International Journal of Current Research and Review
DOI: http://dx.doi.org/10.31782/IJCRR.2021.SP183

Scopus

Research Article

IJCRR
Section: Healthcare
ISI Impact Factor (2019-20): 1.628
IC Value (2019): 90.81
SJIF (2020) = 7.893

Detection of Recovery of Covid-19 Cases using Machine Learning

Joseph V. Raj[1], Joliz V. J. Anton[2], Johnson P. Durai Raj[3]

[1]Head, Research Department of Computer Science, Kamaraj College, (Affiliated to Manonmaniam Sundaranar University), Thoothukudi-628003, Tamil Nadu, India; [2]Assistant Professor in Electronics and Communication Engineering, Holycross Engineering College, (Affiliated to Anna University), Thoothukudi-628851, Tamil Nadu, India; [3]Postulate Info Tech, Thoothukudi-628001, Tamil Nadu, India.

ABSTRACT

Introduction: Classification is one of the most important research and applications of machine learning techniques. Research in the area of human-machine interaction and machine learning contributed to the success of Chatbots.

Objective: This research concentrates on some of the most important developments in machine learning classification research and the issues of Coronavirus Disease 2019 (COVID-19). Since December 2019, COVID-19 has been causing a massive health crisis all over the world resulted in 5,418,237 confirmed and 344,201 death COVID-19 cases to date (24.05.2020). Clinical experts say that COVID-19 patients to be diagnosed in early-stage to save their lives.

Methods: This study attempted to detect COVID-19 patients who can recover from the disease, using machine learning techniques, so that suitable treatment can be given to the patients to save their lives. Support Vector Machines (SVM), Artificial Neural Network (ANN), Decision tree, K- Nearest Neighbors (KNN), Random Forest and Logistic Regression algorithms are used to evaluate the classification performance.

Result and Conclusion: In this paper, a Chatbot was developed using the best algorithm evaluated to serve the society suffering from COVID-19.

Key Words: Machine learning algorithms, Chatbot, Classification, Feature Extraction

INTRODUCTION

On 31st December 2019, COVID-19 caused by coronavirus was reported and began to spread in Wuhan, China.[1] It became an epidemic that turned into a pandemic all over the world. The world was not ready to fight against this virus and the hospitals all over the world were full of corona virus-infected patients who needed treatment and so much care. Despite the policies and safety measures taken to control the spread of COVID-19, it has resulted in more confirmed and death COVID-19 cases all over the world to date possessing a serious health threat to the whole world. Unfortunately, there is no drug to treat COVID-19. Hence, the current treatment of COVID-19 focuses on supportive care and symptom relief.[2]

The COVID-19 particles are spherical in shape and the outer layer is made up of spike protein (lipid) and those spikes bind onto the host cell (human cell). Then it undergoes a structural change allowing the viral membrane to fuse with the cell membrane. This is how viral genes enter the human cell and produce more viruses.

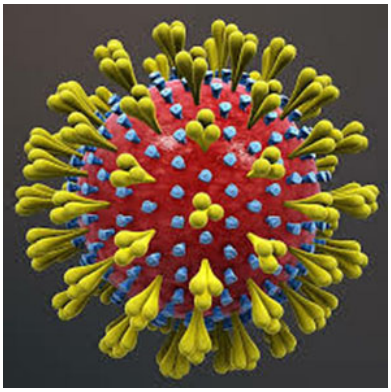

Figure 1: Structure of novel corono virus-2 which cause COVID-19.

Corresponding Author:

Joseph V. Raj, Head, Research Department of Computer Science, Kamaraj College, (Affiliated to Manonmaniam Sundaranar University), Thoothukudi-628003, Tamil Nadu, India; Mobile: +91 9443151625, Fax: +91 4651 250625; Email: v.jose08@gmail.com

ISSN: 2231-2196 (Print) **ISSN:** 0975-5241 (Online)

Received: 11.01.2021 **Revised:** 28.01.2021 **Accepted:** 19.02.2021 **Published:** 30.03.2021

Figure 1 represents the structure of a 2019-nCoV particle which is spherical and has an outer layer made up of spike protein. We need a survival analysis, to find out whether the COVID-19 patient will survive or not to treat them efficiently. Plenty of data related to COVID-19 is available today. Hence, we need machine learning algorithms to make computers learn from data [3] and perform the survival analysis (classify the survivors and non-survivors) for COVID-19 infected patients.

