A

Major Project

On

CORONA VIRUS INFECTION PROBABILITY CLASSIFICATION

(Submitted in partial fulfillment of the requirements for the award of Degree)

BACHELOR OF TECHNOLOGY

in

COMPUTER SCIENCE AND ENGINEERING

By

THUMMA MARY SHREEJA	(177R1A05B6)
MORA SRUJANA	(177R1A0594)
CHILLA SNEHA	(177R1A0569)

Under the Guidance of

Mr. K. Murali

(Associate Professor)

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

CMR TECHNICAL CAMPUS

UGC AUTONOMOUS

(Accredited by NAAC, NBA, Permanently Affiliated to JNTUH, Approved by AICTE, New Delhi)

Recognized Under Section 2(f) & 12(B) of the UGCAct.1956,

Kandlakoya (V), Medchal Road, Hyderabad-501401.

2017-2021

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

CERTIFICATE

This is to certify that the project entitled "**CORONA VIRUS INFECTION PROBABILITY CLASSIFICATION**" being submitted by **THUMMA MARY SHREEJA (177R1A05B6), MORA SRUJANA (177R1A0594)** and **CHILLA SNEHA (177R1A0569)** in partial fulfillment of the requirements for the award of the degree of B. Tech in Computer Science and Engineering to the Jawaharlal Nehru Technological University Hyderabad, is a record of bonafide work carried out by him/her under our guidance and supervision during the year 2020-21.

The results embodied in this thesis have not been submitted to any other University or Institute for the award of any degree or diploma.

Mr. K. Murali
Associate Professor
INTERNAL GUIDE

Dr. A. Raji Reddy
DIRECTOR

Dr. K. Srujan Raju
HOD

EXTERNAL EXAMINER

Submitted for viva voice Examination held on _____

ACKNOWLEGDEMENT

Apart from the efforts of us, the success of any project depends largely on the encouragement and guidelines of many others. We take this opportunity to express our gratitude to the people who have been instrumental in the successful completion of this project. We take this opportunity to express my profound gratitude and deep regard to my guide.

Mr. K. Murali, Associate Professor for his exemplary guidance, monitoring and constant encouragement throughout the project work. The blessing, help and guidance given by him shall carry us a long way in the journey of life on which we are about to embark.

We also take this opportunity to express a deep sense of gratitude to Project Review Committee (PRC) Coordinators: **Mr. J. Narasimha Rao, Mr. B. P. Deepak Kumar, Mr. K. Murali, Dr. Suwarna Gothane and Mr. B. Ramji** for their cordial support, valuable information, and guidance, which helped us in completing this task through various stages.

We are also thankful to **Dr. K. Srujan Raju,** Head, Department of Computer Science and Engineering for providing encouragement and support for completing this project successfully.

We are obliged to **Dr. A. Raji Reddy,** Director for being cooperative throughout the course of this project. We would like to express our sincere gratitude to Sri. **Ch. Gopal Reddy,** Chairman for providing excellent infrastructure and a nice atmosphere throughout the course of this project.

The guidance and support received from all the members of **CMR Technical Campus** who contributed to the completion of the project. We are grateful for their constant support and help.

Finally, we would like to take this opportunity to thank our family for their constant encouragement, without which this assignment would not be completed. We sincerely acknowledge and thank all those who gave support directly and indirectly in the completion of this project.

THUMMA MARY SHREEJA (177R1A05B6)

MORA SRUJANA (177R1A0594)

CHILLA SNEHA (177R1A0569)

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 Decision tree, SVM, Naive Bayes, Neural Network, CNN, and Random Tree.

LIST OF FIGURES

LIST OF SCREENSHOTS

TABLE OF CONTENTS

1. INTRODUCTION

1.INTRODUCTION

1.1 PROJECT SCOPE

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.

1.2 PROJECT PURPOSE

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.

1.3 PROJECT FEATURES

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.

2.LITERATURE SURVEY

2.LITERATURE SURVEY

S. K. Wasan, V. Bhatnagar, and H. Kaur, "The impact of data mining techniques on medical diagnostics," Data Science Journal, vol. 5, pp. 119-126, 2006.

- Medical data mining has great potential for exploring the hidden patterns in the data sets of the medical domain. These patterns can be utilized for clinical diagnosis.
- However, the available raw medical data are widely distributed, heterogeneous in nature, and voluminous.
- These data need to be collected in an organized form. This collected data can be then integrated to form a hospital information system. Data mining technology provides a user-oriented approach to novel and hidden patterns in the data.
- Data mining and statistics both strive towards discovering patterns and structures in data. Statistics deals with heterogeneous numbers only, whereas data mining deals with heterogeneous fields.
- We identify a few areas of healthcare where these techniques can be applied to healthcare databases for knowledge discovery. In this paper we briefly examine the impact of data mining techniques, including artificial neural networks, on medical diagnostics.

J. Wu, P. Zhang, L. Zhang, W. Meng, J. Li, C. Tong, et al., "Rapid and accurate identification of COVID-19 infection through machine learning based on clinical available blood test results," medRxiv, 2020.

- Since the sudden outbreak of coronavirus disease 2019 (COVID-19), it has rapidly evolved into a momentous global health concern. Due to the lack of constructive information on the pathogenesis of COVID-19 and specific treatment, it highlights the importance of early diagnosis and timely treatment.
- In this study, 11 key blood indices were extracted through random forest algorithm to build the final assistant discrimination tool from 49 clinical available blood test data which were derived by commercial blood test equipment’s.
- The method presented robust outcome to accurately identify COVID-19 from a variety of suspected patients with similar CT information or similar symptoms, with accuracy of 0.9795 and 0.9697 for the cross-validation set and test set, respectively.

- The tool also demonstrated its outstanding performance on an external validation set that was completely independent of the modeling process, with sensitivity, specificity, and overall accuracy of 0.9512, 0.9697, and 0.9595, respectively. Besides, 24 samples from overseas infected patients with COVID-19 were used to make an in-depth clinical assessment with accuracy of 0.9167.
- After multiple verification, the reliability and repeatability of the tool has been fully evaluated, and it has the potential to develop into an emerging technology to identify COVID-19 and lower the burden of global public health.
- The proposed tool is well-suited to carry out preliminary assessment of suspected patients and help them to get timely treatment and quarantine suggestion.

S. Kadry, V. Rajinikanth, S. Rho, N. S. M. Raja, V. S. Rao, and K. P. Thanaraj, "Development of a Machine-Learning System to Classify Lung CT Scan Images into Normal/COVID-19 Class," arXiv preprint arXiv:2004.13122, 2020.

- Recently, the lung infection due to Coronavirus Disease (COVID-19) affected a large human group worldwide and the assessment of the infection rate in the lung is essential for treatment planning.
- This research aims to propose a Machine-Learning-System (MLS) to detect the COVID-19 infection using the CT scan Slices (CTS).
- This MLS implements a sequence of methods, such as multi-thresholding, image separation using threshold filter, feature-extraction, feature-selection, feature-fusion, and classification. The initial part implements the Chaotic-Bat-Algorithm and Kapur's Entropy (CBA+KE) thresholding to enhance the CTS.
- The threshold filter separates the image into two segments based on a chosen threshold 'Th'. The texture features of these images are extracted, refined and selected using the chosen procedures. Finally, a two-class classifier system is implemented to categorize the chosen CTS (n=500 with a pixel dimension of 512x512x1) into normal/COVID-19 group. In this work, the classifiers, such as Naive Bayes (NB), k-Nearest Neighbors (KNN), Decision Tree (DT), Random Forest (RF) and Support Vector Machine with linear kernel (SVM) are implemented and the classification task is performed using various feature vectors. The experimental outcome of the SVM with Fused-Feature-Vector (FFV) helped to attain a detection accuracy of 89.80%.

T. Singhal, "A review of coronavirus disease-2019 (COVID-19)," The Indian Journal of Pediatrics, pp. 1-6, 2020.

- There is a new public health crisis threatening the world with the emergence and spread of 2019 novel coronavirus (2019-nCoV) or the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2).
- The virus originated in bats and was transmitted to humans through yet unknown intermediary animals in Wuhan, Hubei province, China in December 2019. There have been around 96,000 reported cases of coronavirus disease 2019 (COVID-2019) and 3300 reported deaths to date (05/03/2020).
- The disease is transmitted by inhalation or contact with infected droplets and the incubation period ranges from 2 to 14 d.
- The symptoms are usually fever, cough, sore throat, breathlessness, fatigue, malaise among others.
- The disease is mild in most people; in some (usually the elderly and those with comorbidities), it may progress to pneumonia, acute respiratory distress syndrome (ARDS) and multi organ dysfunction. Many people are asymptomatic.
- The case fatality rate is estimated to range from 2 to 3%. Diagnosis is by demonstration of the virus in respiratory secretions by special molecular tests. Common laboratory findings include normal/ low white cell counts with elevated C-reactive protein (CRP). The computerized tomographic chest scan is usually abnormal even in those with no symptoms or mild disease.
- Treatment is essentially supportive; role of antiviral agents is yet to be established. Prevention entails home isolation of suspected cases and those with mild illnesses and strict infection control measures at hospitals that include contact and droplet precautions. The virus spreads faster than its two ancestors the SARS-CoV and Middle East respiratory syndrome coronavirus (MERS-CoV) but has lower fatality. The global impact of this new epidemic is yet uncertain.

3.SYSTEM ANALYSIS

3.SYSTEM ANALYSIS

SYSTEM ANALYSIS

System Analysis is the important phase in the system development process. The System is studied to the minute details and analyzed. The system analyst plays an important role of an interrogator and dwells deep into the working of the present system. In analysis, a detailed study of these operations performed by the system and their relationships within and outside the system is done. A key question considered here is, “what must be done to solve the problem?” The system is viewed as a whole and the inputs to the system are identified. Once analysis is completed the analyst has a firm understanding of what is to be done.

