ISSN 2278-3091
Volume 9, No.4, July – August 2020
International Journal of Advanced Trends in Computer Science and Engineering
Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse386942020.pdf
https://doi.org/10.30534/ijatcse/2020/386942020

Machine Learning Approaches for Analysis of Covid-19 Data in India: A Case of Pandemic

Iyyanki Muralikrishna[1], A V Senthil Kumar[2], Amit Dutta[3], Ismail Bin Musirin[4]
[1]Defence Research Development and Organization, India. iyyanki@gmail.com
[2]Hindusthan College of Arts and Science, India.avsenthilkumar@yahoo.com
[3] All India Council for Technical Education, JNU Campus, India amitdutta07@gmail.com
[4]Centre for Electrical Power Engineering Studies, UniversitiTeknologi Mara, Malaysia, ismailbm1@gmail.com

ABSTRACT

Covid -19 has made the whole world upside down with spreading of virus faster in various countries. India cases started in the month of March which panic all the peoples, yet the mortality rate (1.8%) is much less than the other countries. It is believed with native immunity of Indians that surveyed. But thou a dreathful time for the health care centre where the doctors and nurses spent sleepless night treating the cases. The lockdown has made relaxation in spread of the virus. Yet few states showed very high cases with the living culture and spread of virus were due to community spread too. In this study of Covid-19, machine learning techniques were applied to the datasets of twelve states with twelve dates assumed. The results were very promising with SVM, Naïve and DT models with accuracy of 100%. F1-score, precision and recall obtained as 1.0 whereas KNN accuracy was very poor with 60%. The confusion matrix accuracy obtained was 0.0821. CNN prediction is better over LSTM and Hybrid –LSTM – CNN models. Hence, the results proved that implementing ML and DL techniques would help to analysis the cases faster and monitor the region or states in the future any pandemic attacks.

Key words: Covid-19, Decision Tree, k-Nearest Neighbor, Logistic Regression, Naïve, Random Forest Tree, Support Vector Machine, Convolution Neural Network, Long Short Term Memory

1.INTRODUCTION

Arifiticial Intelligence (AI) paved its way with ignorance in human lives. Every act of AI techniques is implemented in everyday work culture in and around the workplace, or be at home. AI occupied the complete structure of the human system without which human cannot exist. Life is made simpler and easy for the existence of the human life. AI techniques are many unnumbered and few are coined as machine learning (ML) techniques. AI techniques have become essential in the healthcare centres. The x-rays, ECG, automation of ICU equipment, EMR monitoring in health centers are supported by AI techniques. Various AI techniques were implemented during lockdown and unlock of Covid-19 period for monitoring and spraying disinfectant to avoid the spread of virus. The work has come to normal with all the precautionary methods followed. In this study of Covid-19 data analysis from April to September 2020, few ML techniques are implemented and found that models like SVM, DT, Naïve Bayes were better over the other techniques as logistic, KNN, and RF. Deep Learning techniques namely RNN and CNN were explored to find the prediction of the cases and the results proved CNN to be best model over LSTM model.

2.LITERATURE SURVEY

Khanday states AI has proved with promising results in health care through its decision making by analyzing the data. Their study included different algorithms for performing classification, it was revealed that logistic regression and multinomial Nai¨ve Bayesian classifier has given excellent results by having 94% precision, 96% recall, 95% f1- score and accuracy 96.2% [1].Al-Turaiki performed Naïve classifier and J48 decision tree to build the models. For maintaining stability models using J48, two attributes were used predicting stability namely symptomatic and age. The model performance was assessed using accuracy, precision, and recall. Overall, the accuracy of the models was between 53.6% and 71.58% [2]. Muhammad implemented DT, SVM, NB, LR, RF, and K-NN on the dataset. It was observed that the model developed with DT algorithm was the most accurate with 99.85% accuracy [3].Jain performed logistic growth development model, the generalized growth development model, generalized logistic growth model on the actual infected population data of CV-19 pandemic. The calculation on confidence intervals

predicted the curve direction and increase in the confirmed cases with 95% accuracy for the month of April 2020. The models predict exponential and sub-exponential spread rate in the number of positive cases in India in April 2020. The findings proved that significant measures are required to control the transmission rate of the virus in the India [4].