In this research, using the data, different methods have been proposed for treating survival analysis as a classification problem. Therefore, different machine learning classification algorithms such as artificial neural network, K-nearest neighbours, decision tree, random forest, logistic regression and support vector machine have been studied and used to construct data-driven modelling, classification and survival analysis. Finally, a Chatbot is developed using a high-performance algorithm.

LITERATURE SURVEY

The study of data (data analytics) using machine learning techniques has become an active research field. To automate the prediction process and to improve the quality of predictions, machine learning plays a great role. Several machine learning algorithms have been used to deal with real-world data. Over the past few years, some machine learning algorithms such as ANN, KNN, decision tree, random forest, logistic regression and SVM have been studied and proved to be effective for data analytics.[4-6] In addition to that, these algorithms have been greatly improved.

Artificial Neural Networks are a promising alternative to various conventional classification methods. Learning and generalization is the most important area in neural network research.[7-12] K-nearest neighbor is the simplest classification algorithm that can be used even there is no or little prior knowledge about the distribution of the data.[13,14] The decision tree is one of the predictive modelling approaches used in machine learning which faster and efficient compared with other machine learning algorithms.[15,16] A group of decision trees functions together as a committee forms the random forest classification algorithm. The concept behind the random forest classification algorithm is simple but powerful. A group of unrelated models (trees) functioning as a committee will outperform any of the individual constituent models.[17] Logistic regression with the help of the maximum likelihood method designs the best-fitting curve to maximize the probability of classifying the recognized data into a proper division.[18] SVM is a nonlinear and effective technique to classify all kinds of datasets. SVM has been proved to be efficient to construct data-driven modelling, classification and fault detection due to its better generalization ability and non-linear classification ability. It has been widely used for fault detection, classification of problems such as the face, object and text detection and categorization, information and image retrieval, etc. Among these, SVM is highly efficient to classify all kinds of data sets.[4] Finally, the need for human-machine interaction combined with machine learning helped to revival the Chatbots.[19] Recently many authors have pointed out the lack of proper comparisons between Machine learning techniques.[20-23]

MATERIALS AND METHODS

Since December 2019, a massive amount of data related to COVID-19 has been collected from a database (Kaggle) which is used to perform this research. Various features such as patient's age, sex, duration of confirmation (in days), systemic weakness, cough, discomfort, headache, fever, fatigue, diabetes, heart disease, respiratory disease, etc. are taken into consideration. The dataset is split into train and test sets. The train set is used to train and the test set is used to test the machine learning algorithms.

COVID-19 DATASET

Dataset-description

The dataset regarding the COVID-19 has been taken from open source Kaggle online repository. This repository consists of various combinations with different dimensions. Our experimented dataset, Admit Symptoms Discharge Summary (ASDS), totally has 13175 patient records. The features consist of age, gender, symptoms, travel exposure, chronic disease history and outcome. The dataset is divided into two major divisions such as patients recovered and died. In this research, this categorization is taken into consideration for the classification task.

These features are used as the learning variables by the classifiers. Before getting into the training task there should be a pre-processing step to make the dataset as standard normally distributed data.

Preprocessing

One of the common requirements for many machine learning estimators is the Standardization of the dataset. Because they might behave badly if the features are not standard normally distributed data (e.g. Gaussian with 0 mean and unit variance). For instance, many elements used in the objective function of a machine learning algorithm assume that all features are centred around 0 and have variance in the same order. If a feature has a variance that is orders of magnitude

larger than others, it might dominate the objective function and make the estimator unable to learn from other features correctly as expected. Standardize features by removing the mean and scaling to unit variance.

The standard score of sample x is calculated as equation 1:

$$z = \frac{x - u}{s} \quad (1)$$

Where, u is the mean of the training samples or zero if with_mean=False and s is the standard deviation of the training samples or one if with std=False.

