Grenze International Journal of Engineering and Technology, Jan Issue

Analysis and Prediction of COVID-19 Disease using Machine Learning

Shabnam Parmar[1], Rinkle Rani[2] and Nidhi Kalra[3]
[1-3]Department of Computer Science and Engineering, Thapar Institute of Engineering and Technology, Patiala, India
Email: sparmar_me20@thapar.edu, raggarwal@thapar.edu, nidhi.kalra@thapar.edu

***Abstract*—In this research, the symptoms and other factors of a patient are utilized to train machine learning algorithms to predict whether the patient would die from or recover from (COVID19). It is probable that the coronavirus (COVID19) will create the highly infectious Coronavirus illness COVID19 (SARSCoV2). By coughing, sneezing, speaking, or inhaling, this virus can be transmitted from an infected person's lips and nose to small liquid particles. The size of these tiny atoms' inhalation droplets and aerosols vary, with bigger droplets being larger than smaller atoms. COVID19 is transferred by inhalation or by touching your eyes, nose, or mouth with your fingers after meeting a contaminated surface. When a big number of people are present, the COVID-19 virus can spread rapidly. We will need to examine the COVID19 dataset to see which models are the most accurate in estimating fatality rates for the virus's most vulnerable victims. Machine learning is used to compute and evaluate the performance of a variety of prediction models. We have used K-nearest neighbor (KNN), Support Vector Machine (SVM) classifier, Gaussian naive Bayesian (GNB), Decision tree (DT), and Logistic regression (LR) for the prediction of death and recovery of symptomatic patients. In this research, a variety of feature selection and extraction strategies were used, and prediction accuracy for feature selection methods for the KNN model and feature extraction methods for the GNB model both reached up to 96 percent. The k-nearest neighbor has performed and predicted high accuracy of 96% in both feature selection and extraction techniques.**

***Index Terms*— COVID19, Decision Trees, Gaussian Naive Bayesian, KNN, Logistic Regression, Machine Learning, SVM, PCA, Fast-ICA, K-Fold, Feature Selection, and Extraction.**

I. INTRODUCTION

Recently, the globe has seen fast technological advancement, which demonstrates the critical role of progressive countries. Today, every aspect of society, including education, employment, trade, military, and media, as well as manufacturing and healthcare, is obsessed with continuing and emerging technological advancements. The centre may be a critical location for the rapid adoption of new technologies, ranging from diagnosis to accurate identification and automated analysis of patients. Coronavirus2 (SARSCoV2) source of severe inhaling pollution and anarchy, in resultant 2019 (COVID19) was identified in humans because of initial cases in the Chinese city of Wuhan in December 2019 [1].

Machine Learning is a promising categorization technology. In general, machine learning is a useful framework for inducing an unknown purpose, relationship, or structure between output and input variables. Generally, all these interactions are extremely difficult to handle using explicit models and machine learning, which is why they are mostly employed to anticipate or forecast the possible number of confirmed cases and hence the number

Grenze ID: 01.GIJET.9.1.577

of impending deaths [2]. Most viruses, except for SARSCoV2, are the virus that causes COVID19, mostly due to temporal changes. Generally, these alterations had little or no effect on the virus's features. Case recognition (COVID19) using a machine learning system is one perspective [3]. There are models available that can assess medical pictures and diagnose COVID19. Human coronaviruses can generate a variety of diseases ranging from the common cold to more serious conditions such as Mideast Respiratory Syndrome (MERS, 34% mortality). This is the eighth known coronavirus to infect humans, known as SARSCoV2, following NL63, MERS-COV, 229E, OC43, and HKU1, as well as the first SARS-COV. On COVID19, we used several automated classification techniques by comparing common feature extraction frameworks that allow machine learning to choose the most accurate feature, which may be a critical learning module. The classification is accomplished by utilizing the information acquired from multiple machine learning models to identify them as COVID19 or other contamination disease cases. Throughout this study, we used Python and its libraries, including SK-learn, Pandas, NumPy, and Machine Learning Classifiers (KNN Classifier, Logistic Regression, and Naive Bayes Classifier) [4]. We used the COVID dataset for this analysis, and we used a spread technique and feature extraction to compare the accuracy of various algorithms. This study is divided into five stages: data pre-processing (cleansing, transformation, and reduction), encoding, feature selection or extraction, visualization, and prediction. We trained several machine learning models or algorithms on the dataset and projected the accuracy score. It is used to estimate the likelihood of a symptomatic patient dying or recovering based on the features. Given a dataset as input to any or all machine learning algorithms, 1 indicates positive and 0 indicates negative [5]. For the accuracy analysis in this work, we used the trained Normal and K-Fold datasets. K-NN, LR, and DT feature selection strategies all produced highly promising results, as did GNB, SVM, K-NN, and LR extraction techniques combined with PCA [6].