3.1 PROBLEM DEFINITION

A detailed study of the process must be made by various techniques like Image processing, feature recognition etc. The data collected by these sources must be scrutinized to arrive to a conclusion. The conclusion is an understanding of how the system functions. This system is called the existing system. Now the existing system is subjected to close study and problem areas are identified. The designer now functions as a problem solver and tries to sort out the difficulties that the enterprise faces. The solutions are given as proposals. The proposal is then weighed with the existing system analytically and the best one is selected. The proposal is presented to the user for an endorsement by the user. The proposal is reviewed on user request and suitable changes are made. This is loop that ends as soon as the user is satisfied with proposal.

3.2 EXISTING SYSTEM

Solutions are always made in a hospital based on intuition and experience of doctors, and not on the rich knowledge data that are hidden in the database. Most hospitals today use decision-support systems, but to get the results of the disease are largely limited. They can answer simple questions such as "clustering the patient ages by any disease", "Prediction of diabetes".

3.3 PROPOSED SYSTEM

Globally Covid-19 pandemic is now one of the major issue for all countries. Peoples in every country across the world are now going through this disaster. Infection spread and death rate ratio is also growing rapidly all over the world. The main challenge for developing this automated

diagnosis system is the proper analysis of the data and accuracy. Numerous machine learning techniques need to use for getting a satisfactory result in various biomedical data analysis. 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. Algorithm's performance is measured in terms of accuracy, precision, and recall.

3.4 FUNCTIONAL REQUIREMENTS

- Admin Login
- Admin Dataset upload
- Admin View Dataset
- Training the Dataset
- Testing Dataset
- User Signup
- Enter Details
- View Results

3.5 NON-FUNCTIONAL REQUIREMENTS

Usability:

Prioritize the important functions of the system based on usage patterns.

Frequently used functions should be tested for usability, as should complex and critical functions. Be sure to create a requirement for this.

Reliability:

Reliability defines the trust in the system that is developed after using it for a period. It defines the likeability of the software to work without failure for a given period. The number of bugs in the code, hardware failures, and problems can reduce the reliability of the software. Your goal should be a long MTBF (mean time between failures). It is defined as the average period the system runs before failing. Create a requirement that data created in the system will be retained for several years without the data being changed by the system. It is a good idea to also include requirements that make it easier to monitor system performance.

Performance:

What should system response times be, as measured from any point, under what circumstances? Are there specific peak times when the load on the system will be unusually high? Think of stress periods, for example, at the end of the month or in conjunction with payroll disbursement.

Supportability:

The system needs to be cost-effective to maintain.

Maintainability requirements may cover diverse levels of documentation, such as system documentation, as well as test documentation, e.g., which test cases and test plans will accompany the system.

3.6 FEASIBILITY STUDY

A feasibility analysis evaluates the project's potential for success; therefore, perceived objectivity is an essential factor in the credibility of the study for potential investors and lending institutions.

3.6.1 ECONOMIC FEASIBILITY

Cost/ benefits analysis of the project as over project is academic project, we will not have only basic cost for learning of the technologies.

3.6.2 TECHNICAL FEASIBILITY

Technical resources need for project Development.

- Windows family Operating System
- Python 3.6 Technology
- PyCharm IDE

3.6.3 OPERATIONAL FEASIBILITY

This assessment involves undertaking a study to analyze and determine whether and how well the organization's needs can be met by completing the project. Operational feasibility studies also examine how a project plan satisfies the requirements identified in the requirements analysis phase of system development.

3.7 HARDWARE & SOFTWARE REQUIREMENTS

3.7.1 HARDWARE REQUIREMENTS:

Hardware interfaces specifies the logical characteristics of each interface between the software product and the hardware components of the system. The following are some hardware requirements.

RAM	4 GB Minimum
Processor	i3 Minimum
Hard disk	500 GB

3.7.2 SOFTWARE REQUIREMENTS:

Software Requirements specifies the logical characteristics of each interface and software components of the system. The following are some software requirements.

Technology	Python 3.6
Operating System	Windows Family
IDE	VS Code
Technology	Python, Django
Database Server	PostgreSQL
Front Design Technology	HTML, CSS, JS

4. ARCHITECTURE

4. ARCHITECTURE

4.1 PROJECT ARCITECTURE

This project architecture shows the procedure followed for predicting corona virus using machine learning algorithms.

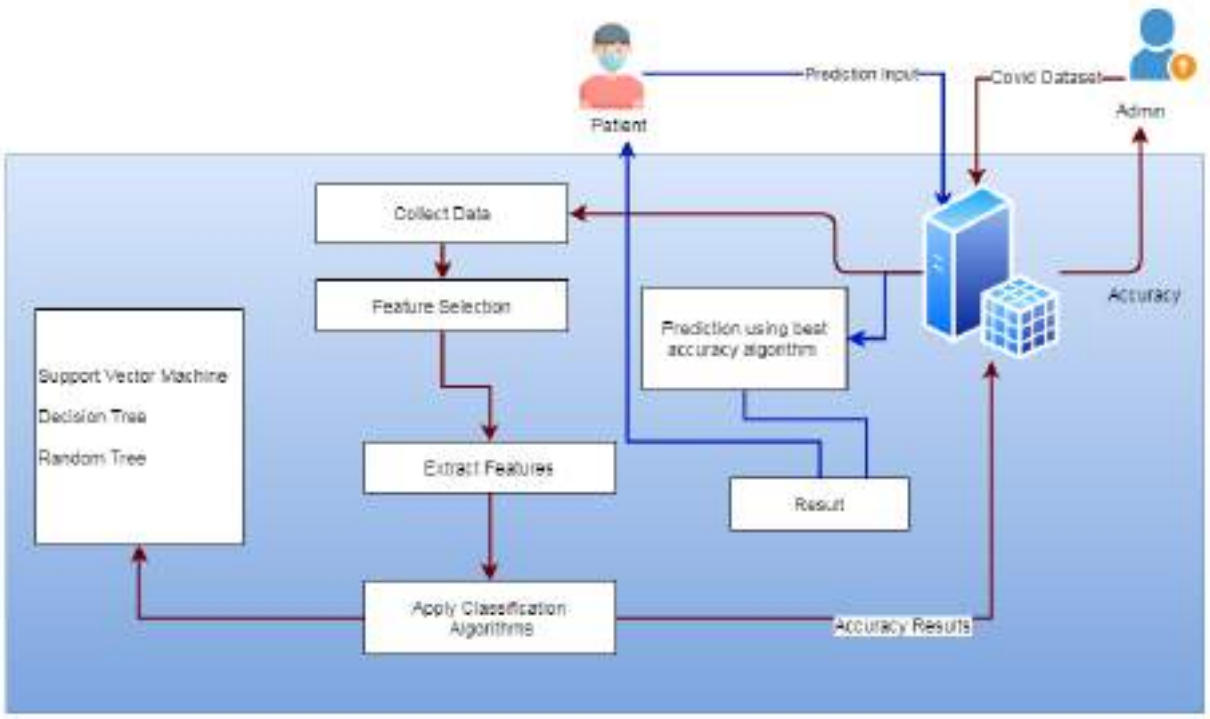


Figure 4.1: Project Architecture of Corona Virus Infection Probability Classification.

4.2 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 Tree.

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.

4.3 UML DIAGRAMS

The System Design Document describes the system requirements, operating environment, system and subsystem architecture, files and database design, input formats, output layouts, human-machine interfaces, detailed design, processing logic, and external interfaces.