Machine Learning Algorithms

Six machine learning algorithms such as artificial neural network, K-nearest neighbour algorithm, decision tree, random forest, logistic regression and support vector machine are used in this research.[12,13]

Artificial Neural Network

Artificial Neural Networks are a promising alternative to various conventional classification methods. The neural networks learn from samples without explicitly stating the rules and being non-linear they solve complex problems efficiently. Learning and generalization is the most important area in neural network research. Learning is the ability to approximate the underlying behaviour adaptively from the training patterns while generalization is the ability to predict beyond the training patterns. The generalization is a desirable and important feature because the common use of a classifier is to produce a good prediction on new or unknown samples.[18,19]

K-nearest neighbours

K-nearest neighbours are the simplest classification algorithm that can be used even there is no or little prior knowledge about the distribution of the data. The performance of a KNN classification algorithm is determined by the choice of k and the distance metric applied.

Decision tree

It is one of the predictive modelling approaches used in machine learning. A decision tree was named after its tree structure, where each node represents a test and each branch represents an outcome of the test. Each leaf (terminal) node represents the class label. The decision tree is faster and efficient compared with other machine learning algorithms. In this research, the decision tree classification algorithm obtains better accuracy, precision, recall and f1-scope comparing with other machine learning algorithms used. Figure 2 represents an example of a decision tree.

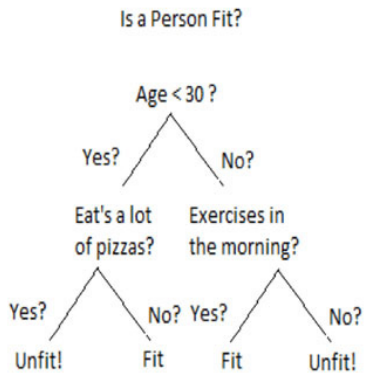

Figure 2: A Decision tree.

Random forest

Random forest, as its name denotes, has a large number of individual decision trees that act as an ensemble. Each decision tree in the random forest has its class prediction and the class with more votes becomes the prediction of the random forest. The concept behind the random forest classification algorithm is simple but powerful. A group of unrelated models (trees) functioning as a committee will outperform any of the individual constituent models.[15-17]

Logistic regression

Logistic regression is used to analyse the dataset and predict binary and categorical outcomes. The outcome is based on one or more independent variables. Logistic regression with the help of the maximum likelihood method designs the best-fitting curve to maximize the probability of classifying the recognized data into a proper division. Logistic regression regulates the impact of various autonomous variables that are conferred at the same time and predicts any one of the two independent categories of variables as an outcome.[17,18]

Support vector machine

SVM is a non-linear and effective technique to classify all kinds of datasets. Fault detection can also be done using SVM as a special classification problem involved in the model-based method and data-based method. With the help of the cross-validation technique, the parameters are optimized and the performance of the classification is enhanced.[19,20]

Performance measure

The performance of these six algorithms has been measured in terms of some performance measures such as accuracy, confusion matrix, recall, precision and f1-score.

Confusion matrix

The confusion matrix is a performance measurement for machine learning. It is a table with four different combinations predicted and actual values as shown in Figure 3.

TP	FP
FN	TN

Figure 3: Confusion matrix

True positive (TP) is what you predicted positive and it is true. True negative (TN) is what you predicted negative and it is true. False-positive (FP) is what you predicted positive and it is false. This is called a type 1 error. False-negative (FN) is what you predicted negative and it is false. This is called a type 2 error. The confusion matrix is used to measure recall, precision and f1 score.

$$\text{Re}call = \frac{TP}{TP + FN} \quad (2)$$

$$\text{Pr}ecision = \frac{TP}{TP + FP} \quad (3)$$

$$F1 - score = \frac{2 * \text{Re}call * \text{Pr}ecision}{\text{Re}call + \text{Pr}ecision} \quad (4)$$

Equation 2,3 and 4 are used to find recall, precision and f1-score values from the confusion matrix. The cross-validation technique is used to optimize the classifiers' hyper-parameters, to increase the performance of the classifiers to their best.[20-22]

Chatbot

The need for human-machine interaction combined with machine learning contributed to the revival of chatbots. Research and development in this field contributed to the introduction of chatbots on social media. In this research, a chatbot is developed using a machine learning algorithm, a decision tree (an algorithm that outperforms the other five algorithms used in this research). The results confirm the effectiveness of the proposed approach to human-machine interaction.[21, 22]

RESULTS AND DISCUSSION

To test the efficiency of different machine learning algorithms, ANN, KNN, decision tree (DT), random forest (RF), logistic regression (LR) and SVM have been used to analyse the COVID-19 related data set. Their performance has been monitored individually and compared in terms of some performance measures such as accuracy, precision, recall and f1-score.