II. Material and Method

SVM, K-NN, Naive Bayes, Logistic regression, and Decision trees were employed in this work with diverse selection and data extraction approaches [7]. Implementation of a machine learning model using the Python programming language and its built-in libraries such as NumPy, Pandas, and sci-kit-learn. Figure 1 depicts the machine learning algorithmic workflow for the COVID-19 Prediction task.

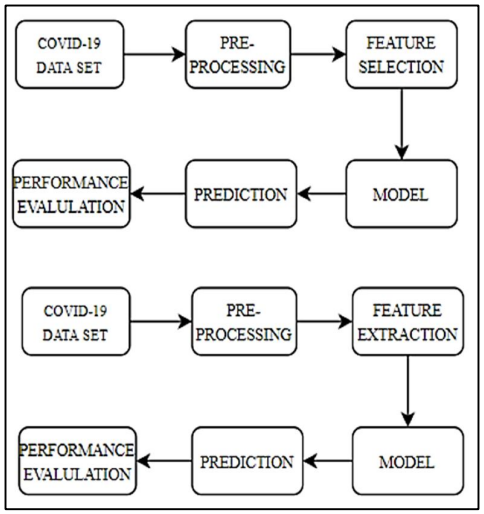

Figure1.Workflow of the proposed model

1) Select the data set and apply the appropriate feature selection technique for expected accuracy.
2) Select good machine learning classifiers to train a model on the data set.

A. Data Set

Data set has been used that consists of various attributes. Using these attributes, we try to find correlations so that a proper analysis of the situation can be made. We have used Python programming language and various in-built libraries of python for the purpose of Data Mining and Analysis. Data set contains different parameters which were registered at different points of time based on the patients identified [8].

1) date – patient identified on a date.
2) place – patient location.
3) nation – identified patient country.
4) gender – patient gender (Male-1, Female-0).
5) age – patient age.
6) visiting Wuhan – have the patient visit Wuhan in recent time
7) Wuhan resident – is symptomatic from Wuhan.
8) death – does because of Covid-19 patient died (0 or 1).
9) recovered – does from Covid-19 the patient recovered (0 or 1).

B. Preprocessing

Data is imported in the juypter notebook and python environment for analysis. The data is a size of 1085 rows and contains 20 columns. We have performed pre-processing normalization on the null values columns and raw data can be reformatted and transformed into the machine, understandable format. We mostly work on the dataset that has labeled one, or more columns and it can be in strings or numbers. Data is made understandable by labeling the training dataset into strings or words. Label Encoding is performed to transform other than numerical values into numerical labels. If labels are numeric, it's easy for machine learning models to decide how those labels must be operated. It's a very important step of pre-processing for the dataset that is structured in super-vised learning.

C. Methodologies

Machine learning has three types of learning supervised, unsupervised, and reinforcement. In this research supervised learning is used during implementation as shown in Fig 2. Simple Algorithms, as well as ensembles techniques, are used to achieve good performance for prediction as ensemble methods use multiple learning algorithms.

