Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 5, 2021, Pages. 1821 - 1826
Received 15 April 2021; Accepted 05 May 2021.

Effective Algorithm for Identification of Covid Infected People Using Machine Learning

Dr.R. Gayathri[1], Dr.T. Prema Kumari[2], M.K. Kamali[3], N. Arthy[4], A. Jayavanth Kumar[5]

[1]Department of Electronics and Communication Engineering, Sona Colleg of Technology, Anna University, Salem, India. E-mail: profgayathri@gmail.com

[2]Department of Electronics and Communication Engineering, Sona College of Technology, Anna University, Salem, India. E-mail: premakumarit@sonatech.ac.in

[3]Department of Electronics and Communication Engineering, Sona College of Technology, Anna University, Salem, India. E-mail: Kamali0699@gmail.com

[4]Department of Electronics and Communication Engineering, Sona College of Technology, Anna University, Salem, India. E-mail: arthynagaraja2712@gmail.com

[5]Department of Electronics and Communication Engineering, Sona College of Technology, Anna University, Salem, India. E-mail: jaiwanth14@gmail.com

ABSTRACT

Our ManKind has Observed various pandemics throughout history where some of them were more disastrous than the others to humans. We are observing a very tough time once again fighting an invisible enemy the novel COVID – 19 coronavirus. Many have lost their lives and their normal lifestyle, due to its rapid spreads across the world because of its unpredictability at the early stages. It has been observed that the Deep Learning Algorithms play an efficient role in predicting the probability that a person gets infected by the virus or not by using the data of the user so that precaution can be taken at the early stage before spreading. The 10000 datasets have been used to train this model which has been collected from the Kaggle website. The deep learning model has been created with different hidden layers besides the best optimization techniques to make the model learn from the data more efficiently. The proposed predicting technique provides 93 % accuracy in coronavirus Prediction.

KEYWORDS

Backpropagation, Adam Optimizer, Hidden Layers, Neurons, Rectified Linear Unit (Relu Function).

Introduction

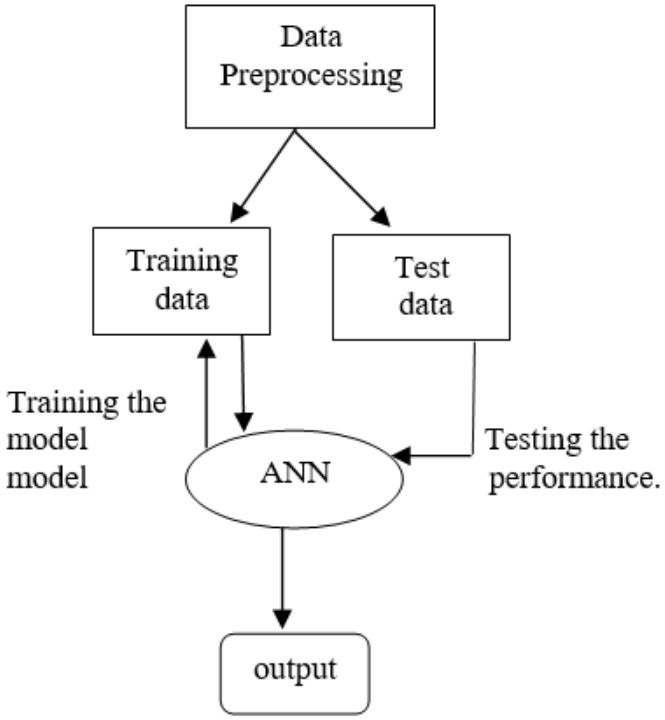
Mankind has faced many pandemic situations over the history right from natural disasters to the deadly viruses, which have taken many lives and we humans overcame each of this pandemic situation with many struggles and now at the start of 2020, we are at another pandemic situation a new deadly virus which has taken many lives right from the start of the year across the world. Coronavirus, the early report of coronavirus infection on animals occurred in the late 1920s and its first occurrence on humans was discovered in the late 1960s. It infects mostly mammals and birds. The virus spread from one to other when one comes in direct contact with the infected one and through the air from the infected one, through which the virus enters into the host body and when the viral spike protein binds to its complementary host cell receptor, infection begins. The receptor-attached spike protein is cleaved and activated by a protease in the host cell after attachment. Cleavage and activation, depending on the host cell protease available, allow the virus to enter the host cell via endocytosis or direct fusion of the viral envelope with the host membrane. On entering into the host cell, the virus releases its genome which acts as messenger RNA which is translated directly by the host cell and starts to create multiple virus cells. This virus on humans mostly infects the epithelial cells of the respiratory tract, causing it difficult to breathe for the host. In the case of animals, it generally infects the epithelial cells of the Digestive tract. Many Human coronaviruses have their origin from bats. This virus causes cold with many other symptoms like fever, sore throat, and even in severe case pneumonia at last death. There are six species of human coronaviruses are known 1. Human coronavirus OC43(HCoV-OC43), 2. Human coronavirus HKU1(HCoV – HKU1), 3.Human coronavirus 229E (HCoV-229E), 4. Human coronavirus NL63(HCoV-NL63),5.The Middle East respiratory syndrome-related coronavirus(MERS-CoV), 5. Severe Acute respiratory syndrome coronavirus(SARS-CoV), 7. Severe acute respiratory syndrome coronavirus(SARS-CoV-2). The main reason for the rapid spread of this virus is due to the unpredictability of its presence at the early stage, only when the

