Analyzing the Performance Factors of Machine Learning Algorithms for COVID'19 Data

R. Venkatesan, S.Nandhagopal, A.Sabari

Abstract*: Machine learning is a branch of Artificial intelligence which provides algorithms that can learn from data and improve from experience, without human intervention. Now a day's many of the machine learning algorithms playing a vital role in data analytics. Such algorithms are possible to apply with the recent pandemic COVID situation across the globe. Machine learning algorithms are classified into 3 different groups based on the type of learning process, such as supervised learning, unsupervised learning, and reinforcement learning. By considering the medical observations on the COVID across the globe it has been discussed and concluded to analyze under the supervised learning process. The data set is acquired from the reliable source, it is processed and fed into the classification algorithms. Since learning behaviors are carried out by knowing the input data and expected output data. The data is labeled and has been classified based on labels. In the proposed work, three different algorithms are used to experiment with the COVID'19 dataset and compared for their efficiency and algorithm selection decision is made.

Keywords: Machine learning, Data analysis, COVID, Algorithm analysis, kNN, SVM, Random Forest, Supervised Learning

I. INTRODUCTION

With the increased availability of data from varied sources, there has been increasing attention paid to the various data-driven disciplines such as analytics and machine learning. Machine learning is a sub-platform of artificial intelligence and mining interesting patterns from a huge dataset. To create a business by applying logical intelligence with an extreme level of dataset the predict the future action by analyzing the historical data. The standard and most popular supervised learning algorithms including linear regression, logistic regression, decision trees, k-nearest neighbor, an introduction to Bayesian learning and the naïve Bayes algorithm, support vector machines and kernels, and neural networks. Also, machine learning extends its platform called deep learning to increase the quality of results with available algorithms. To handle the unlabelled data items it will take another form of learning technique called unsupervised learning method. These methods will deal with various clustering algorithms such a K-Means [15], K-Mediods, etc… All the supervised and unsupervised learning algorithms follow the computational learning theory approach to produce the output. The performance of machine learning algorithms has been analyzed with hypothesis space, overfitting, bias and variance, trade-offs between representational power and learnability, evaluation strategies, and cross-validation techniques. The performance of algorithms is variable one concerning the type of applications.

II. RELATED WORK

Phan Thanh Noi et, al[1] conducted a comparative study on a few machine learning algorithms. They compared these three algorithms with the sentinel-2 image data. The over accuracy level is in the range of 90%-95%. Among the three classifiers and 14 sub-datasets, SVM produced higher overall accuracy and least sensitivity to the training data. Kewen Li and Peng Xie et.al[6] projected the difference between the obtained accuracy level between the AdaBoost algorithm and the improved AdaBoost algorithm with the minority classes. KNN(K-Adaboost) for cutting down the majority class weights which are closer to the minority classes. With this minority classes are considered instead of giving equal weight on both. They projected only the accuracy of KNN with minority class variables. Okfalisa and Mustakim et.al[7] Derrac et al. [14], shown a comparative analysis study on KNN[18, 19] and modified KNN in classification based problems. Here to the accuracy of both the algorithms has been compared by experimenting with cash transactional data and proved that MKNN[20] is giving better performance than KNN. Derrac et al. [16] proposed a classification framework by selecting instances and features using cooperative co-evolution. They used as IS and FS with the help of GA and produced a reduced dataset. The reduced dataset generated by the co-operative co-evolution trained the K-Nearest Neighbour (KNN) classifier. They produced had better accuracies than the contemporary works when experimented with 18 datasets of UCI Machine Learning Repository. All the 18 datasets obtained an average accuracy of 81.64%.

Shaobo Du, Jing Li[8] worked in text classification using the KNN algorithm. Due to the low level of classification accuracy when working with an increased level of complexity in the training data set. To handle such a situation they decided to work on doing improvisation with KNN and experimented with search engine text classification and showed the improvement in time and accuracy of classification.

