An Enhanced Approach to Infer Potential Host of Coronavirus by Analyzing Its Spike Genes Using Multilayer Artificial Neural Network

Kamlesh Lakhwani
dept. of Computer Science & Engineering
Lovely Professional University
Punjab, India
kamlesh.lakhwani@gmail.com

Sandeep Bhargava
dept.of Computer Science & Engineering
Poornima College of Engineering
Jaipur, Rajasthan, India
eng.san83sandy@gmail.com

Devendra Somwanshi
dept of Electronics & Communication
Poornima College of Engineering
Jaipur,Rajasthan,India
imdev.som@gmail.com

Dr. Ruchi Doshi
dept.of Computer Science Technology
Jayoti Vidyapeeth Women's University
Jaipur, Rajasthan,India
doshiruchi18@ieee.org

Kamal Kant Hiran
dept. of Computer Science & Engineering
Sir Padampat Singhania University
Udaipur,Rajasthan,India
kamalhiran@gmail.com

2020 5th IEEE International Conference on Recent Advances and Innovations in Engineering (ICRAIE) | 978-1-7281-8867-6/20/$31.00 ©2020 IEEE | DOI: 10.1109/ICRAIE51050.2020.9358382

***Abstract*—Numerous coronaviruses are capable of transmitting interspecies. In recent years, transmission of coronavirus created a panic situation in the whole world. Therefore it is very important to infer the potential host of coro- navirus. In this research work nineteen parameters computed from the spike genes of coronavirus has been analysed to infer the potential host of coron- avirus. An enhanced multilayer neural network approach is proposed to analyse the data. The proposed model is compared with the other exiting statistical predictors like decision tree predictor, Support vector machine predictor and PNN predictor. All the model shown the higher accuracy such as 82.051 % by SVM predictor, 85.256% by PNN predictor,94.872% by decision tree predictor, and the highest accuracy 95.% is shown by proposed Multilayer Perceptron Predictor.**

***Keywords*—SVM Predictor, Decision Tree Predictor, PNN Predictor, Multilayer Perceptron, Artificial Neural Network.**

I. Introduction

Recently, there have been many life-threatening viruses. These viruses are responsible for causing substantial human causalities, as well as raising significant public health issues around the world. Because of the modern way of life, enormous travel of human and immense transportation of goods and chattels, their epidemic could theoretically be a concern anywhere in the world [1]. Normally the source and the host of the viruses are animals like camels, chimpanzees, and Bates. Other than in transmission from animal to human, the transmission has been recorded from human to human. Generally from pa- tient contaminated to stable person. While no particular treatment for their management has been recommended to date. The development of antiviral vaccines of such virus is in progress but supportive care has been shown to boost outcomes. Such novel viruses pose significant threats for public health in general, and especially for public health and infection control services. Om- nidirectional care, awareness, and education can reduce the spread of such viruses. It is therefore important to have accurate knowledge of their host, its transmission, presenting symptoms approach to their investigation and the best possible management along with preventive measures [1]. In this paper, the focus is given to infer the potential host of the coronavirus by analyzing the nucleotide composition calculated from the spike genes of the virus. For this purpose, an enhanced multilayer perceptron model is proposed. The efficiency shown by the proposed model is compared with the three exist- ing statistic models, SVM, PNN, and decision tree-based prediction models. The further paper has been organized into six subsequent phases, i.e. Basic Concepts, Literature Review, Dataset, Experimental Implementation, Result, and Conclusion respectively.

II. Basic Concept

To diagnose the patients suffering from coronavirus, it is very important to know the potential host of the virus. The potential host of the virus can be identified by analyzing the nucleotide composition of the spike genes. It is a very difficult task to identify the host of the virus by analyzing the spike genes sequence manually. Technology plays an important role in classification and approximation. To understand and to diagnose the disease caused by the in- fection of viruses, it is very important to develop intelligent computer-based models for the classification of the potential host of these viruses. Therefore various literature related to computer-based intelligent classification models has been reviewed in this article. Moreover, a multilayer neural network-based classification model has been proposed to predict the potential host of coronavirus. An appropriate number of literature and classification techniques have been reviewed and presented in the subsequent section.

