SvedbergOpen
DISSEMINATION OF KNOWLEDGE

International Journal of Artificial Intelligence and Machine Learning
Publisher's Home Page: https://www.svedbergopen.com/

Research Paper Open Access

Symptom Driven Predictive Models for Infectious Diseases, Primarily Leveraging Machine Learning Techniques

Md Shadab Hussain[1], Gaurav Agarwal[2], Akash Sanghi[3*]

[1]Research Scholar, Department of Computer Science and Engineering, Invertis University, Bareilly, India, md.shadab93@gmail.com
[2]Department of Computer Science and Engineering, Invertis University, Bareilly, India, gaurav.a1@invertis.org
[3*]Department of Computer Applications, Invertis University, Bareilly, India, sanghiakash@gmail.com

ABSTRACT
An entirely new infectious illness, dubbed "Pandemic X," had a massive effect on worldwide affairs, influencing social, economic and political lives all over the planet. When the virus came back in April 2021 after the initial wave, the world was slowly coming to terms with the "new normal" and once again the consequences were devastating. Widespread testing was not possible due to limited resources and delays in the declaration of results. In this work, we introduce some machine learning predictive models that can predict the positive/negative status of patients using seven common symptoms recommended by the World Health Organization (WHO). Symptoms are Fever, Dry cough, Tiredness, Sore throat, Chest pain, Loss of taste/smell, and Breathing difficulties. Various criteria are used to make a comparative analysis of these models. Furthermore, we test our models with new data not included in our training and test sets in this study.

Keywords: Machine Learning, Classification Problem, Prediction, respiratory illness, Symptom

1. INTRODUCTION

Coronaviruses constitute a viral family capable of inducing symptoms ranging from the common cold to serious respiratory illnesses including Middle East Respiratory Syndrome (MERS) and serious Acute Respiratory Syndrome (SARS). In 2019, a novel coronavirus linked to COVID-19 was identified in Wuhan, China. In some of the previous materials you may see the term "nCoV"; in this context it means the new coronavirus which causes COVID-19.

Most people are infected by the coronavirus through their respiratory system, but some people may have no symptoms. If symptoms occur, the disease may be either mild or moderate and may resemble seasonal influenza and may not require intensive medical care [3].

Symptoms generally encompass fever, cough, respiratory distress, exhaustion, pharyngitis, and anosmia or ageusia. The illness can impact individuals of all ages; however, those aged 70 and above, those who are sedentary, and individuals with chronic health issues such as chronic respiratory disease, liver disorders, diabetes, cancer, cardiovascular ailments, hypertension are at an elevated risk of severe nCoV. Any actions taken to protect vulnerable groups must not lead to discrimination, social isolation, exposure or loss of access to programs and services [2].

The World Health Organization advises for those at elevated risk.:

- When hosting guests at your residence, offer 1m welcome gesture.
- Should you exhibit indications of respiratory sickness, please reach out to your healthcare practitioner via telephone prior to visiting the healthcare institution.
- Prepare a strategic plan in anticipation of a respiratory infection breakout within your community.
- In communal spaces, adhere to the same safety measures as you would at home.
- Stay informed about respiratory diseases by acquiring information from trustworthy sources. [4].

Though all sorts of warnings and guidelines are being given by WHO, Government, NGOs and others, there are still a lot of people being affected by nCoV. The tests procedure available so far to identify the presence of nCoV in a patient such as Antigen, RTPCR, Chest X-Ray etc. are very time consuming.

Therefore, it is crucial to come up with some techniques that can predict the nCoV accurately in patients based on the observed symptoms in less time. In our research we have developed some binary classification models which will predict nCoV in patient according to their present symptoms immediately. In section 2, illustrate a thorough literature review and summarized the previous research conducted by different researchers. Section 3 we describe the Materials and Methods that were used to accomplish our research work. Section 4 of this article contains Results and Discussions [15]. Lastly, section 5 is for the conclusive comments.