Revised Manuscript Received on June 25, 2020.
***** Correspondence Author
R.Venkatesan *, Assistant Professor & Department of Computer Science and Engineering & KSR Institute for Engineering & Technology venkat.ishva@gmail.com
S.Nandhagopal, Assistant Professor & Department of Information Technology & KSR Institute for Engineering & Technology nan.ecs1@gmail.com
A.Sabari, Professor & Head Department of Information Technology & K.S.Rangasamy College of Technology asabari@gmail.com

Chuan-Quan Li et.al [9] worked to enhance the outcome of the random forest algorithm with CPLS for better classification. With the superior inherent characteristics applied in the CPLS method to build a rotation matrix and can promote the predicted performance of their model. The obtained results of different benchmark datasets show that their method is a more efficient classification method than random forest and other variants. V. R. Elgin Christo et.al [10, 17] worked on a clinical dataset using cooperative coevolution and classification using random forest. They used wrapper kind of approach to extract features and instance selection which has been passed to cooperative coevolution and random forest classifier. The obtained result from both the techniques has been suggested as a second opinion for diagnosis and treatment by them. They used Wisconsin Diagnostic Breast Cancer (WDBC), Hepatitis, Pima Indian Diabetes (PID), Cleveland Heart Disease (CHD), Statlog Heart Disease (SHD), Vertebral Column, and Hepatocellular Carcinoma (HCC) datasets and achieved accuracy level up to 97.1%, 82.3%, 81.01%, 93.4%, 96.8%, 91.4%, and 72.2% for datasets WDBC, Hepatitis, PID, CHD, SHD, vertebral column, and HCC, respectively.

Srivathsan Srinivasagopalan et.al[11] used traditional machine learning approaches such as logistic regression, support vector machine, and random forest. They have created three hidden layers in their deep learning model and produced a higher accuracy in diagnosing schizophrenia patients. Ravi Shanker, Mahua Bhattacharya [12] they presented Hierarchical Centroid Shape Descriptor (HCSD) on the already segmented abnormal region by K-Mean clustering and Fuzzy C-Mean clustering methods. The implemented HCSD method selects only interesting regions and defines an abnormal region. The developed skull stripping algorithm improves the segmentation accuracy. They used four different attribute parameters such as the sensitivity, the specificity, the accuracy, and the dice similarity coefficient to define the accuracy of the algorithm.

III. ALGORITHM SELECTION

According to the records of medical observation in the COVID center of our country the health care department has labeled a patient based on the level of infection. With such labeling, we have selected the supervised machine learning algorithms and tried with the training dataset. The World Health Organization ("WHO") given different kinds of the label based on the severity of the infections. Such labels are, deceased, released, and isolated. Where deceased referred to as a new case has been found after the test result, patients are released once it has confirmed by the health care people with the test result, at last, the case will be isolated with the positivity level of the COVID analysis report.

With these considerations, we have selected the three different algorithms such as KNN algorithms, Random Forest, and SVM classifier algorithm to measure the accuracy factor these three algorithms bypassing the training dataset. The supervised learning algorithms are having some unique features. Supervised learning allows you to collect data or produce a data output from the previous experience. It helps you to optimize performance criteria using historical data and helps to solve various types of real-world computation problems. The selected SVM, KNN, and Random Forest having some unique features. SVM is capable of handling nonlinear data elements and possible to apply in high dimension space. KNN algorithm is a robust one to handle noisy training datasets and accept to process a large number of training samples. Whereas Random Forest combines a bunch of decision trees and handles the categorical features as well. Also, it supports to handle high dimensional training data with larger volume samples.

IV. COVID'19 DATA EXTRACTION

WHO released a huge amount of dataset publically over the web. We can get worldwide data from online. Data are available to download from official websites GitHub with hashtag on COVID'19, nCov'19, CORD-19. kaggle.com[5] is another official website to keep updated about the disease in our country. A public AWS COVID-19[2] data lake available – a centralized data warehouse of up-to-date and accurate datasets on or related to the spread and characteristics of the novel coronavirus (SARS-CoV-2) and its associated illness category, COVID-19. Globally, there are several efforts underway to gather this data, and AWS working with partners to make this crucial data freely available and keep it up-to-date. Every data set has been recorded with a basic attribute of patient and feature attributes about the CORONA disease infection. Based on the case severity every patient will be labeled with disease confirmed, no disease, isolated and recovered. With this labeled data item we have three machine learning algorithms to analyze the performance of selected algorithms. Data collection is playing a vital role in data analytics oriented applications. There are various online sources whichever we found that everywhere the format of data is varied somewhere will get structured data somewhere unstructured data. Also, data acquisition can have many possibilities to receive noisy data as well. To handle such relational data items initially we must preprocess those data by holding necessary attributes and removing irrelevant attributes. All kinds of data mining functionalities such as association rule mining, classification, and clustering [13] have been applied to get a smooth training sample.