III. Literature Review

Effective mid-December 2019, Novel coronavirus pneumonia (NCP) has been transmitted from human to human between close contacts. Therefore psycho- logical crisis intervention (PCI) among affected populations should be given greater attention to preventing inestimable harm from a secondary psycho- logical crisis on time [3] [4]. Two novel viruses have recently been involved to be responsible for serious acute disease, Severe Acute Respiratory Syndrome- Corona-Virus (SARS-CoV) and Middle East Respiratory Syndrome-Corona-Virus (MERSCoV) [4] [6]. In January 2016, 1638 human cases were registered by the WHO including 587 deaths from Middle East Respiratory Syndrome (MERS co-V) [6]. The article Al-Hazmi, 2016 reviewed the literature on different aspects such as the transmission, protection, and effectiveness of ther- apies used in patients with MERS-CoV and SARS infections [1]. In recent technological evolution, computer-based intelligent models shown significant performance in classification and approximation. For non-linear classification, SVM can be a suitable approach [2]. SVM uses the kernel level strategy that is rooted in systemic risk minimization and can effectively perform a nonlinear classification [5]. In the article [7] a qualitative diagnosis of the Pima-diabetes disease was made. A multilayer neural network structure trained by the Levenberg Marquardt (LM) algorithm and a probabilistic neural network structure was utilized for this purpose. The study findings were compared with the results published in previous research focusing on diabetes disease diagnosis and using the same UCI machine-learning database. Classification is an essential instrument of de- cision making, specifically in Health Sciences. As per the article [10], there are several classification methods exist, inappropriately many of todays strate- gies struggle to produce satisfactory performance. Over the last years, as an effective classification method, artificial neural networks were suggested [12]. In this article artificial neural network has been used to infer the onset of diabetes in Pima Indian Woman [11]. Neural network modeling capabilities of the proposed method was compared with conventional approaches such as logistic regression and a particular approach called ADAP that was used to predict diabetes. The findings show that neural networks are, indeed, a viable classi- fication method. In the article [5], in terms of breast cancer prediction, the output of single SVM classifier and SVM classifier ensembles obtained using specific kernel functions and various combination methods is analyzed. Infact, a compari- son of two separate scaled data sets is used. Besides, various classifiers are correlated with classification accuracy, ROC, F-measure, and the computa- tional time of the training. In the article [8] the probabilistic Neural Network (PNN) is used as a classification method. Moreover, an iterative method was proposed to train the PNN and the proposed procedure was tested on six sets of repository data. The proposed algorithms predictive potential is measured by measuring the accuracy of the test on 10 percent, 20 percent, 30 percent, and 40 percent of randomly drawn examples from each set of input data. It is demonstrated that the Q(0)-learning based method trained PNN is a classifier that can be regarded as one of the top models in data classification problems. As per the article [9] for classification, the Decision tree algorithm is an effec- tive method because no domain knowledge is needed for building decision tree classifiers. Moreover, hi-dimensional data can be handled through a decision tree model. The steps of decision tree induction in learning and classification are simple and fast. Users quickly assimilate their representation of the ac- quired information in tree form. Classification accuracy shown by the decision tree found better than the other methods. After reviewing the several pieces of related literature is it very clear that var- ious methods like SVM, PNN, Decision-Tree, and Neural-network are shown enhanced result in classifying the diseases. In this paper, a multilayer artificial neural network-based classification method is proposed to infer the potential host of coronavirus. For this purpose, the dataset required for training and testing the model has been described in the subsequence section.

IV. About Dataset

The dataset used to train and test the proposed model was prepared by Tang et al., 2015 during their research work that was based on the nucleotide com- position of spike gene sequences. The available dataset is in the form of a matrix containing twenty-three columns and seven-hundred- seventy-seven rows [2]. The number of rows signifies the total number of in- stances. The number of columns represents dependent and independent variables. Out of twenty-three variables three variables/columns having descriptive data, nineteen columns represent the independent variable and one variable is the dependent variable that is labeled as host. In this research work, the existing dataset has been analyzed by using different analytical approaches. Before the experiment, the dataset has been filtered by removing descriptive columns. The filtered dataset having a total twenty variables (1-dependent (host), 19-independent (nucleotide composition of spike gene sequences)).

A. Data distribution

The dataset instances having numeric values, that represent the nucleotide composition of the spike gene sequence of coronavirus. Each instance of the dataset is associated with their specific host. The dataset consists of a total of eleven different types of hosts. The association of host corresponding to instances are shown in Table-1. The name of columns in the dataset matrix and its associated lower and upper bound values are shown in Table-2. Table-2 consists of nineteen independent and one dependent variable, the value of that dependent variable is the string-type i.e. name of the associated host and its the target input for the neural network model.