Iwendi proposed a fine-tuned RF model boosted by the AdaBoost algorithm. The model has an accuracy of 94% and F1-Score of 0.86 on the dataset used. Iwendi suggested the use of AI to detect and predict pandemics of a massive nature [5].Samuel implementedalgorithms of linear and logistic regress, KNN, NBon Tweets CV data and obtained classification accuracy of 91% for short Tweets, with the Naïve method [6]. Rustamin their study on CV-19 implemented four standard forecasting models, such as linear regression (LR), least absolute shrinkage and selection operator (LASSO), SVM, and exponential smoothing (ES) for forecasting the number of upcoming patients affected by CV-19. The results proved that the ES performs best among all the other models followed by LR and LASSO, while SVM performs poorly in all the prediction on the given dataset [7].Osi performed Linear Discriminant Analysis (LDA), RF, and SVM on CV-19 dataset. The results proved RF was found to be the best algorithm with 100% prediction accuracy in comparison with LDA and SVM with 95.2% and 90.9% respectively [8]. Wei and Zhang implemented a segmented Poisson model on the CV-19 dataset and the results obtained were more accurate in the analysis of CV data [9]. Jain and Kumar applied ML approaches such as SVM, DT, NB, and RF on swine flu, H1N1from tweet data. The evaluation technique produced better results with precision of 0.70 and recall of 0.86 with NB model [10].

Shahid performed forecast models such as ARIMA, support vector regression (SVR), LSTM, Bi-LSTM for time series prediction of CV-19 confirmed cases in ten major countries. The model performance was measured by mean absolute error, root mean square error and r2_score indices [11].Tuli applied ML-based model to predict the potential threat of CV-19 worldwide. Using Generalized Inverse Weibull distribution, a better fit was obtained to develop a prediction framework. Later this was deployed on a cloud computing platform for more accurate and real-time prediction of the growth behavior of the epidemic [12].

3.DISCUSSION AND RESULT

In this study of Covid-19 confirm cases dataset, ML techniques were implemented on twelve states namely Andhra Pradesh (AP), Chhattisgarh (CG), Delhi (DE), Gujurat (GJ), Karnataka (KA), Madhya Pradesh (MP), Maharashtra (MH), Rajasthan (RJ), Tamil Nadu (TN), Telangana (TS), Uttar Pradesh (UP), West Bengal (WB) for twelve different dates April 12, April 28, May 6, May 22, June 7, June 22, July 8, July 24, August 1, August 17, August 26, September 12, 2020. The distribution of data is shown in figure 1in the above mentioned states with the highest peak cases in MH followed by DE in April to July, 2020 whereas AP is presently second position with highest cases 575,079 after MH (1,077,374) as on Sept 14, 2020.