According to COVID'19 dataset collection there are lots many attribute values are collected and that must be preprocessed and cleaned before sending it to the selected algorithms by following the steps.

1. The data filtering unit is assigned to preprocess the collected data before passing to the analysis engine. The analysis engine has been configured with three algorithms namely KNN, Random Forest and SVM Classifier.
2. The preprocessed data has been given input to these algorithms to begin analyzing the data and interpret the result factors each one.
3. The output of each algorithm has been compared with accuracy, precision, recall and F1-Measure.
4. Comparison table will summarize the result and shows the best algorithm by considering performance factors.

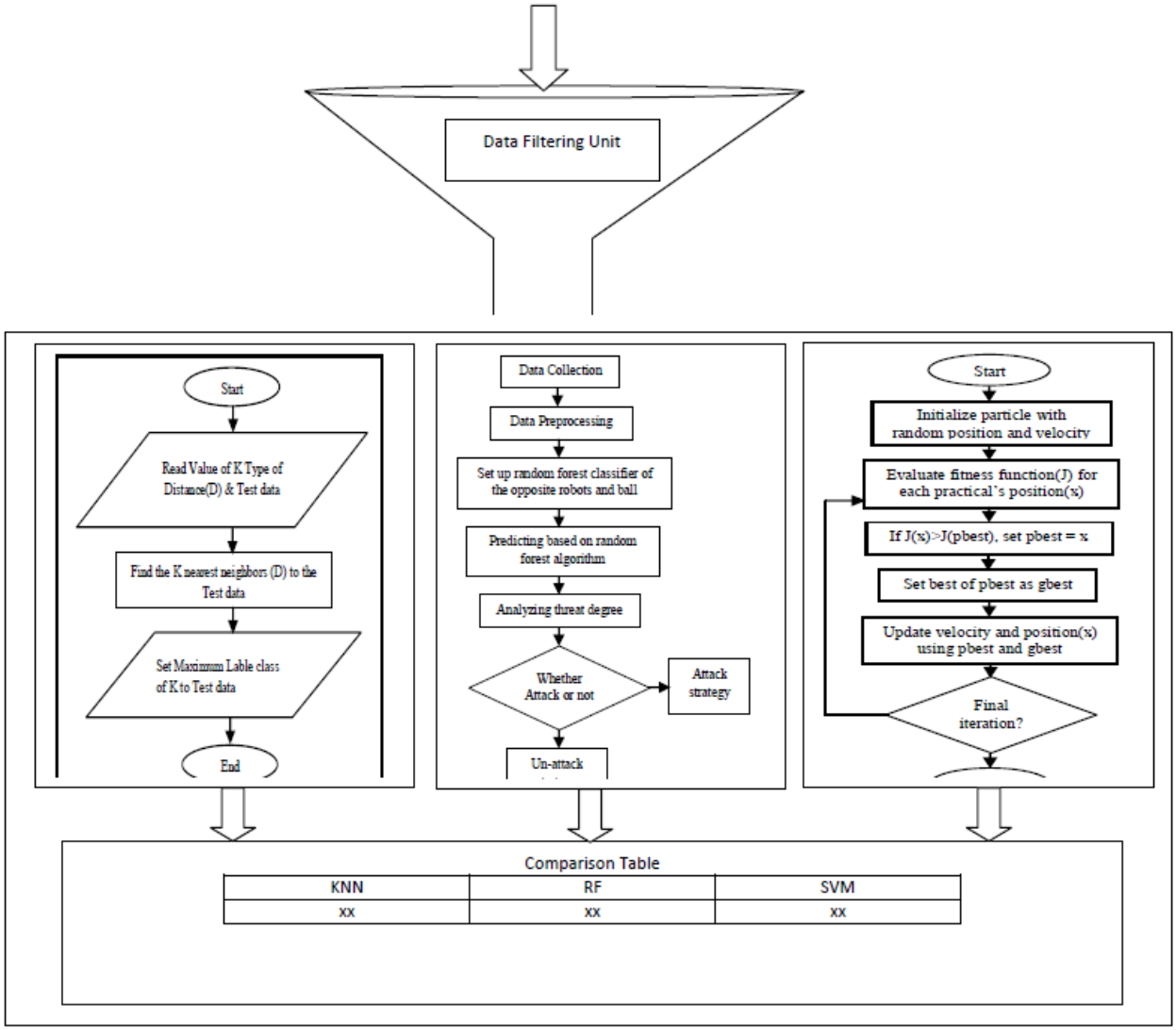

Figure.4.1. Architecture Diagram

In general, the COVID dataset has many NULL value attributes as well such as latitude, longitude, geo_resolution, date_onset_symptoms, etc… Among these attributes, the analysis engine will select only necessary attributes from the training sample and will pass to the algorithm one by one.
Each of the algorithms has been executed separately. At last, the result of each algorithm is recorded and compared with one another in a tabular format.

V. EXPERIMENTATION OF ALGORITHM WITH COVID'19 DATASET

KNN Algorithm

KNN simply defines that similar things are near to each other with closer proximity. Here K denotes that number of training data points lying in proximity to the test data point which is going to use to find the class. It doesn't require to build the model as before and not necessary to tune several parameters. It is versatile, simple and easy to implement. This algorithm was fair in interpreting output, distance calculation time and predictive power. This algorithm can be applied in various applications like banking, calculating credit ratings, prediction. By considering such application it has been selected to apply with the COVID'19 data set. Since basically it produce high accuracy among various supervised learning algorithm.

1. Prepare the training data and test data to begin KNN algorithm
2. Define the value of K
3. For each data point in test data
   - Calculate the Euclidean distance to all training data points
   - Store the determined Euclidean distances in a list and sort it
   - Select the first K points
   - Assign a class to the test point based on the majority of classes present in the selected points.
4. End