2. Literature Review

Authors have published the use of machine learning tools, hand-picked 3 biomarkers that can forecast the survival probability for individual individuals with almost ninety percent accuracy: dairy product dehydrogenase (LDH), white cell and high-sensitivity CRP (hs-CRP). In particular, the high LDH levels seem to be most important in the differentiation of most cases requiring urgent medical treatment [5].

Individuals displaying nCoV symptoms are detected at their places of employment through the application of eight unique deep learning techniques, specifically DenseNet201, VGG16, InceptionResNetV2, ResNet15V2 ResNet50, MobilenetV2, VGG19 and NasNetMobile using four hundred CT scan images and four hundred chest X-ray images [16]. It was determined that NasNetMobile could achieve an accuracy of eighty-two.Ninety-four percent in the CT scan and ninety-three.Achieving 94% accuracy in chest X-ray datasets surpasses all competing models [6].

The authors' research endeavor has cultivated a machine learning methodology that was trained on the data of fifty-one thousand eight hundred thirty-one individuals, among whom 4,769 were verified as having nCoV. The test set data encompassed information from the subsequent week, with 47,401 individuals tested, of whom 3,624 were diagnosed with nCoV. In conclusion, with the assistance of the Israeli Ministry of Health, we created a model proficient in identifying nCoV situations using user-friendly interfaces with straightforward inquiries [7].

Writers frequently bestow an artificial intelligence methodology primarily reliant on a deep convolutional neural network (CNN) for the identification of nCoV patients utilizing real-world statistics [14]. Their technique evaluates chest X-ray pictures to detect such patients. Their research suggests that this associative analysis is proficient in nCoV identification as X-rays are readily accessible in a timely manner and at minimal cost [8].

The researchers indicated in their experimental findings that LSTM is more well suited for forecasting the total confirmed cases, whereas the Elman neural network and SVM employed in this research will project the trends of total confirmed cases, fatalities, and recoveries [9].

This research is an effort to study the prediction of Corona Virus malady 2019 (nCoV) based on the confirmed, deceased and recovered cases. This forecast can help prepare resources, validate government policy, and provide immunity passports to survivors and use constant plasma for treatment. In this analysis, data, such as location wise confirmed, deceased, recovered nCoV, meridian and latitude were gathered around Gregorian calendar month twenty-two, 2020 to Gregorian calendar month twenty-five, 2020, globally [10].

This study aims to identify the nCoV-related applications and algorithms for investigations and multiple other functions of machines. The studies printed throughout the year 2020 and related to this subject were reviewed by using the words: Unsupervised Learning, Supervised Learning, Machine Learning, nCoV in Springer, Science Direct, MDPI and Hindawi [11].

This research introduces the preliminary identification of the Coronavirus (nCoV), recognized by the WHO, via machine learning techniques. The detection technique was applied to abdominal CT images. The CT images were able to give the skilled radiologists information that nCoV exhibits totally different behaviors than other viral infections. The highest classification accuracy was achieved using the methodology of GLSZM feature extraction and 10-fold cross validation, which was 99.68% [12].

In this study, the authors select some popular deep convolution neural networks (DCNNs), including VGGNet, DenseNet, Xception, MobileNet, InceptionV3, ResNet, InceptionResNetV2 and NASNet, as candidates for feature extraction to obtain the correct and most widespread feature, which is essential for learning.

3. METHODOLOGY

Various materials have been described in this section that can be used to carry out our research work and the complete 13 steps of research methodology have also been described.

3.1 Dataset

A Google form was used to collect the data for the study. This form included 8 questions, with two possible answers (Yes = 1 and No = 0). The first seven questions were input variables and the eighth question was used for model development. 200 responses were collected. Questions included were: "Do you have a fever", "Have any of you had a dry cough", "Have any of you been tired, "Have you had a sore throat?", "Have you had chest pain or pressure?", "Have you lost your ability to smell or taste?", "Have you experienced breathing difficulties or dyspnea?" and "What was the outcome of the test?"

