IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 9, Issue 11, November 2020

DOI 10.17148/IJARCCE.2020.91105

A Survey on Detection of COVID-19 Through Machine Learning Approaches

Deepti Chauhan[1], Chetan Agrawal[2], Bhavana Verma[3]

Research Scholar, CSE, Radharaman Institute of Technology & Science Bhopal, India[1]

Assistant Professor, CSE, Radharaman Institute of Technology & Science Bhopal, India[2,3]

Abstract: Advances in technology have a swift impact on any area of life, whether medical or other. Through its analysis of data, artificial intelligence has shown promising results for health care. Over 100 countries have been affected by COVID-19 in no time. People worldwide are vulnerable in the future to its impacts. A control system for the detection of corona virus is imperative. The detection of disease using different AI methods may be one solution to manage the current catastrophe. This article classified textual clinical reports by using classical and ensemble machine algorithms into four classes. The study has used the concept of Natural language processing in which reports are classified using machine learning. In this work we have performed the classification using Naïve Bayes, Support Vector Machine, Logistic regression and Decision tree and we have observed decision tree has outperformed other state of an algorithms with an accuracy of 97.8%. Before implementing the classification, feature engineering has also applied.

Keywords: Coronavirus, Naïve Bayes, Artificial Intelligence, Natural Language Processing, Precision.

I. INTRODUCTION

In Wuhan, China, the community was recognized in 2019 (COVID-19), patients infected with a new coronavirus disease. COVID-19 contagions have spread worldwide since then. COVID-19 has numerous consequences for citizens [1]. Mainstream symptoms (e.g., fever, fatigue, dry coughing), and other symptoms (e.g., aches and pains, nasal congestion, runny nose, moose throat, or diarrhea) can be usual in most infected patients [2]. COVID-19 has exposed the health system gap in many countries, and distress has been caused by health systems' failure to treat patients. The lack of precision in clinical detection methods was a key factor behind COVID-19's rapid propagation [3]. Molecular methods are required and commonly employed for the detection of COVID-19, for example, quantitative Reverse Transcription-Polymerase Chain Reactions (rRT-PCR) [4], as well as serological methods [5] or viral throat pin test [6]. However, studies have revealed queer x-rays [7] and a Chest Tomography (CT) scan [8] that may be helpful and that abnormalities of various lung diseases, including COVID-19, have been revealed. The primary scanning method used for CT-scan and X-ray testing may be COVID-19, track infected patients' emergency, and predict progression from COVID-19 [9]. However, time for emergencies is always limited and does not allow for the use of modern conventional manual diagnostics to be performed for these studies. These procedures require a professional doctor and can be checked or read and interpreted for human error effects inappropriate in sensitive cases. In light of the latest spread of COVID-19, many patients in a hospital are either better or more serious (dying). In this case, CT and x-ray tests should be conducted to save as many lives as possible with optimum speed and reliability [9]. In the diagnostic and classification processes, the role of intelligent technologies will help effectively [10].

In various fields, particularly in medical detection, Artificial Intelligence (AI) has increased. AI is commonly used to achieve more reliable detection outcomes and decrease the health system burden [11]. The decision-making time for the identification of conventional approaches can be reduced. A significant factor in enhancing prediction, prevention, and identification of potential global health dangers is AI technology's advancement to recognize the risks of infectious disease [12]. A few investigators with actual COVID-19 datasets with various case studies and goals have documented multiple AI classifiers [9]. While the diagnosis and classification of COVID-19 can benefit from AI technology, it isn't easy to select an effective AI technique, leading to reliable results. The wide variety of AI techniques available creates difficulties in deciding which ones to use, especially when no dedicated AI technique is much better than the other in developing the diagnosis and classification of COVID-19.

Moreover, most of these methods are poorly effective and efficient in computing. On the other hand, the dynamic factor is associated with assessment and comparison because of the various assessment parameters and the disagreement between them. The complexity is also increased. AI techniques are essential to testing and benchmarking when obtaining a method that can deliver the best results. A similar process is crucial because it can lead to the death and dissemination of the virus, among others suspected of COVID-19 and medical organizations. Several criteria ensure the reliability of such techniques since they are related to patients ' lives, to test and benchmark AI classification techniques that can be used in the identification of COVID-19 medical images.