The Mathematical model of KNN algorithms is given below,

Euclidean Distance(x,xi) = sqrt(sum((xj - xij)^2)
Where,
new point (x)
existing point (xi)
all input attributes j

The training data set has been executed with the KNN algorithm and retrieved performance level concerning precision-recall f1-score and support measure. Other than this parameter accuracy, macro average and the weighted average has been taken to measure the overall performance of an algorithm. The algorithms has produced the overall accuracy of 75% with the support measure of 53 in macro average and weighted average as well.

Table 1.KNN Algorithm Result

State	Precision	Recall	F1-Score	Support
0	0.61	0.78	0.68	18
1	0.87	0.74	0.80	35
Accuracy			0.75	53
Macro Avg	0.74	0.76	0.74	53
Weighted Avg	0.78	0.75	0.76	53

Random Forest Algorithm

The Mathematical term used in Random Forest algorithms is given below, RF is a kind of classification algorithm and will generate smaller kinds of decision trees and each decision tree has been solved individually and the final result has been aggregated as a whole to produce the output. Random forest is an ensemble model by aggregating the many decision trees using bootstrapping, random subsets of features, and average voting to make predictions. RandomForestClassifier class from the sklearn.ensemble library and takes 2 kinds of a parameter such as n_estimators and criterion. These two parameters are used to take care of the overfitting issue and to calculate the entropy for information gain.

Table 2 Random Forest Algorithm Result

State	Precision	Recall	F1-Score	Support
1	0.58	0.71	0.56	48
Accuracy			0.64	37
Macro Avg	0.67	0.69	0.60	48
Weighted Avg	0.64	0.69	0.63	48

Information Gain(S,a) = H(S) – H(S | a)

Where IG(S, a) is the information for the dataset S for the variable a for a random variable, H(S) is the entropy for the dataset before any change and H(S | a) is the conditional entropy for the dataset gave the variable a.

1. Randomly choose “k” features from total “m” features in given data set.

Where k << m

2. Among the “k” features, calculate the node “d” using the best split point.
3. Split the node into daughter nodes using the best split.
4. Loop step 1 to 3 steps until “l” number of nodes has been reached.

Build forest by repeating steps 1 to 4 for “n” number times to create “n” number of trees.

The beginning of random forest algorithm starts with randomly selecting “k” features out of total “m” features. In the image, you can observe that we are randomly taking features and observations.