3.2 Technology and framework used

Supervised machine learning classification techniques [17] [18] employed for model construction are executed utilizing the following Python libraries:
Seaborn – utilized for the creation of heat maps.
Scikit-learn — for the execution of algorithms.
Pandas — for manipulating datasets.
Matplotlib – ROC and AUC graph.
The stacked bar graph is generated in MS Excel.

3.3 Phases of the Research

The many stages we undertake to finalize our study endeavour include a total of 13 processes. To conclude, the next 13 steps are enumerated here sequentially:

Step 1: Problem Identification

In this step we formulate the research problem of predicting respiratory illness based on seven popular symptoms as per WHO.

Step 2: Feature selection

In this step we identify and select popular symptoms responsible for respiratory illness. The seven symptoms selected are used as input parameter for building the models.

Step 3: Design the Google form to collect data.

In this step a Google form is designed. There are 8 questions on this form. The response to these questions is categorical in nature i.e. Yes or No. Yes, corresponds to 1 and No, corresponds to 0.

Step 4: Data collection

In this step, the link of Google form is sent to many populations via e-mail, WhatsApp group and personal contacts. After repetitive reminders, about two hundred responses are received in about fifteen days.

Step 5: Data cleansing

In this step data cleansing is performed on the collected data. In few responses some of the fields are found null even if we have made every field compulsory. These data were omitted from final data set.

Step 6: Identification and selection of classification algorithms

In this step we reviewed the literature and try to find out binary classification algorithms suitable for our study. Finally, we chose algorithms such as Random Forest (RF), K-Nearest Neighbour (KNN), Logistic Regression (LR), Support Vector Machine (SVM), Decision Tree (DT), Neural Network (NN), Stochastic Gradient Descent (SGD), and Gaussian Naive Bayes (GNB) algorithms for implementing the classification model in Python.

Step 7: Identify and select tools for implementation

In this step, we recognize and at last, choose the tools that we should use for our study depending on our suitability and expertise. This is because the tools used have already been explained in the tools used section above.

Step 8: Execution of algorithms and creation of categorization models via training.

Here we will execute all the 8 classification algorithms in sequence for the training data set using the various libraries of Python using Google Collaboratory cloud service.

Step 9: Evaluating the models and predicting their precision

For this step, each of the 8 developed models are tested in turn against the test set. The accuracy score of all the models is also predicted.

Step 10: Test the accuracy of the models on new data.

This step is to validate the models with fresh data that weren't used in training and testing.

Step 11: Model performance evaluation.

In this phase, the efficacy of the models is assessed based on multiple criteria. The specifics of this phase are outlined in the results and discussion segment.

Step 12: The analysis of results is used to prioritize the input feature in.
In this step we decide the priority of the input features through DT generated by DT classifier. The details of this step are presented in results and discussion section.

Step 13: Evaluation of precision and efficacy of the models.
In this step a comparative study of accuracy and performance of all the eight developed models are done. The details of this step are presented in results and discussion section through various figures and tables.

4. RESULTS AND DISCUSSION

The results of our research, presented in six tables and two figures, with the accompanying discussions will be discussed in this section.
Table 1 represents the Selected Algorithm, Target Class, Training dataset and Test dataset split percentage out of the total dataset available. The algorithms chosen are widely used and found in the machine learning literature. The train and test set split is taken as 80% and 20% respectively, justifiable due to most of the literature available.

Table 1: Selected Algorithm, Target Class, Training dataset and Test dataset.

S. No	Selected Algorithm	Target Class	Training Dataset	Test Dataset
1	GNB	Observation	80 %	20 %
2	KNN	Observation	80 %	20 %
3	SVM	Observation	80 %	20 %
4	SGD	Observation	80 %	20 %
5	RF	Observation	80 %	20 %
6	DT	Observation	80 %	20 %
7	LR	Observation	80 %	20 %
8	NN	Observation	80 %	20 %

Table 2 shows the Selected Algorithm, Random State and Accuracy Score. Random state is the simulation hyper parameter to optimize the accuracy score. In the remark column either the default or manual parameters are selected with an objective to optimize the result. The best two results are shown in bold face.