4.3.1 USE CASE DIAGRAM

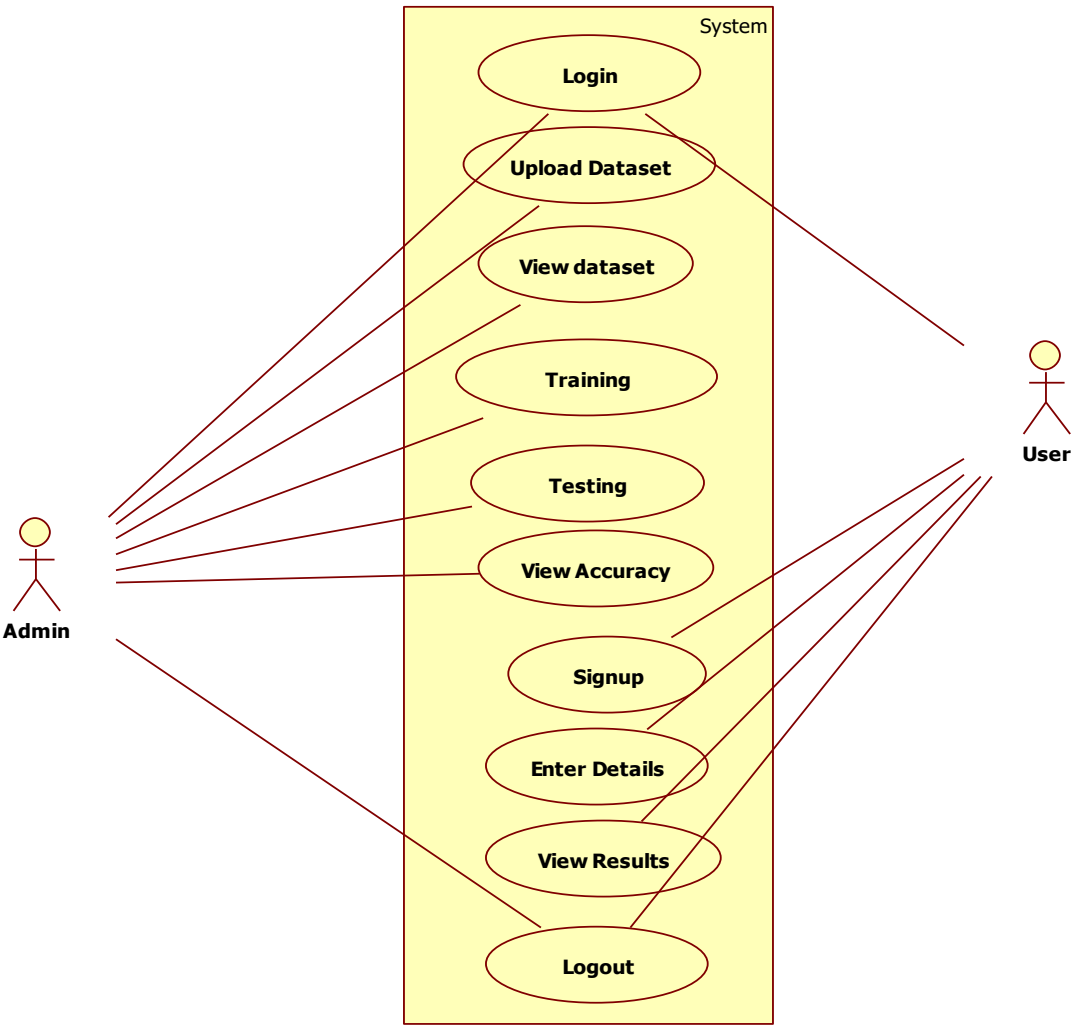


Figure 4.3.1: Use Case Diagram for admin and user

4.3.2 COLLABORATION DIAGRAM

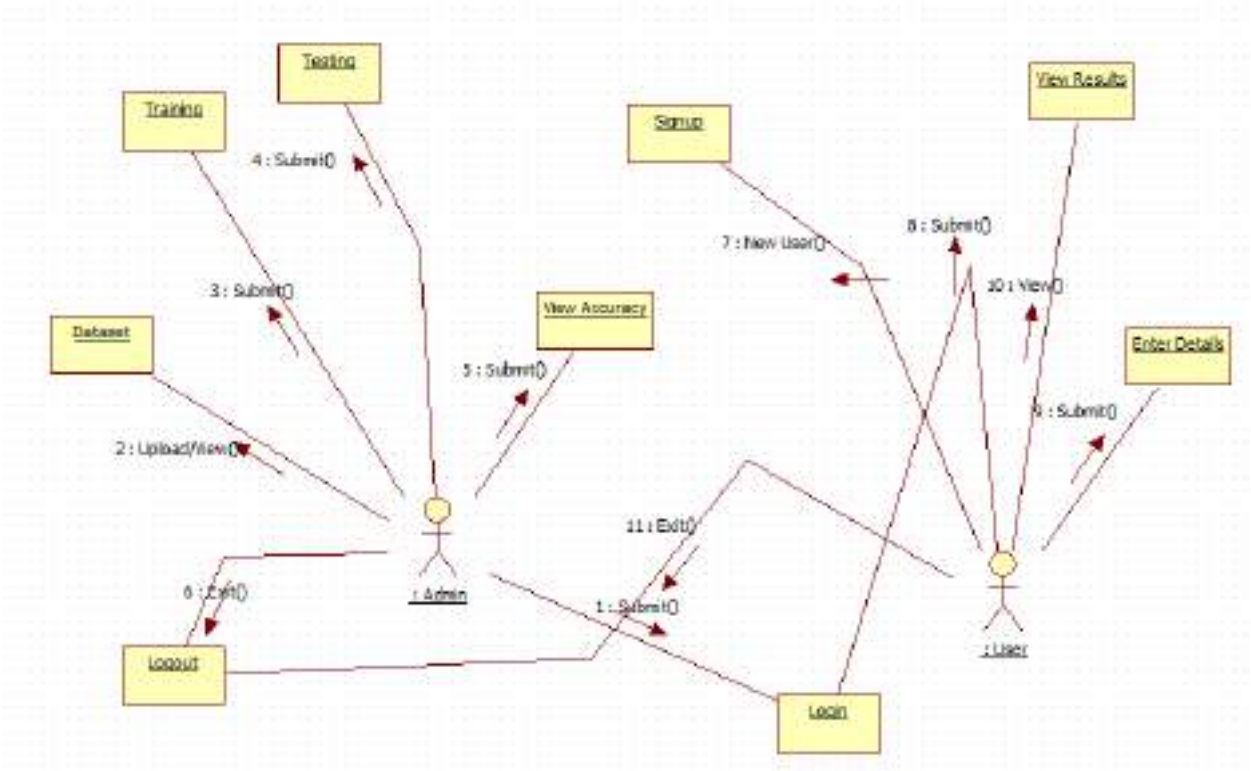


Figure 4.3.2: Coloration Diagram

4.3.3 SEQUENCE DIAGRAM

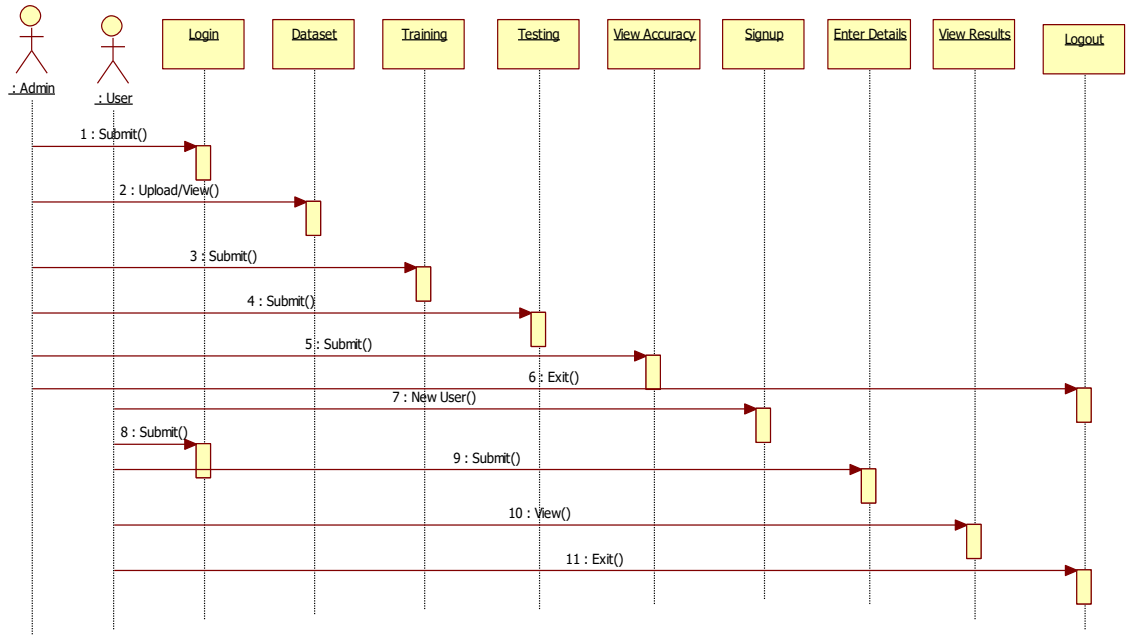


Figure 4.3.3: Sequence Diagram for Corona Virus Infection Probability Classification.

4.3.4. ACTIVITY DIAGRAM

It describes about flow of activity states.