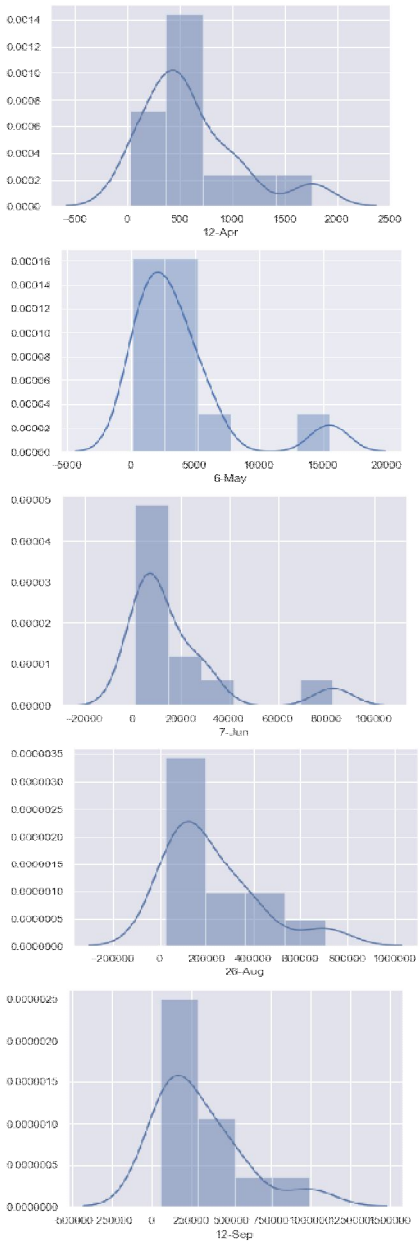


Figure 1: Distribution plot of Covid-19 data of 12 states in India

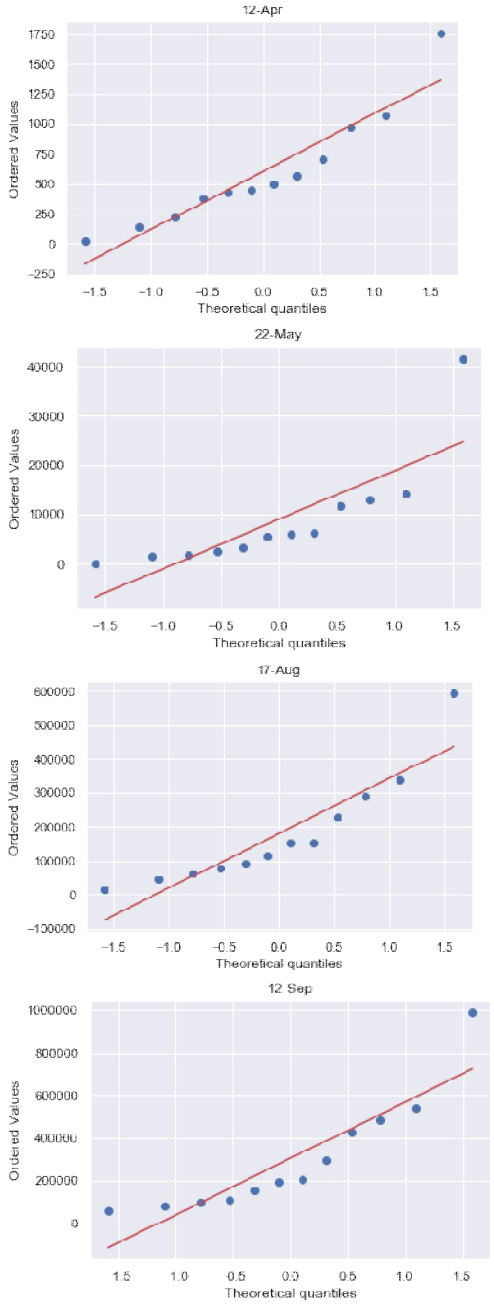


Figure 2: Probability Plot of Covid-19 data of 12 states in India

Probability plot calculates quantiles of a specified theoretical distribution (normal distribution) in figure 2. The plot calculates a best-fit (red line) and dots for the Covid-19 dataset, plots it. Note this plot is not similar as Q-Q plot or normal plot.

3.1 Confusion Matrix

The matrix shows the performance of a classification model while exposed to unseen data. The matrix helps to identify how the model is performing on test set. There are two classes: Class 1 and Class 2 and four groups (TP, TN, FP, FN).
Class 1:Positive
Class 2: Negative

T.P. (True Positive): where truth and prediction both are positive
T.N.(True Negative): truth and prediction both are negative
F.P.(False Positive): truth is negative but prediction is positive
F.N.(False Negative): truth is positive but prediction is negative. Calculation of precision and recall and F1-score from confusion matrix is possible.

Table 1: Precision and Recall for Covid-19 dataset.

Label	Precision	Recall
1	0.0527	0.0002
2	0.0013	0.0003
3	0.1136	0.0051
4	0.1204	0.0258
5	0.0246	0.0043
6	0.0325	0.0412
7	0.3361	0.0609
8	0.0302	0.0908
9	0.1718	0.1305
10	0.0425	0.1797
11	0.0754	0.2306
12	0.0531	0.3079

Accuracy is calculated by the given equation (1)
Acc. = (TP + TN) / (TP + TN + FP + FN)(1)

From the confusion matrix table 1, *total precision* is 0.0878 and
totalrecall is 0.0898 were obtained with the *accuracy* of 0.0821.