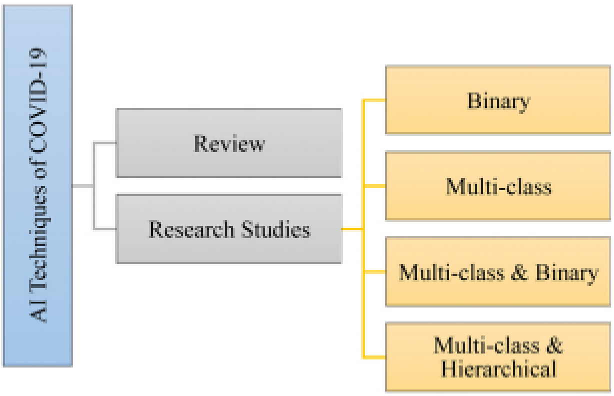

Figure 1. Classification of COVID-19 dataset

In this phase, however, two key issues can be addressed. First of all, what parameters should be used in the assessment? Secondly, what is the best benchmarking protocol to use, among other things, in choosing a proper AI technique?

II. BACKGROUND STUDY

The present research builds on a systematic literature review (SLR), recognized as providing a proper understanding of the subject of interest SLR also emphasizes its remarkable organized study methods and the ability to adapt various studies from different scientific disciplines. During the process, the relevant literature is extracted from different academic digital databases, including ScienceDirect, which provides a variety of scientific literature throughout all disciplines; Scopus, which provides sufficient coverage of literature from all disciplines; IEEE, whose scientific reliability is recognized by its multidisciplinary science & engineering and its co-operation. These databases are considered adequate for COVID-19 diagnosis and identification in the current and most accurate literature.

The studies collected in these databases are essential and accurate to understand intelligent systems' functions and participation (i.e., AI) and scientists in the COVID-19 efforts. In 10 years between 2010 and May 5, 2020, literature searching on the five central databases took place extensively. The databases' selection was based on its scientific soundness, scientific reliability, and literature coverage from different fields of more in-depth learning and COVID-19. During the process, boolean operators were used to obtaining as much material as possible. The first group of keywords was intended for smart systems and their related keywords, with the second group for the COVID-19 keywords.

One of the main tasks in text mining is classification and can be carried out using various algorithms. Sarwar et al. [13] diagnosed diabetes using computer and ensemble learning approaches showed that ensemble technology offered 98.60 percent accuracy. This can be helpful if COVID-19 is to be diagnosed and predicted. Firm and reliable COVID-19 diagnostics will save millions of lives and provide vast amounts of data to train machine learning models. ML may be useful for making a diagnosis based on clinical text, radiography images, etc. in this respect.

Verma et al. [14] examined the Indian government's feelings using the Lexicon dictionary to execute projects. The machine learning changed the diagnostic outlook by delivering outstanding results for diseases such as diabetes and epilepsy. Kumar et al. [15] performed a SWOT analysis of various text classification algorithms for instructible mining data. Kumar et al. Sentiment tests, fraud detections, and spam detection are the different text classification applications. Mainly for election, advertising, business, etc., opinion mining is used.

Machine learning and deep learning, according to Bullock et al. , can replace people by making a specific diagnosis. A perfect diagnosis can save time for radiologists and can be economical than regular COVID-19 tests. For training the machine learning model, X-rays and CT scans can be used. In this respect, several initiatives are in progress. Electroencephalogram (EEG) signals, identification of normal and epileptic conditions using artificial neural networks (ANN) was used by Gaizo et al. [16] observed epilepsy with computer teaching approaches. Yan et al. [17] used computer training for developing a predictory prognosis algorithm to predict the death risk of an infected person. Jiang et al. [18] suggested a machine learning model that can indicate an affected individual with COVID-19, which can evolve ARDs. 80% of the accuracy of the proposed model.

COVID-Net, an extensive convolution neural network, has been developed by Wang and Wong [19] and can diagnose COVID-19 from chest X-ray images. The question is whether and how intensively this person is affected when COVID-19 is identified in one person. Not everyone needs special attention to COVID-19 positive patients. The ability to predict who will be more critically impacted will help coordinate health care and organize medical services distribution and use. Using data from (just) 29 patients in Tongji Hospital in Wuhan, China.

III. CLASSIFICATION METHODS

The W.H.O. announced a pandemic of coronavirus as a health emergency.

