Bayes and Logistic Regression can accurately predict whether a person is infected with coronavirus based on a range of symptoms a person experiences. This will help to identify those with COVID-19 in populations that are experiencing limited clinical testing.

We established a machine learning approach that trained on records from 5618 tested individuals

[1] https://www.worldometers.info/coronavirus, (accessed on 11th August, 2020)

[2] World Health Organization, (2020), Coronavirus disease 2019 (COVID-19): situation report, 98

(of whom 610 were confirmed COVID-19 positive cases). Clearly here data classes are imbalanced as it is unevenly distributed S. Daskalaki [2006]. Most standard machine learning algorithms work well with balanced dataset where both the classes (binary classification) are equally distributed, but they face challenges when the dataset classes are imbalanced S. Boughorbel [2017]. In such situation, classifiers tend to be biased towards the majority class and the results are wrongly interpreted. As the data imbalance is more common in some real life applications such as fraudulent transaction, detection of disease, performance of the classifier must be carried out using adequate metrics to pay more attention to the minority class. Thus, one of the good approaches to deal with this issue is to incorporate other metric to handle data imbalance. Since Matthews Correlation Coefficient (MCC) is widely used in Bioinformatics as a performance metric, Matthews [1975], we incorporated this metric along with F-measure, ROC for the algorithms. Experimental performance of all the three algorithms is compared on various measures and achieved good accuracy.

After presenting background in section 1, related work is described in section 2. Section 3 presents the methodology about the model used in this study. Section 4 covers data characteristics, experimental results and model evaluation. Conclusion is presented in section 5.

2. RELATED WORK

R. Sujath [2020] presented a model to predict the spread of COVID-19. Authors performed Linear regression, Vector autoregression and Multilayer perceptron method for COVID-19 data to anticipate the epidemiological example of the ailment and pace of COVID-19 cases in India.

A. Z. Khuzani and Shariati [2020] hypothesized that machine learning-based classifiers can reliably distinguish the Chest-X-Ray (CXR) images of COVID-19 patients from other forms of pneumonia. A dimensionality reduction method is used by author to generate a set of optimal features of CXR images. Later an efficient machine learning classifier is built that can distinguish COVID-19 cases from non-COVID-19 cases with high accuracy and sensitivity.

A ubiquitously deployable AI-based preliminary diagnosis tool for COVID-19 was presented by A. Imran [2020] using cough sound via a mobile app. This research work proposed and developed a tri-prongedmediator centered AI-engine for the cough-based diagnosis of COVID-19, named AI4COVID-19. The results show that the AI4COVID-19 app is able to diagnose COVID-19 with negligible misdiagnosis probability. A. Waheed [2020] proposed an ACGAN (Auxiliary Classifier Generative Adversarial Network) based model called CovidGAN that generates synthetic CXR images to enlarge the dataset and to improve the performance of CNN in COVID-19 detection. Authors demonstrated that the synthetic images produced from CovidGAN can be utilized to enhance the performance of CNN for COVID-19 detection.

C. Iwendi [2020] proposed a fine-tuned Random Forest model boosted by the AdaBoost algorithm. The model used the COVID-19 patients travel, health, geographical and demographic data to predict the severity of the case and the possible outcome, recovery, or death.

Md. M. Ahamad [2020] developed a supervised machine learning model by examining the details of the individuals e.g. age, gender, observation of fever, history of travel, and clinical details such as the severity of cough and incidence of lung infection. Our work is different from the existing literature as this study proposed three machine learning algorithms to predict COVID-19 positive patients by considering some symptoms like cough, fever, sore throat, shortness of breath etc.