F1-score = (2 * Recall * Precision) / (Recall + Precision) (2)

Now calculating using (2), F1-score
= (2*0.0898*0.0878)/ (0.0898 +0.0878)

= 0.0157/ 0.1776 = 0.0884

3.2 Pair Plots

Pair plots helps to quickly explore distributions and relationships in a dataset. A pairs plot provides with a comprehensive structure look at the data by differentiating the severity (1) and non-severity (0) in the dataset. In this case study, severity of CV cases was where more than 10000 cases exists and less than 10000 cases were non-severity. Hue variable in the syntax adds a semantic mapping and changes the default marginal plot to a layered kernel density

estimate (KDE). Datasets under real-time study contain many variables. In such cases, the relation between each and every variable is analyzed. The diagonal plots are kernel density plots where the other plots are scatter plots shown in figure 3. As the cases gradually increased more severity was seen in almost all the states with the prevention precautions. And unlock1.0 started fromJune 1, 2020 and presently unlock4.0 is declared upto Sept 30, 2020.

3.3 Machine Learning Models Implemented on Covid-19 dataset

3.3.1 Logistic Regression (LR)

One of ML classification algorithm is used to predict the probability of a categorical dependent variable. In logistic regression, the dependent variable is a binary variable that contains data coded as 1 (yes) or 0 (no). It predicts the probability of occurrence of a binary event utilizing a logit function. The equation (3) for LR is given as follows

$$y = \beta 0 + \beta 1X1 + \beta 2X2 + \cdots + \beta nXn \quad (3)$$

where, y is dependent variable and x1, x2 ... and Xn are explanatory variables.

3.3.2 Support vector Machine (SVM)

In supervised learning models, SVM algorithm is used for classification and prediction as well on the labelled datasets. The vector points in the dataset that are closest to the hyperplane are known as the support vector points because these points contribute to the outcome of the algorithm. The purpose of hyperplane is to distinguish the two or more classes in the given dataset. With the set of points x, the hyperplane is written as in equation (4)

$$w*x - b = 0 \quad (4)$$

3.3.3 k-Nearest Neighbor Classifier (KNN)

KNN algorithm is used for both classification and regression predictive models. By selecting the nearest neighbor, using the Euclidean distance (as in equation (5)) calculate the distance between two points (neighbors). It is calculated by

$$distance = sqrt\{(x1 - y1)^2 + (x2 - y2)^2 + .. + (xn - yn)^2\} \quad (5)$$

3.3.4 Naïve Bayes classifier (NB)

Naive Bayes is a technique for constructing classifiers, where the models are assigned class labels to problem instances, represented as vectors of feature values. All naive classifiers assume that the value of a particular feature is independent of the value of any other feature, given the class variable.
When characteristic values are continuous in nature then an assumption is made that the values linked with each class are dispersed according to Gaussian that is Normal Distribution. It is calculated with equation(6).

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

3.3.5 Decision Tree (DT)

This are used for classification and regression predictive models. DT can handle high dimensional data with good accuracy. It works on the principle of attribute selection method. Information gain, and entropy are calculated, entropy is impurity of the system. The higher gain for an attribute, is placed as root node and other nodes follows, it is calculated by Gain = Entropy class – Entropy attribute.

$$Entropy = -\frac{P}{P+N} log_2 \left(\frac{P}{P+N}\right) - N(P + N) log_2 \left(\frac{N}{P+N}\right) \quad (7)$$

Figure 3. Pair plots for Covid-19 datasets

where P and N represent the positive and negative outputs of the predicted column. Information Gain (equation (8)) and Entropy of an attribute (equation (9)) is given below.

$$I(P_i,N_i) = -\frac{P}{P+N}log_2\left(\frac{P}{P+N}\right) - \frac{N}{P+N}log_2\left(\frac{N}{P+N}\right) \quad (8)$$

$$\text{Entropy attribute} = \sum \frac{P_i+N_i}{(P+N)} . I(P_i,N_i) \quad (9)$$

3.3.6 Random Forest Tree Classifier (RF)

In supervised ML algorithm, RF is used for classification and regression models. Select N random record from the dataset and build DT based on these records and takes the average to improve the predictive accuracy of that dataset. The problem of overfitting is prevented with the greater number of trees in the forest that leads to higher accuracy. It handles large datasets with high dimensionality.