3.1 Collection of Data - Researchers and hospitals have open access to pandemic results. We also obtained data from available sources GitHub.1 data archive, in which approximately 212 data patients with corona virus and other viruses have been processed. Data consists of roughly 24 attributes, namely Patient I d, Opposite, Sex, Age, Finding, Survival, Incubating, Went iqu, Necessity Supplemental O2, Augustine, Temperature, Saturation, Leukocyte count, Licensed Notes and other clinical notes.

3.2 Important dataset - We have gathered clinical reports and results as our study includes text mining. The clinical notes consist of text, while the discovery of the attribute consists of a text mark. There were roughly 212 reports used, and their duration was estimated. The papers written in English are considered only. The stable distribution of clinical reports written in English is given in Figure 3. The clinical statements are identified according to their respective grades. We have four groups in our dataset: COVID, ARDS, and SARS (COVID, ARDS).

3.3 Preprocessing Text is unstructured so that machine learning can be conducted. In this process, several steps are followed, and the needless removal of the document cleans the document. Punctuation and lemmatization were carried out to improve data processing. Stopwords, icons, URLs, links are omitted to allow a better classification.

3.4 Feature engineering Different features are extracted by semantization and transformed in probabilistic values from the preprocessed clinical records. For the extraction of relevant features we use TF/IDF technique. Word bags were also taken into account, as well as unigrams, bigrams. We have listed 40 features that can achieve the classification. The same volume and input is provided to the machine learning algorithms by providing the corresponding weight.

The classification is carried out to classify the text into four distinct virus groups. The four virus types, COVID (one with coronavirus), ARDS, SARS, and both (one with both corona viruses and ARDS) are used. Different supervised algorithms are used in order for the text to be classified in these categories. In order to accomplish this task the algorithms for machine learning such as vector support (SVM), multicultural Naı ̈ve Bayes (MNB), logistic regression, tree judgement, the random forest, bagging, adaboost and stochastic gradient boosting were used.

3.5.1 Conventional algorithms for machine learning

Naïve Bayes It is a classification technique with an assumption of independence among predictors based on Bayes ' Theorem. In short, a Naive Bayes classificator believes that there is no relation between the existence of a particular characteristic in a class and any other characteristic. For instance, if the fruit is red, circular and approximately 3 " in diameter, it can be considered an apple. While these characteristics depend on each other or the other's life, all these properties contribute independently to the likelihood that this fruit is an apple and is known as 'Naive' therefore.

Model Naive Bayes for very large data sets is simple to construct. Naive Bayes is considered to be beyond even the most advanced classification methods in addition to simplicity. The theorem Bayes offers a way to measure the P(a|y) P(a), P(y) and P(y|a) probability. See the following equation:

$$P(a|y) = (P(y|a) * P(a)) / P(y) \quad (1)$$

$$P(a|Y)=P(y_1|a)* P(y_2|a)*\ldots\ldots.* P(y_n|a)* P(a) \quad (2)$$

Support vector machine: The goal of the support vector machine algorithm is to find a hyperplane that separately classifies the data points in an N-dimensional space (N — the number of characteristics).

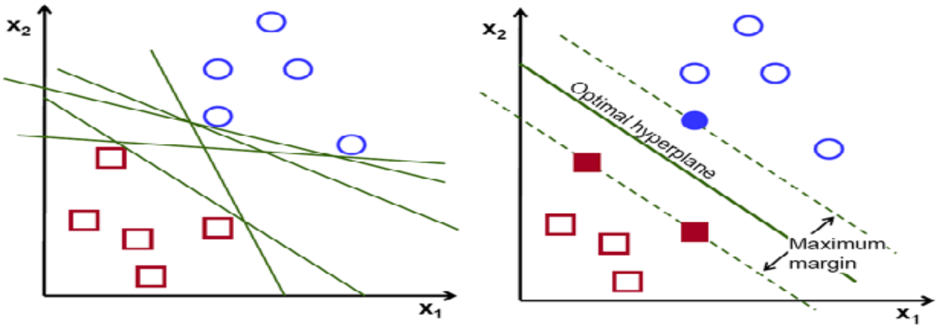

Figure 2. Possible Hype planes

There are several potential hyperplanes to select to differentiate the two types of data points. Our goal is to find a target with the greatest range, i.e., the most significant distance from both groups of data points. Maximizing the margin distance provides strengthening to enhance confidence in the future data points.

