An Autoencoder based Decision Support System for predicting Covid-19 - A Non Clinical based approach

Dr.M.Subathra[1*], Dr.V.Umarani[2]
[1] *Assistant Professor, Department of Computer Applications, PSG College of Technology , Coimbatore, Tamilnadu, India*
[2] *Assistant Professor, Department of Computer Applications, PSG College of Technology , Coimbatore, Tamilnadu, India*
[1]msa.mca@psgtech.ac.in, [2]vur.mca@psgtech.ac.in

Abstract

The detection of covid-19 is significant for improving the exactness of the medicinal service sectors. Enormous efforts and works have been undertaken towards developing deep learning techniques in solving the covid-19 crisis. In this study, a deep learning based characterization framework for covid-19 identification is proposed. This framework utilizes multinomial regression with auto encoder to settle on a choice about the covid-19. Anova f-test and mutual information is employed for feature selection. The proposed auto encoder with highlights determination is utilized for dimensionality reduction. The experiments are tested using the covid-19 pre-condition open dataset that are available in kaggle repository. The experimental results reveal that the proposed framework improves the normal exactness of the covid-19 prediction with the average of 99% in contrast with that of the estimated accuracy.

Keywords: *Autoencoder, multinomial regression, Anova f-test, mutual information*

1. Introduction

In recent times, machine Learning and deep learning techniques have been considered to design automatic diagnosis system for Covid-19.The rapid growth of the availability of healthcare related data poses challenges such as the extraction of important and valuable data from those information. And there exists an urgent need in the healthcare industry to predict the possibility of the disease and to reduce the amount of cumbersome tests on patients that serves as an efficient computational analysis tool. The result of these strategies would support the doctor and the clinical researchers to foresee the chance of the infection.

Corona virus (Covid-19) is an infection caused by a virus that can spread from person to person. Covid-19 is a new corona virus that has rapidly increasing throughout the world. Its symptoms can range from mild to severe illness. As per World Health Organization [14], the quantity of individuals with Corona -19 has prolonged throughout the year.

Many models have been created that have enough intelligence to properly classify the patent's record as either Positive or Negative in regard to Covid-19 detection. Out of the deep learning algorithms, Auto encoder is one of the techniques that is used to learn a compressed, distributed representation for a set of data. Hence the objective of this proposed work is to address the problem by employing an auto encoder with logistic regression algorithms for the prediction of covid-19 affected patient.

The rest of the paper is organised as follows. Section 2 describes related works in deep learning algorithms that are available in the literature towards prediction and diagnosis of covid-19. Section 3 discusses the problem description. Section 4 deals with applied methodology. Section 5 shows the experimental analysis. Section 6 is devoted to results and discussions. Finally, Section 7 concludes the paper.

[*]M.Subathra. E-mail: msa.mca@psgtech.ac.in

2. Related works

There have been plenteous researches that have been carried out in the healthcare domain mainly due to the large amount of healthcare related data that are available. Due to the availability of large volume of data it is important to process those data's to gain some valuable information. With paramount growth of data in healthcare sectors, accurate analysis of medical data aids in early disease detection of disease and timely patient care. However, the analysis and accuracy is reduced when the quality of medical data is incomplete. Moreover, different regions show signs of unique characteristics of certain regional diseases, which may perhaps weaken the prediction of disease outbreaks.

Mallick et al [1] have demonstrated wonderful performance in classification and segmentation task. Carrying this idea into consideration, in this paper, a technique for image compression using a deep wavelet, which blends the basic feature reduction property of along with the image decomposition property of wavelet transform, is proposed

Macías et al [2] have used s based neural networks to improve the quality of the data for developing immune histo chemistry signatures with prognostic value in breast cancer.

Mienye et al [3] have developed parse auto encoder based Artificial Neural Network (ANN) is proposed to aid the prediction of heart disease. The sparse auto encoder was utilized to learn the best representation of the data whereas ANN was used to make predictions based on the learned records. Adam algorithm was used for optimization and their model was able to achieve 90% on test data.

