Comparing the effects of applying Genetic Algorithms to classification models in order to detect SARS/COVID

Maitreyi Singh

The Australian National University,
Canberra, Australia
Maitreyi.Singh@anu.edu.au

Abstract. With the rapid increasing COVID19 cases in various countries and consequently increasing stress on the medical facilities of the countries, it would be of a great help if we could identify the patients of COVID19 before hand just by the symptoms. Usually, the dataset consists of many physical features or symptoms. Applying models on all of these features consumes a lot of time and is highly inefficient. In this paper, the effort has been made to identify important feature set using Genetic Algorithm (GA) and then just feed them to model to analyse if the symptoms classify into one of the: high blood pressure, pneumonia, SARS/COVID and normal. This makes the computation much more efficient and accurate. In this paper various models have been compared with each other with and without the application of GA. The models included for classification tasks are: Decision Tree, Linear Support Vector Machine (SVM), Radial Basis Function SVM (RBF SVM), Random Forest and Artificial Neural Network(ANN). Almost all the models showed improvements to a varying degree after feeding the feature set obtained applying genetic algorithm. Out of all these, Decision Tree showed a significant rise in accuracy of about 25 percent.

Keywords: Decision Tree, Linear Support Vector Machine, Radial Basis Function Support Vector Machine, Random Forest, Artificial Neural Network, Genetic Algorithm, Fuzzy c-means clustering, Silhouette analysis, SARS, COVID

1 Introduction

The rapid increase in the COVID19 cases in various parts of the world has put immense pressure on the medical systems of these countries. There could be a way to alleviate this problem by trying to identify if a person is displaying any symptoms of COVID19. In that case, such people could be asked not to visit hospitals, instead medical team can go to their respective houses. This would reduce the spread of COVID19 and help in curtailing the chain effect.

In this paper dataset having values for various physical symptoms have been used to classify the symptoms into either of the four categories: high blood pressure, pneumonia, SARS/COVID and normal. Singh, M. (2021)[1] uses classification methods like Decision Tree, Maximum Likelihood Estimation, and Neural Network to classify the dataset. However, the dataset was modified a little and the fuzzy nature of the inputs was not taken into consideration. On the other hand, in this paper the fuzzy nature has been used and 5 major classification methods have been used: Decision Tree, Linear SVM, RBF SVM, Random Forest, Artificial Neural Network. The idea for choosing these methodologies was taken from the paper by Sharma and Gedeon (2013)[2]. The paper applies GA to ANN, SVM and to a combination of SVM and ANN. This paper uses these models except the last combination, and in addition few more methods for better comparison and analyse the effects of genetic algorithm on each of them. Sharma and Gedeon (2013)[2] conclude that the GA helps in improving the accuracy of the models but the maximum was observed in the case of SVM applied with GA.

In this paper, all the models mentioned above were initially used separately to calculate the accuracy of classification and then were applied in combination with GA. As a result of application of GA, a considerable improvement was noticed in almost all the models. The maximum improvement was observed in Decision Tree. The accuracy has been showed in the form of confusion matrix for each of the models.

2 Data

The dataset was formed from 4 different files corresponding to 4 different diseases and their symptom values. These files corresponded to high blood pressure, pneumonia, SARS/COVID and normal categories. Each of the files consists of 8 features namely: Temperature at 8 AM, temperature at 12 PM, temperature at 4 PM, temperature at 8 PM, systolic blood pressure, diastolic blood pressure, nausea and abdominal pain. The first seven attributes or features are further categorized into three classes slight, medium and high. Abdominal pain is bifurcated into yes and no columns.

Each of the files had 1000 data points. As a part of pre-processing and creating a usable dataset, a target column named 'disease' was added to all the four files corresponding to the disease they were related to. All these data points were then combined into one file such that they were stacked on top of one another. To make the values of the target column usable, it was label encoded.

The values in the entire dataset are normalized to a value between 0 and 1, so this dataset didn't require any further normalization and was good to be fed to the model directly. Before applying this dataset to the models, a correlation matrix (Fig. 1) was formed between all the features. As can be seen from the correlation matrix, many features have strong, perfect positive correlation with each other (squares shown in darkest blue colour), while some have perfect negative correlation (squares shown in light yellow colour).

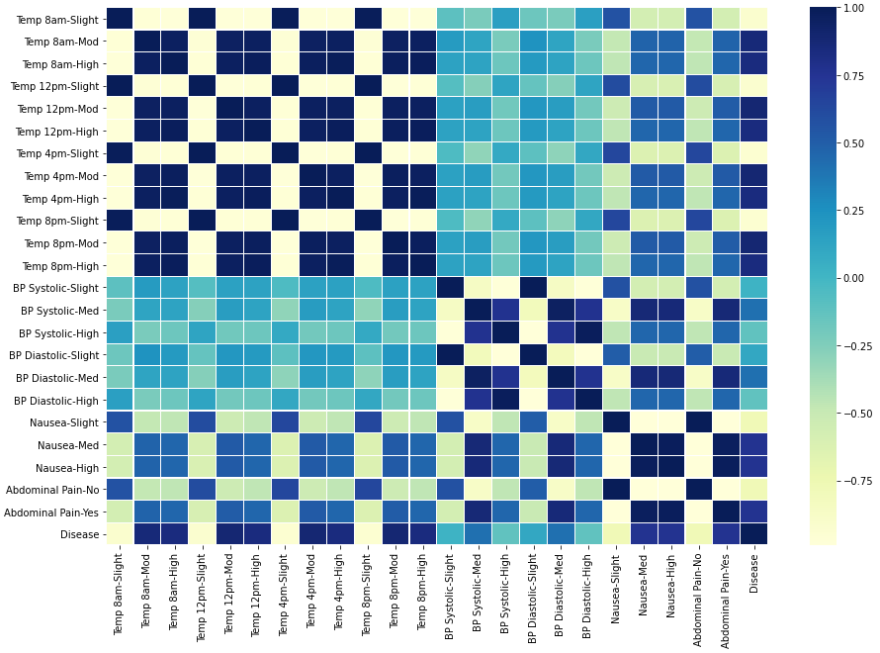

Fig. 1. Correlation matrix between features of the dataset

Reducing the number of features by finding the optimum features using GA for each of the models help in preventing overfitting. As the number of data points is very low, it's much easier for the models to overfit. Thus, only using the important features and discarding the rest can help in generalizing better.