Hyperplanes and Support Vectors

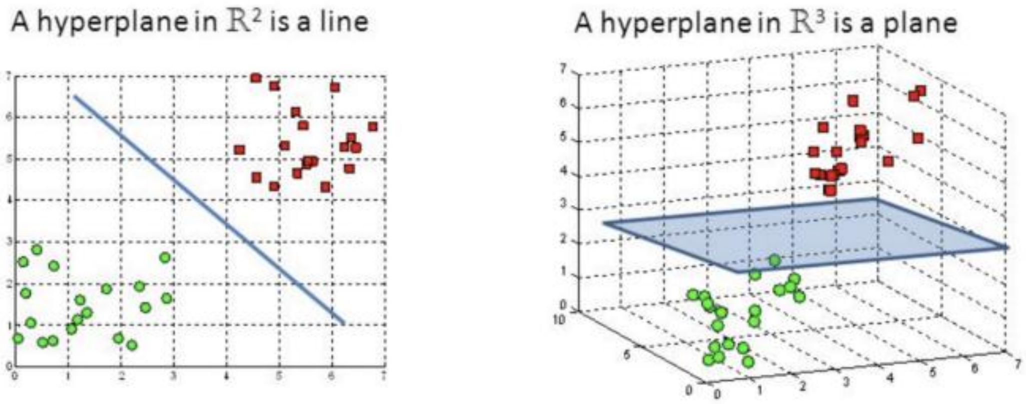

Figure 3. Hyperplanes in 2D and 3D feature space

Hyperplanes are decision limits that help to categorize data points. There may be different classes of data points that fall on either side of the hyperplane. The size of the hyperplane also depends on the number of characteristics. The hyperplane is only a line if the number of input features is 2. If the number of input features is 3, then the hyperplane becomes a two-dimensional plane. When the number of functions exceeds 3, it isn't easy to imagine.
Supporting vectors are data points closer to the hyperplane and influence hyperplane position and orientation. We optimize the margin of the classifier using these support vectors. The elimination of the support vectors shifts the hyperplane location. These are the issues that help us to develop our SVM.

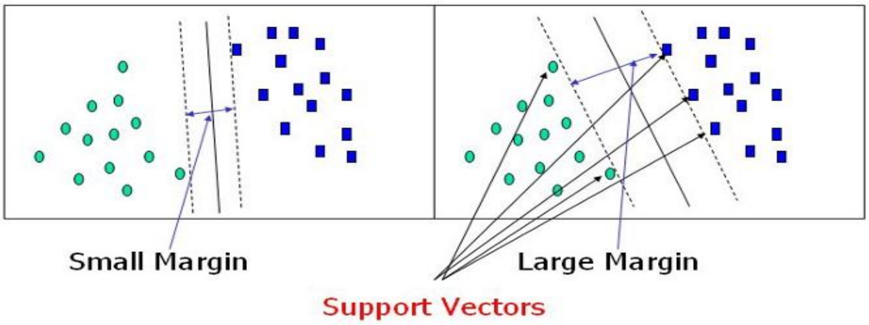

Figure 4. Support Vectors

Decision tree – Decision Trees are supervised Machine Learning where data is continuously split by specific parameters (that is, you clarify what the input is and what the appropriate output is in the training data). The tree can be defined by two individuals, namely decision nodes and leaves. The decisions or the final results are the leaves and where the data is divided, the decision nodes. In a decision tree, the decision node and the leaf node are two nodes. Decision nodes are used for decision-making and multiple branches, while the Leaf nodes are the output and do not contain any other branches.

The choices or the evaluation are based on the unique characteristics of the dataset. It represents a graphic representation of all possible alternatives based on certain criteria to a decision / problem. The decision tree is named since, like the tree, the root node begins with the root tree growing on additional branches and forming a tree-like structure. A decision tree simply asks a question, and based on the answer (Yes / No), it further split the tree into subtrees. The general structure of a decision tree is shown in the following diagram:

DOI 10.17148/IJARCCE.2020.91105

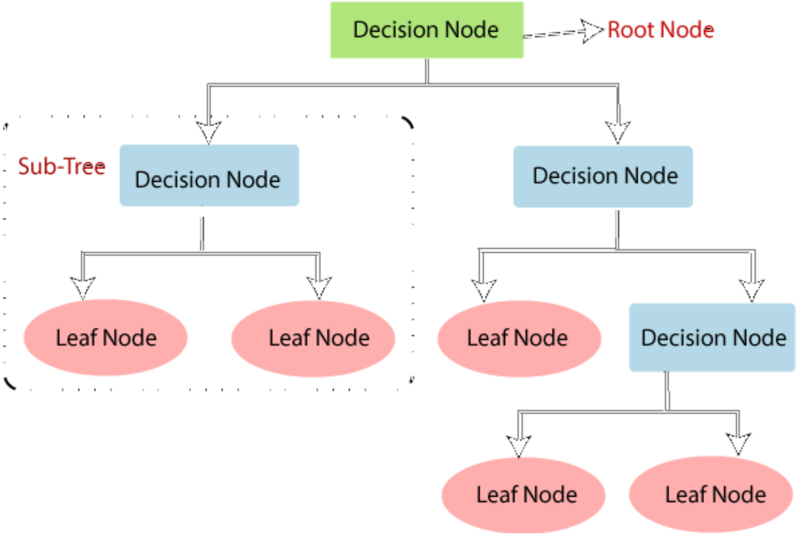

Figure 5. Decision Tree

Logistic regression - Logistic regression is one of the most common algorithms for machine learning, which is supervised learning. It is used with an independent set of variables to predict the category dependent variable. Logistic regression estimates the output of a categorical dependent variable. The result must either be categorical or confidential. It can be either yes or no; 0 or 1, true or false, but it gives the probabilistic values of 0 to 1 instead of defining the exact value as 0 and 1. Logistic regression is much like linear regression except how it is used. Linear Regression is used for solving Regression problems, while Logistic regression is used for solving the classification problems. We have an "S" formed logistic function that predicts the maximum two values (0 or 1) in the logistic regression, instead of fitting a regression line. The Logistic Function curve shows the probability that a mouse is obese or not depending on its weight. The cell is cancerous or not. Logistic regression is an essential algorithm for machine learning because it can provide possibilities and classify new data through constant and distinct datasets. Logistic regression can classify observations using different data types and quickly determine the most useful classification variables. The picture below shows the logistic feature:

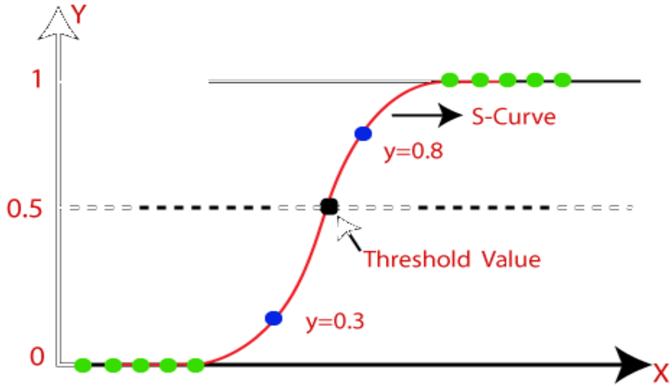

Figure 6. Logistic Regression

Equation for logistic regression:
The Equation for Logistic Regression can be derived from the Linear Regression equation. The logistic regression equations are described below in the mathematical steps:
We know the straight line equation can be written as follows:

$y = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + \cdots . + b_nx_n$ (3)

In logistic regression, only 0 and 1 may be used so that the above equation can be divided into (1-y):

$\frac{y}{1-y}; 0 \; for \; y = 0, and \; infinity \; for \; y = 1$ (4)

However, we have to range from -[infinity] to + [infinity], so to take the equation logarithm it will become:

$\log \frac{y}{1-y} = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + \cdots . + b_nx_n$ (5)

The above equation is the final equation for Logistic Regression.

Vol. 9, Issue 11, November 2020

DOI 10.17148/IJARCCE.2020.91105

IV. RESULTS AND DISCUSSION

We used a windows system with 4 GB Ram and 2.3 GHz processors for performing this work. Scikit learn is used to perform machine learning classification with the help of various libraries like NLTK, STOPWORDS, etc. for improving the accuracy of all the machine learning algorithms pipeline. After performing the statistical computation, deeper insights about the data were achieved. The data is being split into 70:30, where 70% of information is being used for training the model, and 30% is used to test the model. We have clinical text reports of 212 patients that are labeled into four classes. The classification was done using machine learning algorithms by supplying them with features extracted in the feature engineering step. To explore our model's generalization from training data to unseen data and reduce the possibility of overfitting, we split our initial dataset into separate training and test subsets. The tenfold cross-validation strategy was conducted for all algorithms. This process was repeated five times independently to avoid the sampling bias introduced by randomly partitioning the cross-validation dataset. Table 1 gives a comparative analysis of all the classical machine learning methods used to perform this task. Table 2 provides a comparative analysis of all the classical machine learning and Ensemble learning methods used to classify the clinical text into four classes.