SVM Classifier

SVM Classifier finds a hyperplane in an N-dimensional data space that distinctly classifies the data points. Through the maximization of the margin distance provides some reinforcement so that future data points can be classified with more confidence. The dimension of hyperplane is depended upon the number of features that have been selected to perform the classification operation. Here the system selected features like sex, ages, country, province, city, infection_case, state to measure the performance level of a classifier. The performance will be measured in range(0,1) either small size of margin or larger size of margin has been set with the selected hyperplane. Hypothesis function of SVM has been defined as,

$$h(x_i) = \begin{cases} +1 \; ; if \; w.x + b \geq 0 \\ -1 \; ; if \; w.x + b < 0 \end{cases}$$

The point is greater or on the hyperplane will be classified as class +1, and the point below the hyperplane will be classified as class -1. Computing the soft-margin of SVM classifier amounts to minimizing an expression of the form

$$\left[\frac{1}{n}\sum_{i=1}^{n} \max\left(0, 1 - y_i\left(w.x_i - b\right)\right)\right] + \lambda\|w\|^2$$

Here soft-margin classifier since choosing a sufficiently small value for lambda yields the hard-margin classifier for linearly-classifiable input data.

1. Initialize particle with random position and velocity
2. Evaluate fitness function(J) for each practical’s position(x)
3. If J(x)>J(pbest), set pbest = x
4. Set best of pbest as gbest
5. Update velocity and position(x) using pbest and gbest
6. Loop it step 2 to step 5 until all the training data points are lies any one of margin
7. Stop the algorithm once all data points are assigned to margin place.

With the SVM classifier identifying the right hyper-plane is the challenging one and segregating the two difference classes. SVM algorithm having specific feature to select and apply with the COVID data set. SVM is effective in high dimensional space. More effective when number of dimensions is greater than the number of samples has been taken. SVM works well with a clear margin of separation.

Table 3 SVM Classifier Result

State	Precision	Recall	F1-Score	Support
0	0.60	0.58	0.63	18
1	0.78	0.70	0.75	32
Accuracy			0.70	53
Macro Avg	0.72	0.71	0.68	53
Weighted Avg	0.70	0.69	0.72	53

VI. ANALYZING THE ACCURACY FACTOR

The considered training sample dataset[3][4] has been given here in tabular form in below. Initially the raw data set has been taken which has many attribute which holds values of some attribute and few attributed are NULL at the beginning state.

Among the different set of attributes we have considered some attribute which are strongly associated with values. Null value attributes are avoided and given importance to the fields which are majorly enough to work on data. Some of the NULL value attributes which are not considered such as, travel_history_dates,

Table 4 Sample Training Dataset

sex	ages	country	province	city	infection_case	state
male	50s	Korea	Seoul	Gangseo-gu	overseas inflow	released
male	30s	Korea	Seoul	Jungnang-gu	overseas inflow	released
male	50s	Korea	Seoul	Jongno-gu	contact with patient	released
male	20s	Korea	Seoul	Mapo-gu	overseas inflow	released
female	20s	Korea	Seoul	Seongbuk-gu	contact with patient	released
female	50s	Korea	Seoul	Jongno-gu	contact with patient	released
male	20s	Korea	Seoul	Jongno-gu	contact with patient	released
male	20s	Korea	Seoul	etc	overseas inflow	released
male	30s	Korea	Seoul	Songpa-gu	overseas inflow	released
female	60s	Korea	Seoul	Seongbuk-gu	contact with patient	released
female	50s	China	Seoul	Seodaemun-gu	overseas inflow	released
male	20s	Korea	Seoul	etc	overseas inflow	released
male	80s	Korea	Seoul	Jongno-gu	contact with patient	released
female	60s	Korea	Seoul	Jongno-gu	contact with patient	released
male	70s	Korea	Seoul	Seongdong-gu	Seongdong-gu APT	isolated
male	70s	Korea	Seoul	Jongno-gu	contact with patient	released
male	70s	Korea	Seoul	Jongno-gu	contact with patient	released
male	20s	Korea	Seoul	etc	etc	isolated
female	70s	Korea	Seoul	Jongno-gu	contact with patient	released
female	70s	Korea	Seoul	Seongdong-gu	Seongdong-gu APT	isolated
male	80s	Korea	Seoul	Jongno-gu	contact with patient	released
male	30s	Korea	Seoul	Seodaemun-gu	Eunpyeong St. Mary's Hospital	isolated
male	50s	Korea	Seoul	Seocho-gu	Shincheonji Church	isolated
male	40s	Korea	Seoul	Guro-gu	contact with patient	released
male	60s	Korea	Seoul	Gangdong-gu	Eunpyeong St. Mary's Hospital	isolated
male	30s	Korea	Seoul	Seocho-gu	etc	released
male	50s	Korea	Seoul	Gangseo-gu	overseas inflow	released
female	70s	Korea	Seoul	Jongno-gu	Eunpyeong St. Mary's Hospital	released
female	20s	Korea	Seoul	Jongno-gu	Eunpyeong St. Mary's Hospital	released

travel_history_location, reported_market_exposure, additional_information, chronic_disease_binary, chronic_disease, source sequence_available, outcome, date_death_or_discharge.

