CORONA VIRUS INFECTION PROBABILITY CLASSIFICATION USING MACHINE LEARNING ALGORITHMS

K. Murali[1], Thumma Mary Shreeja[2], Mora Srujana[3], Chilla Sneha[4]
[1]*Associate Professor, CSE, CMR Technical Campus, Hyderabad, India*
[2]*Student, CSE, CMR Technical Campus, Hyderabad, India*
[3]*Student, CSE, CMR Technical Campus, Hyderabad, India*
[4]*Student, CSE, CMR Technical Campus, Hyderabad, India*

***Abstract*: This project is titled as "Corona Virus Infection Probability Classification". Due to the unexpected outbreak of COVID-19 disease, the world is facing a major epidemic in current days. The infection as well as the death rate is growing rapidly in every country. The world economic status is also decreasing due to this disaster. It is more essential to detect the infected people at an early stage to make a break in spreading of virus. Machine learning techniques will be very useful for this purpose due to its automatic data analysis and classification ability. In the proposed work, authors have classified samples having chance of infection. A set of randomly generated data is considered for the classification purpose. The dataset contains 1200 samples with five types of COVID symptoms. By analyzing the body temperature, age, body pain, runny nose status, and breathing problem. Based on experiment result on data set, found best algorithm among the three classifiers: Decision tree, SVM, Naive Bayes, Neural Network, CNN, and Random Forest.**

***Keywords*: SVM, Naive Bayes, Neural Network, Decision Tree, Random Forest, CNN.**

1. Introduction

In the information technology (IT) society, knowledge is the most important asset for any organization. It also plays a significant role in the healthcare sector. As the progress of IT in healthcare domain is growing, people's expectation is also gradually increasing for better treatment with minimum expenses. With the wide application of the automatic computerized system in the healthcare sector, the generation of data is also increasing day-by-day. These data may be information about diseases, electronic patient records, hospital resources, diagnosis methods, etc. Extraction of useful information from these complex data is an important task for clinical decision making and it can be done by applying different data mining techniques in medical data. Data mining is a process of extracting useful information from a large amount of data. It has the great potential to extract useful and hidden knowledge from the datasets available in the medical domain.

2. Proposed System

It is a challenging task to detect and start the diagnosis process of this awful disease at an early stage. Various controlling and diagnosis techniques are applied by different countries for creating a break in the COVID-19 infection chain. Automatic analysis of different symptoms can reduce the diagnosis time as well as human interference in COVID-19 treatment. Machine learning and data mining frameworks can classify different disease by analyzing the pathological reports. These techniques will be very useful for classification and detection of
COVID-19 infection probability. Different data mining techniques can be taken for this automatic symptoms data classification system. The main challenge for developing this automated diagnosis system is the proper analysis of the data and accuracy. Numerous machine learning techniques were also used by the researchers for getting a satisfactory result in various biomedical data analysis. Support vector machine (SVM) and its variants are one of the most popular data mining technique and have shown astonishing performance for binary classification problems. The main advantage behind using SVM is that it can be paired with the kernel function.

The main objective of this work is to develop an automated COVID-19 infection probability classification system by using a machine learning technique for early detection of COVID-19. A support vector machine-based classification system is proposed for classifying infection probability by analyzing different symptoms. The performance of the proposed SVM classifier is measured for four different types of SVM kernel functions.

3. Methodology

A. System Analysis

This application is developed using XML, Java, Python. JetBrains PyCharm Community contains

many libraries. Android programs are written in Java and XML and run through a JVM that is optimized for devices. Here java is used for the backend and XML is used for the frontend.

This application is integrated with ML Algorithms. JetBrains PyCharm Community is used to write ML code (Embedded C).

Block diagram representing the architecture of Corona Virus infection probability classification shown in Fig 1.