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 5, 2021, Pages. 1821 - 1826
Received 15 April 2021; Accepted 05 May 2021.

person gets infected by the virus and after the outcome of the symptoms, one comes to know that he is infected, within that it starts spreading from the infected one to the society. This paper proposes a method through which using the user data we can predict the chances that a person will get affected by the virus or not. A deep learning model is created and trained on the dataset which contains more than 10,000 data consist of both infected and non-infected persons, such as blood sugar level, Haemoglobin level, platelet level, etc.

Proposed Methodology

The proposed method involves two stages, the training stage, and the test stage. The training stage is used to train the created model and the test stage is used to test the created model performance on new data. The data which is used taken from the Kaggle Website which contains 10,000 data with 32 attributes.



1. Data Preprocessing

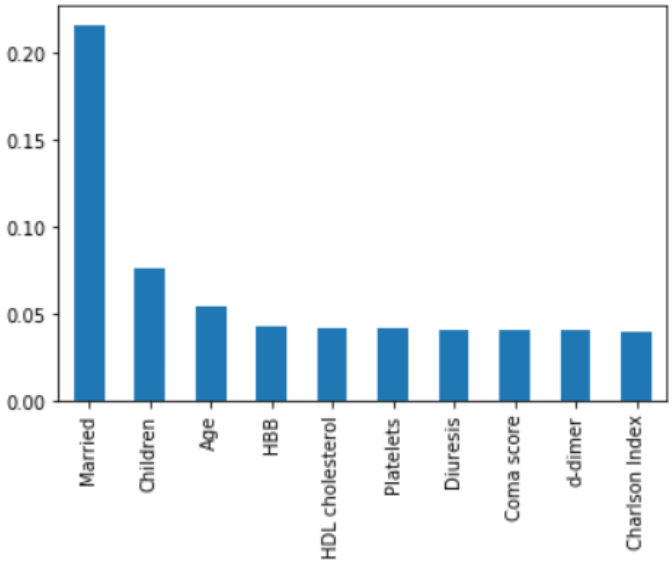
The data Preprocessing method involves several methods such as data cleaning, data encoding for categorical data, data scaling. The Data is cleaned by checking if there are any null values or not and if it then replacing the null values with its mean value for numerical attribute and categorical attribute replacing the null value with its most frequent value.

After cleaning the data, the categorical data are encoded using the LabelEncoder and OneHotEncoder. The encoding process is required because the model can only learn from numerical data more efficiently.

2. Data Analysis

After completion of the Data cleaning phase, the analysis is done on the data, to know more about the data i.e how much each attribute correlated with the output. And Plotting the top 10 features which contribute most to the output among the other 37 features.

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 5, 2021, Pages. 1821 - 1826
Received 15 April 2021; Accepted 05 May 2021.



From the above chart, it is clear that those who are married and have children contribute most to the output. By analyzing each feature with the output, the non-correlated feature with the output is removed. So that model performance is improved.