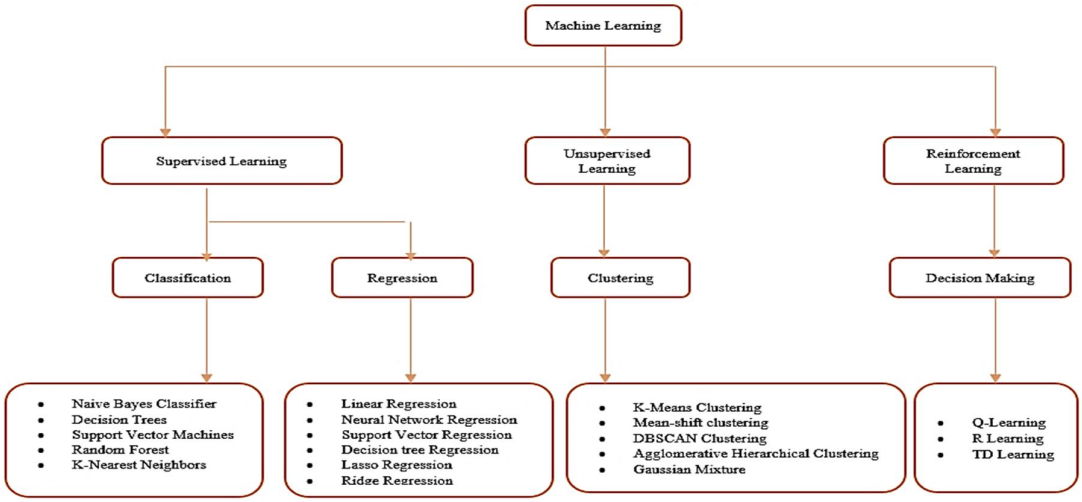

Figure 2. Types of machine learning

D. Attribute Selection

Different types of supervised and unsupervised learning models were performed to check the accuracy as mentioned below. Attribute selection is a selection of features by reducing the no of input variables for developing a machine learning model called feature selection. Which is performed on the data set before applying to the machine model, it is a method used for dimension reduction to predict better results. In this study, we have used different selection and extraction techniques to reduce the number of features the of data set.

Feature Selection Techniques

1) Select K-Best: - Select features according to the k=5 highest scores. This is the Feature selection technique where we choose those attributes in our data that will contribute most to the target variable.
- f_regression- f_classif- ch2

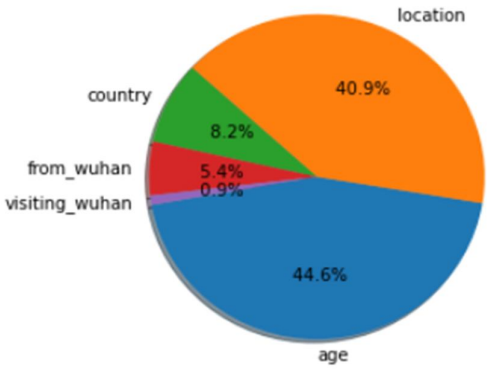

Figure 3. Top 5 features selected with Chi2

2) Forward Feature Selection: - This decision is based on the iteration process, and it began with none of the model's characteristics. At every iteration, it keeps on adding features that better enhance models until adding a new attribute does not increase model performance.
3) Backward Feature Selection: - When creating a machine learning model, the approach picks this feature. This is used to filter away characteristics that have no bearing on the output's dependent attribute or prediction.
4) Random Forest Importance: -The feature importance (variable importance) describes which features are relevant.

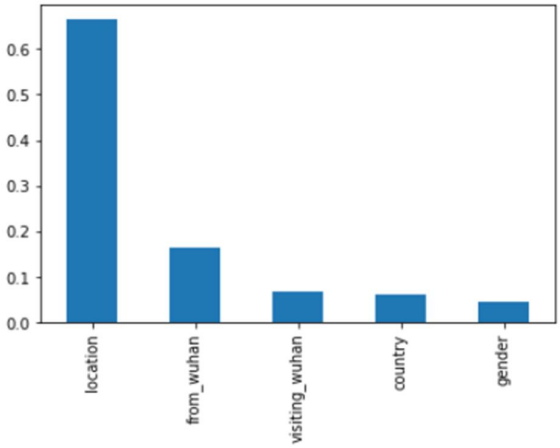

Figure 4. Top 5 features selected with Random Forest Importance

5) Recursive Feature Elimination (RFE): -It is a feature selection strategy for fine-tuning a model by removing the weakest features until the desired number of features is obtained.