The results showed that logistic regression and Multinomial Naı¨ve Bayes Algorithm shows better results than all other algorithms by having precision 94%, recall 96%, F1 score 95%, and accuracy 96.2%. Different algorithms like random forest and gradient boosting also showed promising results by having an accuracy of 94.3%. The visualized comparative analysis of all the algorithms used in our work is shown in Fig. 7. Since we all know, the COVID-19 data is least available. To get the real accuracy of the model, we experimented with it in two stages. In the first stage, we took 75% of the available data, and it shows less accuracy than the stage in which whole data was used for experimentation. So we can conclude that if more information is supplied to these algorithms, there are chances of performance improvement. As we face a severe challenge in tackling the deadly virus, our work will help the community by analyzing the clinical reports and taking necessary actions. Also, it was diagnosed that the COVID-19 patients report length is much smaller than other classes, and it ranges from 125 characters to 350 characters.

Table 1. Comparison of traditional machine learning algorithms

Algorithm	Accuracy	Precision	Recall	F1 score
Naïve Bayes	97.1	0.95	0.96	0.95
Support Vector Machine	91.1	0.81	0.91	0.86
Decision Tree	97.8	0.91	0.91	0.91
Logistic Regression	96.3	0.94	0.95	0.94

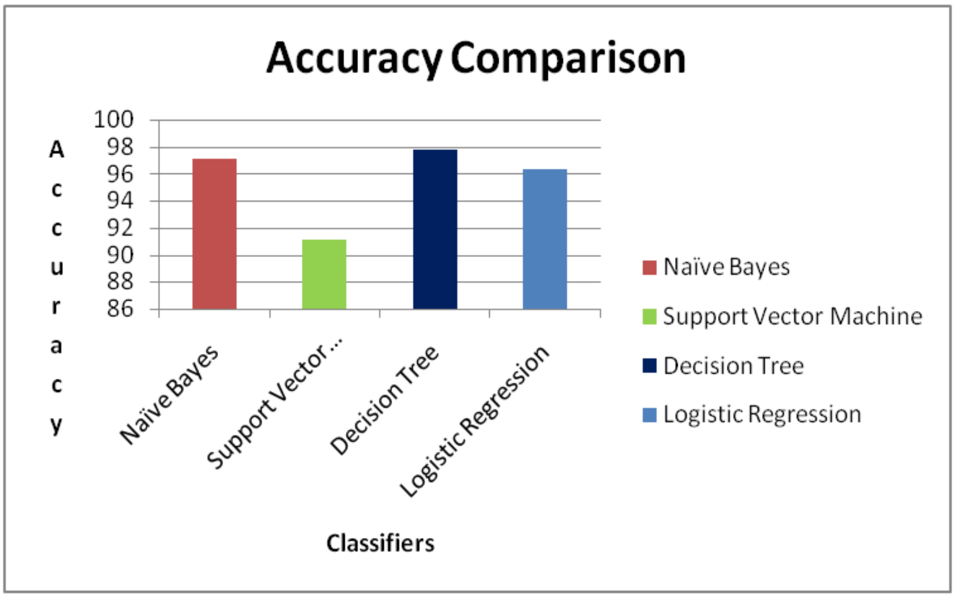

Figure 7. Accuracy comparison of classifiers

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 9, Issue 11, November 2020

DOI 10.17148/IJARCCE.2020.91105

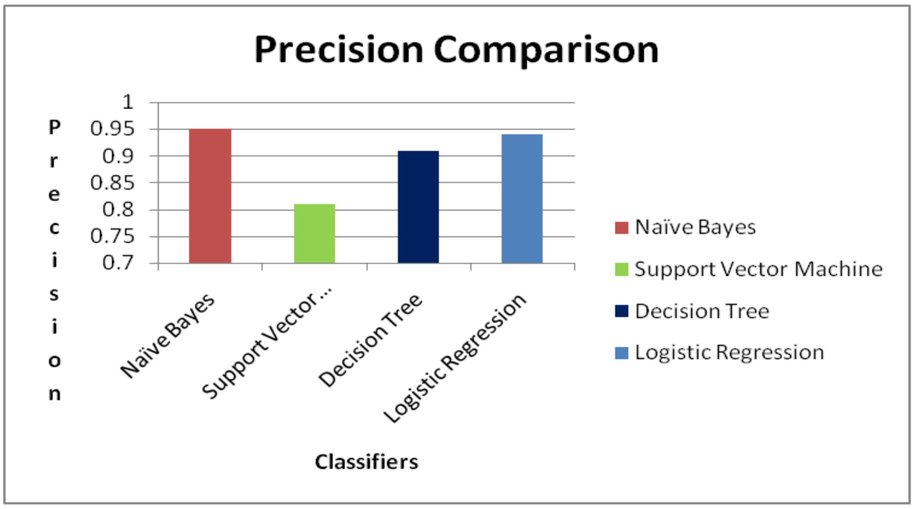

Figure 8. Precision comparison of classifiers

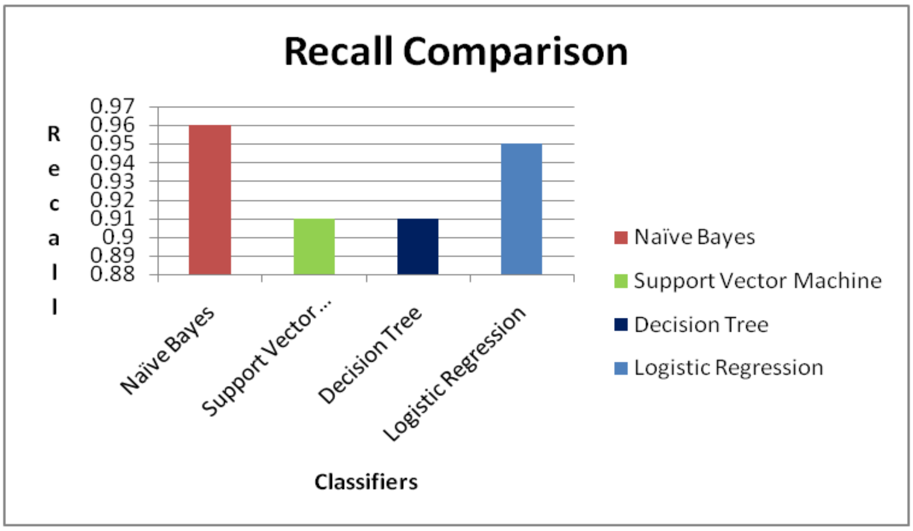

Figure 9. Recall comparison of classifiers

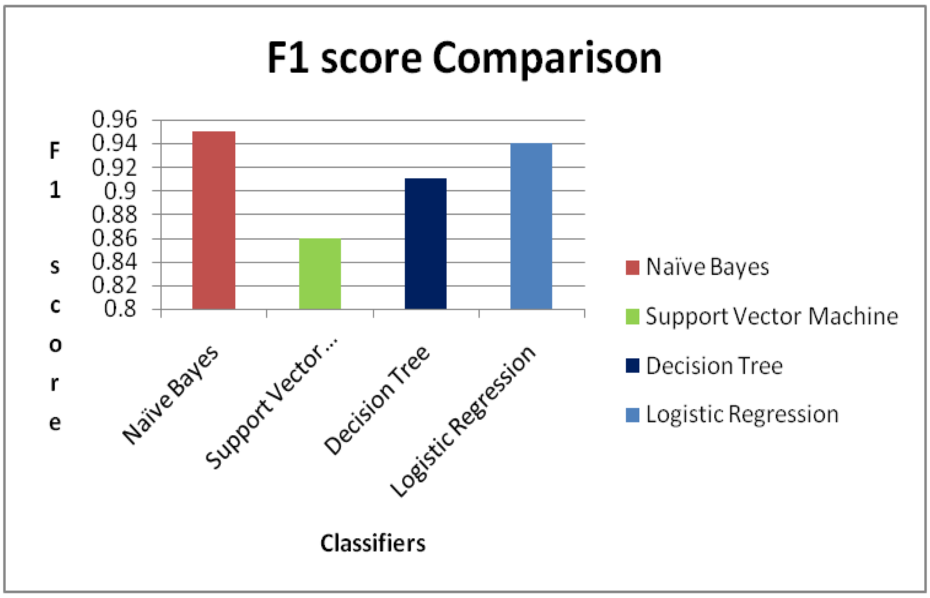

Figure 10. Accuracy comparison of classifiers

DOI 10.17148/IJARCCE.2020.91105

V. CONCLUSION AND FUTURE WORK