Table 1: Analysis of different machine learning algorithms

	Accuracy	Precision	Recall	F1-Score
SVM	0.8	0.64	0.8	0.71
ANN	0.7	0.75	0.7	0.72
DT	**0.8**	**0.9**	**0.8**	**0.82**
KNN	0.6	0.6	0.6	0.6
LR	0.7	0.75	0.7	0.72
RF	0.8	0.64	0.8	0.71

Table 1 represents the analysis of different machine learning algorithms. For this COVID-19 survival analysis, the accuracy of SVM, ANN, decision tree, KNN, logistic regression and random forest are 0.8, 0.7, 0.8, 0.6, 0.7, 0.8 respectively. The precision values for SVM, ANN, decision tree, KNN, logistic regression and random forest are noted to be 0.64, 0.75, 0.9, 0.6, 0.75, 0.64 respectively. The recall values for SVM, ANN, decision tree, KNN, logistic regression and random forest are 0.8, 0.7, 0.8, 0.6, 0.7, 0.8 respectively. The f1-score values for SVM, ANN, decision tree, KNN, logistic regression and random forest are 0.71, 0.72, 0.82, 0.6, 0.72, 0.71 respectively.

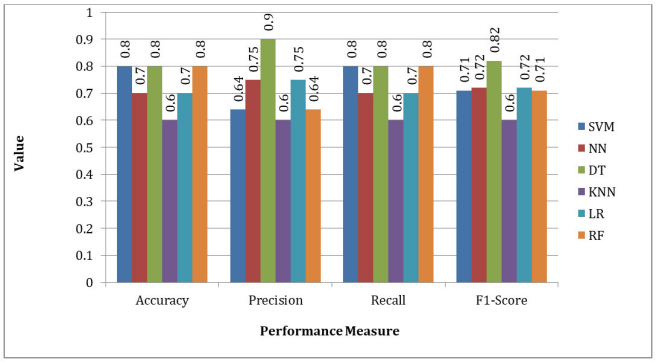

Figure 4: Performance of different machine learning algorithms.

Figure 4 represents the performance of different machine learning algorithms. From figure 4, it is clear that for this COVID-19 survival analysis problem, the decision tree classifier algorithm outperforms KNN, ANN, SVM, random forest classifier and logistic regression. From the result, it is concluded the decision tree classifier algorithm is of good predictive ability.

Future works

- Machine learning algorithms can be applied to study and analyse the COVID'19 related data and medicines to find better medicine for the corona virus
- Design of an efficient machine learning algorithm
- Improvement on Decision Tree
- Improvement of COVID'19 Database.

CONCLUSION

World Health Organization has characterized the COVID-19 situation as a pandemic. Hence, Preventive measures should be taken as soon as possible. This research has presented a focused study of several important recent developments in Machine learning techniques for COVID'19 related classification problems. These include Support Vector Machines, Artificial Neural Network, Decision Tree, K-Nearest Neighbours, Random Forest and Logistic Regression algorithms. It is proved that the performance of the Decision Tree is better than the other techniques in solving classification problems related to COVID-19. The Chatbot is designed using the best Machine learning technique, the Decision Tree. The administrators and Medical experts who have been fighting against COVID-19 can use the Chatbot to judge the condition of the patients to treat them suitably, save their lives and avoid misclassification treatment. In this research various Machine learning techniques are systematically evaluated and compared with each other. It is strongly believed that the multidisciplinary nature of Machine learning classification research produces more research activities and brings more fruitful results.

ACKNOWLEDGEMENT

I thank Kamaraj College, Holycross Engineering College and Postulate Info Tech for encouraging us to carry out this research. Financial contribution is given by the authors (Dr V. Joseph Raj, V. J. Joliz Anton and Dr P. Johnson Durai Raj) of this paper. Dr V. Joseph Raj and V. J. Joliz Anton conceived of the presented idea. V. J. Joliz Anton collected the data and performed the analysis. Dr V. Joseph Raj and Dr. P. Johnson Durai Raj contributed analysis tools and verified the analysis. V. J. Joliz Anton wrote the paper. All authors discussed the results and contributed to the final manuscript.