3. METHODOLOGY

Classification is a two-step process, learning step and prediction step, in machine learning. In the learning step, the model is developed based on given training data. In the prediction step, the model is used to predict the response for given data. Proposed procedure of the study is summarized in Figure 1 in the form of model diagram. To build the model and predict COVID-19 disease we proceed as follows. The entire study is divided into following steps: (i) Preprocess: Preprocessing is required in order to build a better model with good prediction. Here missing

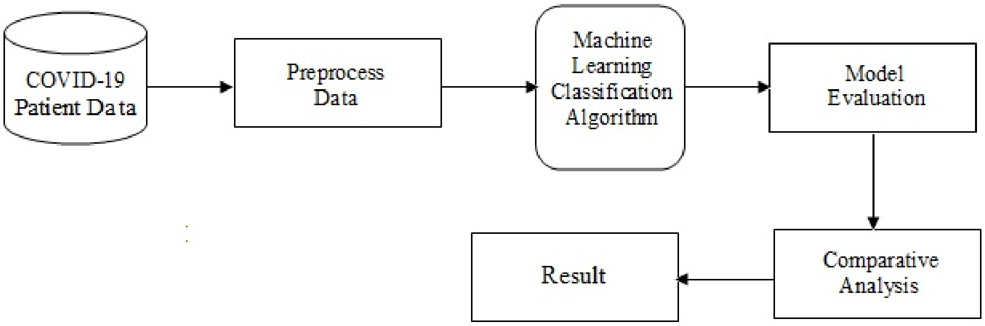


Figure. 1: Proposed Model Diagram

values, duplicate records, encoding categorical variables are taken care using preprocessing tools. (ii) Data is split into training and testing set (iii) We train the models with 75% of the samples (4213 instances) and test with the remaining 25% (1405 instances). (iv) Models are evaluated using statistical score.

3.1 Algorithm used for classification

In this article Decision Tree, Naive Bayes and Logistic Regression Flach [2012] are used as classifier. A confusion matrix is a table that is used to describe the performance of a classification model on a set of test data for which the true values are known. Outcome of Confusion Matrix is used to indicate the summary of prediction results including correct and incorrect on a classification problem. In addition, this is used to not only errors but also types of errors. Lets understand the segments of the confusion matrix below.

—True Positives (TP): COVID-19 cases which are predicted yes (they have the disease), and they do have the disease.

—True Negatives (TN): COVID-19 cases which are predicted no, and they do not have the disease.

—False Positives (FP): COVID-19 cases which are predicted yes, but they do not actually have the disease (Type I error).

—False Negatives (FN): COVID-19 cases which are predicted no, but they actually do have the disease (Type II error).

A large research work discussed on how to assess the quality of classifiers on imbalance dataset, and a large set of metrics proposals that address one or more of the problems. As we discussed above, an important aspect of imbalanced problems is that misclassifying a positive case (a false positive) should be costlier than misclassifying a negative case (a false negative). Accuracy does not take those differences into consideration. MCC is a correlation coefficient between target and predictions. Thus, its value lies between -1 and +1. -1 interprets when there is perfect disagreement between actual value and predicted value, 1 when there is a perfect agreement between actual and predictions. 0 represents when the prediction may be random with respect to the actuals. As confusion matrix involves values of all the four quadrants, it is considered as a balanced measure. Accuracy, Precision, Recall, F-Measure, ROC (Receiver Operating Curve) and MCC measures are used for the classification of this work. Table I defines accuracy measures below:

3.2 Algorithm

Inputs: COVID-19 symptom Database for patients
Outputs: Classification of Covid-19 disease

(i) Read the data

Table I: Accuracy Measures

Measures	**Definition**	**Formula**
Accuracy (A)	Accuracy represents the accuracy of the algorithm in predicting instances	A=$\frac{(TP+TN)}{TotalNo.ofSamples}$
Precision (P)	Classifiers correctness / accuracy is measured by Precision	$P = \frac{TP}{(TP+FP)}$
Recall (R)	Recall is used to measure the classifiers completeness or sensitivity	$R = \frac{TP}{(TP+FN)}$
F-Measure (F)	F-Measure is the weighted average of precision and recall	$F = \frac{2*(P*R)}{(P+R)}$
ROC	ROC (Receiver Operating Curve) curves are used to compare the usefulness of tests	
Matthews Correlation Coefficient (MCC)	The MCC describes how changing the value of one variable will affect the value of another and returns a value between -1 and 1	$MCC = \frac{(TP \times TN)-(FP \times FN)}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$

(ii) Split data for Training and Testing

(iii) Apply Classification Algorithm (Decision Tree, Naive Bayes and Logistic Regression) on Training dataset

(iv) Test Classifier Algorithm using Test data

(v) Generate Confusion matrix and classification report

(vi) Evaluate the Performance of the model

(vii) Stop

This paper will focus on 3 metrics F-Measure, ROC and MCC to evaluate the model. In the following sections algorithms and results of confusion matrix are presented.

