An Artificial Intelligence System for Classification of COVID-19 Suspicious Person using Support Vector Machine (SVM) Classifier

Nitin M. Shivale
Department of Computer Engineering JSPM'S BSIOTR Wagholi, Pune

Gauri Virkar
Department of Computer Engineering JSPM'S BSIOTR Wagholi, Pune

Tejas L. Bhosale
Police Patil Shivajinagar Grampanchyat
Tal Khandala Dist Satara

ABSTRACT
The outbreak of corona virus disease 2019 (COVID-19), caused by severe acute respiratory syndrome (SARS) corona virus 2 (SARS-CoV-2), has till date(April 2020) killed over 825 people, 5939 recovered and infected over 26,496 in India and elsewhere in the world, resulting in destruction for humans. However, COVID-19 has lower severity and mortality than SARS but is much more transmissive and affects more elderly individuals, youth and more men than women. In response to the rapidly increasing number of infected count of the emerging disease in urban area and people in urban areas have no jobs due to lockdown so they start migration from urban to rural area which may create lots of problem in rural area even though the lower density of rural areas may help keep transmission rates of the disease down. This research claims to provide better accuracy since the data received is verified by the reliable source. Further, this paper attempts to provide an Artificial Intelligence System for classification of COVID-19 suspicious person using different machine algorithms to break the chain of novel corona virus outbreak. The rural areas can be kept secured from getting infected once the chain of transmission is broken. Although many questions still require answers, this paper helps in the identifying the suspicious person and eradication of the threatening disease.

Keywords
Coronavirus, COVID-19, Classifiers, Support Vector Machine (SVM), K-NN

1. INTRODUCTION
In December 2019, cluster of people affected with an unknown disease was reported in the city of Wuhan, China. This previously unknown virus is now known as 2019 Novel Corona virus (2019-nCoV). A novel corona virus (nCoV) is a new strain that has not been previously identified in humans. World Health Organization(WHO) identified that this virus belonged to the family of Severe Acute Respiratory Syndrome corona virus(SARS-CoV) first identified in China in 2003 and the Middle East Respiratory Syndrome Corona virus (MERS-CoV) that was first identified in Saudi Arabia in 2012[1][2]. According to current evidence, COVID-19 spreads mainly through the droplets of a COVID-19 infected person produced when he/she coughs or sneezes. This virus can also spread if the droplets of infected person fall on the surface and then if normal healthy person comes in touch of such surfaces or objects. Novel corona viruses (nCoV) have varying abilities to infect people. For COVID-19, each person with the virus can go on to infect around 2.5 people. If each of those people go about their day as normal, and infect another 2.5 people, within a month, 406 people would be infected just from that first infection [3]. All these factors, made World Health Organization (WHO) announced the outbreak of this disease to be Public Health Emergency of International Concern on 30 January 2020 and recognized it as a pandemic on 11 March 2020. [4]. More than 212 countries and territories around the globe are affected with this deadly disease leading the death toll rate of more than 2.5 lakhs people. Whereas in India around 29,435 confirmed cases of COVID-19 were reported at the end of April 2020 and more than 900 people lost their lives fighting against this disease [5]. It has become the need of the day to control this pandemic. Various preventive measures are been given Center of Disease Control (CDC) against this virus. In the absence of treatment or a vaccine, ceasing most human contact is really one of the vital measures to stop the spread of the virus. Essentially, the less contact people have with each other, the less the virus can spread. Social distancing is one such primary preventive measure.

In order to curb the widespread of this pandemic disease, the Government had imposed a strict lockdown in India as a preventive measure to implement social distancing. According to the 2011 census of India, 68.84% of Indian population lives in 640,867 different villages [6]. One of the good strategies can be preventing the villages from getting infected by this deadly disease. If the villagers are protected from this infectious disease, we can control large population of India (around 833.1 million people) from this pandemic. With this viewpoint, a small survey was been conducted in two small corona-free villages of India in Maharashtra. While conducting this survey, two main problems were observed in these villages:

1. It was becoming difficult to keep track of people coming in / out of the city every day.
2. Villagers were reluctant to visit a doctor if they suffered through any fever, cough/cold out of fear.

