ISSN 2347 - 3983
Volume 8. No. 5, May 2020
International Journal of Emerging Trends in Engineering Research
Available Online at http://www.warse.org/IJETER/static/pdf/file/ijeter117852020.pdf
https://doi.org/10.30534/ijeter/2020/117852020

Analysis, Prediction and Evaluation of COVID-19 Datasets using Machine Learning Algorithms

Kolla Bhanu Prakash[1], S. Sagar Imambi[2], Mohammed Ismail[3], T Pavan Kumar[4], YVR Naga Pawan[5]
[1,2,3,4] Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Green Fields, Vijayawada, INDIA, drkbp@kluniversity.in
[5] Department of Computer Science and Engineering, Anurag Engineering College, Ananthagiri (V&M), Suryapet (Dist). Telangana, India, ynpawan@gmail.com

ABSTRACT

COVID-19, Corona Virus Diasease-2019, belongs to genus of Coronaviridae. A virus with no vaccine creating unpredictable havocs in the human lives and financial and economic systems in every country throughout the world. It is precariously halted everything in the society mercilessly. An analysis on COVID-19 datasets to understand which age group is mostly effected due to COVID-19. Different prediction models are built using machine learning algorithms and their performances are computed and evaluated. Random Forest Regressor and Random Forest Classifier outperformed the other machine learning models like SVM, KNN+NCA, Decision Tree Classifier, Gaussian Naïve Bayesian Classifier, Multilinear Regression, Logistic Regression and XGBoost Classifier.

Key words : COVID-19, Decision Tree Classifier, Gaussian Naïve Bayesian Classifier, KNN+NCA, Logistic Regression, Machine Learning,, Multilinear Regression, SVM, XGBoost Classifier

1. INTRODUCTION

Covid-19 epidemic occurred in December 2019 in Wuhan China, which is caused by a novel Extreme Acute Respiratory Syndrome Corona Virus 2 virus (SARS-CoV-2). SARS-CoV-2 is the source of Coronaviris Disease 2019 (COVID-19). World Health Organization (WHO) declared on January 30, 2020 the outbreak as an emergency and pandemic for public health.COVID-19's clinical symptoms are respiratory disorder, fatigue, dry cough, tiredness, etc. while 80 percent of patients heal without any care. [31] Elderly men, children, men who already have cardiovascular disease, obesity , and diabetes are vulnerable to COVID-19. [30] COVID-19's clinical symptoms are respiratory disorder, fatigue, dry cough, tiredness, etc. while 80 percent of patients heal without any care.[29] Elderly men, children, men who already have cardiovascular disease, obesity, and diabetes are vulnerable to COVID-19.

The best way to prevent and slow down transmission is maintaining social distance. We have to protect our self and others from infection by washing our hands or using sanitizers and avoid touching face. The number of COVID-19 cases in India are 67,161 and the death toll is 2,212 by 11th MAY 2020, as per the *Worldometer* data. Worldwide 4,180305 people have been attacked by virus and the total number of deaths caused by disease now are 283,865.

There are very less number of COVID-19 test kits available in hospitals which are not at all sufficient for the increasing cases. Hence, it is required to implement an automatic detection system to prevent COVID-19 spreading among people. Artificial Intelligence is actually dominant tool in the fight against the COVID-19 crisis. [21, 22] AI has subdomains like Machine Learning, Deep Learning. [23, 24] It has several application in the area of Natural Language Processing and Computer Vision applications.[15, 16, 17] It helps in diagnose and predict COVID-19. Deep learning and ML Techniques are useful in tracking COVID cases, predicting, Generating dashboards, Diagnose and treatments, Generating alerts to maintain social distance and for other possible control mechanism. [18, 19, 20]

2. SARS-CoV-2

.SARS-CoV-2 is a single stranded Ribonucleic Acid (RNA) Virus. It is contagious in humans and created pandemic throughout the world. It is endangering millions of human beings in the world and also lead to economic disruption. SARS-CoV-19 invades human cells and bind to ACE2, a protein present in cells of several human bodies, after one of its proteins. Coronaviruses are a member of the genus Coronaviridae. Coronaviridae is a single stranded, enveloped family of RNA viruses. The etiology of SARS-CoV-2 is explained in [4] by Yang et al.