Here is a summary of three difference algorithms has been selected for measuring the performance factor as follows according to the state 0 state 1 and state 2, where 0 refers “released” , 1 refers “isolated” and 2 refers “deceased”. Based on these states only the future decisions were taken to save the public and it has been simulated as a prototype.

Experimental Result and Comparisons

Algorithm performance has been measured by using the four different metrics such as precision, recall, F1-Score and based on these metrics value the accuracy factor has been calculated. It has been simulated using python code and uses the method called accuracy_score which is exist within the sklearn package. The whole data set has been split into 70% of training sample and 30% of test sample and passed to the algorithms. Precision and recall are two most important metrics to measure the classification operations using machine learning algorithms. Precision is a measure to calculate the percentage of relevant result, recall refers the percentage of total relevancy of result correctly classified the algorithm.

Precision supports when the costs of false positives are high, recall supports when the cost of false negatives are high. The both precision and recall has been calculated as follows,

$$Precision = \frac{TP}{TP + FP} \; ;defines\ the\ proportion\ of\ positive\ identification$$

$$Recall = \frac{TP}{TP + FN} \; ;defines\ the\ proportion\ of\ actual\ positives\ was\ identified$$

F1-Score is the harmonic mean of both precision and recall and it has been calculated as follows,

$$F1 - Score = 2 * \frac{P * R}{P + R} \begin{cases} ; = 1, Perfect \\ ; = 0, Total\ Failure \end{cases}$$

$$Precision = \frac{TP}{TP + FP} \; ;defines\ the\ proportion\ of\ positive\ identification$$

$$Recall = \frac{TP}{TP + FN} \; ;defines\ the\ proportion\ of\ actual\ positives\ was\ identified$$

F1-Score is the harmonic mean of both precision and recall and it has been calculated as follows,

$$F1 - Score = 2 * \frac{P * R}{P + R} \begin{cases} ; = 1, Perfect \\ ; = 0, Total\ Failure \end{cases}$$

A good F1-Score defines that the algorithm predicts low false positives and low false negatives, and helps to identify the real threats not distributed by false values.

Accuracy can tell us immediately whether a model is being trained correctly and how it may perform generally. The term accuracy will consider both positive and negative sample among the overall training data set. It can be calculated as follows.

$$Accuracy = \frac{TP + TN}{Total\ Samples}$$

All these four metrics are closely bind with the algorithms used with COVID'19 data analysis and produced the result has been tabulated based on the states has been taken as 0, 1 and 2. Here precision and recall are the directly proportional to each other, generally if it is required to achieve higher precision rate then it needs to restrict the positive predictions to those with highest certainty in data model.

Table 5. State 0 Comparative Results
State 0 = Released

Algorithm	Precision	Recall	F1-Score	Accuracy
KNN	0.61	0.78	0.68	0.75
RF	0	0	0	0
SVM	0.6	0.58	0.63	0.7

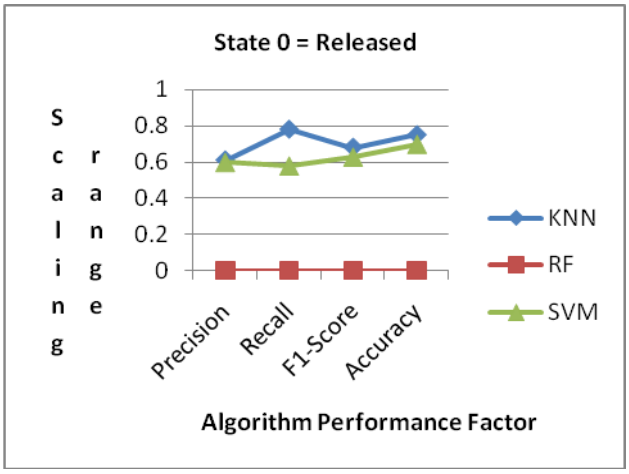

Figure.6.1. State 0 Graphical Analysis

Table 6.State 1 Comparative Results
State 1 = Isolated

Algorithm	Precision	Recall	F1-Score	Accuracy
KNN	0.87	0.74	0.8	0.75
RF	0.58	0.71	0.56	0.64
SVM	0.78	0.7	0.75	0.7