With an aim to address the above issues, a simple methodology has been proposed and implemented to observe, identify and predict the suspicious person using Machine Learning algorithms and report about the same to the Government or higher authorities thereby preventing others from getting infected and also starting the early treatment for the infected ones.

2. PROBLEM STATEMENT
Large number of population which lives in urban areas gets feed by Indian villagers (rural area) so the population which lives in villages (rural area) are the backbone of urban area in India. If these villages are not prevented from getting infected

by this deadly disease, it may collapse the development as well as economical cycle of India. So to make the Indian villages safe and secure, the chain of COVID-19 outbreak must be broken by analyzing and then predicting the suspicious person suffering through COVID-19 using various Machine Learning algorithms.

3. PROPOSED METHODOLOGY

The Proposed System architecture consists of three main components.

A. Data Provider: Villagers who are responsible to provide the asked data about themselves.

B. Data Verifier: They are the village coordinators appointed by Police Patil(Police Incharge) of Grampanchyat who ensures and verify the data provider's filled data.

C. Grampanchyat Officer/Hospital: These are the ones who are notified about COVID-19 suspicious cases if found.

Basically the proposed architecture as shown in Fig.1 is divided into following modules:

3.1 Data Acquisition

3.2 Data Verification and Finalization

3.3 Data Pre-processing

3.4 Result Prediction

3.5 Reporting

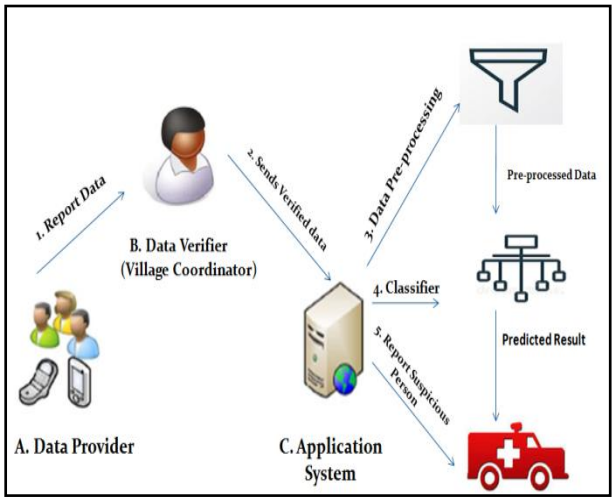

Fig 1: Proposed System Architecture

3.1 Data Acquisition

Basically in this module, the data is collected from the people of two corona-free villages in the form of questionnaires. These questionnaires are shared to them through a web link where their details such as: 1.Personal Details (Name, Age, Gender, Profession, etc.), 2. Medical history or any recent symptoms of any ailments are enquired, 3. The person's where about(Travel History), 4. Any contact direct/indirect with COVID-19 Patient or whether any public event, attended etc. such kind of information is acquired from the subjects of these villages.

The following images show the glimpse Google Form of these questionnaires.

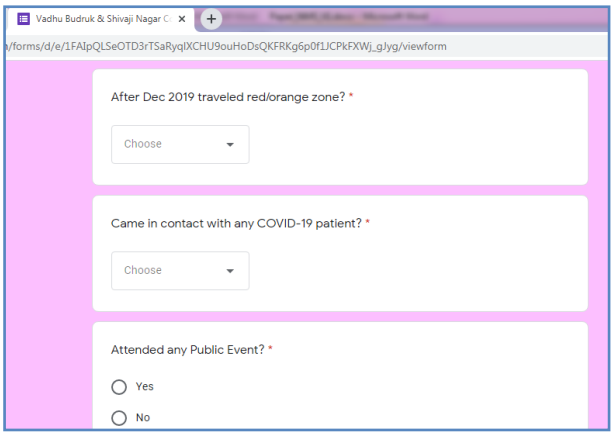

Fig 2: Snapshots of Google Questionnaire to be filled by the Villagers

All such information has to be filled by each individual of the selected village. If in case due to any reason a person is unable to fill this Google form, the Village coordinator fills it for the

3.2 Data Verification & Finalization

In the proposed system, against every fifteen families of the village, one coordinator is appointed who is responsible for following activities:

1. To ensure if all the people under his surveillance, fills the Google Form. If not, it is his responsibility to assist them.
2. To verify, the people are providing right information.
3. To observe and report if he finds any susceptible person of COVID-19.
4. To observe and report to the Village chief if any newcomer is coming to the village.