V. Experimental Implementation

The proposed Multi-layer perceptron predictor has been implemented by using an open-source tools KNIME analytics

TABLE I
HOST CLASSES CORRESPONDING INSTANCES

Serial No.	Host Name	Count
1	alpaca	1
2	avian	173
3	bat	77
4	bovine	77
5	Camelus dromedarius	9
6	human	197
7	human and bovine	1
8	murine	28
9	palm civet	30
10	porcine	183
11	raccoon dog	1

TABLE II
DEPENDENT AND INDEPENDENT VARIABLE WITH CORRESPONDING LOWER(LB) AND UPPER BOUND(UB)

index	Variable Name	Type	LB	UB
1	G_frequency	Independent	0.17	0.23
2	C_frequency	Independent	0.14	0.27
3	T_frequency	Independent	0.28	0.43
4	AG_bias	Independent	0.83	1.12
5	AC_bias	Independent	0.94	1.5
6	AT_bias	Independent	0.76	1.04
7	GA_bias	Independent	0.61	1.01
8	GC_bias	Independent	1.01	1.39
9	GT_bias	Independent	0.85	1.25
10	CA_bias	Independent	0.96	1.5
11	CG_bias	Independent	0.25	0.78
12	CT_bias	Independent	0.95	1.25
13	TA_bias	Independent	0.67	1.2
14	TG_bias	Independent	1.1	1.49
15	TC_bias	Independent	0.58	0.97
16	AA_bias	Independent	0.73	0.95
17	GG_bias	Independent	0.77	1.07
18	CC_bias	Independent	0.65	1.02
19	TT_bias	Independent	0.68	0.86
20	Host	dependent	Target String	Target String

TABLE III
CONFIGURATION OF PROPOSED NEURAL NETWORK MODEL

Serial No.	Property	Value
1	Neural network function name	MLP_Predictor
2	Algorithm name	RProp
3	Activation Function	logistic
4	Normalisation method	none
5	Number of hidden layers	03
6	Number of neurons in output layer	11
7	Number of neurons in input layers	19
8	Number of hidden neurons per layer	18
9	Maximum number of iteration	1000
10	Size of Training dataset	80%
11	Size of Testing dataset	20%

TABLE IV
ALGORITHM OF PROPOSED MULTILAYERED ANN

Algorithm 1:	Multilayer Artificial Neural Network (MANN)
Input:	Matrix 0f twenty parameters with seven hundred rows
Output:	Learning model to predict the Host
Step 1:	Start
Step 2:	Set Nodes in Input layer := vector (x) Set Number of Hidden layers: three Set output nodes for output layer:= Vector (y) Set seed as random initialization Random set of weight and bias w and b Set Activation Function: logistic: l
Step 3:	Fun_MANN() : Training of MANN $\gamma = (w_2 l(w_1 x + b_1) + b_2) + b_3$
Step 4:	Perform feedforword for predicted y Output for feedforword calculus : $\gamma = w_3 l(w_2 l(w_1 x + b_1) + b_2) + b_3$
Step 5:	Apply Loss function: calculate SumOfSquereErrorSSE SSE= $\sum_{i=1}^{n} (y - \gamma)^2$
Step 6:	Apply backpropagation to minimize loss; Perform$_backpropagation(gradientdescent)and$ update weight and bias; LOSS= $\sum_{i=1}^{n} (y - \gamma)^2$ $\delta \frac{LOSS(y,\gamma)}{\delta \times w} = \delta \frac{Loss(y,\gamma)}{\delta \gamma} \times \frac{\delta \gamma}{\delta \beta} \frac{\delta \beta}{\delta w}$ where $\beta = wx + b$ Update weight
Step 7:	Repeat Step 6: to get minimum loss.
Step 8:	Stop.

platform. In medical sciences, classification is an important decision-making tool.

Artificial neural networks (ANN) can play an important role in classification, analysis and decision making. Es- pecially in the medical sciences, ANN is suggested as an alternative tool for classification. The configuration of proposed model is shown in table 3.

An ANN inspired multilayer perceptron predictor is created and analyzed in the subsequent section. The implementation of the proposed neural network architecture is shown in algorithm 1, and the Diagrammatic representation of proposed neural network architecture is depicted in the figure-1.

VI. TRAINING ERROR

The graphical view of the training-error of the proposed model is shown in Fig-2. The X-Axis of the graph shown in Fig.2 shows the number of iterations during the training of data whereas The Y-Axis of the graph shows the training error rate. From the figure-2, it can be observe that after 600 iteration the training error is lowest and it seems to be static. After the training the proposed model has been tested with reaming unseen dataset, and result has been compared with other predictors i.e. SVM, Decision Tree, and PNN based predictors.