Table 2: Comparative table for the models implemented on the Covid-19 data

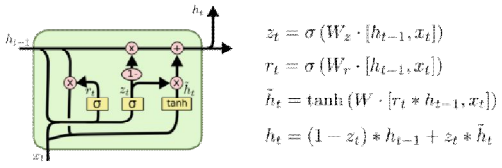
Model	Train score	Precision	Recall	F1-score	Jaccard similarity score (%)	Accuracy (%)
LR	85.72	0.79	0.86	0.86	85.72	85.72
SVM	100.0	1.0	1.0	1.0	100	100
KNN	60.0	0.50	0.60	0.53	60	60
NB	100.0	1.0	1.0	1.0	100	100
DT	100.0	1.0	1.0	1.0	100	100
RF	85.71	0.79	0.86	0.86	85.71	85.71

On comparison of the above algorithm, SVM, Naïve, decision tree has the good score 100% and jaccard similarity score over the other models. KNN model produced very poor results among all the models.

4. LONG SHORT TERM MEMORY AND CONVOLUTION NEURAL NETWORK

4.1 Long Short Term Memory

It is arecurrent neural network that is capable of learning order dependence in sequence prediction problems. Here, in this RNN model, the output from the final step is fed as input in the current step. LSTMs areof two types namely bidirectional and sequence-to-sequence to the field. The structure and equations are mentioned in the figure 4.



$$z_t = \sigma(W_z \cdot [h_{t-1}, x_t])$$
$$r_t = \sigma(W_r \cdot [h_{t-1}, x_t])$$
$$\tilde{h}_t = \tanh(W \cdot [r_t * h_{t-1}, x_t])$$
$$h_t = (1 - z_t) * h_{t-1} + z_t * \tilde{h}_t$$

Figure 4: LSTM Architecture and equations

** Source http://colah.github.io/posts/2015-08-Understanding-LSTMs/img/LSTM3-var-GRU.png

4.2 CNN

It consists of an input and an output layer, with multiple hidden layers. These hidden layers consist of a series of conv layers that convolve with a multiplication or other dot product. The activation function is a RELU layer, followed by additional conv such as pooling layers, fully connected layers and normalization layers, final layer is again a conv layer.

Table 3: Actual and predicted values of confirm cases

States	Actual Cases	Predicted Values		
		LSTM	Hybrid LSTM - CNN	CNN
AP	557587	540345	537815	549520
CG	61763	55750	56528	56047
DE	214069	206953	195769	206235
GJ	112174	116631	108850	110137
KA	449551	435042	427301	432652
MP	85966	88697	79921	82060
MH	1037765	982093	991727	995398
RJ	100705	94547	99368	99794
TN	497066	496676	496693	487879
TS	157096	155623	153376	153362
UP	305831	285215	292484	293224
WB	199493	193174	194309	193812

The predicted values obtained using LSTM, Hybrid–LSTM-CNN and CNN algorithmare very nearest to the actual values of the cases of the twelve different states. Changing the epochs, filters in Conv layer and dense layer varies the values produced. The graph for the predicted cases is shown for all twelve states in figure 5. Thus CNN model is the better model among the three models that gave the nearest value of the actual values.