2.1 Symptoms

The most common symptoms of COVID-19 are flu-like symptoms [1] [2] [3]. The details are tabulated in Table 1. Due to mild and unspecific symptoms, it is becoming difficult to identify and quarantine.

Table 1: Symptom of COVID-19

Most Common	Moderate	Severe
Tiredness, Fever and dry cough.	Conjunctivitis, headache, diarrhea, aches and severe pains sore throat, lack of taste or unable to smell, skin rashes, or fingers or toes discoloration, Chills	difficulty in breathing or shortness of breath, pain in chest or pressure, loss of speech or movement

2.2 Diagnosis

Viral tests notifies about infection with SARS-CoV-2, the virus which triggers COVID-19. If a test results as positive indicates the person is infected. The diagnostic test is dependent on the affected person's geographic location [5]. Rapid Diagnostic Test (RDT) tests for the existence of proteins of the virus, called antigens, developed by the SARS-CoV-2 virus in the respiratory tract of a person. Usually within 30 minutes, if the SARS-CoV-2 antigen exists in sufficient concentrations in the collected sample, it can bind to numerous antibodies attached to a paper strip in plastic case. It generates a signal which is easily detectable. The RDT tests are used for diagnosing the acute or early infections of SARS-CoV-2, as the developed antigens are released only when the virus replicates successfully. These tests are considered as reliable for diagnosing of COVID-19 [6]. Another specific form of RDT advertised for COVID-19; a test that measures the existence of antibodies in the blood of those suspected to have been COVID-19 infected. Antibodies grow within days to weeks after an infection with the virus. The recommended method for COVID-19 case evaluation and laboratory testing is molecular analysis (e.g., PCR) of respiratory tract samples [6].

2.3 Treatment

The COVID-19 infected patients have no defined treatment. The medication is given based on symptoms. It may include pain relievers, cough syrup, rest and fluid intake. If the patients have mild symptoms, they may stay at home and take treatment in isolation. Otherwise, treatment in the hospital is evident [6].

3. RESULTS

The Machine Learning Techniques [12] [13] [14] are used to understand the COVID-19 affecting people, its confirmation and recovery predictions. The Figure 1 shows the various age groups and percentage of cases obtained from kaggle dataset [8]. The age groups of 20-50 are highly probable of getting infected with COVID-19.

The two datasets Covid-19-India [9] and Covid-19-Data [10] are used to analyze their features and to build ML models for performance assessment. The Figure 2 and Figure 3, shows the correlation matrices of the datasets.

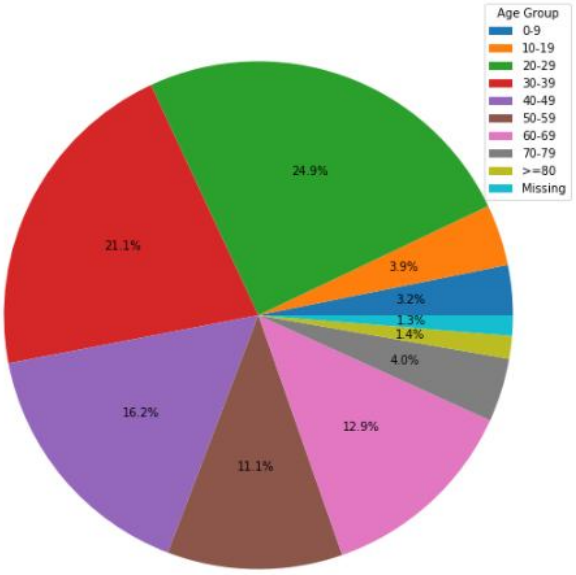

Figure 1: Percentage of COVID-19 cases as per Age Group.