3.2.1 Decision Tree Algorithm

Decision Tree is one of the easiest and popular classification algorithms to understand and interpret. It belongs to the family of supervised learning algorithms. The goal of using a Decision Tree is to create a training model that can use to predict the class or value of the target variable by learning simple decision rules inferred from prior data (training data). In Decision Trees, for predicting a class label for a record we start from the root of the tree, then the values of the root attribute are compared with the record's attribute. On the basis of comparison, we follow the branch corresponding to that value and jump to the next node. The evaluated performance of Decision Tree technique using Confusion Matrix is shown in Table II.

Table II: Confusion Matrix of Decision Tree

Confusion Matrix	Class	Precision	Recall	F-Measure	Support
[[1260 0] [136 9]]	0	0.90	1.00	0.95	1260
	1	1.00	0.06	0.12	145
	accuracy			0.90	1405
	macro avg	0.95	0.53	0.53	1405
	weighted avg	0.91	0.90	0.86	1405

3.2.2 Naive Bayes Classifier

Naive Bayes Classifier is a probabilistic classifier, which means it predicts on the basis of the

probability of an object. Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature. It defines that status of a specific feature in a class does not affect the status of another feature. Since it is based on conditional probability it is considered as a powerful algorithm employed for classification purpose. Naive Bayes is a machine learning classifier which employs the Bayes Theorem. Bayes theorem provides a way of calculating the posterior probability, $P(A|B)$ from P(A), P(B) and $P(B|A)$. Therefore, $P(A|B) = \frac{(P(B|A)P(A))}{P(B)}$ where $P(A|B)$ is the posterior probability of class (A, target) given predictor (B, attributes)
P(A) is the prior probability of class
$P(B|A)$ is the likelihood which is the probability of predictor given class
P(B) is the prior probability of predictor
The evaluated performance of Naive Bayes algorithm using Confusion Matrix is presented in Table III.

Table III: Confusion Matrix of Naive Bayes Classifier

Confusion Matrix	Class	Precision	Recall	F-Measure	Support
[[1243 17] [135 10]]	0	0.90	0.99	0.94	1260
	1	0.37	0.07	0.12	145
	accuracy			0.89	1405
	macro avg	0.64	0.53	0.53	1405
	weighted avg	0.85	0.89	0.86	1405

3.2.3 Logistic Regression Classifier

Logistic Regression is a statistical and machine-learning technique classifying records of a dataset based on the values of the input fields. It predicts a dependent variable based on one or more set of independent variables to predict outcomes. It can be used both for binary classification and multi-class classification. The evaluated performance of Logistic Regression using Confusion Matrix is shown in Table IV.

Table IV: Confusion Matrix of Logistic Regression Classifier

Confusion Matrix	Class	Precision	Recall	F-Measure	Support
[[1255 5] [141 4]]	0	0.90	1.00	0.95	1260
	1	0.44	0.03	0.05	145
	accuracy			0.90	1405
	macro avg	0.67	0.51	0.50	1405
	weighted avg	0.85	0.90	0.85	1405

4. EXPERIMENT

The proposed system is implemented for COVID-19 diagnosis using Python 3.8.2 programming language with a processor of Intel Core i5-8300H CPU @ 2:30GHz and RAM of 8 GB running on Windows 10. The main aim of this study is the prediction of the patient affected by COVID-19 using Python by using the database. Table V shows a brief description of the dataset. All the data used in this study were retrieved from the Israeli Ministry of Health Website [3]. The dataset was downloaded, translated into English, and can be accessed at [4]. However, this dataset has been pre-processed further to meet the needs of this study. This dataset is licensed and available for general usage. For the model architecture, a total of 5618 patients are used with 4213 (75%) patients for training and the remaining 1405 (25%) in testing.