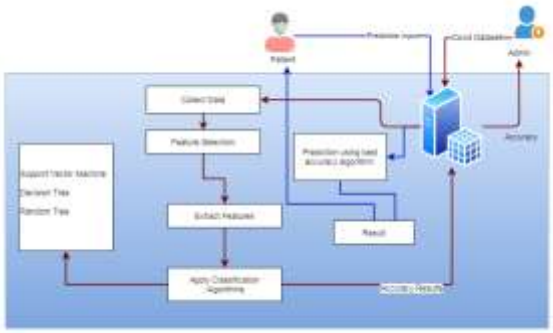

Fig. 1: Block diagram

B. Modules Description

Admin:

Admin use the dataset of covid patient's data history for classifications. Admin, train the dataset with following three algorithms.

1. Decision tree
2. SVM
3. Random forest
4. CNN
5. Neural network
6. Naive bayes.

After training the dataset, admin will test the accuracy by test data. Then admin will find the best classification algorithm. Admin also can see the graph of the accuracy of the three algorithms and find accuracy scores of all algorithms.

User:

User is end user of the application; our application will help to the user by prediction Covid disease by train the previous patient's dataset with best accuracy algorithm. User can register with own details and after login user can enter details of his/her medical parameter like fever, cough etc. User can get result with prediction of best accuracy algorithm.

System:

Our system is developed in python with PyQt5 interface components with user friendly. System will interact with database and process every action of user and admin inputs.

C. Database

In any system storing of data is very important part. In this application for the storing data, database is provided. In addition to this SQLyog is used to store some data in user internal storage. SQLyog is an application to store data, documents like text, images, video file, PDFs, tables etc.

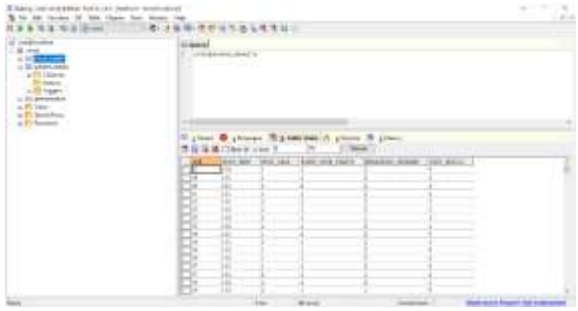

Fig. 2: SQLyog Database

D. Machine Learning Algorithms

SUPPORT VECTOR MACHINE (SVM):

Support Vector Machine is an extremely popular supervised machine learning technique (having a pre-defined target variable) which can be used as a classifier as well as a predictor. For classification, it finds a hyper-plane in the feature space that differentiates between the classes. An SVM model represents the training data points as points in the feature space, mapped in such a way that points belonging to separate classes are segregated by a margin as wide as possible. The test data points are then mapped into that same space and are classified based on which side of the margin they fall.

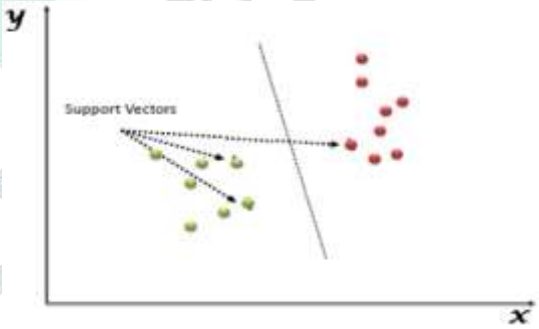

Fig. 3. SVM Algorithm Diagrammatic Representation

DECISION TREE:

This algorithm belongs to the family of supervised learning algorithms. Unlike other supervised learning algorithms, decision tree algorithm can be used for solving regression and classification problems too. The general motive of using Decision Tree is to create a training model which can use to predict class or value of target variables by learning decision rules inferred from prior data (training data).

Fig. 4: Decision tree Algorithm Diagrammatic Representation.