VII. RESULT

In this work the dataset used in proposed model is also used for train and test the existing models i.e. SVM, PNN, and Decision-tree predictors. After testing the models classification accuracy has been calculated and the result has been stored and compared. The performance analysis between the SVM, PNN, Decision-tree and proposed model is shown in table-4. To visualise the result of accuracy and error rate, among the models used in this work graph has been plotted and shown in figure-3 and figure-4.

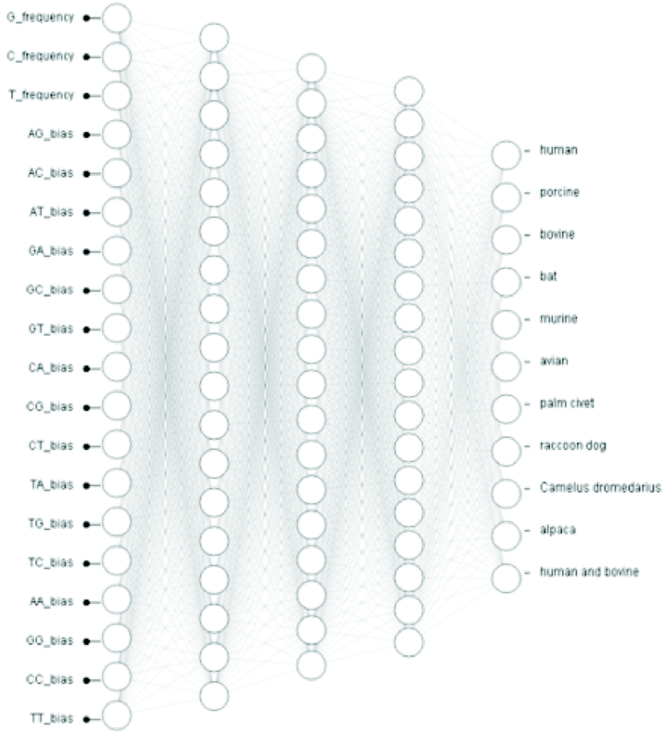

Fig. 1. Layered architecture of proposed neural network

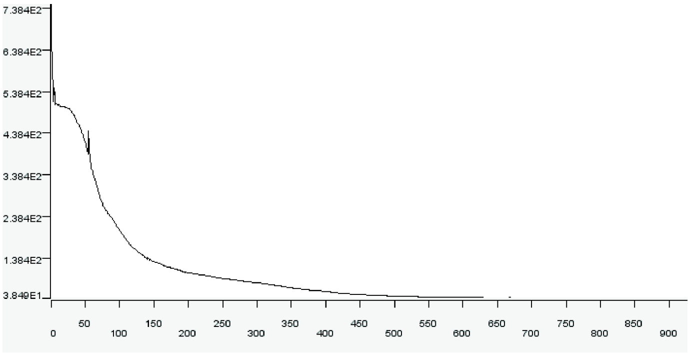

Fig. 2. Analysis of training error

VIII. Conclusion

This paper presents an enhanced approach (a multilayer perceptron predic- tor) to infer the potential host of coronavirus by analysing its nucleotide com- position of spike gene sequence. The proposed model was trained and tested with dataset-ratio 80% and 20% respectively. The testing result was compared with the three existing statistical predictors SVM, PNN, and Decision Tree predictors respectively. The accuracy, and error rate of all these predictors are calculated and compared for the same dataset. The result and outcome of this work is concluded in the following points:

- Classification is an important tool that can be used for analysis, especially in the medical field artificial intelligence and machine learning based clas- sifiers having a special significance.