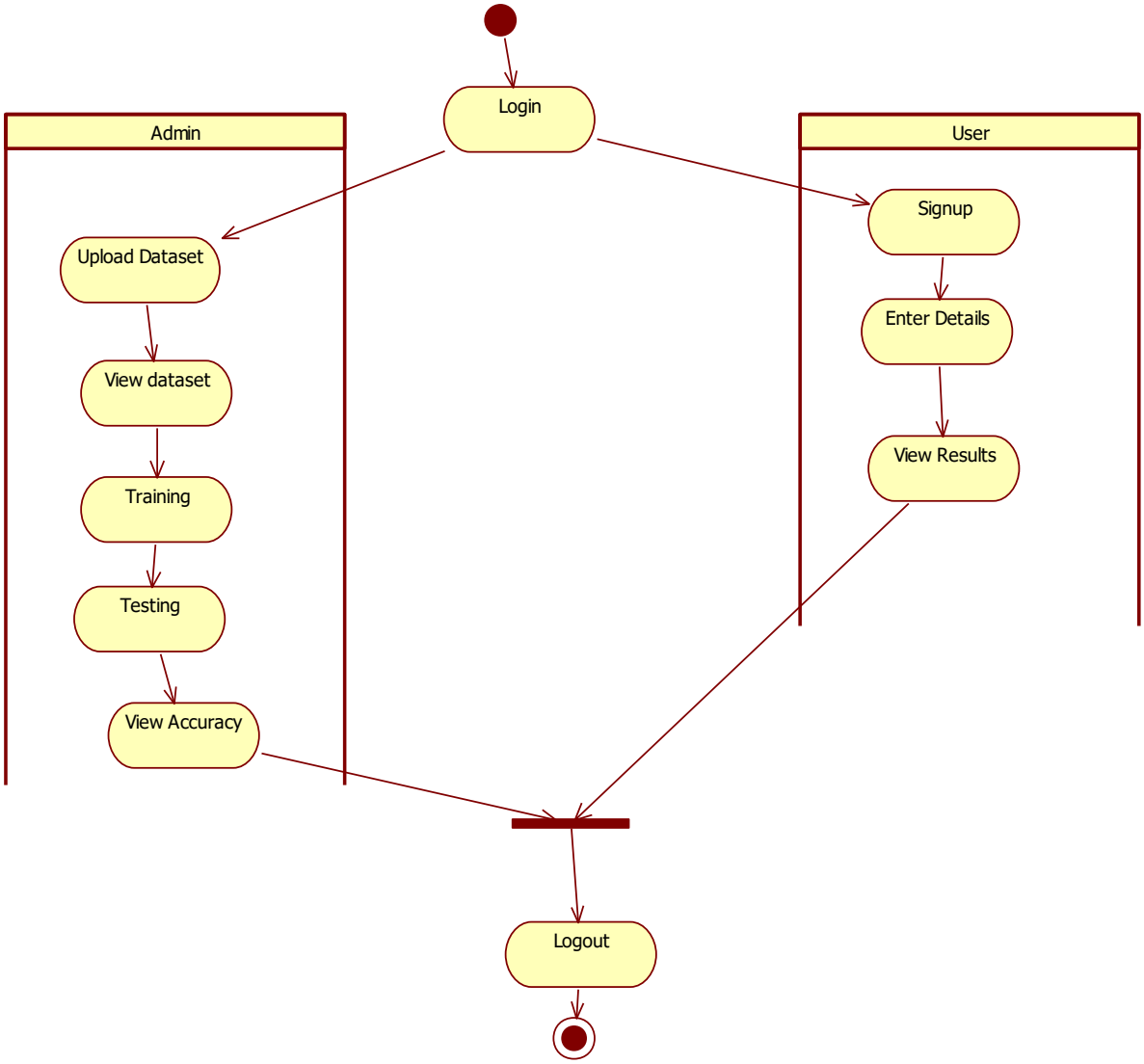


Figure 4.3.4: Activity Diagram for Admin and User for Corona Virus Infection Probability Classification

5.IMPLEMENTATION

5.IMPLEMENTATION

5.1 SAMPLE CODE

DBconn.py:

```
import mysql.connector
import sys
class DBConnection:
    @staticmethod
    def getConnection():
        try:
            database = mysql.connector.connect(host="localhost", user="root", password="root", db='covid')
            return database
        except Exception as e:
            print("Error=" + e.args[0])
            tb = sys.exc_info()[2]
            print(tb.tb_lineno)
if __name__=="__main__":
    print(DBConnection.getConnection())
```

5.2 ALGORITHMS

5.2.1 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.

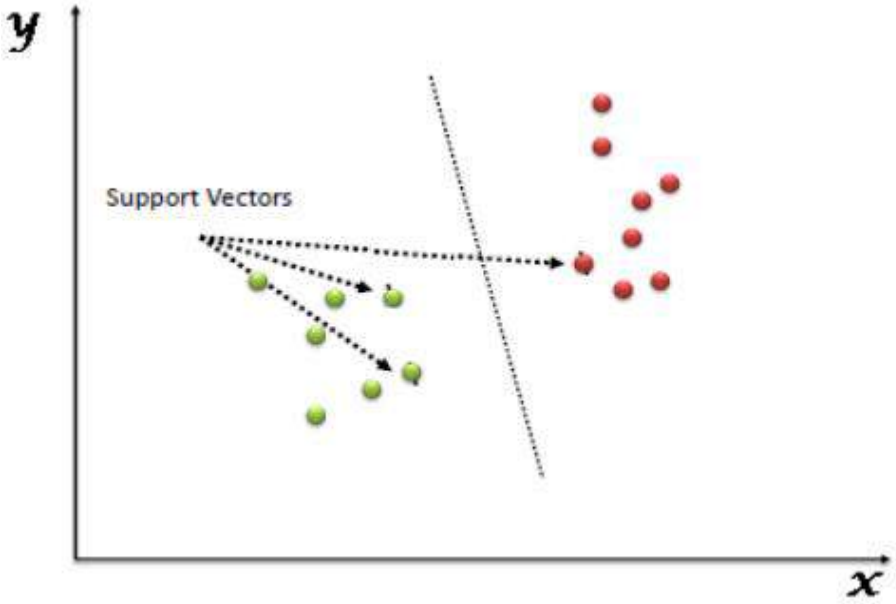


Figure 5.2.1: SVM

5.2.2 DECISION TREE:

Decision Tree 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).

Figure 5.2.2: Decision Tree

5.2.3 NAIVE BAYES ALGORITHM

Naive Bayes 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 naive bayes classification.

Bayes theorem: -

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

- 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.

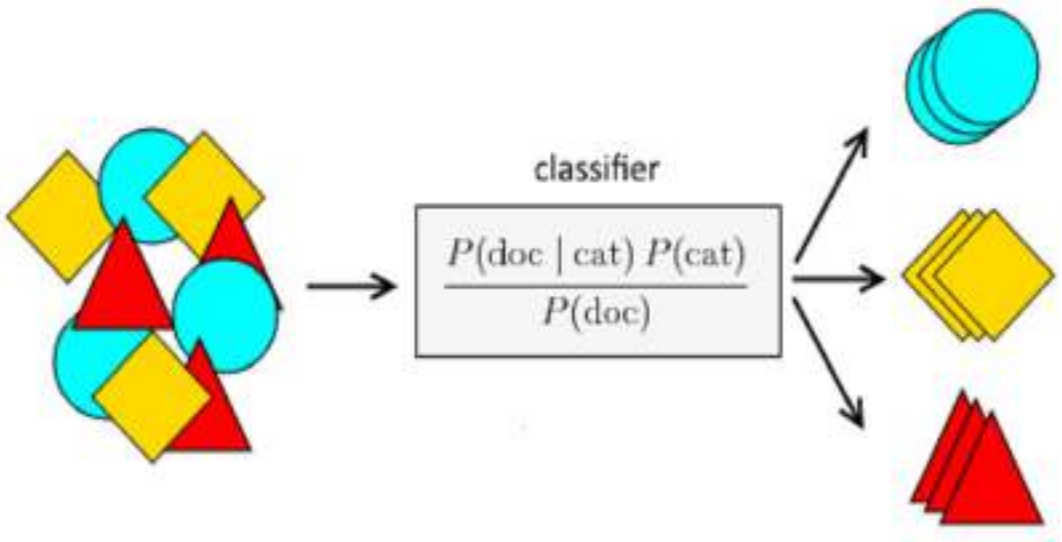


Figure 5.2.3: Naive Bayes

5.2.4 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.

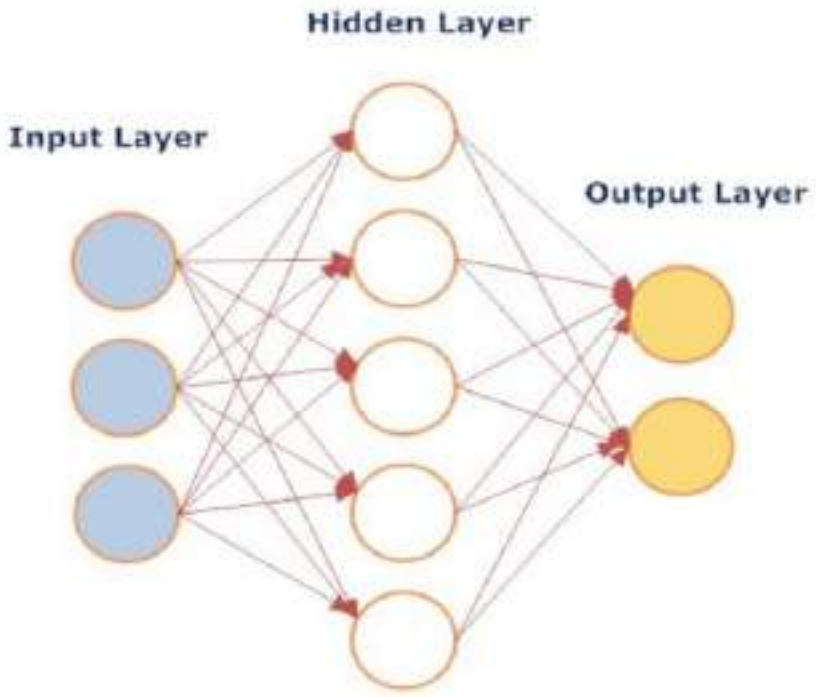


Figure 5.2.4: Neural Network

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.

5.2.5 CONVOLUTIONAL NEURAL NETWORK (CNN)

In deep learning, a convolutional neural network (CNN, or ConvNet) is a class of deep neural network, most commonly applied to analyze visual imagery. They are also known as shift invariant or space invariant artificial neural networks (SIANN), based on the shared-weight architecture of 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.