3.3 Data Pre-Processing

Once the data is verified by the village co-coordinators, the available data can be used as a dataset for predicting and classifying the susceptible person. The dataset contains integer and categorical type of 20 attributes. In order to get accurate results, several pre-processing techniques such as **Data Cleaning** (handling missing values/ null values), **Data Reduction**(Attribute Selection- Attributes that do not have any significance on the output variables are deleted from use during pre-processing (E.g.:-Profession))[7]. **Data Transformation** techniques such as label encoding are used for handling categorical data.

After the initial pre-processing, this pre-processed dataset is provided as an input to the various supervised machine learning algorithms. K-fold Cross Validation technique is been used with k= 10 is selected in the implementation since the accuracy obtained using these values were the best [8] .

3.4 Result Prediction

In this study, following six different classification algorithms are considered in order to find the classifier with most accurate results [9].

i. Support Vector Machine (SVM),

ii. K-NN,

iii. Logistic Regression,

iv. Decision Tree,

v. Linear Discriminant Analysis

vi. Gaussian NB

Mean value, accuracy and standard deviation values for these classifiers were calculated.

3.5 Reporting

The classifier predicts the susceptible person for COVID-19 from the given dataset. If the system finds any susceptible person, immediately the alert notification about the same is given to the Grampanchyat authorities - Gram Sevak, Police Patil and Village Coordinator. This suspected person is then screened for COVID-19 test and checks if his/her results are positive or not. This early notification can be a great aid to curb down the number of corona virus infected cases.

4. EXPERIMENTAL RESULT

For performance evaluation, Support Vector Machine (SVM), K-NN, Logistic Regression, Decision Tree, Linear Discriminate Analysis and Gaussian NB classification algorithms are considered. Mean and Standard Deviation of these algorithms were calculated [10][11]. The Fig.3 shows the screenshot of this classification metrics:

```
#Calculating Classification Accuracy
for model_name, model in models:
    k_fold_validation =model_selection.KFold(n_splits=10, random_state=random_seed)
    results = model_selection.cross_val_score(model, X, Y, cv=k_fold_validation, scoring='accuracy')
    outcome.append(results)
    model_names.append(model_name)
    output_message = "%s| Mean=%f STD=%f" % (model_name, results.mean(), results.std())
    print(output_message)
```

```
LogReg| Mean=0.966667 STD=0.100000
SVM | Mean=0.966667 STD=0.066667
DecTree| Mean=0.933333 STD=0.133333
KNN| Mean=0.783333 STD=0.236291
LinDisc| Mean=0.983333 STD=0.050000
GaussianNB| Mean=0.966667 STD=0.100000
```

Fig. 3 Screenshot of Classification Metrics

In order to evaluate these predictions, the above obtained result values are converted into percentage. The following table shows the results in percentage.

Table 1. Table indicating Mean and Standard error deviation.

Sr. no	Classifier	Mean(%)	STD(%)
1	Logistic Regression	96.67	10
2	Support Vector Classifier	97.78	6.67
3	Decision Tree	93.33	13.333
4	K-NN Classifier	78.33	23.629
5	Linear Discriminate Analysis	98.33	5
6	Gaussian NB	96.66	10
Average Accuracy and Std. Deviation		93.51	11.43

The following figure shows the comparative analysis of different classifiers using graph with respect to Mean value (%). It can see from these results that, Linear Discriminate Analysis (LDA) shows the best results with an accuracy of 98.33% followed by Support Vector Classifier (97.78% accuracy), Logistic Regressor, Gaussian NB, Decision Tree and then K-Nearest Neighbors Classifier with accuracy score of 96.67%, 96.66%, 93.33%, 78.33% respectively. Overall average accuracy of 93.51% is obtained.