3. Data Scaling and Encoding

The scaling is performed on the pre-processed data. The scaling is the process used to normalize the range of Independent variables so that model can able to learn and train more efficiently and fastly. Many Normalization techniques are available in this paper MinMaxScaler is used.

$$X_{new} = \frac{X_i - \min(X)}{\max(x) - \min(x)}$$

After Scaling the categorical data are encoded using two Encoding Techniques, LabelEncoder and OneHotEncoder. First, the categorical features are LabelEncoder, but there is a disadvantage with LabelEncoder that there are different numbers in the same column, the model will misunderstand the data to be in some kind of order. To overcome this problem after LabelEncoder, One Hot Encoder is performed.

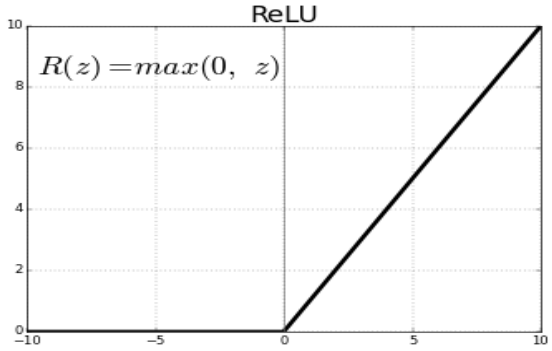
4. Data Splitting

The Data is then split into training data as well as test data. The training data is used to train the model, while the test data is used to evaluate its performance. The training and test data split is always an 8:2 ratio. The training data should be more because more data is required to train the data.

5. ANN Model Creation

The Artificial Neural Network mimics the human brain. Like the human brain, each node in Artificial Neural Network represents a neuron and every neuron is connected with a randomly initialized weight just like the human brain. There are three layers of a neuron's input layer, hidden layer, and output layer. Before the data is transferred from one neuron to another neuron it is passed through the activation function. There are several activation functions such as ReLu, sigmoid, etc. In this proposed work ReLu activation function is used since we are predicting the probability. The ReLu activation function is a Linear Function.

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 5, 2021, Pages. 1821 - 1826
Received 15 April 2021; Accepted 05 May 2021.



The hidden layer is used to extract data characteristics that aid the learning of the model, so if it is needed to extract more features then the number of hidden layers is increased.

Then the data is fed to the model for training and the loss function which is used mean squared error to measure the loss during the training phase and to decrease the loss function there are various optimization techniques such as gradient descent, Stochastic Gradient Descent, Adam optimizer, etc. In this proposed word we have used Adam Optimizer.

6. Adam Optimizer

RMSprop and Stochastic Gradient Descent with momentum are combined in Adam Optimizer. It employs an adaptive learning rate approach, which determines individual learning rates for various parameters. The Adam optimizer adjusts the learning rate for each weight of the neural network by estimating the first and second moments of the gradient. The N-th moment of a random variable is defined as the expected value of that variable to the power of n

$$m_n = E[X^n]$$

To estimates the moments, Adam employs exponential moving averages, which are calculated based on the gradient of a current mini-batch.

$$m_t = \beta_1 m_{t-1} + (1 - \beta_1) g_t$$
$$v_t = \beta_2 v_{t-1} + (1 - \beta_2) g_t^2$$

where m and v are moving averages, g is a gradient on current mini-batch and betas newly introduced hyperparameters of the algorithm. The vectors of moving averages are initialized with zeros at the first iteration. The Excepted values of the estimators should equal the parameters we are trying to estimate.i.e,.

$$E[m_t] = E[g_t]$$
$$E[v_t] = E[g^2_t]$$

And it moves on for each step and the final formula for moving average is

$$m_t = (1 - \beta_1) \sum_{i=0}^{t} \beta_1 g_i^{t-i}$$

$$E[m_t] = E[(1 - \beta_1) \sum_{I=1}^{t} \beta_1^{t-i} g_i]$$

$$= E[g_i] ((1 - \beta_1) \sum_{i=1}^{t} \beta_1^{t-I} + \delta$$

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 5, 2021, Pages. 1821 - 1826
Received 15 April 2021; Accepted 05 May 2021.