Table 2: Selected Algorithm, Random State and Accuracy Score.

S. No	Selected Algorithm	Random State	Accuracy Score
1	GNB	82	0.95
2	KNN	21	0.90
3	SVM	82	0.92
4	SGD	14	0.85
5	RF	03	0.925
6	DT	81	0.875
7	LR	82	0.95
8	NN	3	0.90

Table 3 contains the Selected Algorithm, Confusion Matrix and its corresponding Heat Map. These results were generated by various classifiers developed through implementation using Python. The best two results are shown in bold face.

Table 3: Selected Algorithm and their Confusion Matrix

S. No	Selected Algorithm	Confusion Matrix
1	GNB	$\begin{bmatrix} 19 & 0 \\ 2 & 19 \end{bmatrix}$
2	KNN	$\begin{bmatrix} 19 & 1 \\ 3 & 17 \end{bmatrix}$
3	SVM	$\begin{bmatrix} 19 & 0 \\ 3 & 18 \end{bmatrix}$

4	SGD	$\begin{bmatrix} 15 & 2 \\ 4 & 19 \end{bmatrix}$
5	RF	$\begin{bmatrix} 12 & 2 \\ 1 & 25 \end{bmatrix}$
6	DT	$\begin{bmatrix} 13 & 1 \\ 4 & 22 \end{bmatrix}$
7	LR	$\begin{bmatrix} 19 & 0 \\ 2 & 19 \end{bmatrix}$
8	NN	$\begin{bmatrix} 11 & 3 \\ 1 & 25 \end{bmatrix}$

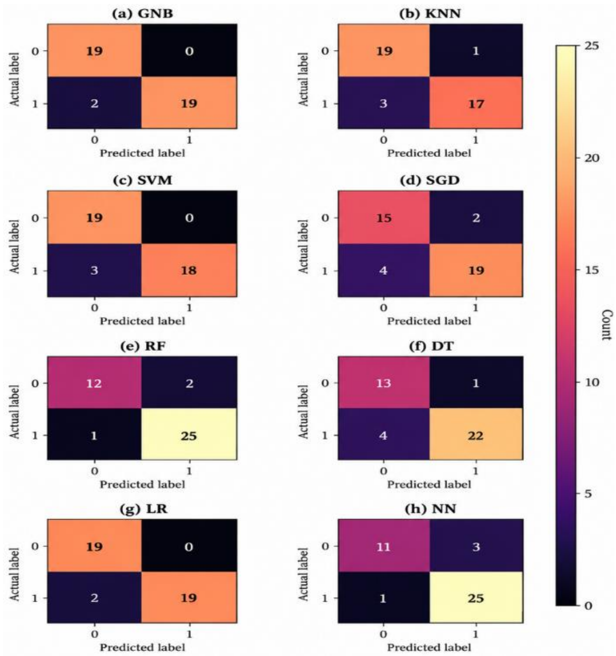


Figure 1: Confusion matrices of the eight evaluated classification algorithms: (a) GNB, (b) KNN, (c) SVM, (d) SGD, (e) RF, (f) DT, (g) LR, and (h) NN.

Table 4 depicts Selected Algorithm and Classification Report generated by the classifiers. The best two results are shown in bold face.

Table 4: Selected Algorithm and Classification Report.

S. No.	Algorithm	Class	Precision	Recall	F1-score	Support
1	GNB	0	0.9	1	0.95	19
		1	1	0.9	0.95	21
2	KNN	0	0.86	0.95	0.9	20
		1	0.94	0.85	0.89	20
3	SVM	0	0.86	1	0.93	19
		1	1	0.86	0.92	21
4	SGD	0	0.79	0.88	0.83	17
		1	0.9	0.83	0.86	23
5	RF	0	0.92	0.86	0.89	14
		1	0.93	0.96	0.94	26
6	DT	0	0.76	0.93	0.84	14
		1	0.96	0.85	0.9	26
7	LR	0	0.9	1	0.95	19
		1	1	0.9	0.95	21
8	NN	0	0.92	0.79	0.85	14
		1	0.89	0.96	0.93	26