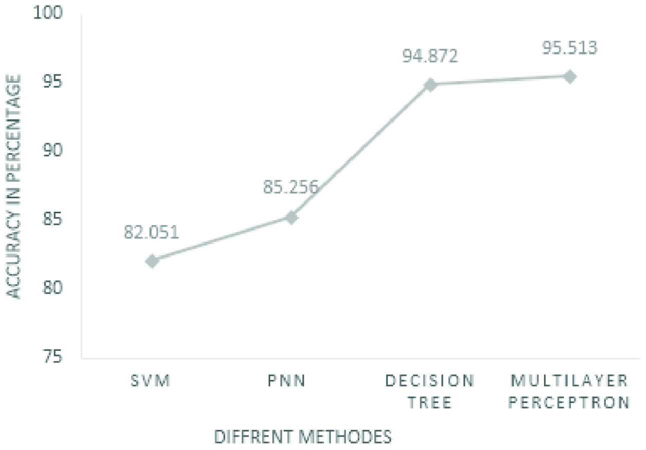

Fig. 3. Accuracy analysis of different prediction approaches

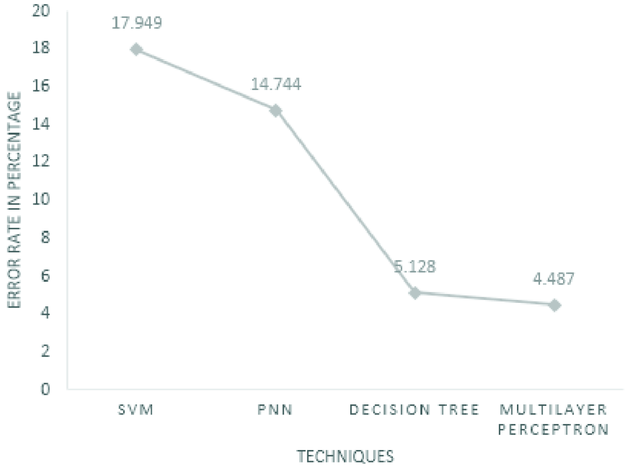

Fig. 4. Error rate analysis of different prediction approaches

- It is found that accuracy of SVM, PNN, Decision Tree and Multilayer Perceptron is 82.051%, 85.256%, 94.872%, 95.513% respectively.
- Although every predictor shown a good efficiency but multilayer perceptron shown highest accuracy than others.

To analyse the nucleotide composition of spike genes sequence of coron- avirus to identify its host, Multilayer perception model is the most suitable predictor and it shows the 95.513% accuracy.

References

[1] Al-Hazmi, A. (2016) Challenges presented by MERS corona virus, and SARS corona virus to global health, Saudi Journal of Biological Sciences. King Saud University, 23(4), pp. 507511. doi: 10.1016/j.sjbs.2016.02.019.

[2] Tang, Q. et al. (2015) Inferring the hosts of coronavirus using dual statistical models based on nucleotide composition, Scientific Reports, 5(7). doi: 10.1038/srep17155.

[3] Corman, V.M.et.al. (2013) Detection of a novel human coronavirus by real-time reverse- transcription polymerase chain reaction, Middle East Respiratory Syndrome Coronavirus (MERS-CoV), 12, 30.

[4] Meyer, B., Garc a-Bocanegra, I., Wernery, U., Wernery, R., Sieberg, A., Mu ller, M.A., Eckerle, I., 2015. Serologic assessment of possibility for MERS-CoV infection in equids. Emerg. Infect. Dis. 21 (1), 181.

[5] Huang, M.-W. et al. (2017) SVM and SVM Ensembles in Breast Cancer Prediction, PLOS ONE. Edited by E. Hernandez-Lemus. Public Library of Science, 12(1), p. e0161501. doi: 10.1371/journal.pone.0161501.

[6] Jiang, X. et al. (2020) Psychological crisis intervention during the outbreak period of new coronavirus pneumonia from the experiences of Shanghai Elsevier B.V., p. 112903. doi: 10.1016/j.psychres.2020.112903.

[7] Temurtas, H., Yumusak, N. and Temurtas, F. (2009) A comparative study on diabetes disease diagnosis using neural networks, Expert Systems with Applications. Elsevier Ltd, 36(4), pp. 86108615. doi: 10.1016/j.eswa.2008.10.032.

[8] Kusy, M. and Zajdel, R. (2014) Probabilistic neural network training procedure based on Q(0)-learning algorithm in medical data classification, Applied Intelligence, 41(3), pp. 837854. doi: 10.1007/s10489-014-0562-9.

[9] Patel, B. N. (2012) Efficient Classification of Data Using Decision Tree, Bonfring Inter- national Journal of Data Mining, 2(1), pp. 0612. doi: 10.9756/bijdm.1098.

[10] Shanker, M. S. (1996) Using neural networks to predict the onset of diabetes mel- litus, Journal of Chemical Information and Computer Sciences, 36(1), pp. 3541. doi: 10.1021/ci950063e.

[11] Temurtas, H., Yumusak, N. and Temurtas, F. (2009) A comparative study on diabetes disease diagnosis using neural networks, Expert Systems with Applications. Elsevier Ltd, 36(4), pp. 86108615. doi: 10.1016/j.eswa.2008.10.032.

[12] Bhargava, S. and Choudhary, S., 2018, May. Behavioral Analysis of Depressed Sentimental Over Twitter: Based on Supervised Machine Learning Approach. In Proceedings of 3rd International Conference on Internet of Things and Connected Technologies (ICIoTCT) (pp. 26-27).