Result of Different Selection Techniques

A tabular representation of the top 5 feature selections with different feature selection techniques is shown below in TABLE I. And these different selection techniques will be used to select these top features on different models to compare accuracy.

Feature Extraction Techniques

1) Principal Component Analysis (PCA): - PCA should be performed on the dataset to limit it to only two dimensions. PCA is an unsupervised learning approach utilized in machine learning to cut back dimensionality. it's a method that uses an orthogonal transformation to convert observations of correlated characteristics into a group of linearly uncorrelated features, as shown in Figure 5.
2) Linear Discriminant Analysis (LDA): - LDA was used to limit our data set to a single feature. This is a supervised classification strategy used in the development of competitive machine learning models.

TABLE I. DIFFERENT FEATURE SELECTION TECHNIQUES AND RESULTS

Feature Selection	
Feature Selection Name	Results
1) Univariate Feature Selection	
• Select Kbest with f_regression	['location', 'country', 'age', 'visiting_wuhan', 'from_wuhan']
• Select Kbest with f_classif	['location', 'country', 'age', 'visiting_wuhan', 'from_wuhan']
• Select Kbest with Chi2	['location', 'country', 'age', 'visiting_wuhan', 'from_wuhan']
2) Forward Feature selection	['location', 'country', 'age', 'visiting_wuhan', 'gender']
3) Backward Feature Elimination	['location', 'country', 'age', 'visiting_wuhan', 'gender']
4) Random Forest Importance	['location', 'country', 'gender', 'visiting_wuhan', 'from_wuhan']
5) Recursive Feature Elimination	['age', 'country', 'gender', 'visiting_wuhan', 'from_wuhan']

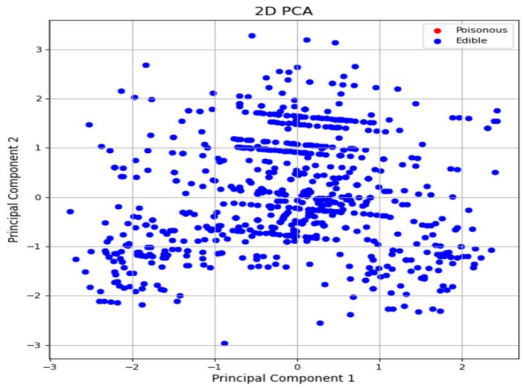

Figure 5. Extraction of Feature with PCA

Dimensionality reduction in this category is employed in fields such as picture recognition and predictive analysis in marketing.

3) Independent Component Analysis (ICA): -Using ICA, we can reduce our data set to just four features. Fast ICA is a right and famous set of rules for independent component analysis invented through Aapo Hyvärinen at Helsinki University of Technology.
4) Locally Linear Embedding (LLE): We can run LLE on our data set to reduce our data dimensionally to 3 dimensions. It is an unsupervised learning algorithm that generates low dimensional embeddings of high dimensional input, relating each learning instance to its nearest neighbor.

E. Algorithms

Following are the algorithms used on different features selected shown in TABLE I. with selection techniques. The different algorithms used are KNN, SVM, LR, DT, and GNB [9].

1) K-Nearest Neighbor (KNN): - K-Nearest Neighbor is one of the foremost basic supervised learning machine learning algorithms. During the training phase, this algorithm saves the info set, and when new data is received, it classifies it into a sort that's highly like the new data [10]. This approach can resolve both regression and classification problem statements. The letter "K" denotes the closest neighbor number of a brand-new unknown variable that has got to be predicted or rated.
2) Support Vector Machine (SVM)-: Support vector machine classification is a non-parametric clustering approach that makes no assumptions regarding the number or form of the data's clusters. This works well

for little data, therefore if your data is large, such as from a disadvantage step, utilize principal components analysis.