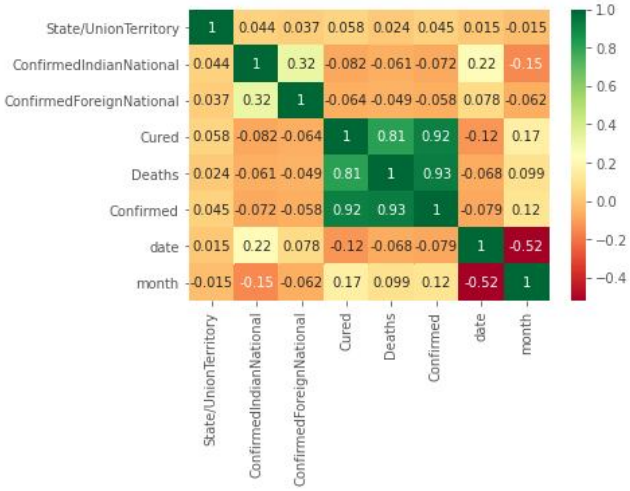

Figure 2(a):. Correlation Matrix for Covid-19-India Dataset

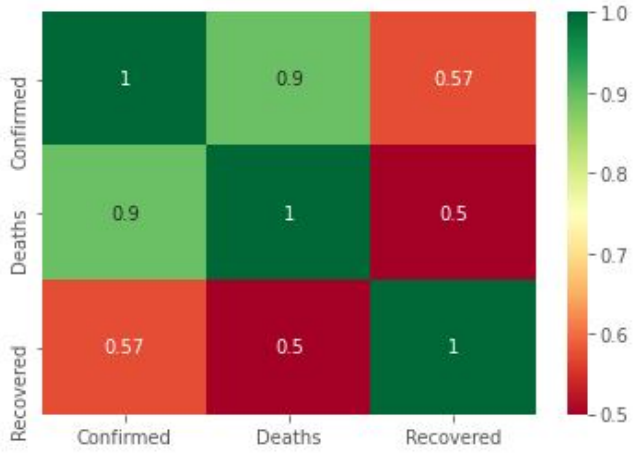

Figure 2(b): Correlation Matrix for Covid-19-Data Dataset

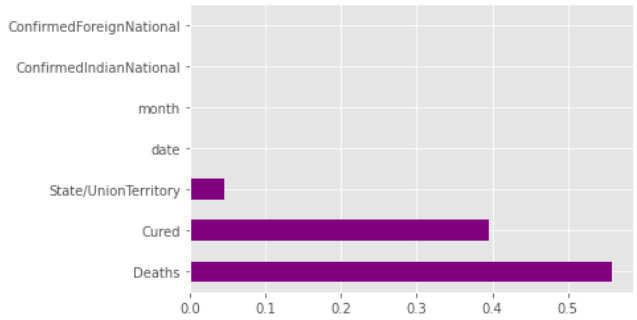

Figure 3(a): Feature Importance using DT Classifier

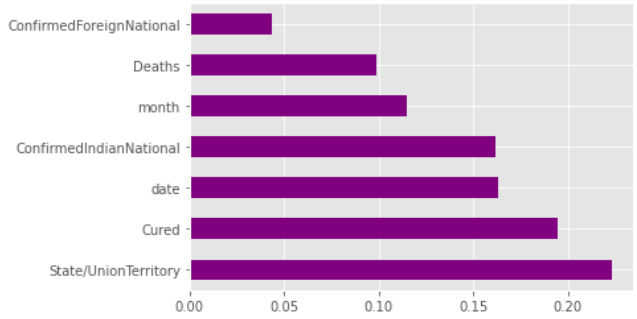

Figure 3(b): Feature Importance using Radom Forest Classifier

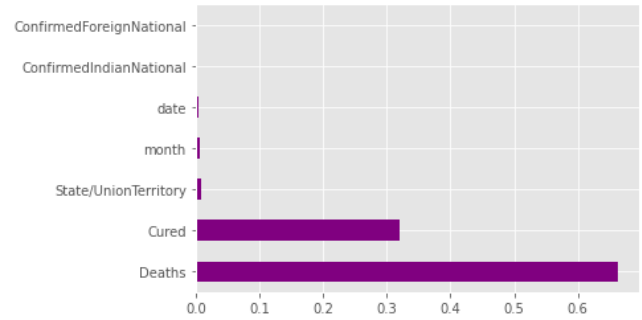

Figure 3(c): Feature Importance using Random Forest Regressor

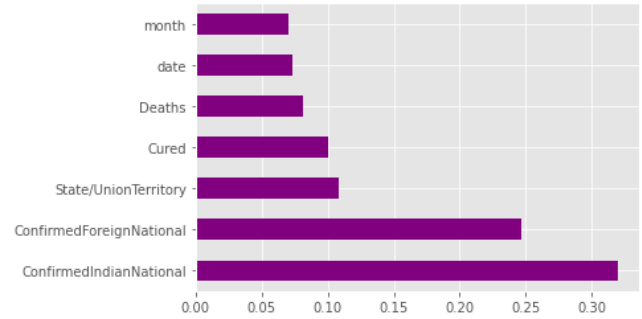

Figure 3(d): Feature Importance using XGBoost Classifier

The ML models based on the algorithms like SVM, KNN+NCA, Decision Tree Classifier, Gaussian Naïve Bayes Classifier, Multilinear Regression, Logistic Regression, Random Forest Classifier, and XGBoost Classifier are built using the two datasets [9] [10]. The R-Squared (coefficient of determination) regression score and accuracy are computed with train and test dataset ratio as 70:30. The feature importance for Covid-19-India Dataset is shown in the figure Figure 3(a) – Figure 3(d).

The Figures 4(a) and 4(b) shows the Coefficient of Determination (CoD), also called R-Squared, and Accuracy for the models built on COVID-19-India Dataset. The Figures 5(a) and 5(b) shows the CoD and Accuracy for the models built on COVID-19-Data Dataset. The results show that the RandomForest Classifier and the Random Forest Regressor outperformed the other ML Models. [25, 26, 27, 28]

Table 2: Coefficient of Determination & Accuracy

Machine Learning Model	Covid-19-India		Covid-19-Data	
	Coefficient of Determination	Accuracy	Coefficient of Determination	Accuracy
SVM	0.72128	0.11704	0.41006	0.96667
KNN+NCA	0.88327	0.37522	-2.04628	0.93333
DT Classifier	0.14316	0.158348	0.99448	0.96667
GNB Classifier	0.12211	0.11876	0.40759	0.83333
Multilinear Regression	0.12211	0.11876	-2.04628	0.93333
Logistic Regression	0.12210	0.11876	-2.04628	0.93333
Random Forest Classifier	0.92442	0.41824	0.99448	0.96667
Random Forest Regressor	0.96843	---	0.77839	---
XGB Classifier	0.46803	0.42513	-2.04628	0.93333