Siddique et al [4] have demonstrated the performance of deep and Neighbourhood Components Analysis(NCA) dimensionality reduction methods with K-Nearest Neighbors(KNN), Extended Nearest Neighbor(ENN) and Support Vector Machine(SVM), Extended nearest neighbor with deep exhibited the highest accuracy of 91.9, when the NCA dimensionality reduction technique is adapted, SVM generally outperforms both KNN and ENN classifiers and exhibited the highest classification accuracy of 94.52%

Vinay et al [5] developed a forecasting model of Covid-19 outbreak in Canada using state-of-the-art Deep Learning models. They have proposed a novel research where they have evaluated the key features to predict the trends and possible stopping time of the current Covid-19 outbreak in Canada and around the world. The authors have presented the LSTM networks, a DL approach to forecast the future Covid-19 cases and have showed the trends of different countries and compared them with Canadian data to predict the future infections.

Xayasouk et al [6] have developed models to predict fine Particulate Matter (PM) concentrations using Long Short-Term Memory (LSTM) and deep auto encoder methods, and compared the model results in terms of root mean square error. These proposed models effectively predicted fine PM concentrations, with the LSTM model showing slightly better performance.

Xie et al [7] have proposed deep learning model which is a regression-based predictive model based on the MultiLayer Perceptron and Stacked denoising Auto-Encoder (MLP-SAE). The model is trained with a help of stacked denoising auto-encoder for feature selection and a multilayer perceptron framework for backpropagation.

Shahin et al [8] proposed an autoencoder based semi-supervised learning methodology to extract the infected legions in chest X-ray manifestation of Covid-19 and other Pneumonia-like diseases was considered. Highly-tailored deep architecture was used to extract the relevant features for training a classifier to perform the task of automatic diagnosis.

Schonecker et al [9] have proposed multinomial logistic regression model that was able to accurately classify 97.1% parkinsonian patient's syndromes. On the other hand, diagnostic accuracy is essential for estimating the accuracy of prognosis and in the process of optimizing patient care, and allocation to therapeutic trials.

2. Problem Description

Classification is a supervised machine learning technique [11] that creates an intelligent program to classify the new instances by assigning the correct label with the help of labeled instances. The accuracy of the program is calculated with the help of the properly classified instances.

Each instance is defined as P(X, Y), where X consists of a set of features of the instance, and Y consists of labels. In Covid-19detection, every record is considered as an instance $P=\{X_1, X_2, ..,X_d\}$, where d is a dimension of the vector. The vector is denoted by the score values of the feature set.

It is a binary classification problem defined by a problem space S over X × Y, where X= {X1, X2, XN), where X1= $\{f_1, f_2, f_3,.. ,f_{14}\}$, Y= {0, 1} and N is the number of instances. The classifier classifies the given sample N, so that it belongs to the class using the classifier threshold value. The objective of this

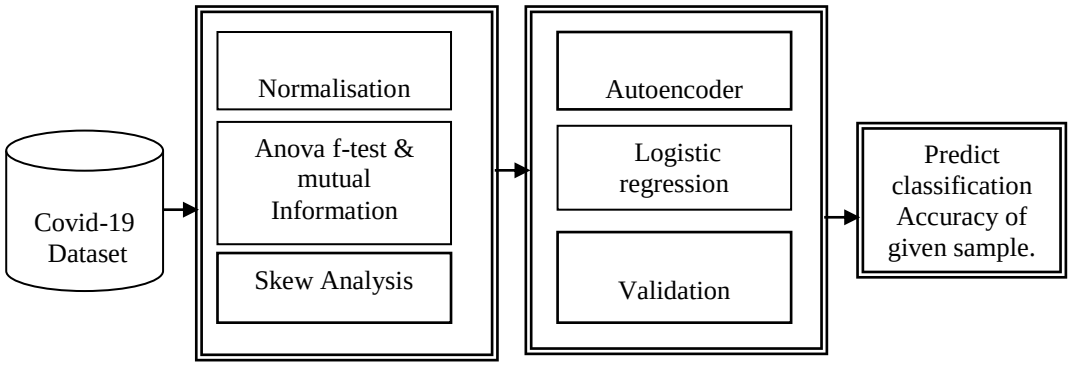
work is to improve a classifier C: $X \rightarrow Y$ that minimizes the classification error on S and enriches the performance using reduced representation of features.

3. Applied Methodologies

4.1 Auto encoders

Auto encoders are widely used for Dimensionality Reduction[15], Feature extraction, Image Compression, Image Generation and Image De-noising. A neural network is trained to attempt to copy its input to its output. Its architecture consists of three main parts viz., an Encoding architecture, Decoding architecture, and a Latent view Representation. It is a special case of feed forward network and trained with all the same technique. The network may be viewed as consisting of two functions: an encoder function for input hidden layer(h)=f(x) and a decoder that produces a reconstruction r=g(h). In this work, uses deep Auto encoder for Covid-19 detection.