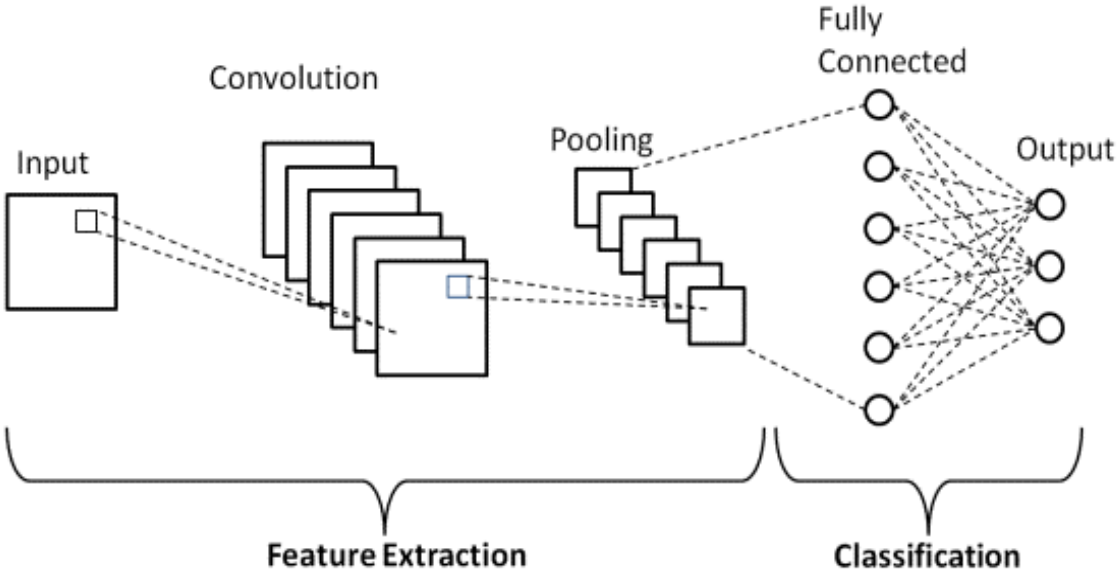


Figure 5.2.5: CNN

5.2.6 RANDOM FOREST

In our experiment, we use random forest as a classifier. The popularity of decision tree models in data mining is owed to their simplification in algorithm and flexibility in handling different data attribute types. However, single-tree model is possibly sensitive to specific training data and easy to overfit. Ensemble methods can solve these problems by combine a group of individual decisions in some way and are more accurate than single classifiers. Random forest, 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. The capacity of random forest not only depends on the strength of individual tree but also the correlation between different trees. The stronger the strength of single tree and the less the correlation of different tress, the better the performance of random forest. The variation of trees comes from their randomness which involves bootstrapped samples and randomly selects a subset of data attributes.

Below is the step-by-step Python implementation. ...

Step 2: Import and print the dataset.

Step 3: Select all rows and column 1 from dataset to x and all rows and column 2 as y.

Step 4: Fit Random forest regressor to the dataset.

Step 5: Predicting a new result.

Step 6: Visualizing the result.

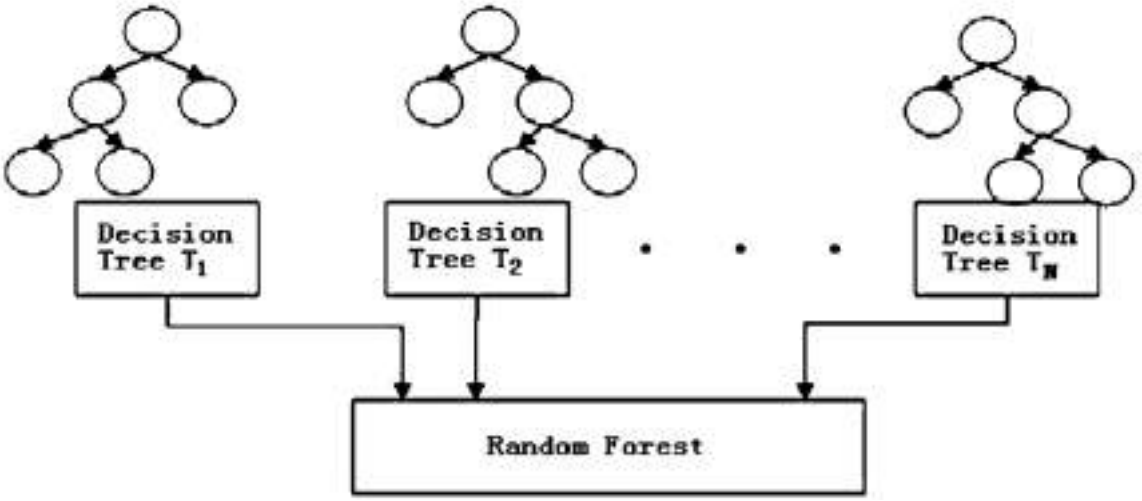
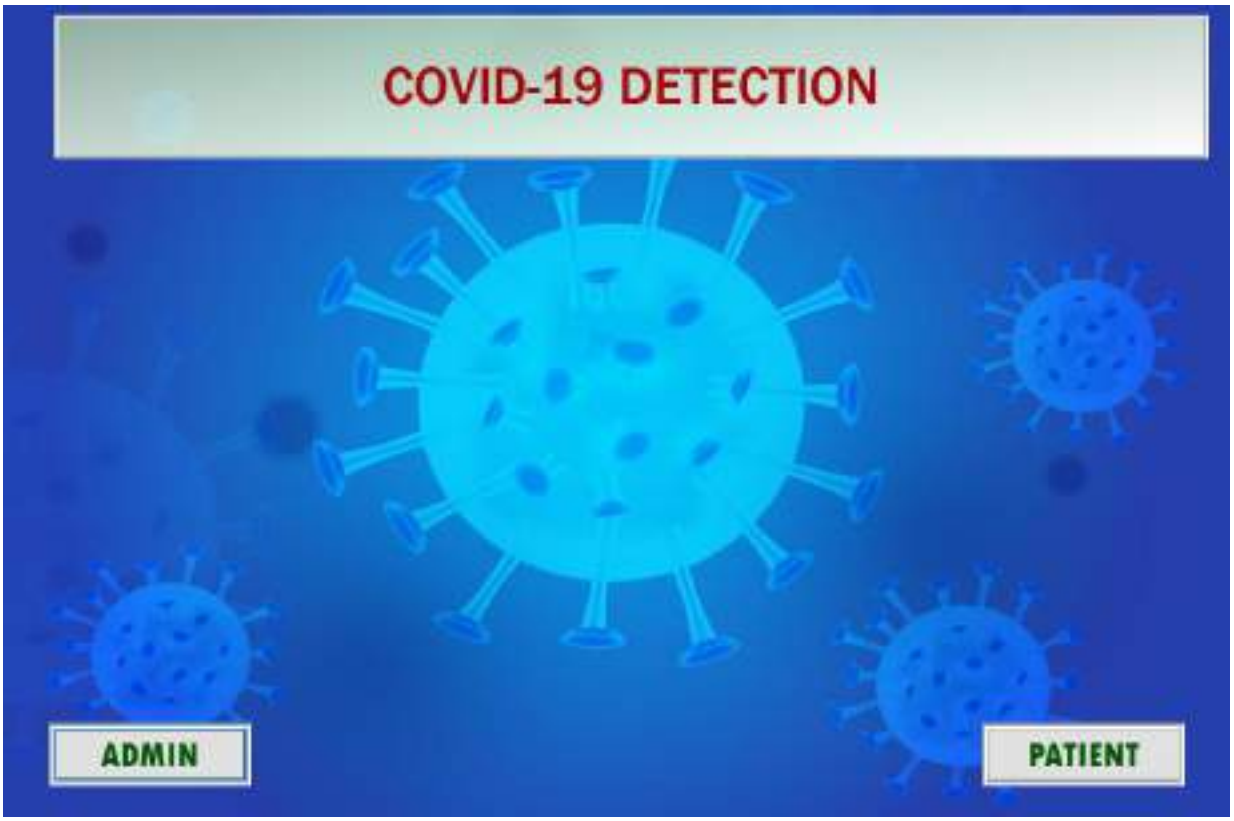


Figure 5.2.6: Random Forest

6.SCREENSHOTS

6.SCREENSHOTS

6.1 HOME PAGE RESULT



Screenshot 6.1: Home page result for covid probability.

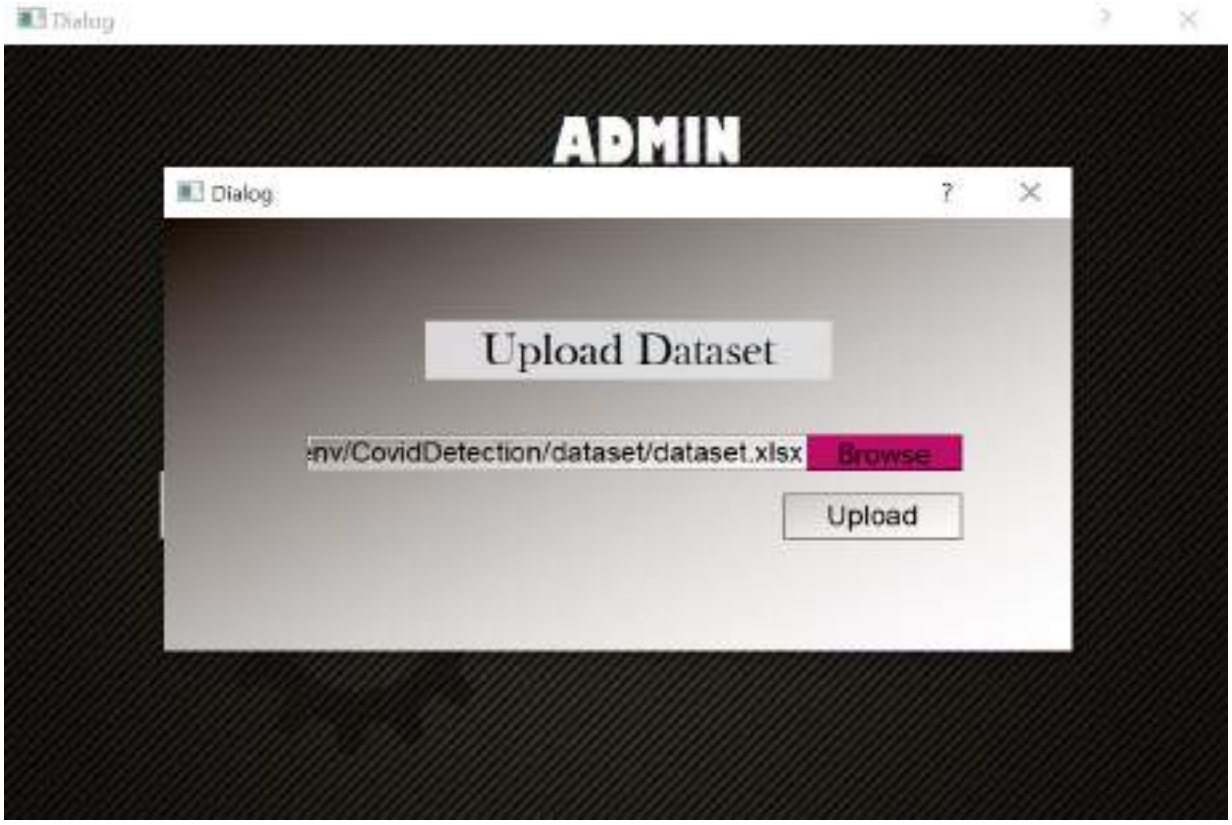
6.2 ADMIN LOGIN PAGE RESULT

Screenshot 6.2: Admin login page result of covid probability.