[3] https://data.gov.il/dataset/covid-19

[4] https://github.com/nshomron/covidpred

4.1 Data Characteristics

The proposed methodology is evaluated on COVID-19 data collected from a publicly available dataset. The dataset provides the COVID patients' information. Data was collected in CSV file uploaded in Jupyter notebook and analysed with Python 3.8.2 software. Datasets also contain some missing values. Thus, NA/N values are converted to zero or they may be neglected. This dataset comprises of medical detail of 5,618 instances which include female and male patients. The dataset also comprises numeric-valued 9 attributes where value of one class "0" treated as negative and value of another class "1" is treated as positive for symptoms of COVID-19 disease. Table V consists of data type information about the attributes. Fever, cough, cold, sore throat, headache, shortness of breath are the most common symptoms that are noticed in patients whose data is available in this dataset and are shown in Table VI. Figure 2 shows the frequency distribution of age of patients. We could see that maximum number of patients' age lies between 50 to 60. Figure 3(a) shows the correlation heatmap between attributes. Figure 3(b) shows the correlation between attributes and the outcome i.e. whether patient is COVID positive or negative. As Figure 3(b) illustrates, some attributes like shortness of breath and sore throat are the top features with high correlation to the patient's COVID positivity risk. Table VII represents different performance values of all classification algorithms calculated on various measures. Maximizing precision will minimize the number of false positives, whereas maximizing the recall will minimize the number of false negatives. Sometimes, we want excellent predictions of the positive class. We want high precision and high recall. This can be challenging, as increases in recall often come at the expense of decreases in precision.

Table V: Data type of attributes

Column	Type
Patient Id	Numeric
Age	Numeric
Gender	Categorical
Cough	Categorical
Fever, n (%)	Categorical
sore throat, n (%)	Categorical
shortness of breath, n (%)	Categorical
headache, n (%)	Categorical
Chronic Disease, n (%)	Categorical
SARS-Cov-2 positive result	Categorical

Table VI: Data description and attribute information

Number of instances (5618)	Number of Attributes (9)
Attribute	Characteristic
Basic Information	
Age	Mean age =52.54
Gender: Male, n (%)	2955 (52.6%)
Gender: Female, n (%)	2663 (47.4%)
Symptom (0 / 1)	Class (1)
Cough, n(%)	54(0.96%)
Fever, n (%)	44(0.78%)
sore throat, n (%)	28(0.50%)
shortness of breath, n (%)	23(0.41 %)
headache, n (%)	30(0.53%)
Chronic Disease, n (%)	225(4.00%)
SARS-Cov-2 positive result	610(10.86%)

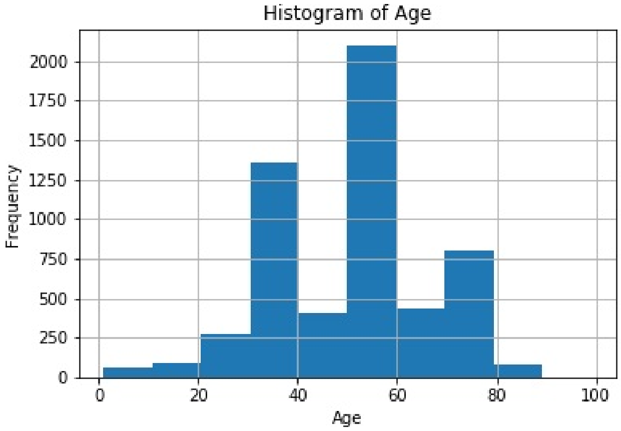
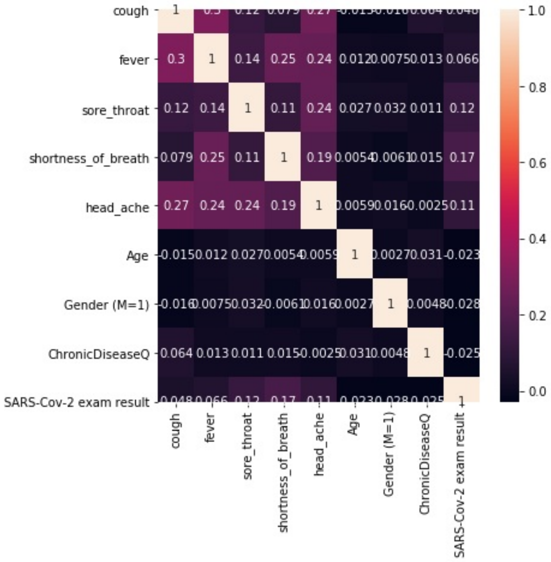