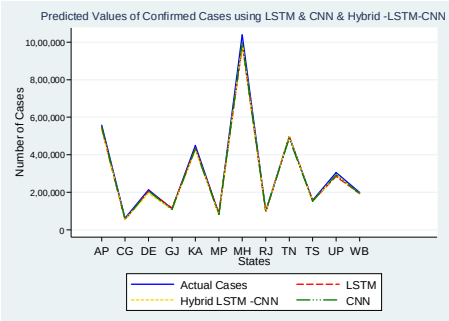


Figure 5: The graph of actual and predicted values of cases

5. CONCLUSION

Machine learning models proved in analyzing the Covid-19 data for predicting and produced better precision, recall and F1-score with SVM, naïve bayes and DT. The prediction value obtained by CNN is much better than LSTM and Hybrid –LSTM and CNN. In this way ML and DL techniques can be proposed for the study of Covid-19 data analysis and prediction helps in various ways in making the required and necessary arrangement to tackle the situation in the future for any pandemic or epidemic issues.

REFERENCES

[1] M. U.D. Khanday, S. T. Rabani, Q.R. Khan, N. Rouf and M.M.U.Din. **Machine learning based approaches for detecting COVID-19 using clinical text data.***Int. j. inf. tecnol*. Vol 12(3): pp 731–739. Sept 2020.

[2] Al-Turaiki, M. Alshahrani and T. Almutairi. **Building predictive models for MERS-CoVinfections using data mining techniques**. *Journal of Infection and Public Health*. Vol 9, pp 744—748. July 2016

[3] L. J. Muhammad, M. Islam, S. S. Usman, and S. I. Ayon. **Predictive Data Mining Models for Novel Coronavirus (COVID‑19) Infected Patients' Recovery.***SN Computer Science* 1:206. June 2020.

[4] M. Jain, P. K. Bhati, P. Kataria and R. Kumar.**Modelling Logistic Growth Model for COVID-19 Pandemic in India.***Proceedings of the Fifth International Conference on Communication and Electronics Systems.* IEEE Xplore ISBN: 978-1-7281-5371-1. Sept 2020.

[5] Iwendi, A. K. Bashir, A. Peshkar, R. Sujatha, J. M.Chatterjee, S. Pasupuleti, R. Mishra, S. Pillai, and O. Jo. **COVID-19 Patient Health Prediction Using Boosted Random Forest Algorithm.***Frontiers in Public Health*. Vol (8). July 2020

[6] J. Samuel, G. G. N. Ali, M. Rahman, E. Esawi, and Y. Samuel. **COVID-19 Public Sentiment Insights and Machine Learning for Tweets Classification.***Information*, 11, 314; 2020

[7] F. Rustam, A. A. Reshi, A. Mehmood, S. Ullah, B. On, W. Aslam, and G. S. Choi. **Covid-19 Future Forecasting Using Supervised Machine Learning Models**. *IEEE Access.* Vol 8 pp-101489 -1010499. May 2020.

[8] A. A. Osi, M. Abdu, U. Muhammad, A. Ibrahim, L. A. Isma'il, A. A.Suleiman, H.S. Abdulkadir, S. S. Sada, H.G.Dikko, and M. Z. Ringim. **A Classification Approach for Predicting COVID-19 Patient's Survival Outcome with Machine Learning Techniques**. *medRxiv preprint* 2020

[9] W. Wei and X. Zhang **An updated analysis of turning point, duration and attack rate of COVID-19 outbreaks in major Western countries with data of daily new cases.** *Data Article*. Vol (30). 2020

[10] V. K. Jain and S. Kumar. **An Effective Approach to Track Levels of Influenza-A (H1N1) Pandemic in India Using Twitter.***Procedia Computer Science* Vol 70 pp 801 – 807. 4th International Conference on Eco-friendly Computing and Communication Systems, ICECCS 2015

[11] F. Shahid, A Zameer, and M. Muneeb. **Predictions for COVID-19 with deep learning models of LSTM, GRU and Bi-LSTM**. *Chaos, Solitons and Fractals.* Vol 140 .110212 (2020).

[12] S. Tuli, S. Tuli , R. Tuli, and S. Gill. **Predicting the growth and trend of COVID-19 pandemic using machine learning and cloud computing.***Internet of Things.* Vol 11. 100222 (2020).