6.3 ADMIN DASHBOARD RESULT

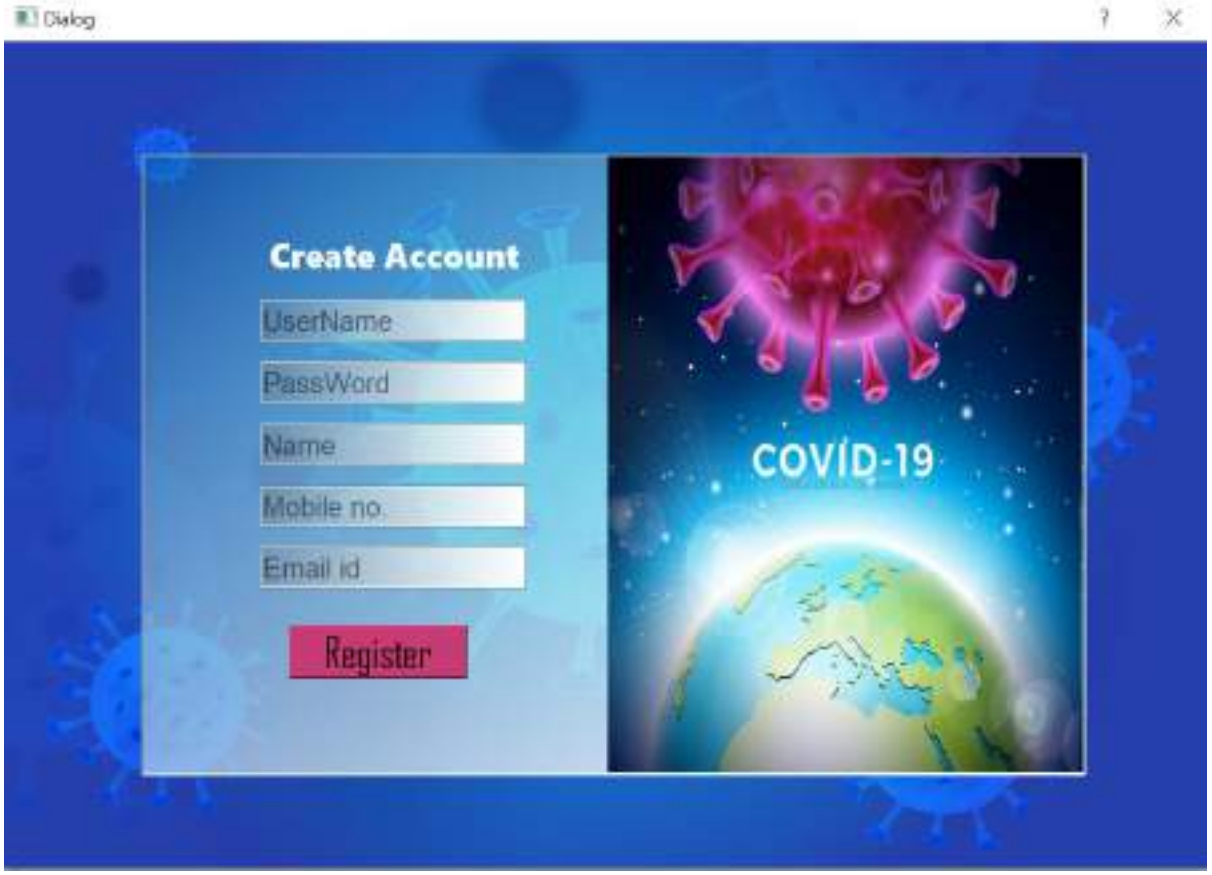
Screenshot 6.3: Admin Dashboard Result

6.4 DATASET UPLOAD



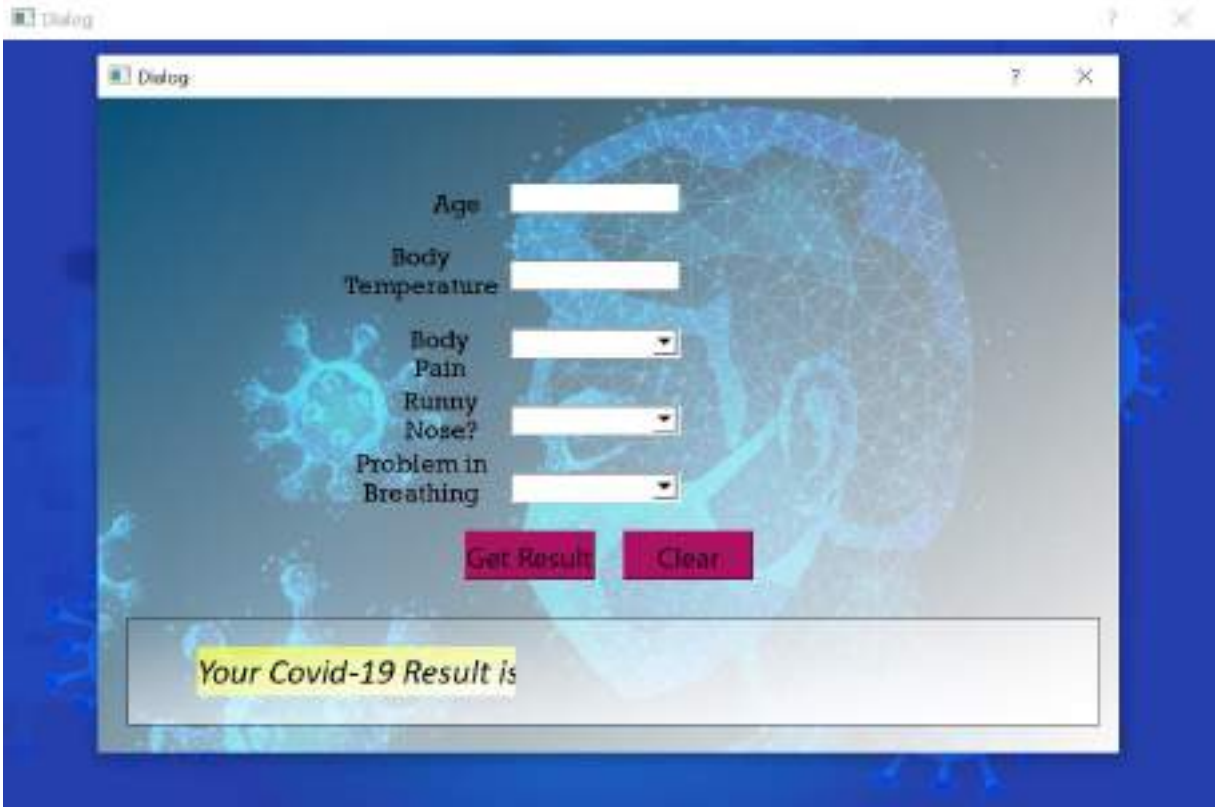
Screenshot 6.4: Dataset Upload

6.5 USER REGISTRATION



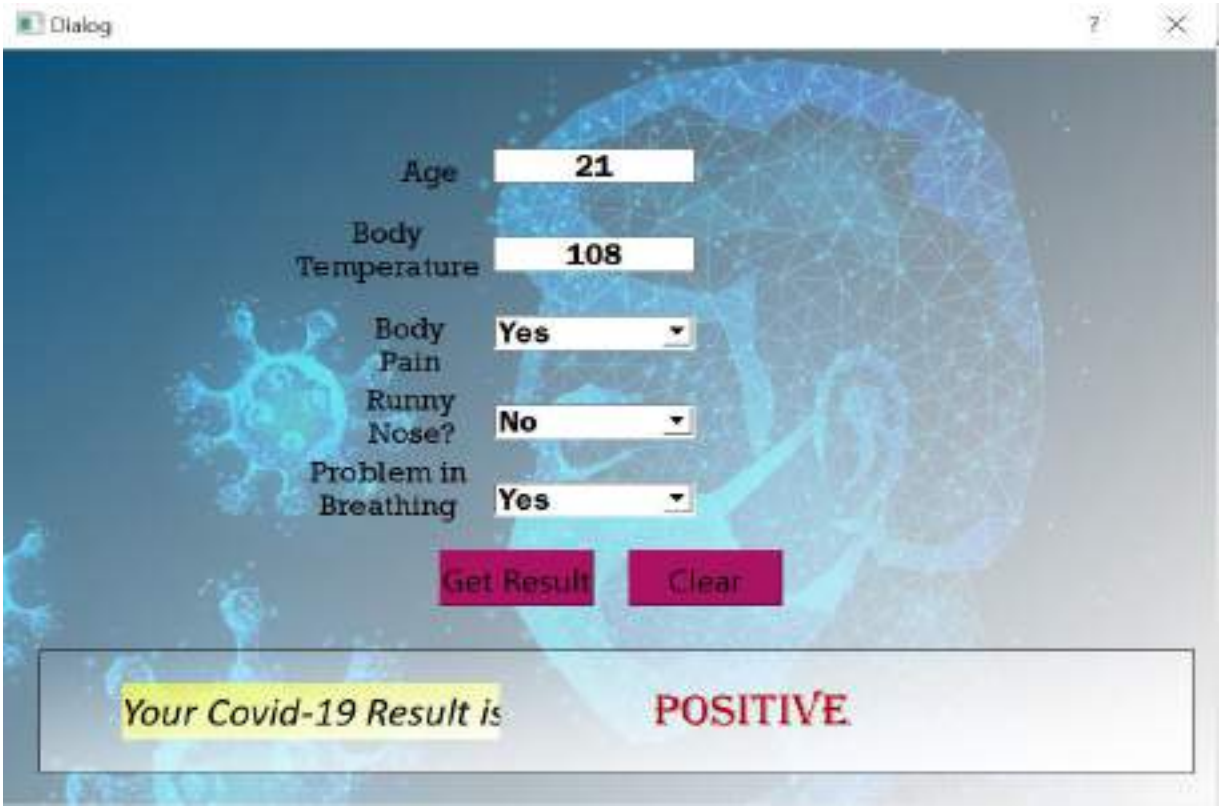
Screenshot 6.5: User Registration Page.