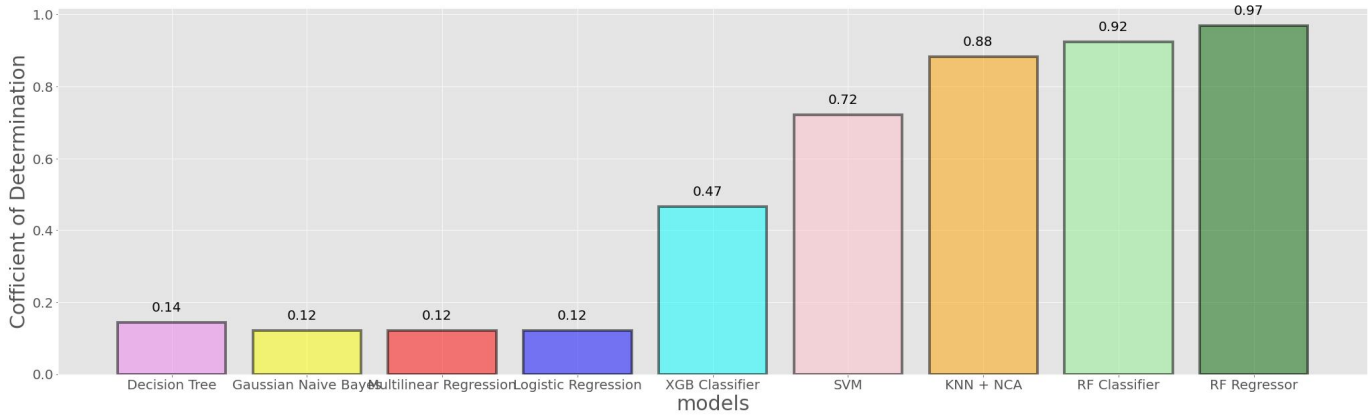

Figure 4(a): Coefficient of Determination for COVID-19-India

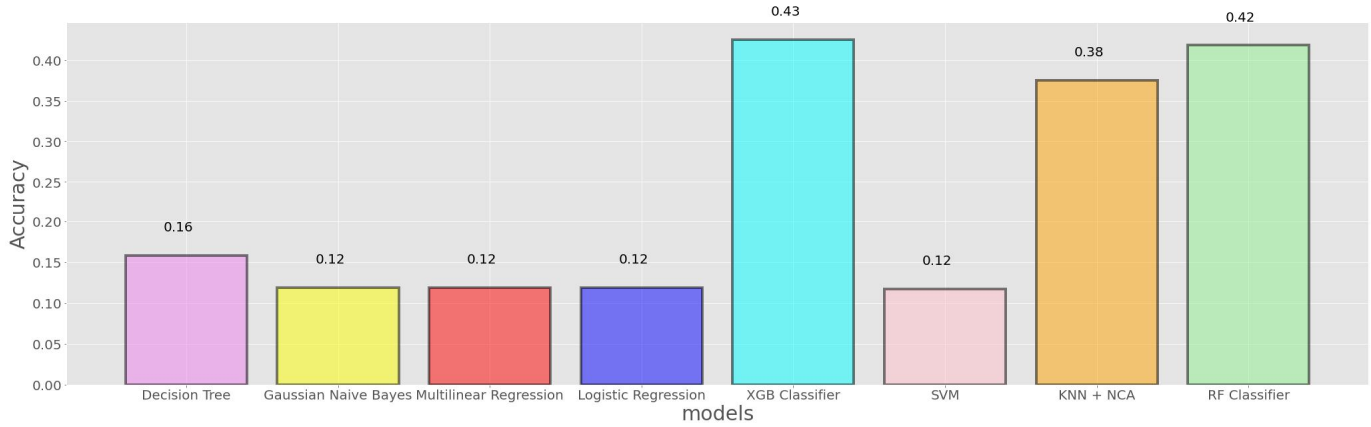

Figure 4(b): Accuracy for COVID-19-India

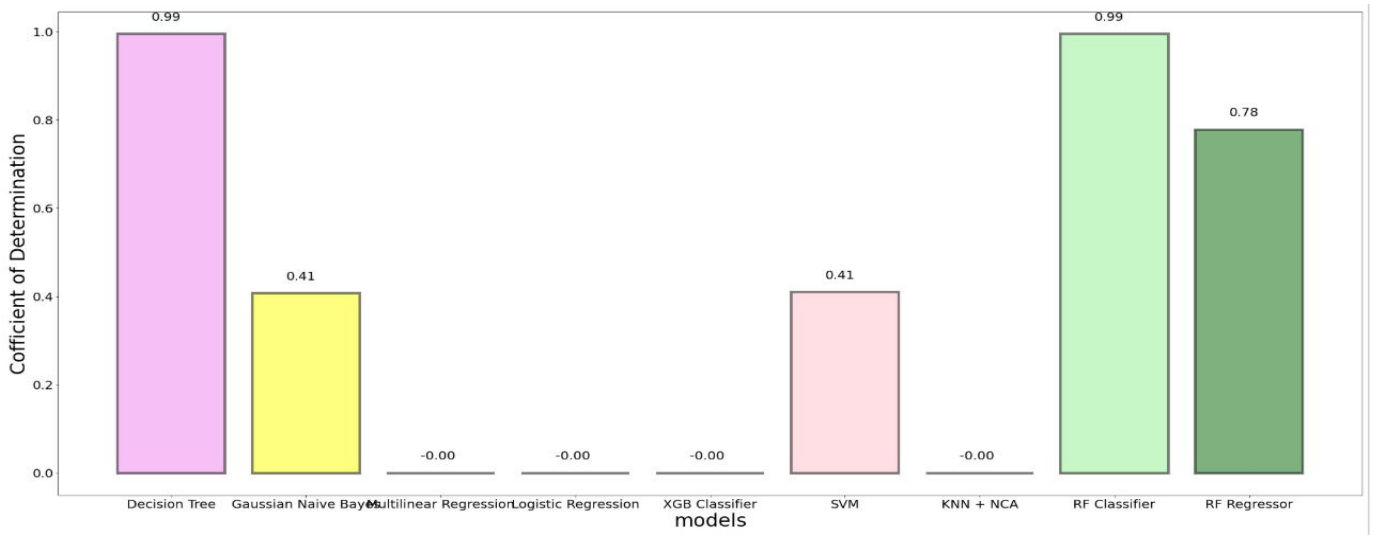

Figure 5(a): Coefficient of Determination for COVID-19-Data

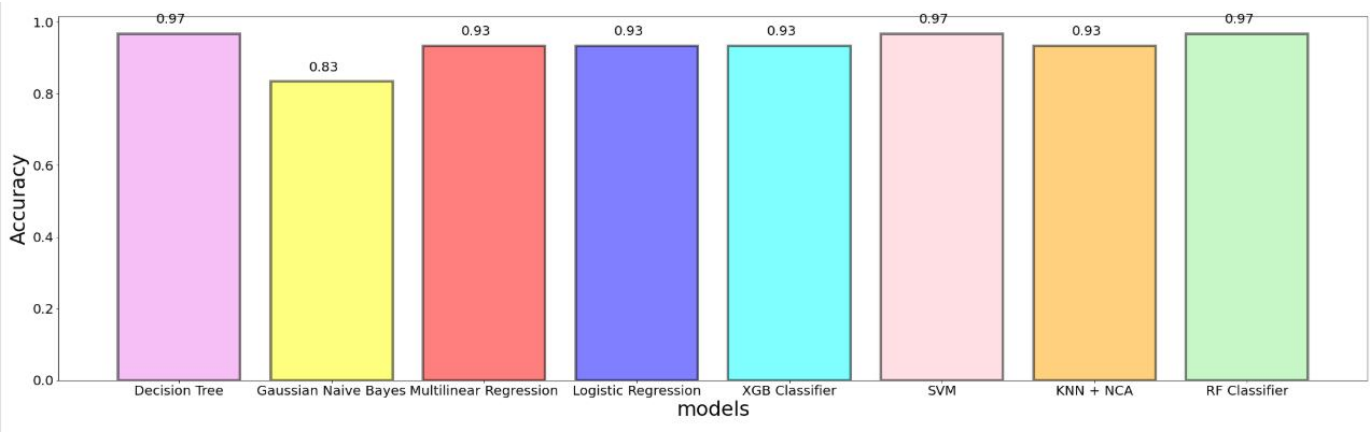

Figure 5(b): Accuracy for COVID-19-Data

4. CONCLUSION

The experiments reveal the persons of age groups 20-30, 30-40 and 40-50 are suffered with COVID-19. The correlation matrices are built to understand the relationship between the features of the datasets. The feature importance is computed for the classifiers built. Along with the classifiers and repressors are also built for prediction. The results show that the Random Forest Regressor and Random Forest Classifier has outperformed other models in terms of CoD and Accuracy. [22] In future, more ML classifiers and Regressors are evaluated on the evolving COVID-19 datasets. [24]