4.1.1 Modeling with Auto encoder: The System is modeled using Auto encoder and logistic regression is shown in the Figure 1. This is trained in an unsupervised manner in order to learn the extremely low level representations of the input data, so that these low level representations are then deformed back to project the actual data. The model type that the system will be using is Sequential. Sequential allows build a model layer by layer. Every layer has weights that correspond to the layer that follows it. Dense is the layer type. Dense is a standard layer type. In a dense layer, each and every node in the previous layer connects to the nodes in the current layer. Activation is the activation function that corresponds to the layer. An activation function allows models to consider the nonlinear relationships. The activation function used is ReLU or Rectified Linear Activation and tanh. In this work, AE has the input equal to the output in the hidden layer that has one or less the kind of input units.



In the Covid-19 detection model as in shown in Figure 2, uses tangent function or “tanh” and rectified linear unit “reLU” function is used to encode and decode the methods because it achieves a high level of Accuracy. The system uses keras, as high-level neural API implemented using python in parallel processing to get confusion matrix. In the Keras method, 6 hidden layers with 3 encoders and 3 decoders where designed. Every hidden layer used was the “Tanh” and “reLU” activation function. The dataset is divided as the training set and testing set as 80 and 20 percentages of data to predict covid-19 predictions for logistic regression.The proposed framework and the flow of the system is illustrated in Figure 1 and Figure 2 respectively.

4.2 Multinominal Regression

Logistic Regression is a machine learning model for binary classification that uses the logarithmic function to build a model. The method can handle both the numeric and categorical variable and it is an S-shaped function whose range lies between 0.0 and 1, which makes it useful to model the classification problem and an output close to 1 can indicate that an instance belongs to a certain class. It allows many features and consists of weights for each feature. Given a learned model, the value of the output variable is computed by applying the logistic function to a linear combination of the attribute values and a weight vector, and the resulting learnt model is used to predict the unlabeled instances with high accuracy. The classifier is accurate, when the difference between the estimated and actual probabilities is low.

Estimated probability P ($Y_i = 1/ x_{i,w}$) = $\frac{1}{1+e^{-w^T x_i}}$, where w is weight factor, x_i is i^{th} instance

The cost function of multinomial Logistic Loss w

$$J(\theta) = \left\{ \sum_{1}^{m} y^{(i)} \log h\theta(x^{(i)}) + (1 - y^{(i)}) \log(1 - h\theta(x^{(i)}) \right\}$$

Where i=1,…., m - m is the sample size.

The dataset is divided as the training set and testing set as 80 and 20 percentages of data to predict covid-19 predictions for regression.

Input: Covid data set
Output: classification of cases
Step1: Scaling and normalizing a dataset
Training set and testing set as

$$N = \left\{ \begin{matrix} (x_i, y_j), x_i \varepsilon \ R^n, y_j \varepsilon \, R^m \, , \\ i = 1, \dots \dots, N \ j = 1, \dots, m \end{matrix} \right\}$$

Step2: Design a auto encoder with decoder and encoder with
Constraints:

- regularizer =l1
- Input activation function =Tanh,
- Output activation function =Relu
- Optimizer=adadelta
- Loss=mse
- Hidden layer=sequential
- Data transformation =Min_max scalar
- epochs=100

Step3: For each sample set

- Tune LR with multinomial loss optimizer
- Evaluate a model with LR
- Prediction with trained model

Next

Figure 2. Pseudo code for System Model

5. Experimental analysis

The simulations are performed in python with Google colab on an Intel(R) Core(TM) i3 processor with 4 GB RAM and 3.40 GHz CPU on the platform Microsoft Windows 7.

5.1. Data source

For evaluation of the proposed model, the sample data is fetched as shown in Figure 3 and Table 1, patient pre condition data set was selected from the Kaggle released by Mexico Government[10]. The dataset contains 563201 records of the patients with 22 features and one class variable. The features like as id, sex, patient_type, entry date, date_symptoms, date_died, intubed, pneumonia age, pregnancy, diabetes and others. The 'covid-19_result' is class variable which indicates the positive class and negative class of patient records. After preprocessing , the dataset containing only positive and negative cases are extracted for further processing.