6.6 TEST SCREEN



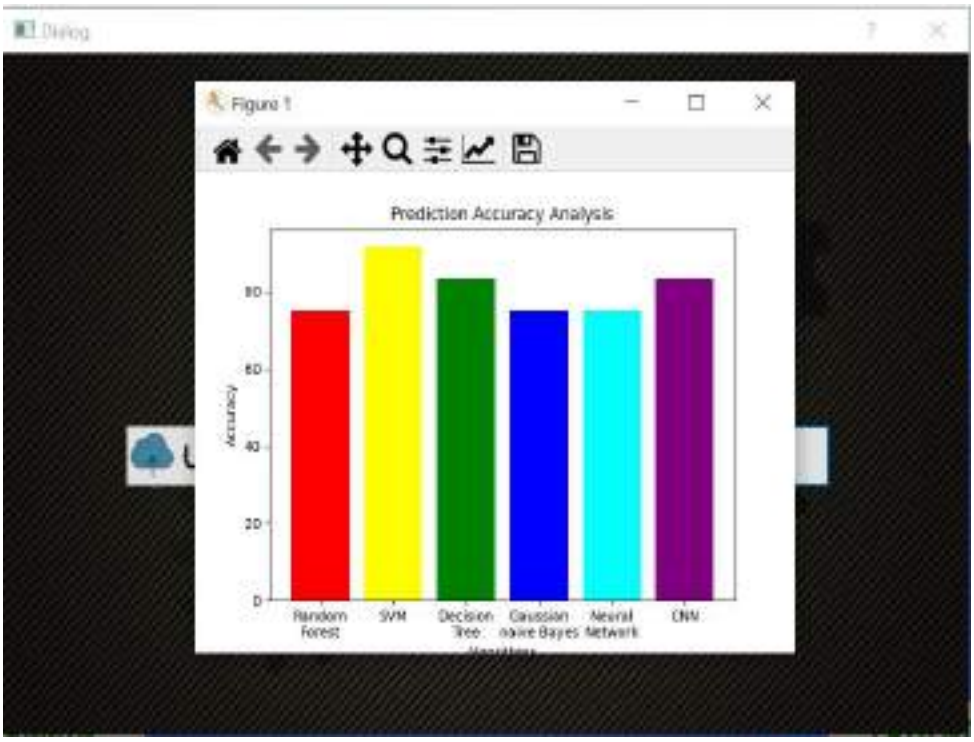
Screenshot 6.6: Test Screen.

6.7 RESULT SCREEN



Screenshot 6.7: Result Screen.

6.8 PREDICTION



Screenshot 6.8: Prediction Accuracy Analysis.

6.9 PERFORMANCE

Algorithm	Accuracy	precision	recall	Fscore
RFC	70.0	0.35	0.5	0.4117647058823529
SVM	88.0	0.8888888888888888	0.6666666666666666	0.6875
DTC	70.0	0.625	0.5952380952380952	0.6
GNB	80.0	0.8888888888888888	0.6666666666666666	0.6875
MN	70.0	0.35	0.5	0.4117647058823529
CNN	60.0	0.75	0.6666666666666666	0.5833333333333333
(NULL)	(NULL)	(NULL)	(NULL)	(NULL)

Screenshot 6.9: Performance of algorithms.

7. TESTING

7.TESTING

7.1 INTRODUCTION TO TESTING

The purpose of testing is to discover errors. Testing is the process of trying to discover every conceivable fault or weakness in a work product. It provides a way to check the functionality of components, subassemblies, assemblies and/or a finished product. It is the process of exercising software with the intent of ensuring that the Software system meets its requirements and user expectations and does not fail in an unacceptable manner. There are various types of test. Each test type addresses a specific testing requirement.

7.2 TYPES OF TESTING

7.2.1 UNIT TESTING

In the unit testing we test each module individually and integrate with the overall system. Unit testing focuses verification efforts on the smallest unit of software design in the module. This is also known as module testing. The module of the system is tested separately. This testing is carried out during programming stage itself. In the testing step each module is found to work satisfactorily as regard to expected output from the module. There are some validation checks for fields also. For example, the validation check is done for varying the user input given by the user which validity of the data entered. It is very easy to find error debut the system.

7.2.2 INTEGRATION TESTING

Data can be lost across an interface, one module can have an adverse effort on the other sub functions, when combined, may not produce the desired major functions. Integrated testing is the systematic testing for constructing the uncover errors within the interface. The testing was done with sample data. The developed system has run successfully for this sample data. The need for integrated test is to find the overall system performance.

7.2.3 FUNCTIONAL TESTING

Functional tests provide systematic demonstrations that functions tested are available as specified by the business and technical requirements, system documentation, and user manuals.

Functional testing is centered on the following items:

Valid Input : identified classes of valid input must be accepted.

Invalid Input : identified classes of invalid input must be rejected.

Functions : identified functions must be exercised.

Output : identified classes of application outputs must be exercised.

Systems/Procedures : interfacing systems or procedures must be invoked.

Organization and preparation of functional tests is focused on requirements, key functions, or special test cases. In addition, systematic coverage pertaining to identify Business process flows, data fields, predefined processes.

7.3 TEST CASES

SNO	Test case	Results
1	Data Set Upload	Yes
2	Prediction	Yes

8.CONCLUSION

8.CONCLUSION

The main objective of our work is to develop an automatic infection probability classification system using machine learning technique. The model is classifying the chance of infection probability of the persons having different COVID-19 symptoms. A Support Vector Machine 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. Support Vector Machine performance is measured and from the result it is observed that Support Vector Machine is performing better as compared to other five. The performance is measured in terms of accuracy, precision, and recall. Around 90% classification accuracy is obtained by using Support Vector Machine. In future some more accuracy can be obtained by modifying the Support Vector Machine as well as with some other machine learning techniques.

GitHub Link

https://github.com/MaryShreeja/Corona-Virus-Infection-Probabilty-Classification

9.BIBILOGRAPHY

9.BIBILOGRAPHY

9.1 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.

9.2 WEBSITES

- https://scholar.google.co.in/scholar?q=covid+prediction+using+machine+learning&hl=en&as_sdt=0&as_vis=1&oi=scholart
- https://www.sciencedirect.com/science/article/pii/S2211379721000012
- https://www.nature.com/articles/s41746020003706#:~:text=We%20established%20a%20machine%2Dlearning,to%20have%20COVID%2D19).

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.

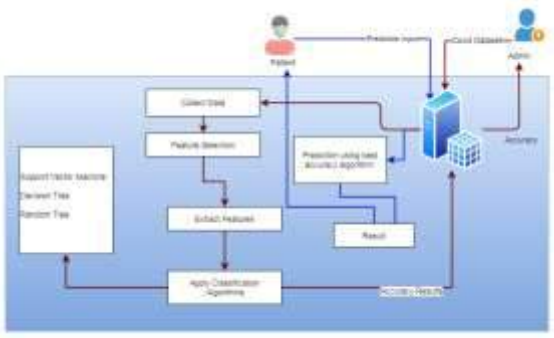


Fig. 1: Block diagram

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.

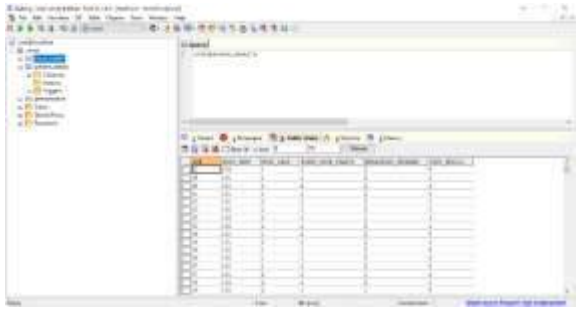


Fig. 2: SQLyog Database

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.

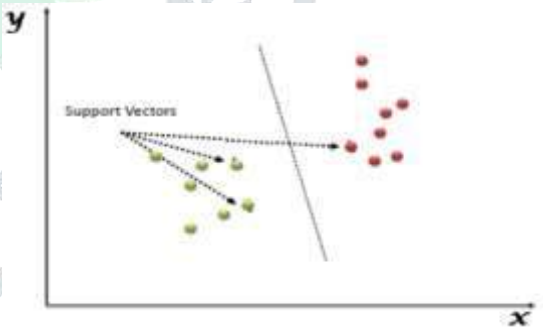


Fig. 3. SVM Algorithm Diagrammatic Representation

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.