3) Logistic Regression (LR)-: It is a machine learning method for classification issues, as well as a predictive analytic technique based on the idea of probability. This is one of the most widely used machine learning algorithms, and it belongs to the category of supervised learning approaches. It is used to forecast the categorical dependent variable based on a set of independent factors. The output of a categorical dependent variable is predicted by this method [11].
4) Decision Tree (DT)-: It is a supervised learning method that may be applied to both classification and regression issues. This approach entails developing a model that predicts the value of an objective variable, using a decision tree that solves the issue using tree representation, with the leaf node corresponding to the class label and the characteristics represented. conducted on the tree's inner node [12].
5) Gaussian Naïve Bayes (GNB)-: GNB is a supervised machine learning classification algorithm family based on the Bayes theorem. The very commercial Naive Bayes classifier necessitates the use of a linear parameter describing the number of variables (predictors/features) in a math problem.

$$P(B/Y) = \frac{P(Y/B)P(B)}{P(Y)} \qquad (1)$$

Algorithm for the Naïve Bayes classifier:

1) Input COVID-19 data set.
2) Create COVID-19 Infection table using the data set.
3) Different Feature selection and extraction techniques passed algorithm as input.
4) Compute the posterior probability of each class calculated by Naive Bayes equations.
5) Output 1 for Positive and 0 for Negative.

F. Analysis

The classification was done in this study utilizing several machine learning classifiers, and the overall goal of this work was to discover the best machine learning classifier for predicting COVID19 fatalities based on accuracy, precision, and recall.

III. RESULTS

Both attribute selection and attribute extraction techniques are based on a maximum accuracy of 96% for the algorithms KNN (K-Fold with SelectK-Best), LR (Normal with SelectK-Best), and SVM (PCA, LDA), KNN (PCA with, K-Fold) as shown in TABLE II & TABLE III.

1) Normal Used to split train/test data manually with test size=0.30 and random state=5.
2) K-Fold is Used to provide training and testing metrics to divide data into training and testing sets. Using shuffle, divide the dataset into k successive folds. shuffle = True, n split = 5, random state = 5. The K-Fold cross-validation technique splits the input dataset into "K" equal-size groups of samples. Folds are the name given to these patterns. The prediction function employs 'k1' folds for each learning set, while the remainder of the folds is utilized for the test set. We divide the dataset into k subsets or folds, then train on all of them while leaving (k-1) one subset for evaluating the trained model. In this strategy, we iterate k times, each time with a different subset reserved for testing purposes.

TABLE II. ACCURACY OF ALGORITHMS ON SELECTION TECHNIQUES

Algorithms	SelectKBest with f_regression		SelectKBest with f_classif		SelectKBest with Chi2		Forward Feature Selection		Backward Feature Elimination		Recursive Feature Elimination	
	Normal	K-Fold	Normal	K-Fold	Normal	K-Fold	Normal	K-Fold	Normal	K-Fold	Normal	K-Fold
	Accuracy		Accuracy		Accuracy		Accuracy		Accuracy		Accuracy	
KNN	**0.95**	**0.96**	**0.95**	**0.96**	**0.95**	**0.96**	**0.95**	**0.96**	**0.95**	**0.96**	0.90	0.94
SVM	0.90	0.93	0.90	0.93	0.90	0.93	0.90	0.93	0.90	0.93	0.90	0.93
LR	0.96	0.95	0.96	0.95	0.96	0.95	0.93	0.95	0.93	0.95	0.93	0.95
DT	0.93	0.96	0.95	0.96	0.93	0.96	0.95	0.96	0.94	0.96	0.90	0.92
GNB	0.90	0.83	0.90	0.83	0.90	0.83	0.93	0.83	0.93	0.83	0.89	0.82