Table1. Sample Data set

Class	count	percentage
0	2290	55.98
1	1801	44.02

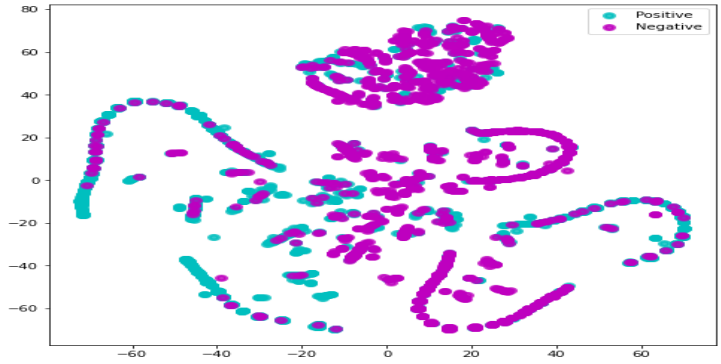


Figure 3. Visualization of Sample dataset

5.2 Correlation Analysis

The Correlation [12] between all pairs of attributes are given in Table 2 with the use of *Pearson's Correlation Coefficient.* A correlation of -1 show full negative and 1 shows that there exists a positive correlation respectively. On the other hand, a value of 0 shows that correlation does not exist at all. A pictorial representation is shown in Figure 4.

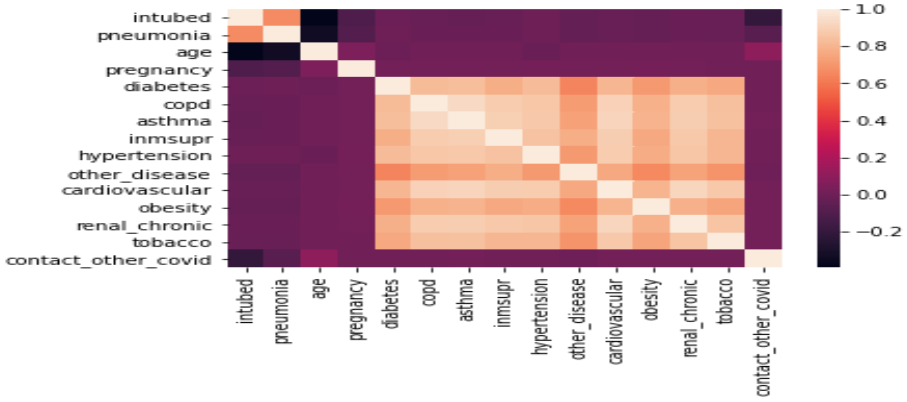


Figure 4. Correlation strength of Features

5.3 Skew Analysis

Skew is a measure of the asymmetry of probability distributions [17]. The normal distribution is a bell shape curve and symmetric. A positive value denotes a right-skewed distribution, and a negative value denotes a left-skewed distribution [16]. The Skew of feature distributions is shown in Table 2 and Figure 5.

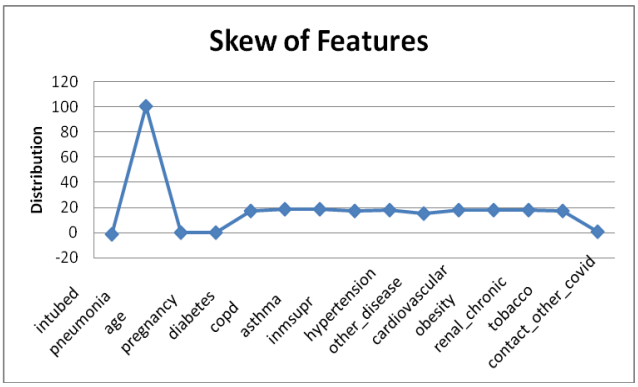


Figure 5. Skew Analysis

Table 2. Skew distribution

Features	Values
intubed	-1.35122
pneumonia	100.3867
age	0.289585
pregnancy	-0.04511
diabetes	17.19396
copd	18.4037
asthma	18.38829
inmsupr	17.14193
hypertension	17.98072
other_disease	14.92971
cardiovascular	18.00393
obesity	18.10148
renal_chronic	18.19368
tobacco	17.53044
contact_other_covid-19	0.817063

5.4 Analysis of Variance

ANOVA is an abbreviation for "analysis of variance" and is a parametric statistical hypothesis test [13] for investigating data by comparing the means of subsets of the data. The results of this test can be employed in selecting the feature where in those features that are independent of the target variable can be detached from the dataset is shown in Figure 6(a) and Figure 6(b). It is returned the features are 1, 2, 3, 4, 9 and 15 are most relevant.