The world's COVID-19 has been surprised by the lack of vaccines or drugs available. Several researchers work for this lethal virus to be defeated. In four classrooms, we have used 212 clinical reports: COVID, SARS, ARDS, and both (COVID, ARDS). Many features are derived from these clinical studies, such as TF / IDF, word bags. For classifying clinical records in four separate groups, the machine learning algorithms are used. After conducting a grade, In this work we have performed the classification using Naïve Bayes, Support Vector Machine, Logistic regression and Decision tree. A logistic and multi-installation regression of Naı ̈ve Bayesian grades showed promising results with 95% precision, 96% recall, 95% f1 scoring, and 97.1% accuracy. and we have observed decision tree has outperformed other state of an algorithms with an accuracy of 97.8%. Before implementing the classification, feature engineering has also applied.

REFERENCES

[1] V. Chamola, V. Hassija, V. Gupta, and M. Guizani, "A Comprehensive Review of the COVID-19 Pandemic and the Role of IoT, Drones, AI, Blockchain, and 5G in Managing its Impact," *IEEE Access*, 2020.

[2] K. Liu *et al.*, "Clinical characteristics of novel coronavirus cases in tertiary hospitals in Hubei Province," *Chin. Med. J. (Engl).*, 2020.

[3] A. I. Khan, J. L. Shah, and M. M. Bhat, "CoroNet: A deep neural network for detection and diagnosis of COVID-19 from chest x-ray images," *Comput. Methods Programs Biomed.*, 2020.

[4] C. Long *et al.*, "Diagnosis of the Coronavirus disease (COVID-19): rRT-PCR or CT?," *Eur. J. Radiol.*, 2020.

[5] S. E. F. Yong *et al.*, "Connecting clusters of COVID-19: an epidemiological and serological investigation," *Lancet Infect. Dis.*, 2020.

[6] F. Yu *et al.*, "Quantitative Detection and Viral Load Analysis of SARS-CoV-2 in Infected Patients," *Clin. Infect. Dis.*, 2020.

[7] F. Ucar and D. Korkmaz, "COVIDiagnosis-Net: Deep Bayes-SqueezeNet based diagnosis of the coronavirus disease 2019 (COVID-19) from X-ray images," *Med. Hypotheses*, 2020.