TABLE III. ACCURACY OF ALGORITHMS ON EXTRACTION TECHNIQUES

Algorithms	PCA		LDA		ICA		LLE	
	Normal	K-Fold	Normal	K-Fold	Normal	K-Fold	Normal	K-Fold
	Accuracy		Accuracy		Accuracy		Accuracy	
KNN	**0.95**	**0.96**	**0.95**	**0.95**	**0.96**	**0.95**	0.92	0.91
SVM	0.96	0.96	0.96	0.96	0.96	0.95	0.93	0.93
LR	0.96	0.95	0.96	0.96	0.93	0.93	0.93	0.93
DT	0.93	0.94	0.94	0.93	0.94	0.94	0.91	0.89
GNB	0.95	0.96	0.95	0.96	0.94	0.95	0.93	0.75

TABLE IV. CLASSIFICATION REPORT OF DIFFERENT ALGORITHMS ON FEATURES SELECTION [12]

Algorithms		SelectKBest f_regression	SelectKBest f_classif	SelectKBest Chi2	Forward feature Selection	Backward feature Elimination	Recursive feature Elimination
KNN	Accuracy	**0.95**	**0.95**	**0.95**	**0.95**	**0.95**	0.90
	Precision	1.00	1.00	1.00	1.00	1.00	0.50
	Recall	0.50	0.50	0.50	0.50	0.50	0.50
	F1 Score	0.67	0.67	0.67	0.67	0.67	0.50
SVM	Accuracy	0.90	0.90	0.90	0.90	0.90	0.90
	Precision	0.00	0.00	0.00	0.00	0.00	0.00
	Recall	0.00	0.00	0.00	0.00	0.00	0.00
	F1 Score	0.00	0.00	0.00	0.00	0.00	0.00
LR	Accuracy	0.96	0.96	0.96	0.93	0.93	0.93
	Precision	1.00	1.00	1.00	1.00	1.00	0.75
	Recall	0.62	0.62	0.62	0.25	0.25	0.38
	F1 Score	0.77	0.77	0.77	0.40	0.40	0.50
DT	Accuracy	0.93	0.93	0.93	0.94	0.94	0.90
	Precision	0.67	0.67	0.67	0.80	0.80	0.50
	Recall	0.50	0.50	0.50	0.50	0.50	0.25
	F1 Score	0.57	0.57	0.57	0.62	0.62	0.33
GNB	Accuracy	0.90	0.90	0.90	0.93	0.93	0.89
	Precision	0.50	0.50	0.50	0.67	0.67	0.45
	Recall	0.62	0.62	0.62	0.50	0.50	0.62
	F1 Score	0.56	0.56	0.56	0.57	0.57	0.53

TABLE IV. CLASSIFICATION REPORT OF DIFFERENT ALGORITHMS ON FEATURES EXTRACTION

Algorithms		PCA	ICA	LDA	LLE
KNN	Accuracy	**0.95**	**0.96**	**0.95**	0.92
	Precision	0.67	0.77	0.67	0.30
	Recall	0.59	0.59	0.59	0.18
	F1 Score	0.62	0.67	0.62	0.22
SVM	Accuracy	0.96	0.96	0.96	0.93
	Precision	0.71	0.80	0.77	0.00
	Recall	0.59	0.47	0.59	0.00
	F1 Score	0.65	0.59	0.67	0.00
LR	Accuracy	0.96	0.93	0.96	0.93
	Precision	0.71	0.00	0.77	0.00
	Recall	0.59	0.00	0.59	0.00
	F1 Score	0.65	0.00	0.67	0.00
DT	Accuracy	0.93	0.94	0.94	0.90
	Precision	0.50	0.53	0.59	0.28
	Recall	0.65	0.47	0.59	0.29
	F1 Score	0.56	0.50	0.59	0.29
GNB	Accuracy	0.95	0.95	0.95	0.93
	Precision	0.67	0.67	0.67	0.00
	Recall	0.59	0.59	0.59	0.00
	F1 Score	0.62	0.62	0.62	0.00