Conflict of interest: Nil

Source of Funding: Nil

REFERENCES

1. Chaolin H, Yeming W, Xingwang L, Lili R, Jianping Z, et al. Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. Lancet 2020;395:497-506.
2. Sen L, Qiang Z, and Zhiying W. Potential Covalent drugs targeting the main protease of the SARS-CoV-2 coronavirus. J Bio Info 2020;23(5):225-229.
3. Hector AC, Juan Z, Pedro AF, and Maria FA. A Machine Learning Approach for the Automatic Classification of Schizophrenic Discourse. Int J Acce 2019;7:45544–45553.
4. Shen Y, Xin G, Hamid RK, and Xiangping Z. Study on Support Vector Machine Bases Fault Detection in Tennessee Eastman Process. Appl Anal 2014;83(7):895-899.
5. Zhang GP. Neural Networks for Classification: A survey. IEEE Tranct Syst Man Cybernet Part C Appl Rev 2000;30(4):451-462.
6. Antonio V, John OS, Betsy WG, and Robert LC. Diagnosis of Ovarian Cancer using Decision Tree Classification of Mass Spectral Data. J Biomed Biotech 2003;5:308-314.
7. Amirikian B, Nishimura H. What size network is good for generalization of a specific task of interest. Int J Neur Netw 1994;7(2):321–329.
8. Baum EB. What size net gives valid generalization. Int J Neural Comput 1989;1:151–160.
9. Sietsma S, Dow R. Creating artificial neural networks that generalize. Int J Neu Netw 1991;4:67–79.
10. White H. Learning in artificial neural networks: A statistical perspective. Neur Compt 1989;1:425-464.
11. Jeneela MM, Joseph Raj V. Fuzzy based Genetic Neural Network for the Classification of Murder Cases Using Trapezoidal and Language Interpolation Membership Function. Int J Appl Soft Comput 2013;13(1):743-754.
12. Jenelle MM, Joseph Raj V. A Maximum Spanning Tree-based Dynamic Fuzzy Supervised Neural Network Architecture for Classification of Murder Cases. Int J Softw Compt 2016;20(6):2353-2365.
13. Sadegh BI, Mohammed B. Application of K-Nearest Neighbour (KNN) Approach for Predicting Economic Events: Theoretical Background. Int J Engg Res Appl 2013;3(5):605-610.
14. Gongde G, Hui W, David B, Yaxin B, and Kieran G. KNN Model-Based Approach in Classification. Int J Engg Res Appl 2003;12(6):986-996.
15. Himani S, Sunil KA. Survey on Decision Tree Algorithms for Classification in Data Mining. Int J Sci Res 2016;5(4):2094-2097.
16. Boris M, Milan M. Prediction and Decision making in Health Care using Data Mining. Int J Pub Hea Sci 2012;1(2):69-78.
17. Alessia S, Antonio C, Aldo Q. Random Forest Algorithm for the Classification of Neuroimaging Data in Alzheimer's Disease: A systematic review. Front Agi Neuro Sci. 2017; 9(5): 173-176.
18. Celine S, Maria Dominic M, and Savitha Devi M. Logistic Regression for Employment Prediction. Int J Inn Tech Expl Engg, 2020; 9 (3): 2471 - 2478.
19. Ievgen S, Galyna K, Pavlo K, and Yuriy K. Peculiarities of Human-Machine Interaction for Synthesis of the Intelligent Dialogue Chatbot. 10th IEEE Intr C Intelligent Systems: Tech Appl. 2019; 18 - 21.
20. Duin RPW. A note on comparing classifiers. Pattern Recognit Lett 1996;17:529-536.
21. Flexer A. Statistical evaluation of neural network experiments: Minimum requirements and current practice. Proc 13th Eur Meeting Cyber Syst Res 1996;1005-1008.
22. Prechelt L. A quantitative study of experimental evaluation of neural network algorithms: Current research practice. Neural Netw 1996;9(3):457-462.