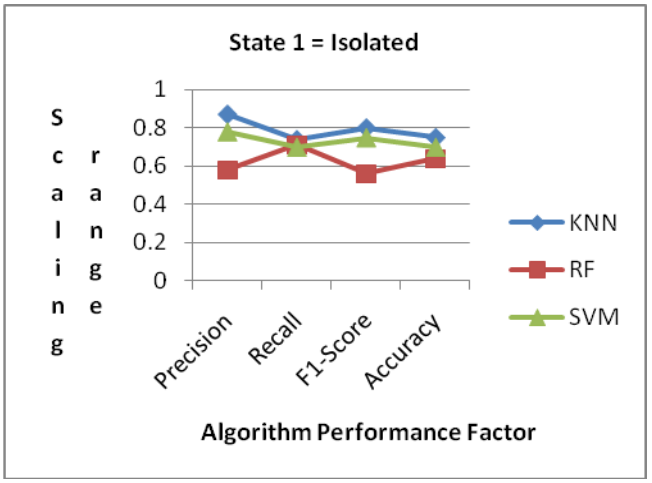

Figure.6.2. State 1 Graphical Analysis

With the training dataset the occurrence of state 2 i.e., "deceased" is accountably poor in the count of training data set, so that all three algorithms have not been used state 2 anywhere in the execution and state 0 is not processed by the Random Forest algorithm since the beginning time releasing action has not been taken so that decision trees and sub trees are not constructed by the algorithm.

VII. CONCLUSION

An investigation result shows that the performance factors of the kNN algorithm produce an accuracy of 0.75 and it has been compared with the other two more algorithms such as Random Forest and SVM classifier. The random forest algorithm has didn't classify the state 0 which refers released case of COVID infected patient. With kNN algorithm, it produces better precision, recall, and F1-Measure rate that has been projected in a chart. So that forecasting of COVID decease in the future it's better to use kNN algorithm to predict whether the percentage of decease will get an increase or will decrease. Also, it will help to classify the regional based alerting based on the severity. We can improve the efficiency in terms of time complexity, by combing the best features of each algorithms(hybrid classification). In future, we can also use the same dataset with deep learning and image analysis. There is a research scope to predict the level of decease positivity or negativity from the scanned lung images. There are possibilities to give accuracy with lung image analysis using deep learning techniques with concern no of deep neural network layers.

REFERENCES

1. Phan Thanh Noi, and Martin Kappas, "Comparison of Random Forest, k-Nearest Neighbor, and Support Vector Machine Classifiers for Land Cover Classification Using Sentinel-2 Imagery", Published in MDPI 2017