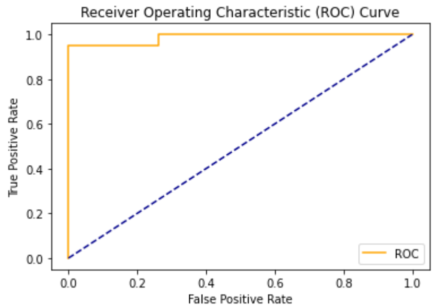
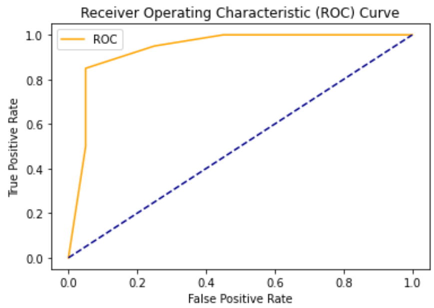
Table 5 explains the Selected Algorithm Mean Squared Error (MSE) and Log loss. The best two results are shown in bold face.

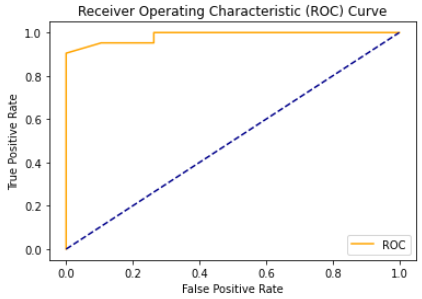
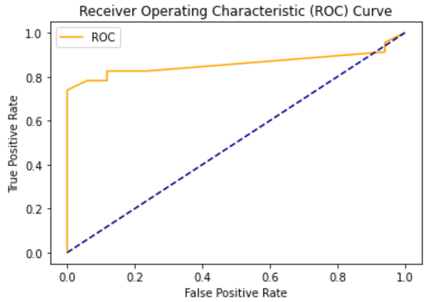
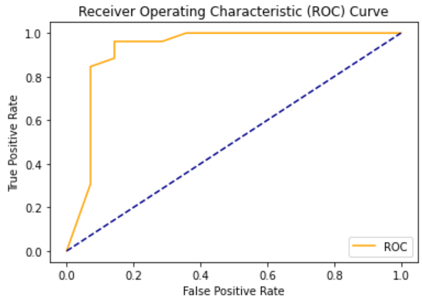
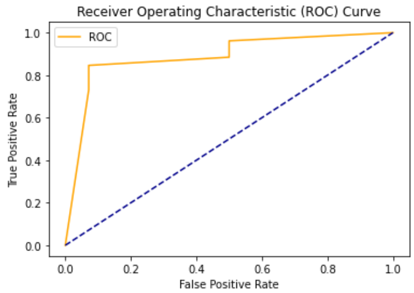
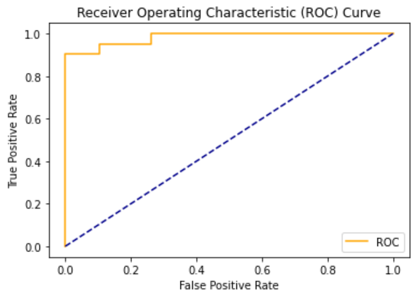
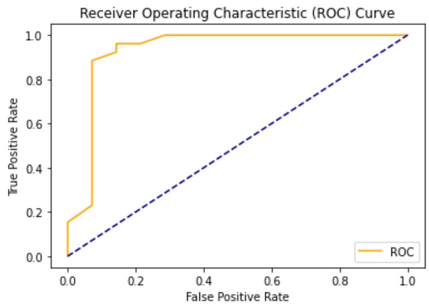
Table 5: Selected Algorithm MSE, Log loss.