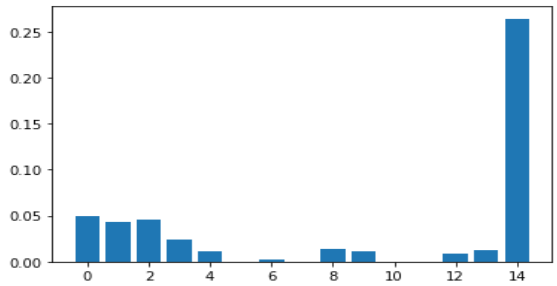


Figure 6(a). ANOVA F-test

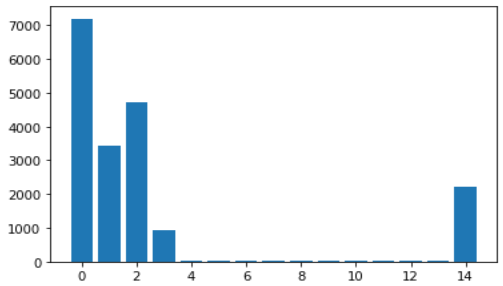


Figure 6(b). Mutual Information

The information theory states that the mutual information is the application of information gain that helps in feature selection process. It is calculated between two variables that measures the reduction in uncertainty for one variable. It is returned the same features as like as F-test as shown in Table 3.

6. Result and Discussion

This section gives the details of the experimental evaluation and the performance of the proposed system. It is analyzed with the Covid-19 dataset [10]. This work presents a hybrid intelligent diagnosis system using auto encoder with regression to facilitate diagnose of covid-19. This diagnosis system is an auxiliary tool to help physician for diagnose the disease. Initially, auto encoder is trained with dataset of

fetched sample. Then, the remaining dataset is used for testing of auto encoder the test dataset. The outputs of classifier represent disease affected and healthy cases are shown in Figure 6.

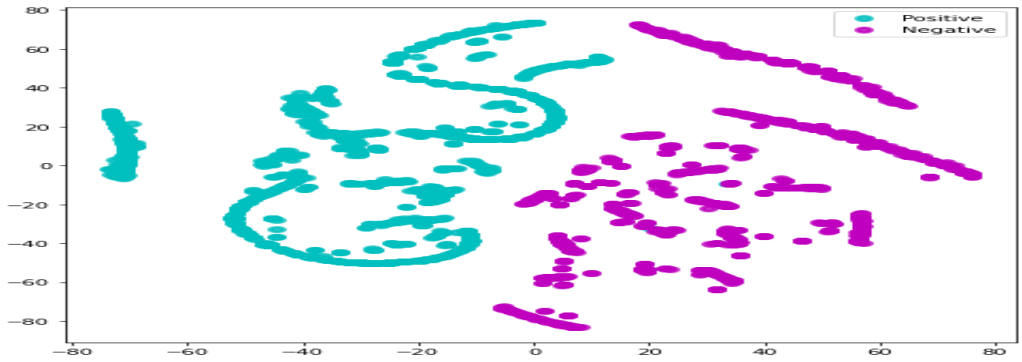


Figure 7. Visualization of sample dataset using Auto Encoder

Table 3. Features and it's importance

Features	Anova F-Test	Mutal Information (IG)
intubed(1)	**267.70**	**0.046**
pneumonia(2)	**155.94**	**0.022**
age(3)	**137.18**	**0.030**
pregnancy(4)	**12.53**	**0.007**
diabetes(5)	4.58	0.000
copd(6)	1.07	0.004
asthma(7)	3.62	0.000
inmsupr(8)	0.80	0.001
hypertension(9)	**56.76**	**0.009**
other_disease(10)	1.49	0.000
cardiovascular(11)	1.11	0.000
obesity(12)	3.17	0.000
renal_chronic(13)	1.00	0.013
tobacco(14)	0.60	0.000
contact_other_covid-19(15)	**2413.89**	**0.260**

Evaluation Metrics

The confusion matrix is one of the ways that clearly indicates in predicting the performance of the classification model while making predictions and provides the user with summary of prediction results on a classification problem. The number of correct and incorrect predictions are summarized along with count values of each class viz., True Positives (TP): These are cases in which the classifier predicted yes, True Negatives (TN): These are cases in which the classifier predicted no, False Positives (FP): These are cases in which the classifier predicted yes .but they are actually covid-19. False positives are also known as a "Type I error." False Negatives (FN): These are cases in which the classifier predicted no, but they are actually Negative cases. False negatives are also known as a "Type II error. It provides details about the types of error that are been made by classifiers. Analyze the performance of the covid-19 cases using the Confusion matrix. The obtained result is given in Figure 7.