NAIVE BAYES ALGORITHM:

This classifier is based on Bayes theorem. It has strong independence assumption. It is also known as independent feature model. It assumes the presence or absence of a particular feature of a class is unrelated to the presence or absence of any other feature in the given class. Naïve bayes classifier can be trained in supervised learning setting. It uses the method of maximum similarity. It has been worked in complex real-world situation. It requires small amount of training data. It estimates parameters for classification. Only the variance of variable need to be determined for each class not the entire matrix. Naïve bayes is mainly used when the inputs are high. It gives output in more sophisticated form. The probability of each input attribute is shown from the predictable state. Machine learning and data mining methods are based on naïve bayes classification.

$$\frac{P(H|X) = P(X|H)\,P(H)}{P(X)}$$

- Where P (H|X) is posterior probability of H conditioned on X
- P(X|H) is posterior probability of X conditioned on H.
- P(H)is prior probability of H P(X) is prior probability of X.

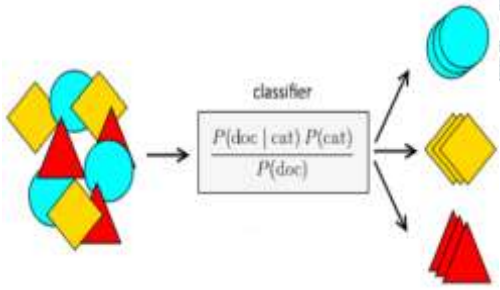

Fig. 5: Naive bayes Algorithm Diagrammatic Representation.

NEURAL NETWORKS:

A neural network is a series of algorithms that endeavors to recognize underlying relationships in a set of data through a process that mimics the way the human brain operates. In this sense, neural networks refer to systems of neurons, either organic or artificial in nature. Neural networks can adapt to changing input; so, the network generates the best possible result without needing to redesign the output criteria.

f(x) = f_3(f_2(f_1(x))) where:

f_1(x): Function learned on first hidden layer.

f_2(x): Function learned on second hidden layer.

f_3(x): Function learned on output layer.

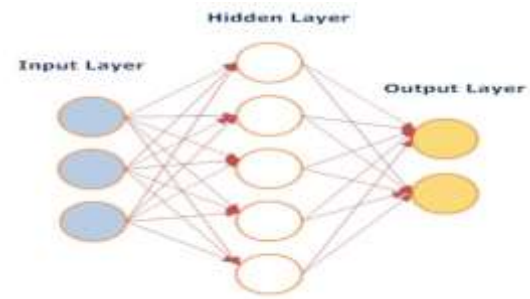

Fig. 6: Neural network Algorithm Diagrammatic Representation.

CONVOLUTIONAL NEURAL NETWORK (CNN):

The convolution kernels or filters that slide along input features and provide translation equivariant responses known as feature maps. Counter-intuitively, most convolutional neural networks are only equivariant, as opposed to invariant, to translation. They have applications in image and video recognition, recommender systems, image classification, image segmentation, medical image analysis, natural language processing, brain-computer interfaces, and financial time series.

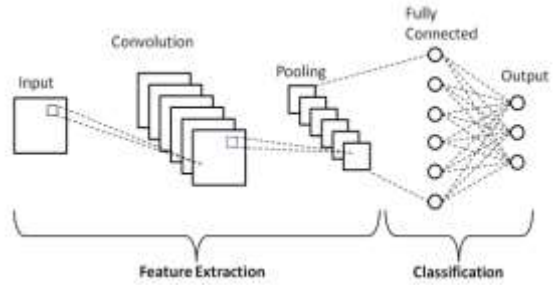

Fig. 7: CNN Algorithm Diagrammatic Representation.

RANDOM FOREST:

It is one of ensemble methods, is a combination of multiple tree predictors such that each tree depends on a random independent dataset and all trees in the forest are of the same distribution.

1. Import and print the dataset.
2. Select all rows and column 1 from dataset to x and all rows and column 2 as y.
3. Fit Random forest regressor to the dataset.
4. Predicting a new result.
5. Visualizing the result.