Now we take the summation term out because it does not depends on i

$$= E[g_i] (1- \beta_1) + \delta$$

The final formula for estimator will be

$$m_t = \frac{m_t}{1- \beta^t_1}$$

$$V_t = \frac{v_t}{1- \beta^t_2}$$

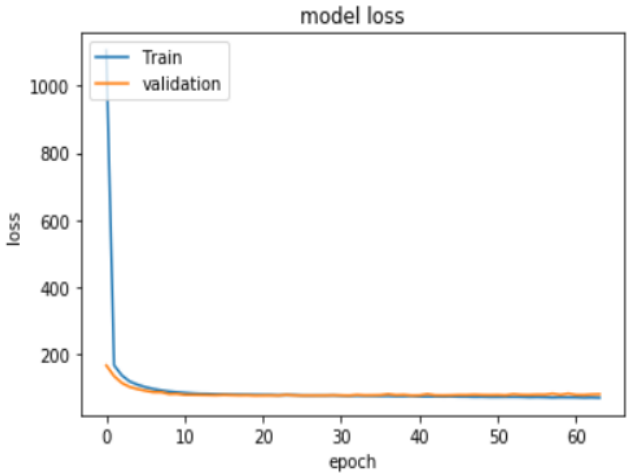
and the weight updation is done as follow

$$w_t = w_{t-1} - \eta \frac{m_t}{\sqrt{v_t + \epsilon}}$$

Where w is model weights is the step size.

Result

The model is trained on the preprocessed data for different epochs to reduce the mean Squared error function for which the adam optimizer is used. The final result of the loss function of train data concerning the validation loss decreases gradually as shown in the below graph.



After the training phase, is completed. The model doesn't have any prior information regarding the test data. The test data is then given to the model and prediction is made. From the predicted reading we calculate the accuracy of the model using the MSE(mean squared error)

$$MSE = 1/n \sum(y - y\hat{})^2$$

Where y is the actual value, y^ predicted value from the model. The calculated accuracy of the model is 93 %.

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 5, 2021, Pages. 1821 - 1826
Received 15 April 2021; Accepted 05 May 2021.

Future Scope

Our paper proposed a method to predict the probability that a person will get infected by the virus or not. This methodology helps to effectively stop the spread of the virus. Further, the method performance can be involved by using various AI techniques, and in the future, it can be developed as a GUI which can be used as a real-time application.

References

[1] Artificial Neural Networks and their Application National Conference on 'Unearthing Technological Developments and their Transfer for Serving Masses' GLA ITM, Mathura India 17- 18 April 2005 by Nitin Mali.

[2] A view of Artificial Neural Network, 2014 *International Conference on Advances in Engineering and Technology Research (ICAETR-2014).*

[3] Dionysis Goularas, Sani Kamis, "Evaluation of Deep Learning Techniques in Sentiment Analysis from Twitter Data", 2019 *Internation Conference on Deep Learning and Machine Learning in Emerging Applications (Deep-ML).*

[4] J.Koay, C.Herry, M.Frize, "Analysis of breast thermography with an Artificial Neural Network" *The 26th Annual International Conference of the IEEE Engineering in Medicine and Biology Society.*

[5] L. Ozyilmaz, T. Yildirim, "Artificial Neural Networks for diagnosis of Hepatitis disease", *Proceedings of the International Joint Conference on Neural Networks 2003 (IEEE).*

[6] Mohammed Ishaque, Ladislav Hudec, "Feature extraction using Deep Learning for Intrusion Detection System", *2019 2nd International Conference on Computer Applications and Information Security (ICCAIS).*

[7] Oludare Isaacu Abiodun, Aman Jantan, Abiodun Ester Omolara, "Comprehensive Review of Artificial Neural Network Application to patter Recognition, 2019 Journal Article Volume 7.

[8] Salvatore Caorsi, Claudio Lenzi, "Skin Removal Techniques for breast cancer radar detection based on Artificial Neural Networks", 2015 *IEEE 15th Mediterranean Microwave Symposium (MMS).*

[9] T.Ae, R. Aibara, Y. Nishioka, "A Memory Based Artificial Neural Network", *1991 IEEE International Joint Conference on Neural Networks.*

[10] Zaiying Wang, Bohao Song, "Research on hot news classification algorithm based on deep learning", *2019 IEEE 3rd Information Technology, Networking,Electronic and Automation Control Conference (ITNEC).*