III. Conclusion and Future Scope

Covid-19 is a disease that has affected the entire World. Covid -19 has spread to more than 85 countries as of May 2022. In this present research work, a literature survey has been conducted to identify a suitable algorithm for the prediction of Covid-19 in patients. Both the essential approach and thus the composite algorithm provide proper results. Studies utilizing machine learning methods vary by country and author but are all relevant to the Covid-19 pandemic. Among these publications, regression techniques (logistic) are used, and each one indicates promising findings in Covid-19 healthcare applications and implications. Our conclusion is that by employing various models and algorithms, ML applications produce promising outcomes with high accuracy, precision, f1 score, and specificity. During this investigation, we discovered that patients above the age of 55 have a stronger risk of dying than recovering because they are weak in their internal immunity system. The model performance or data accuracy is additionally enhanced if additional data are provided, in line with the concept. Using the data from different health authorities, a model predicting outcomes has been created. While it's true that predictions aren't always 100% correct, they'll still be utilized as an adjustment and a security net. Future work can take pride in the addition the of latest components and algorithms. Since we've demonstrated in TABLE II & TABLE III, that with attribute selection strategies, Algorithms like KNN or LR, and with attribute extraction approaches, Algorithms like KNN or SVM or LR or GNB are giving accuracy up to 96%. We may infer that these techniques are accurate up to 96%. Using feature extraction and selection strategies, KNN performed consistently up to 96%.

	Accuracy(%)
KNN	96

References

[1] Y. Zoabi, S. Deri-Rozov, N. Shomron "Machine learning-based prediction of COVID-19 diagnosis based on symptoms," *npj digital medicine*, vol. 4, pp. 1-5,2021.

[2] F. T. Fernandes, T. A. De Oliveira, C. E. Teixeira, A. F. Batista, G. Dalla Costa, A. D. Chiavegatto Filho "A multipurpose machine learning approach to predict COVID-19 negative prognosis in São Paulo, Brazil," *Scientific reports*, vol. 11, pp. 1-7,2021.

[3] R. Majhi, R. Thangeda, R. P. Sugasi, N. Kumar "Analysis and prediction of COVID□19 trajectory: A machine learning approach," *Journal of Public Affairs*, vol. 2537, pp. 1-21,2021.

[4] K. B. Prakash,S. S. Imambi, M. Ismail, T. P. Kumar, Y. N. Pawan "Analysis, prediction and evaluation of covid-19 datasets using machine learning algorithms," *International Journal*, vol. 8, pp. 2199-204,2020.

[5] M. Satu, K. C. Howlader, M. Mahmud, M. S. Kaiser,S. M. Shariful Islam,J. M. Quinn, S.A. Alyami, M. A. Moni "Short-term prediction of COVID-19 cases using machine learning models," *Applied Sciences*, vol. 4266, pp. 1-11,2021

[6] R. A. Sujath,J. M. Chatterjee,A. E. Hassanien "A machine learning forecasting model for COVID-19 pandemic in India," *Stochastic Environmental Research and Risk Assessment*, vol.959, pp. 1-72,2020.

[7] Y. Tian, I. Luthra, X. Zhang "Forecasting COVID-19 cases using Machine Learning models," *MedRxiv*, 2020.

[8] S. Mohan, A. Abugabah,S. Kumar Singh ,A. Kashif Bashir ,L. Sanzogni "An approach to forecast impact of Covid□19 using supervised machine learning model," *Software: Practice and Experience*, vol.52, pp. 824-40,2022.

[9] R. K. Mojjada, A. Yadav, A. V. Prabhu, Y. Natarajan "Machine learning models for covid-19 future forecasting. Materials Today," *Proceedings*, 2020.

[10] S. Senupta, S. Mugde, G.Sharma, "Covid-19 pandemic data analysis and forecasting using machine learning algorithms," *medRxiv*,2020.

[11] E. Dong, L. Gardner "An interactive web-based dashboard to track COVID-19 in real time," *The Lancet infectious diseases*, vol. 20, pp. 533-4.

[12] F. Y. Cheng, H. Joshi, P. Tandon, R. Freeman, D. L. Reich,M. Mazumdar, R. Kohli-Set, M. A. Levin, P. Tamsin, A. Kia "Using machine learning to predict ICU transfer in hospitalized COVID-19 patients," *Journal of clinical medicine*, vol. 1668, 2021.