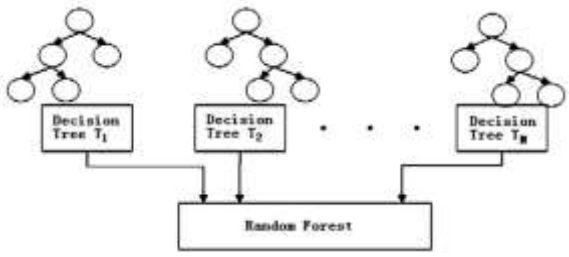

Fig. 7: Random forest classifier Diagrammatic Representation.

4. Results and Discussion

After execution we will get a home page dialog box which contains admin and patient logins.

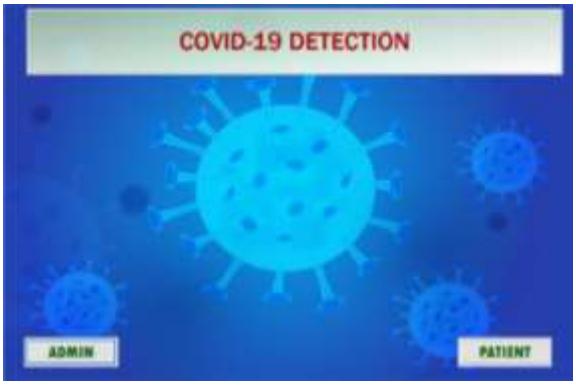

Fig. 8: Home page result for covid probability.

After clicking the admin button, we will get admin login page.

Fig. 9: Admin login page.

When we enter correct admin id and password then only, we can be able to login otherwise it shows an error called please enter valid credentials.

Fig. 10: Admin login validation page.

After validating the login credentials, we will get admin dashboard page.

Fig. 11: Admin dashboard page.

By clicking on the upload, we can upload the dataset which is in the form of excel sheet. It contains rows and columns. Columns are symptoms of corona virus infection i.e., age, body temperature, body pains, running nose, breathing problem and covid result.

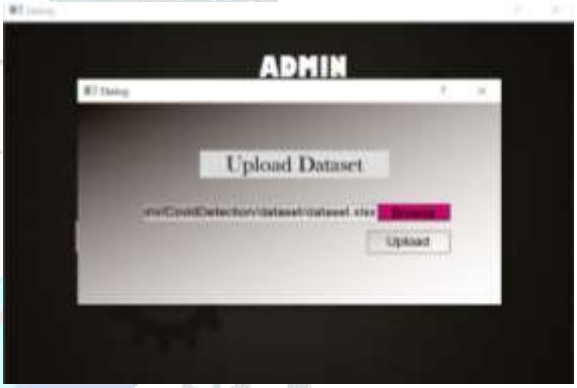

Fig. 12: Uploading dataset page.

By clicking accuracy on the admin dashboard page, we will get the prediction accuracy analysis of the data from the uploaded dataset.

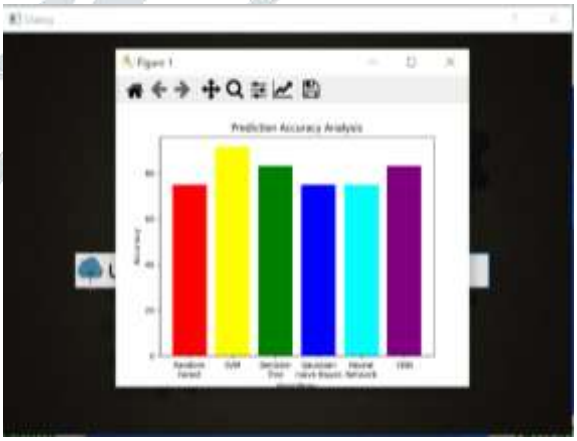

Fig. 13: Prediction accuracy analysis page.

By clicking analysis on the admin dashboard page, we get the performance of the algorithms.