S.No	Selected Algorithm	MSE	Log loss
1	GNB	0.05	1.7269388197455349
2	KNN	0.1	3.453897629426824
3	SVM	0.075	2.590408229618302
4	SGD	0.15	5.180856439108114
5	RF	0.075	2.590448209489812
6	DT	0.125	4.317367039299591
7	LR	0.05	1.7269388197455349
8	NN	0.1	3.453937609298334

Table 6 describes the Selected Algorithm, AUC and its corresponding ROC curve. The best result is highlighted in bold face.

Table 6: Selected Algorithm, AUC and ROC curve.

S. No.	Selected Algorithm	AUC	ROC Curve
1	GNB	0.99	
2	KNN	0.94	

3	SVM	0.98	
4	SGD	0.86	
5	RF	0.92	
6	DT	0.89	
7	LR	0.98	
8	NN	0.93	

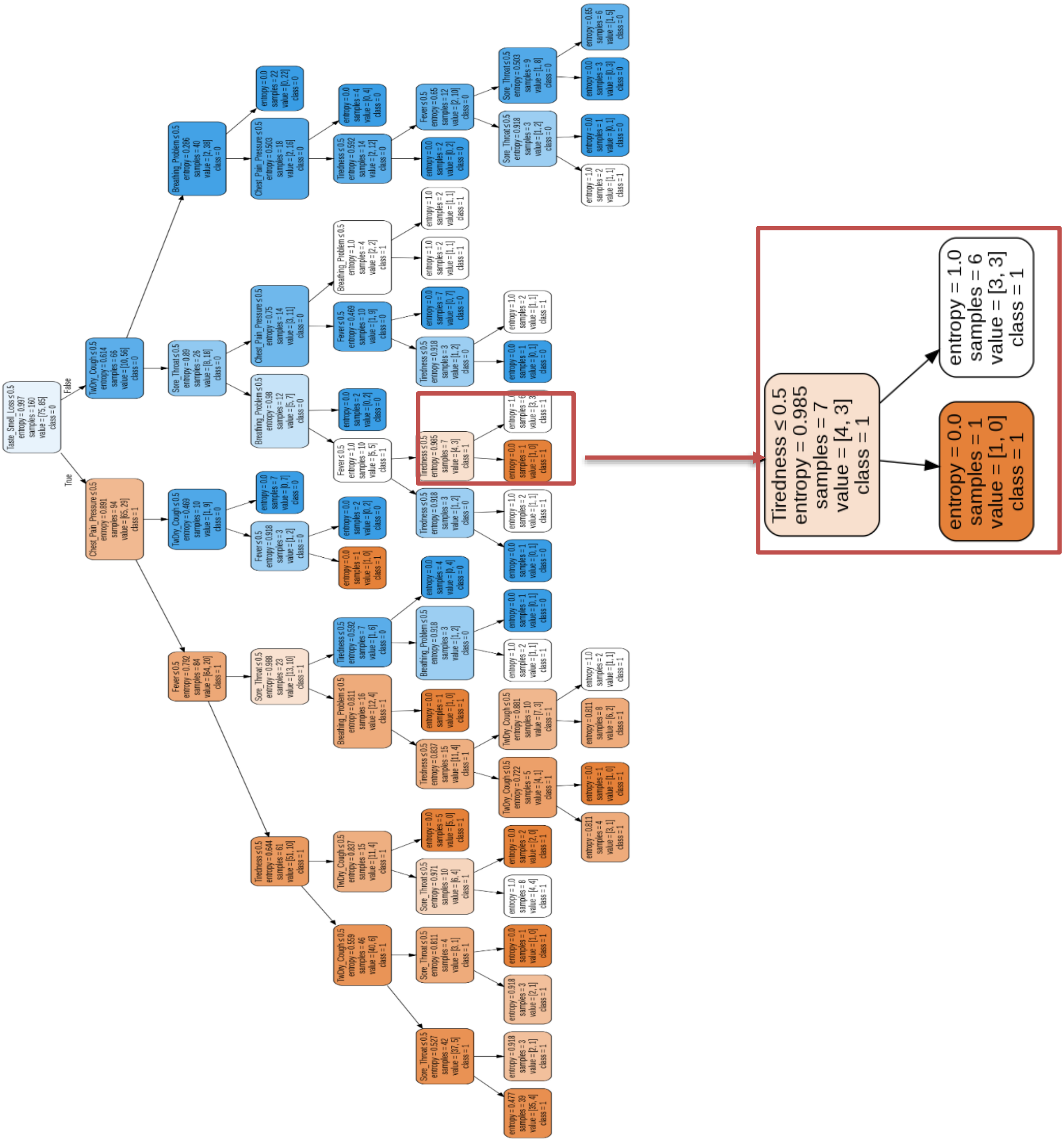


Figure 2: Decision Tree

Figure 2 shows the DT generated by DT classifier. Taste/smell loss is at the root. Hence, it is the most significant symptom. A dry cough and chest pain are positioned at the next tier in the hierarchy, thereby categorizing them as secondary priority symptoms.