REFERENCES

1) Menni, C., Valdes, A.M., Freidin, M.B. et al. Real-time tracking of self-reported symptoms to predict potential COVID-19. Nat Med (2020). https://doi.org/10.1038/s41591-020-0916-2.
2) What are the symptoms of COVID-19?, https://www.who.int/emergencies/diseases/novel-coronavirus-2019/question-and-answers-hub/q-a-detail/q-a-coronaviruses#:~:text=symptoms (Last Accessed: 11.05.2020)
3) Symptoms of Coronavirus, https://www.cdc.gov/ coronavirus/2019-ncov/ symptoms-testing/ symptoms.html (Last Accessed: 11.05.2020)
4) Yang, P., Wang, X. COVID-19: a new challenge for human beings. Cell Mol Immunol 17, 555–557 (2020). https://doi.org/10.1038/s41423-020-0407-x
5) Coronavirus disease 2019 (COVID-19), https://www.mayoclinic.org/diseases-conditions/coronavirus/diagnosis-treatment/drc-20479976, (Last Accessed 12.05.2020)
6) Advice on the use of point-of-care immunodiagnostic tests for COVID-19, https://www.who.int/news-room/commentaries/detail/advice-on-the-use-of-point-of-care-immunodiagnostic-tests-for-covid-19, (Last Accessed 12.05.2020)
7) Viral Testing Data in the U.S., https://www.cdc.gov/coronavirus/2019-ncov/cases-updates/testing-in-us.html, (Last Accessed 12.05.2020)
8) ICMR, COVID-19 Testing, https://main.icmr.nic.in/sites/default/files/ upload_documents/Revised_Advisory_Rapid_Anibody_blood_tests.pdf, (Last Accessed 12.05.2020)
9) COVID-19 in India, https://www.kaggle.com/ sudalairajkumar/covid19-in-india?select=AgeGroupDetails.csv, (Last Accessed 12.05.2020)
10) COVID-19 in India, https://www.kaggle.com/sudalairajkumar/ covid19-in-india? select=covid_19_india.csv, (Last Accessed 12.05.2020)
11) Novel Corona Virus 2019 Dataset, https://www.kaggle.com/sudalairajkumar/novel-corona-virus-2019-dataset?select=covid_19_data.csv, (Last Accessed 12.05.2020)
12) Balika J. Chelliah, S. Kalaiarasi, Apoorva Anand, Janakiram G, Bhaghi Rathi, Nakul K. Warrier, Classification of Mushrooms using Supervised Learning Models, IJETER, Vol 6(4), April 2018, ISSN 2454-6410, pp 229-232.
13) D. Shona, A.Shobana, Fast and Effective Network Intrusion Detection Technique Using Hybrid Revised Algorithms, IJETER, Vol 4(11), November 2016, ISSN 2454-6410, pp 42-46.
14) Detecting Central Nervous System Disorder Using Machine Learning Technique (XGB Classifier), Sri Lasya Dharmapuri, Pavan Kumar Dandamudi, Vinoothna Manohar Botcha and Bhanu Prakash Kolla, IJETER, Vol 8(4), April 2020, ISSN 2454-6410, pp 1142-47.
15) K.B., Prakash. "Information extraction in current Indian web documents." International Journal of and Technology (UAE), Vol. 7(2.8), 2018:68-71. https://doi.org/10.14419/ijet.v7i2.8.10332
16) Kolla B.P., Dorairangaswamy M.A., Rajaraman A. "A neuron model for documents containing multilingual Indian texts." 2010 International Conference on Computer and Communication Technology, ICCCT-2010,2010: 451-454.