[8] E. Article, "CT SCAN," *Diagnostic Radiol. Radiother.*, 2018.

[9] J. Chen *et al.*, "Clinical progression of patients with COVID-19 in Shanghai, China," *J. Infect.*, 2020.

[10] R. Nair and A. Bhagat, "Genes expression classification using improved deep learning method," *Int. J. Emerg. Technol.*, 2019.

[11] L. Shafner, A. Hanina, and A. Kalali, "Using artificial intelligence on mobile devices to measure and maximize medication adherence in CNS trials," *Neuropsychopharmacology*, 2016.

[12] U. K. Patel *et al.*, "Artificial intelligence as an emerging technology in the current care of neurological disorders," *Journal of Neurology*. 2019.

[13] A. Sarwar, M. Ali, J. Manhas, and V. Sharma, "Diagnosis of diabetes type-II using hybrid machine learning based ensemble model," *Int. J. Inf. Technol.*, 2020.

[14] P. Verma, A. M. U. D. Khanday, S. T. Rabani, M. H. Mir, and S. Jamwal, "Twitter sentiment analysis on Indian government project using R.," *Int. J. Recent Technol. Eng.*, 2019.

[15] A. Kumar, V. Dabas, and P. Hooda, "Text classification algorithms for mining unstructured data: a SWOT analysis," *Int. J. Inf. Technol.*, 2018.

[16] J. Del Gaizo *et al.*, "Using machine learning to classify temporal lobe epilepsy based on diffusion MRI," *Brain Behav.*, 2017.

[17] L. Yan *et al.*, "Prediction of criticality in patients with severe Covid-19 infection using three clinical features: a machine learning-based prognostic model with clinical data in Wuhan," *medRxiv*, 2020.

[18] X. Jiang *et al.*, "Towards an artificial intelligence framework for data-driven prediction of coronavirus clinical severity," *Comput. Mater. Contin.*, 2020.

[19] L. Wang, Z. Qiu, and A. Wong, "COVID-Net: un diseño de red neuronal convolucional profunda a medida para la detección de casos de COVID-19 a partir de imágenes de rayos X de tórax," *preimpresión de arXiv arXiv*, 2020.