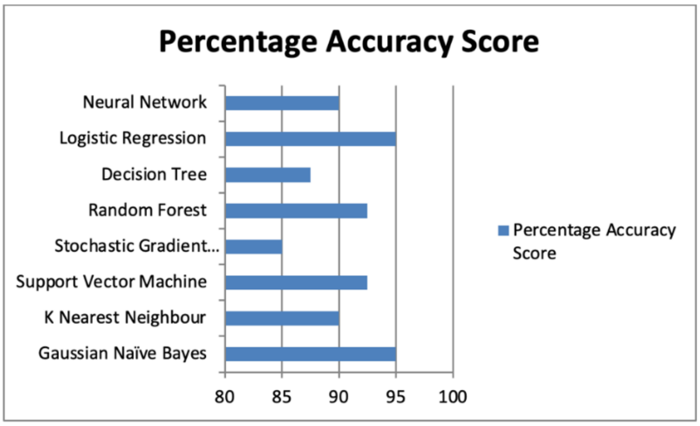


Figure 3: Percentage Stacked Bar Graph accuracy score comparison of classifiers.

Figure 3 represents a Percentage Stacked Bar Graph, accuracy score comparison of all eight classifiers developed in the research. GNB and Logistic Regression show the best accuracy score of 95%.

5. CONCLUSION

From Table 2, 3, 5 and Figure 3 it is concluded that GNB and LR have the highest accuracy 95%. From Table 4 it is observed that f1-score is highest in GNB and LR which is 0.95. From Table 6 it is observed that AUC of GNB is 0.99 and that of LR and SVM is 0.98. Thus, at the end, we determine that GNB is the most effective algorithm in terms of performance and accuracy, followed by LR and SVM. From Figure 2 it is concluded that Taste/smell loss is the most significant symptom followed by Dry cough and Chest pain.