17) Kolla B.P., Raman A.R. "Data Engineered Content Extraction Studies for Indian Web Pages." Advances in Intelligent Systems and Computing, 2019: 505-512.
18) Prakash K.B., Dorai Rangaswamy M.A. "Content extraction studies using neural network and attribute generation." Indian Journal of Science and Technology, 2016: 1-10.
19) Prakash K.B., Dorai Rangaswamy M.A., Raman A.R. "Text studies towards multi-lingual content mining for web communication." Proceedings of the 2nd International Conference on Trendz in Information Sciences and Computing, TISC-2010, 2010: 28-31. https://doi.org/10.1109/TISC.2010.5714601
20) Prakash K.B., Rangaswamy M.A.D. "Content extraction of biological datasets using soft computing techniques.", Journal of Medical Imaging and Health Informatics, 2016: 932-936.
21) Prakash, K.B., Rajaraman, A., Perumal, T. & Kolla, P. 2016, "Foundations to frontiers of big data analytics", Proceedings of the 2016 2nd International Conference on Contemporary Computing and Informatics, IC3I 2016, pp. 242 https://doi.org/10.1109/IC3I.2016.7917968
22) Kolla, B.P. & Raman, A.R. 2019, Data Engineered Content Extraction Studies for Indian Web Pages, Advances in Intelligent Systems and Computing, 711, pp. 505-512.
23) Prakash, K.B. 2017, "Content extraction studies using total distance algorithm", Proceedings of the 2016 2nd International Conference on Applied and Theoretical Computing and Communication Technology, iCATccT 2016, pp. 673.
24) Prakash, K.B., Kumar, K.S. & Rao, S.U.M. 2017, "Content extraction issues in online web education", Proceedings of the 2016 2nd International Conference on Applied and Theoretical Computing and Communication Technology, iCATccT 2016, pp. 680. https://doi.org/10.1109/ICATCCT.2016.7912086
25) Prakash, K.B., Rajaraman, A., Perumal, T. & Kolla, P. 2016, "Foundations to frontiers of big data analytics", Proceedings of the 2016 2nd International Conference on Contemporary Computing and Informatics, IC3I 2016, pp. 242.
26) Ismail, M., Prakash, K.B. & Rao, M.N. 2018, "Collaborative filtering-based recommendation of online social voting", International Journal of Engineering and Technology(UAE), vol. 7, no. 3, pp. 1504-1507.
27) Prakash, K.B., Rajaraman, A. & Lakshmi, M. 2017, "Complexities in developing multilingual on-line courses in the Indian context", Proceedings of the 2017 International Conference On Big Data Analytics and Computational Intelligence, ICBDACI 2017, pp. 339.
28) Mohammed Ismail .B, Dr. T. Bhaskara Reddy, Dr. B. Eswara Reddy "Spiral Architecture Based Hybrid Fractal Image Compression" IEEE International Conference on Electrical, Electronics, Communication, Computer and Optimization Techniques (ICEECCOT)" Dec 2016 https://doi.org/10.1109/ICEECCOT.2016.7955179
29) Ghousia Anjum Shaik, T.Bhaskara Reddy, Mohammed Ismail.B Mansooor Alam and Mansour Tahernezhadi "Variable Block Size Hybrid Fractal Technique for Image Compression" *IEEE Proceedings 2020 6th International Conference on Advanced Computing & Communication Systems (ICACCS)* March 2020
30) Mohammed Ismail. B,B. Eswara Reddy,T. Bhaskara Reddy "Cuckoo Inspired Fast Search Algorithm for Fractal Image Encoding" Journal of King Saud University Computer and Information Sciences volume 30 issue 4,pp 462–469 2018 https://doi.org/10.1016/j.jksuci.2016.11.003
31) M. Ismail.B, B. Eswara reddy "Improved Fractal Image Compression Using Range Block Size" IEEE Proc (CGVIS) 2015 pp: 284 – 289 https://doi.org/10.1109/CGVIS.2015.7449938