Figure. 2: Histogram of Age



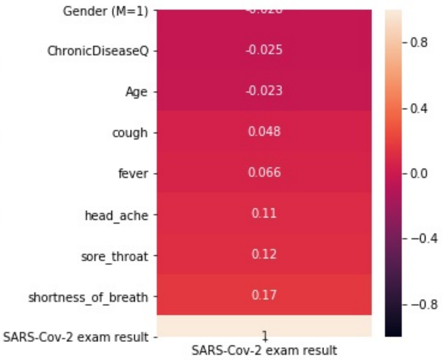


Figure 3(a):Correlation heatmap between attributes

Figure 3(b):Correlated heatmap for the most correlated attributes

Accuracy is used when the True Positives and True negatives are more important while F-Measure is used when the False Negatives and False Positives are crucial. Accuracy can be used when the class distribution is similar while F-Measure is a better metric when there are imbalanced classes as in the above case. In most real-life classification problems, imbalanced class distribution exists and thus F-Measure is a better metric to evaluate our model on. From

Table VII: Comparative Performance of Classifiers on Various Measures

Classification Algorithm	F-Measure	Accuracy %	ROC	MCC	Correctly Classified Instances	Incorrectly Classified Instances
Decision Tree	0.95	90	0.733	0.24	1269	136
Naive Bayes Classifier	0.94	89.2	0.63	0.123	1253	152
Logistic Regression	0.95	89.6	0.60	0.09	1259	146

Table VII it is analyzed that Decision Tree shows the maximum accuracy. So the Decision Tree machine learning classifier can predict the chances of COVID-19 with more accuracy as compared to other classifiers. According to these classified instances, accuracy is calculated and analyzed. Performance of individual algorithm is evaluated on the basis of Correctly Classified Instances and Incorrectly Classified Instances out of a total number of instances.
Figure 4 shows ROC curve of three classifiers used in prediction phase of COVID cases.