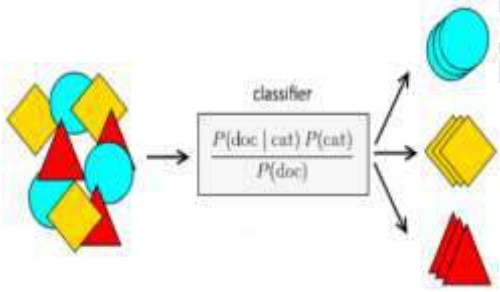


Fig. 5: Naive bayes Algorithm Diagrammatic Representation.

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.

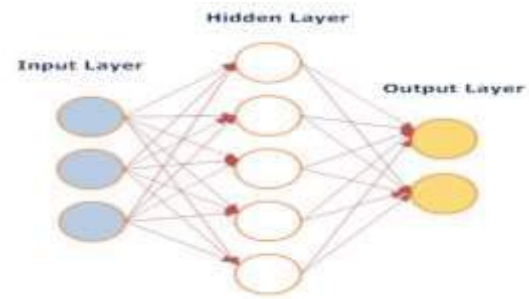


Fig. 6: Neural network Algorithm Diagrammatic Representation.

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.

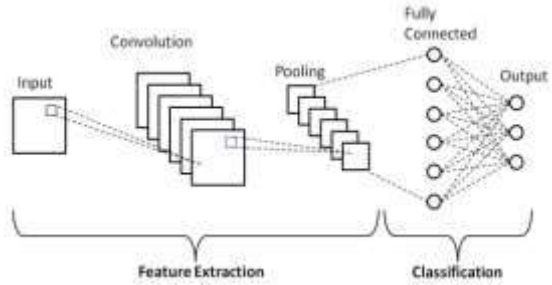


Fig. 7: CNN Algorithm Diagrammatic Representation.

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.

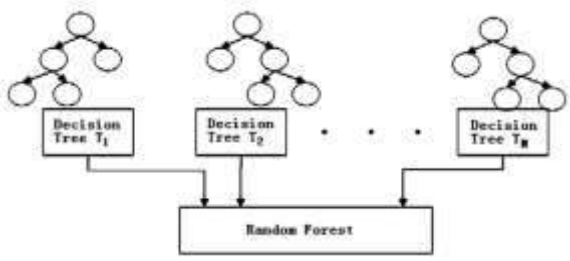


Fig. 7: Random forest classifier Diagrammatic Representation.

4. Results and Discussion

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

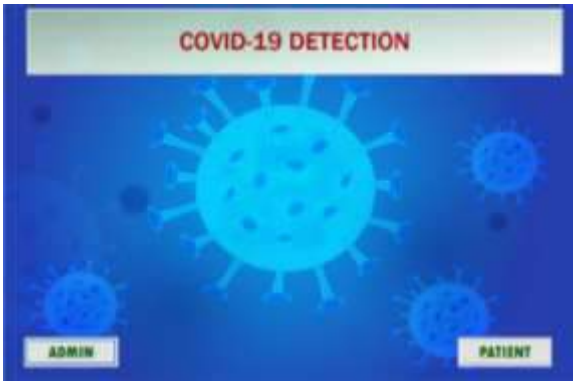


Fig. 8: Home page result for covid probability.

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.

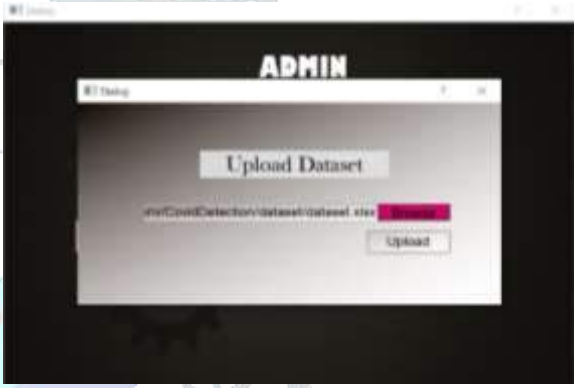


Fig. 12: Uploading dataset page.

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

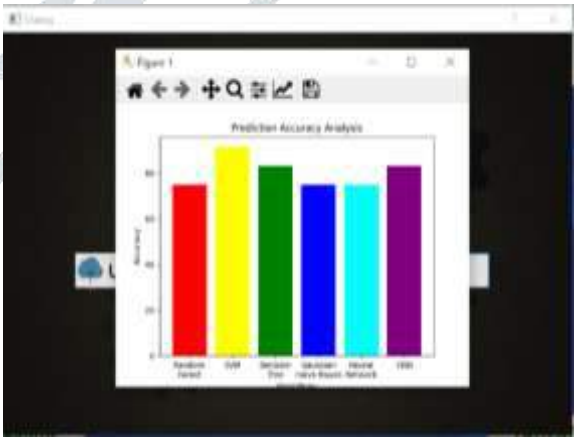


Fig. 13: Prediction accuracy analysis page.

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

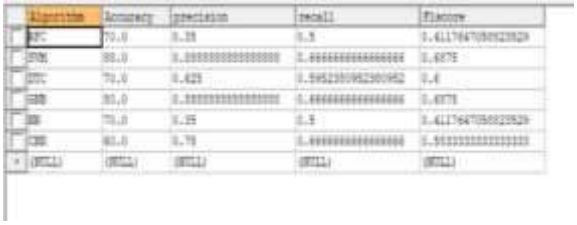


Fig. 14: Performance of algorithms.

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.

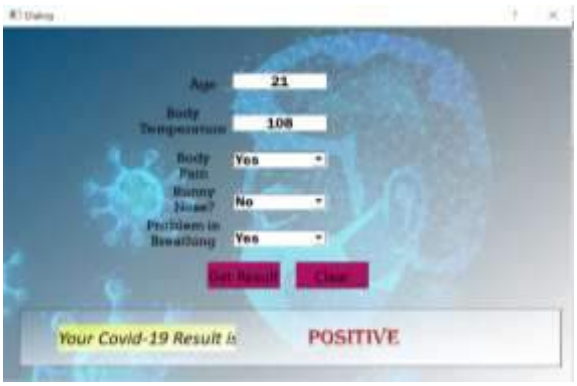


Fig. 18: Result page.

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.

Journal of Emerging Technologies and Innovative Research
An International Open Access Journal Peer-reviewed, Refereed Journal
www.jetir.org | editor@jetir.org An International Scholarly Indexed Journal

Certificate of Publication

The Board of

Journal of Emerging Technologies and Innovative Research (ISSN : 2349-5162)

Is hereby awarding this certificate to

Thumma Mary Shreeja

In recognition of the publication of the paper entitled

CORONA VIRUS INFECTION PROBABILITY CLASSIFICATION USING MACHINE LEARNING ALGORITHMS.

Published In JETIR (www.jetir.org) ISSN UGC Approved (Journal No: 63975) & 7.95 Impact Factor

Published in Volume 8 Issue 6 , June-2021 | Date of Publication: 2021-06-12

Parisa P
EDITOR

EDITOR IN CHIEF

JETIR2106221 Research Paper Weblink http://www.jetir.org/view?paper=JETIR2106221 Registration ID : 310536

Journal of Emerging Technologies and Innovative Research
An International Open Access Journal Peer-reviewed, Refereed Journal
www.jetir.org | editor@jetir.org An International Scholarly Indexed Journal

Certificate of Publication

The Board of

Journal of Emerging Technologies and Innovative Research (ISSN : 2349-5162)

Is hereby awarding this certificate to

Mora Srujana

In recognition of the publication of the paper entitled

CORONA VIRUS INFECTION PROBABILITY CLASSIFICATION USING MACHINE LEARNING ALGORITHMS.

Published In JETIR (www.jetir.org) ISSN UGC Approved (Journal No: 63975) & 7.95 Impact Factor

Published in Volume 8 Issue 6 , June-2021 | Date of Publication: 2021-06-12

Parisa P

EDITOR

EDITOR IN CHIEF

JETIR2106221 Research Paper Weblink http://www.jetir.org/view?paper=JETIR2106221 Registration ID : 310536

Journal of Emerging Technologies and Innovative Research
An International Open Access Journal Peer-reviewed, Refereed Journal
www.jetir.org | editor@jetir.org An International Scholarly Indexed Journal

Certificate of Publication

The Board of

Journal of Emerging Technologies and Innovative Research (ISSN : 2349-5162)

Is hereby awarding this certificate to

Chilla Sneha

In recognition of the publication of the paper entitled

CORONA VIRUS INFECTION PROBABILITY CLASSIFICATION USING MACHINE LEARNING ALGORITHMS.

Published In JETIR (www.jetir.org) ISSN UGC Approved (Journal No: 63975) & 7.95 Impact Factor

Published in Volume 8 Issue 6 , June-2021 | Date of Publication: 2021-06-12

Parisa P
EDITOR

EDITOR IN CHIEF

JETIR2106221 Research Paper Weblink http://www.jetir.org/view?paper=JETIR2106221 Registration ID : 310536

Journal of Emerging Technologies and Innovative Research
An International Open Access Journal Peer-reviewed, Refereed Journal
www.jetir.org | editor@jetir.org An International Scholarly Indexed Journal

Certificate of Publication

The Board of

Journal of Emerging Technologies and Innovative Research (ISSN : 2349-5162)

Is hereby awarding this certificate to

K. Murali

In recognition of the publication of the paper entitled

CORONA VIRUS INFECTION PROBABILITY CLASSIFICATION USING MACHINE LEARNING ALGORITHMS.

Published In JETIR (www.jetir.org) ISSN UGC Approved (Journal No: 63975) & 7.95 Impact Factor

Published in Volume 8 Issue 6 , June-2021 | Date of Publication: 2021-06-12

Parisa P

EDITOR

EDITOR IN CHIEF

JETIR2106221 Research Paper Weblink http://www.jetir.org/view?paper=JETIR2106221 Registration ID : 310536