2. A public data lake for analysis of COVID-19 data from amazon data lake.
3. https://data.gov.in/major-indicator/covid-19-india-data-source-mohfw
4. https://www.coronavirus.gov/
5. https://www.kaggle.com/allen-institute-for-ai/CORD-19-research-challenge
6. Kewen Li, Peng Xie, Jiannan Zhai, Wenying Liu, "An Improved Adaboost Algorithm for Imbalanced Data Based on Weighted KNN" published in IEEE 2nd International Conference on Big Data Analysis – 2017
7. Okfalisa, Mustakim, Ikbal Gazalba, Nurul Gayatri Indah Reza, "Comparative Analysis of K-Nearest Neighbor and Modified K-Nearest Neighbor Algorithm for Data Classification" published in 2nd International Conferences on Information Technology, Information Systems and Electrical Engineering (ICITISEE) – 2017
8. Shaobo Du, Jing Li, "Parallel Processing of Improved KNN Text Classification Algorithm Based on Hadoop" published in The 7th International Conference on Information, Communication and Networks – 2019
9. Chuan-Quan Li, You-Wu Lin, Qing-Song Xu, "An enhanced random forest with canonical partial least squares for classification" published in Communications in Statistics – Theory and Methods – Jan 2020
10. V. R. Elgin Christo, H. Khanna Nehemiah, J. Brighty, Arputharaj Kannan, "Feature Selection and Instance Selection from Clinical Datasets Using Co-operative Co-evolution and Classification Using Random Forest" published in IETE Journal of Research – Jan 2020
11. Srivathsan Srinivasagopalan, Justin Barry, Varadraj Gurupur, Sharma Thankachan, "A deep learning approach for diagnosing schizophrenic patients", published in Journal of Experimental & Theoretical Artificial Intelligence – Dec 2018
12. Ravi Shanker, Mahua Bhattacharya, "Brain tumor segmentation of normal and lesion tissues using hybrid clustering and hierarchical centroid shape descriptor" published in Computer Methods In Biomechanics And Biomedical Engineering: Imaging & Visualization – Feb 2019
13. J. Han, J. Pei, and M. Kamber. Data Mining: Concepts and Techniques. Elsevier, 2011.
14. J.Derrac, S. Garcia, and F. Herrera, "Ifs-coco: instance and feature selection based on cooperative coevolution with nearest neighbor rule," Pattern Recognit., Vol. 43, no. 6, pp. 2082–2105, 2010.
15. Maksoud EA, Elmogy M. 2016. 3D Brain Tumor Segmentation Based on Hybrid Clustering Techniques using Multi-Views of MRI. In: Dey N, Bhateja V, Hassanien A, editors. Springer medical imaging in clinical applications. Cham: Springer. p. 81–104.
16. J.Derrac, S. Garcia, and F. Herrera, "Ifs-coco: instance and feature selection based on cooperative coevolution with nearest neighbor rule," Pattern Recognit., Vol. 43, no. 6, pp. 2082–2105, 2010.
17. N. Leema, H. K. Nehemiah, and A. Kannan, "Neural network classifier optimization using differential evolution with global information and back propagation algorithm for clinical datasets," Appl. Soft. Comput., Vol. 49, pp. 834–844, 2016.
18. MA Ying, ZHAO Hui, CUI Yan. Parallel processing of improved KNN classification algorithm based on Hadoop platform [J]. Journal of Changchun University,2018,39(05):484− 489.
19. B. Ma.A New Kind of Parallel K_NN Network Public Opinion Classification Algorithm Based on Hadoop Platform[J], Applied Mechanics and Materials, Vols. 644-650, pp. 2018-2021, 2014.
20. Parvin, Hamid, Alizadeth, Hoseinali, and M. Behrouz, "A Modification on K-Nearest Neighborn Classifier". Global Journal of Computer Science and Technolgy. 2010. Vol. 10 No. 14, pp. 37-41.

AUTHORS PROFILE

Venkatesan R is having 2 years of research experience of 6 years of teaching. He has received B.E., in CSE and M.E., in Software Engineering from the Anna University Chennai in 2012 and 2014. He is currently working as an Assistant Professor in the Department of Computer Science and Engineering at KSR Institute for Engineering and Technology, Tamilnadu from 2016 to till date, published 4 papers in international journal, presented 5 papers in international and national conferences and provisionally filed a patent of a societal application product. He is interested in applications of data mining, recommender systems, and information retrieval.

Nandhagopal S is having 5 years of teaching experience. He has received the Bachelor's Degree in Information Technology and M.E., in Computer Science and Engineering from the Anna University Chennai in 2011 and 2014. He is currently working as an Assistant Professor in the Department of Information Technology at KSR Institute for Engineering and Technology Tamilnadu, from 2017 to till date. He has published 2 papers in International journals, presented 6 papers in international and national conference and provisionally filed a patent of a societal application product He is interested in image processing, applications of machine learning and Data mining.

A. Sabari is having 15 years of teaching experience. He has received B.E. in CSE from Madras University, Tamil Nadu, in 2000 M.Tech in CSE from Vishweshraiya Technological University, Karnataka, in 2002 and completed Ph.D. in Anna University, Chennai in 2011. He is currently working as a Professor in Information Technology at K. S. Rangasamy College of Technology, Tamil Nadu, from 2002 to till date. Published 38 papers in international journals, 14 papers in conferences, produced 10 research scholars and supervising 5 research scholars under Anna University, Chennai. He is interested in AdHoc networks, WSN, Network Security and Data Mining.