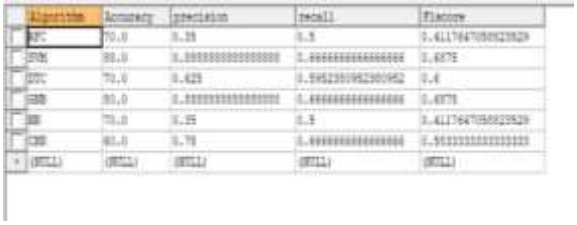

Fig. 14: Performance of algorithms.

Coming to the patient button on the home page, if you are not registered earlier then you need to register by clicking register here.

Fig. 15: User login page.

To register we need to enter username, password, name, mobile number, email id.

Fig. 16: User registration page.

After successfully creating an account we need to login, after login we should enter the symptoms of the user.

Fig. 17: Test page.

After entering the symptoms of the patient, by clicking on get result we will get the result whether it's a covid positive or negative and by clicking on clear all the details entered will be cleared.

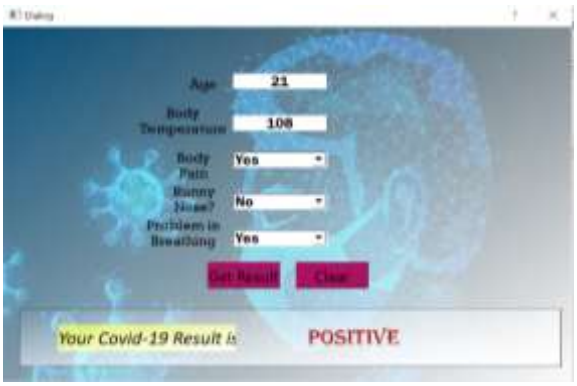

Fig. 18: Result page.

This result is predicted by SVM algorithm.

5. Conclusion

Our Study on this project revealed that the model is classifying the chance of infection probability of the persons having different COVID-19 symptoms. A SVM algorithm is designed that classify the infection probability by considering five types of input features. Total 1200 random generated sample data is considered for validating the model performance. SVM performance is measured with three different types of kernels and from the result it is observed that SVM is performing better as compared to other five algorithms. The performance is measured in terms of accuracy, precision, and recall. Around 90% classification accuracy is obtained by using SVM. In future some more accuracy can be obtained by modifying the kernels of the SVM as well as with some other machine learning techniques.

6. References

1. N. Zhu, D. Zhang, W. Wang, X. Li, B. Yang, J. Song, et al., "A novel coronavirus from patients with pneumonia in China, 2019," New England Journal of Medicine, 2020.
2. T. Bhatnagar, M. V. Murhekar, M. Soneja, N. Gupta, S. Giri, N. Wig, et al., "Lopinavir/ritonavir combination therapy amongst symptomatic coronavirus disease 2019 patients in India: Protocol for restricted public health emergency use," Indian Journal of Medical Research, vol. 151, p. 184, 2020.
3. C. Cai, L. Han, X. Chen, Z. Cao, and Y. Chen, "Prediction of functional class of the SARS coronavirus proteins by a statistical learning method," Journal of proteome research, vol. 4, pp. 1855-1862, 2005.

4. A. Rajkumar and G. S. Reena, "Diagnosis of heart disease using datamining algorithm," Global journal of computer science and technology, vol. 10, pp. 38-43, 2010.
5. K. S. Durgesh and B. Lekha, "Data classification using support vector machine," Journal of theoretical and applied information technology, vol. 12, pp. 1-7, 2010.
6. L. Chen, H. Liu, W. Liu, J. Liu, K. Liu, J. Shang, et al., "Analysis of clinical features of 29 patients with 2019 novel coronavirus pneumonia," Zhonghua jie he he hu xi za zhi= Zhonghua jiehe he huxi zazhi= Chinese journal of tuberculosis and respiratory diseases, vol. 43, pp. E005-E005, 2020.