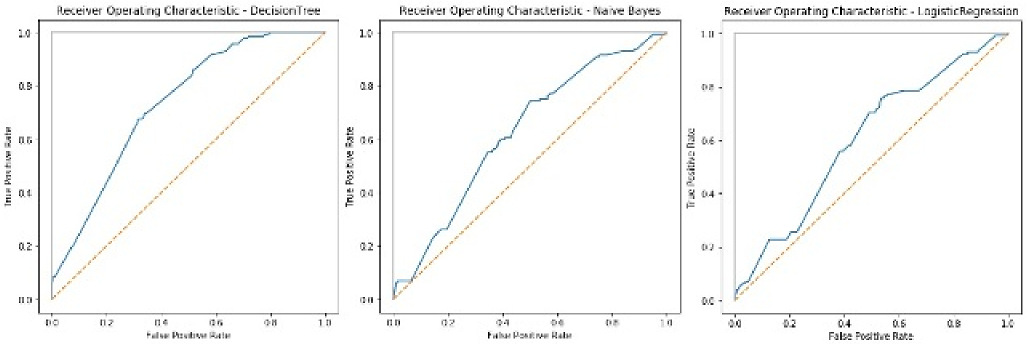


Figure. 4: ROC curve of three classifiers

5. CONCLUSION AND FUTURE WORK

The COVID-19 pandemic poses a historical challenge across the world. The most important problem is to fight with COVID-19 pandemic by detection of disease at its early stage. In this study, systematic efforts are made in designing a model which results in the prediction of COVID-19. During this work, three machine learning classification algorithms are studied and evaluated on various measures. Experiments are performed on COVID-19 database of patients. Model evaluation using Decision Tree classification algorithm yielded better accuracy when compared to other algorithms. Since imbalanced data set pose important challenges to existing approaches to predictive modeling, in future, the better hybrid system with machine learning classification algorithms can be used to predict or diagnose other diseases. The work can be extended and improved for the automation of COVID-19 analysis including some other machine learning algorithms.

References

A. Callahan, N. H. S. 2017. Machine learning in healthcare key advances in clinical informatics: Transforming health care through health information technology. *Elsevier*.

A. Imran, I. Posokhova, H. N. Q. U. M. S. R. K. A. C. N. J. I. H. M. N. 2020. Ai4covid-19: Ai enabled preliminary diagnosis for covid-19 from cough samples via an app. *Informatics in Medicine*.

A. Waheed, M. Goyal, D. G. A. K. F. A. P. R. P. 2020. Covidgan: Data augmentation using auxiliary classifier gan for improved covid-19 detection. *IEEE Access*.

A. Z. Khuzani, M. H. and Shariati, A. 2020. Covid-classifier: An automated machine learning model to assist in the diagnosis of covid-19 infection in chest x-ray images. *Preprint. medRxiv.2020.*.

C. Iwendi, A. K. Bashir, A. P. R. S. J. M. C. S. P. R. M. S. P. O. J. 2020. Covid-19 patient health prediction using boosted random forest algorithm. *Front. Public Health*.

Flach, P. 2012. Machine learning: The art and science of algorithms that make sense of data, intelligent systems laboratory. *University of Bristol, UK, Cambridge University Press*.

J. D. Kelleher, B. M. Namee, A. D. 2015. *Fundamentals of Machine Learning for Predictive Data Analytics: Algorithms.* MIT Press.

M. A. Ahmad, C. Eckert, A. M. 2018. Interpretable machine learning in healthcare. In *Proceedings of ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics.* pp.559560.

Matthews, B. 1975. Comparison of the predicted and observed secondary structure. *Biochimica et biophysica acta 405,* No.2, 442451. pmid:1180967.

Md. M. Ahamad, S. Aktar, M. R. M. S. U. P. L. H. X. M. A. S. J. M. Q. M. A. M. 2020. A machine learning model to identify early stage symptoms of sars-cov-2 infected patients. *Expert Syst Appl*.

P. Chowriappa, S. Dua, Y. T. 2013. Introduction to machine learning in healthcare informatics. *Intelligent Systems Reference Library book series ISRL, volume 56.*

R. Sujath, J. M. Chatterjee, A. E. H. 2020. A machine learning forecasting model for covid-19 pandemic in india. *Stoch Environ Res Risk Assess 34, Springer*, 959972.

S. Boughorbel, F. Jarray, M. E.-A. 2017. Optimal classifier for imbalanced data using matthews correlation coefficient metric. *PLOS ONE*.

S. Daskalaki, I. Kopanas, N. A. 2006. Evaluation of classifiers for an uneven class distribution problem. *Applied artificial intelligence*.

Dr. Jhimli Adhikari received Master of Computer Application and Ph D in Computer Science from Jadavpur University, Kolkata and Goa University, respectively. At present, she is Associate Professor in the Department of Computer Science, Narayan Zantye College, Goa, India. She has twenty three years of teaching experience. Her areas of research interest include data mining and knowledge discovery, decision support systems and data science. She is coauthor of two research monographs and published seven international journal papers, three international conference papers, one book chapter and one book review. She is regular reviewer of Pattern Recognition Letters, Elsevier and Editorial board member of International Journal of Image Processing and Pattern Recognition, JournalsPub.

Adhikari.jpg