References

1. World Health Organization. "Introduction to emerging respiratory viruses, including novel coronavirus" Geneva: World Health Organization; 2020 [cited 2021 May 7]. Available from: https://openwho.org/courses/ introduction-to-ncov
2. Aronson JK. Coronaviruses—"A general introduction Oxford: Centre for Evidence-Based Medicine, University of Oxford"; 2020 Mar 25 [cited 2021 May 9]. Available from: https://www.cebm.net/covid-19/coronaviruses-a-general-introduction/
3. Moses R. COVID-19: "Respiratory physiotherapy on-call information and guidance Lancashire: Lancashire Teaching Hospitals NHS Foundation Trust"; 2020 [cited 2021 May 11]. Available from: https://www.acprc.org. uk/media/gomdrs0i/covid19 respiratoryphysiotherapyoncallinformationandguidancev2.pdf
4. World Health Organization. Coronavirus disease 2019 (COVID-19): Situation Report No. 51 "Geneva: World Health Organizatio"; 2020 Mar 11 [cited 2021 May 14]. Available from: https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200311-sitrep-51-covid-19.pdf
5. Yan L, Zhang HT, Goncalves J, Xiao Y, Wang M, Guo Y, et al. "A machine learning-based model for survival prediction in patients with severe COVID-19 infection. medRxiv "2020. doi: 10.1101/2020.02.27.20028027.
6. Ahsan MM, Gupta KD, Islam MM, Sen S, Rahman ML, Hossain MS. "COVID-19 symptoms detection based on NasNetMobile with explainable AI using various imaging modalities. Machine Learning and Knowledge Extraction". 2020;2(4):490–504. doi: 10.3390/make2040027.
7. Zoabi Y, Deri-Rozov S, Shomron N. "Machine learning-based prediction of COVID-19 diagnosis based on symptoms. npj Digital Medicine". 2021; 4:3. doi: 10.1038/s41746-020-00372-6.
8. Alazab M, Awajan A, Mesleh A, Abraham A, Jatana V, Alhyari S. "COVID-19 prediction and detection using deep learning". International Journal of Computer Information Systems and Industrial Management Applications. 2020; 12:168–181.
9. Hao Y, Xu T, Hu H, Wang P, Bai Y. "Prediction and analysis of Corona Virus Disease 2019". PLOS ONE. 2020;15(10): e0239960. doi: 10.1371/journal.pone.0239960.
10. Singh V, Poonia RC, Kumar S, Dass P, Agarwal P, Bhatnagar V, et al. "Prediction of COVID-19 coronavirus pandemic based on time-series data using support vector machine. Journal of Discrete Mathematical Sciences and Cryptography". 2020;23(8):1583–1597. doi: 10.1080/09720529.2020.1784535.
11. Kwekha-Rashid AS, Abduljabbar HN, Alhayani B. "Coronavirus disease (COVID-19) cases analysis using machine-learning applications". Applied Nanoscience. 2023; 13:2013–2025.

doi: 10.1007/s13204-021-01868-7.

12. Barstugan M, Ozkaya U, Ozturk S. "Coronavirus (COVID-19) classification using CT images by machine learning methods". 2020. arXiv:2003.09424. Available from: https://arxiv.org/abs/2003.09424
13. Kassani SH, Kassasni PH, Wesolowski MJ, Schneider KA, Deters R. "Automatic detection of coronavirus disease (COVID-19) in X-ray and CT images: A machine learning-based approach". arXiv . 2020. arXiv:2004.10641. Available from: https://arxiv.org/abs /2004.10641
14. Sambasivarao, K.V., & Roopa Devi Bhima, A.S. (Eds.). (2026). "Artificial Intelligence, Computational Intelligence and Inclusive Technologies" (1st ed.). CRC Press. https://doi.org/10.1201/9781003740100
15. Maurya, L. S., Hussain, M. S., & Singh, S. (2021). "Developing Classifiers through Machine Learning Algorithms for Student Placement Prediction Based on Academic Performance". Applied Artificial Intelligence,35(6), 403–420. https://doi.org/10.1080/ 08839514. 2021.1901032.
16. Ahsan, Md Manjurul, Kishor Datta Gupta, Mohammad Maminur Islam, Sajib Sen, Md. Lutfar Rahman, and Mohammad Shakhawat Hossain. 2020. "COVID-19 Symptoms Detection Based on NasNetMobile with Explainable AI Using Various Imaging Modalities" Machine Learning and Knowledge Extraction 2, no. 4: 490-504. https://doi.org/10.3390/make2040027
17. D. K. Singh and A. Sanghi, "Advancing Heart Attack Prediction Through AI: A Review of ML, DL, and Hybrid Models," 2025 14th International Conference on System Modeling & Advancement in Research Trends (SMART), Moradabad, Uttar Pradesh, India, 2025, pp. 499-505, doi: 10.1109/SMART66937.2025.11389447.
18. A. Sanghi, R. Agarwal, S. Pandey, A. Kumar and K. Johari, "A Comprehensive Machine Learning Framework for Diabetes Prediction: Incorporating Unique Performance Metrics and K-Fold Cross-Validation," 2025 Seventh International Conference on Computational Intelligence and Communication Technologies (CCICT), Sonepat, India, 2025, pp. 564-569, doi: 10.1109/CCICT65753.2025.00091.