Table 4. Accuracy score of sample runs

Sample Run	TP	TN	FP	FN	Accuracy score
#1	500	450	1	0	0.99
#2	504	446	1	0	0.99
#3	505	445	1	0	0.99
#4	505	446	0	0	1.00
#5	502	445	4	0	0.99
#6	488	463	0	0	1.00
#7	491	458	2	0	0.99
#8	504	445	2	0	0.99
#9	572	445	2	0	0.99
#10	501	450	0	0	1.00

It may be observed that not all the result of medical test are absolutely true. When something is concluded true and it is actually false, represents a **false positive** or type I error. On the other hand, when something is false and it is actually true, represents a **false negative** or type II error. The table 4 is shown the sample runs achieved the efficiency of this model.

7. Conclusion

This work proposes a deep learning framework model based on auto encoder with regression to diagnose covid-19. The proposed intelligent diagnosis model has advantages such as reduced false negative and generalization capability over conventional neural networks with back propagation. The experimental results reveal that the intelligent diagnosis system for covid-19 achieves considerable classification accuracy in identifying Corona-19 affected patients. It is concluded that these results obtained will be of utmost importance and useful for physicians in decision making during diagnosing process.

References

[1] Mallick, P. K., Ryu, S. H., Satapathy, S. K., Mishra, S., Nguyen, G. N., & Tiwari, P, "Brain MRI image classification for cancer detection using deep wavelet -based deep neural network", IEEE Access(2019), 7, pp. 46278-46287.

[2] Macías-García, L., Luna-Romera, J. M., García-Gutiérrez, J., Martínez-Ballesteros, M., Riquelme-Santos, J. C., & González-Cámpora, R, "A study of the suitability of s for preprocessing data in breast cancer experimentation", Journal of biomedical informatics, (2017),72, pp.33-44.

[3] Mienye, I. D., Sun, Y., & Wang, Z, "Improved sparse based artificial neural network approach for prediction of heart disease", Informatics in Medicine Unlocked,(2020), 100307.

[4] Siddique, Md, et al,"Performance Analysis of Deep and NCA Dimensionality Reduction Techniques with KNN, ENN and SVM Classifiers." arXiv preprint,(2019), arXiv:1912.05912.

[5] Vinay Kumar Reddy Chimmula, Lei Zhang, "Time series forecasting of COVID-19 transmission in Canada using LSTM networks", Chaos, Solitons & Fractals,Vol.135,109864.

[6] Xayasouk, T., Lee, H., & Lee, G, "Air Pollution Prediction Using Long Short-Term Memory (LSTM) and Deep (DAE) Models", Sustainability,(2020),12(6), 2570.

[7] Xie, R., Wen, J., Quitadamo, A., Cheng, J., & Shi, X,"A deep auto-encoder model for gene expression prediction", BMC genomics,(2017), 18(9), 845.

[8] Shahin Khobahi, Chirag Agarwal, Mojtaba Soltanalian,"CoroNet: A Deep Network Architecture for Semi-Supervised Task-Based Identification of COVID-19 from Chest X-ray Images" ,(2020) medRxiv.

[9] Schönecker, S., Brendel, M., Palleis, C., Beyer, L., Höglinger, G.U., Schuh, E., Rauchmann, B.S., Sauerbeck, J., Rohrer, G., Sonnenfeld, S. and Furukawa, K.,"PET imaging of astrogliosis and tau facilitates diagnosis of parkinsonian syndromes", Frontiers in aging neuroscience,(2019), 11, p.249.

[10] https://www.kaggle.com/tanmoyx/covid19-patient-precondition-dataset.

[11] T.Mitchell, Machine Learning, McGrawHill, New York, 1997.

[12] https://towardsdatascience.com/end-to-end-data-science-example-predicting-diabetes-with-logistic-regression-db9bc88b4d16.

[13] https://machinelearningmastery.com/feature-selection-with-numerical-input-data.

[14] https://www.who.int/emergencies/diseases/novel-coronavirus-2019/situation-reports.

[15] https://en.wikipedia.org.

[16] https://gist.github.com/SoumenAtta.

[17] http://www.fusioninvesting.com/2010/09/what-is-skew-and-why-is-it-important.