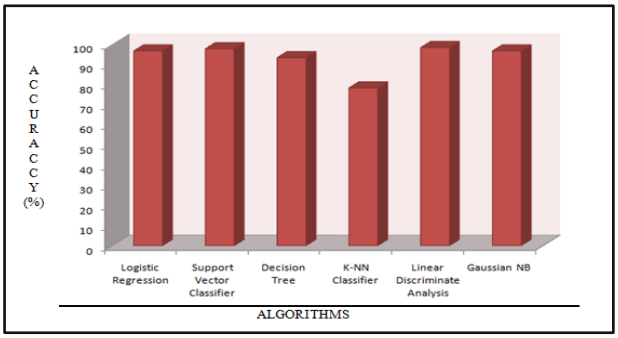

Fig 3: Comparative analysis of different classifiers

It can be observed that both Linear Discriminate Analysis (LDA) and Support Vector Machine (Linear Support Vector Classifier) shows marginal difference and hence any of them can be selected as the main classifier that can be used in this system for prediction. In our system we have selected Support Vector Machine for the prediction [12][13]. This implemented Linear Support Vector Classifier provides an average of 83.33% of accuracy.

5. CONCLUSION

Identifying and predicting the susceptible person of COVID-19 is one of the challenging tasks. The problem becomes more difficult especially in collecting reliable data. Only if we have reliable data, then we can predict accurate results of suspicious person. In order to deal with this problem, we have proposed human intervention in this research. A village coordinator, one of the village inhabitants helps here to ensure and verify whether the data provided by the people is valid or not. This helps us in giving more precise results since we get data verified by the village coordinator. In order to break the chain of this pandemic disease, it is essential to make early prediction of susceptible cases. The SVM algorithm provides an average of 97.78% of accuracy in predicting such cases whereas overall average accuracy using different selected classification algorithms attained is around 95.31%. This will enable to forecast the suspicious person early can surely act as a life-saver hack in controlling this deadly disease and thereby decreasing the rate of spread of this disease amongst the uninfected people. Currently, in this research data of only two villages is considered. For future scope, it can be extended to larger geographic locations.

6. REFERENCES

[1] World Health Organization. 2020. Novel Coronavirus – China. [online] Available at: <https://www.who.int/csr/don/12-january-2020-novel-coronavirus-china/en/> [Accessed 6 May 2020].

[2] 2020. [online] Available at: <http://wjw.wuhan.gov.cn/front/web/list2nd/no/710> [Accessed 6 May 2020].

[3] Gavi.org. 2020. Why Is Coronavirus Lockdown Necessary?. [online] Available at: <https://www.gavi.org/vaccineswork/why-coronavirus-lockdown-necessary> [Accessed 6 May 2020].

[4] En.wikipedia.org. 2020. COVID-19 Pandemic. [online] Available at: <https://en.wikipedia.org/wiki/COVID-

19_pandemic> [Accessed 6 May 2020].

[5] Worldometers.info. 2020. Coronavirus Update (Live): 3,778,012 Cases And 261,243 Deaths From COVID-19 Virus Pandemic - Worldometer. [online] Available at: <https://www.worldometers.info/coronavirus/>

[6] En.wikipedia.org. 2020. Village. [online] Available at: https://en.wikipedia.org/wiki/Village

[7] Wen, Z.; Li, B.; Kotagiri, R.; Chen, J.; Chen, Y.; and Zhang, R. 2016. Improving efficiency of SVM k-fold cross-validation by alpha seeding. Technical Report arXiv:1611.07659

[8] Z. Nematzadeh, R. Ibrahim and A. Selamat, "Comparative studies on breast cancer classifications with k-fold cross validations using machine learning techniques," 2015 10th Asian Control Conference (ASCC), Kota Kinabalu, 2015, pp. 1-6, doi: 10.1109/ASCC.2015.7244654.

[9] Chang, C.-C., and Lin, C.-J. 2011. LIBSVM: a library for support vector machines. TIST 2(3):27.

[10] Chu, B.-Y.; Ho, C.-H.; Tsai, C.-H.; Lin, C.-Y.; and Lin, C.-J. 2015. Warm start for parameter selection of linear classifiers. In Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 149-158. ACM

[11] Schott, J. R. 2005. Matrix analysis for statistics.

[12] Wen, Z.; Li, B.; Kotagiri, R.; Chen, J.; Chen, Y.; and Zhang, R. 2016. Improving efficiency of SVM k-fold cross-validation by alpha seeding. Technical Report arXiv:1611.07659 [cs.LG], arXiv

[13] International Journal of Modern Trends in Engineering & Research, 2017. Comparative Analysis of Linear Regression, Multilayer Perceptron and Support Vector Machines for its utilization in Stock Price Prediction. 4(6), pp.7-12.