Also, Fuzzy c-means Clustering was used just to see how the datapoints are arranged, and if there's a considerable overlap between the clusters. For this Silhouette analysis was performed. According to Wang et al. (2017)[3], "Silhouette[4] analyses the distances of each data point to its own cluster and its closest neighbouring cluster (defined as the average distance of a data point to all the other data points in its own cluster and that to all the data points in the neighbouring cluster nearest to the data point)". The values are referred to as Silhouette Coefficients. A value of +1 for the coefficient suggests that the sample is far away from the neighbouring clusters while a value of 0 represents that the sample is close to the decision boundary and hence near the neighbouring clusters. Negative values show that the samples might have been assigned to a wrong cluster. The width of the bars is representative of number of data points corresponding to each of the cluster.

Applying the Silhouette analysis, the coefficient obtained was around 0.88 which suggests that the clustering is of good quality. There is considerable inter-cluster difference and intra-cluster similarity. An attempt was made to test the application of GA in Fuzzy c-means clustering as well, the coefficient value rose to almost 0.91 showing even better clustering effects. The value of coefficients can be seen by the dotted lines in Fig. 2 and Fig. 3. Also, as the dataset consists of 1000 data points for each condition, the width of the bars also appear to be same representing that they have been almost equally divided into 4 clusters.

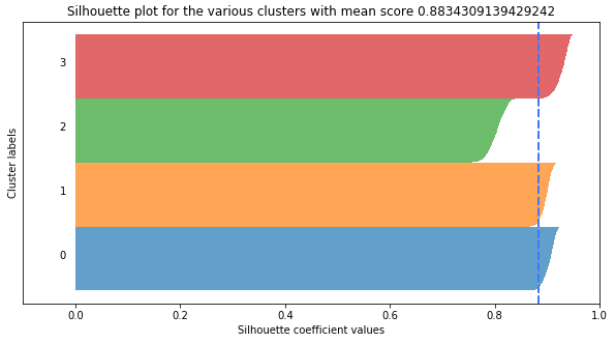

Fig. 2. Silhouette plot for Fuzzy c-means Clustering

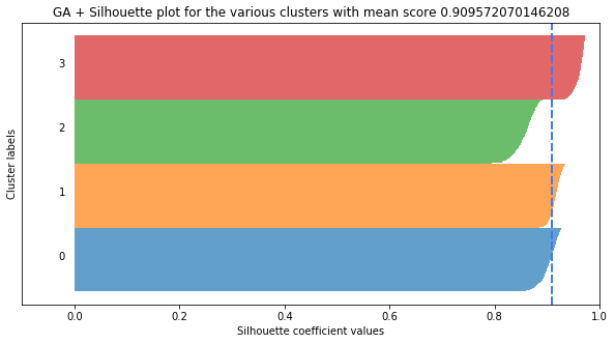

Fig. 3. Silhouette plot for Fuzzy c-means Clustering paired with GA

3 Method

The models included in this paper are mentioned as below:

3.1 Decision Tree

A Decision Tree is a supervised learning technique which seems like a flowchart where the uppermost node is called the root. It is generally used for classification. Each internal node, except for the leaf (terminal node), represents a test of the feature. The leaf represents the class for the object. Each branch is the outcome of that test. In the decision process, the sample (population) is split into two or more sub-populations sets of maximal, which is decided by the most significant splitter or differentiator in the input variables. The ultimate goal is to create a predictive model that can take observations about a sample and make accurate conclusions about the sample's target value. [5]

3.2 Support Vector Machine (SVM)

An SVM is also a supervised learning technique which works on defining a hyperplane which partitions data into classes or labels. It performs non-linear mapping of the training samples so that they get transformed to a higher dimension. Support vectors, samples that provide maximum margin from themselves to the hyperplane, help in determining optimal hyperplane. SVMs are known to produce global solution. As the entire model depends only on the support vectors and not the entire dataset, it generalizes well and there are less chances of overfitting.
This paper has used two types of SVM: Linear SVM and RBF SVM.

3.3 Random Forest

It is an ensemble method which can be used for classification. Random forest contains a large number of small decision trees, which are called estimators. These small decision trees make their own predictions individually. Random Forest combines these predictions and make its own more accurate prediction.[6]

3.4 Artificial Neural Network (ANN)

ANN is a model which tries to mimic brain's biological neural network. It is based on collection of thousands of hidden neurons (node) organized into multiple hidden layers. Each hidden neuron is connected to various other hidden neurons from the previous layer and to the next layer. This network is a feed-forward network. The node allots a weight to its incoming connections. This node multiplies these connections with their corresponding weights and add them (linear operation). The value thus obtained is passed through an activation function. The result from the activation function is a number. If this number is less than the threshold of the neuron, then the neuron doesn't fire. When the number is more than the threshold of neuron, the neuron fires and the data is sent forward. This process is repeated for all the neurons in all the layers till the very end. Then the loss is calculated and through back propagation, the weights of nodes are adjusted. This continues till the model achieves a required accuracy.

3.5 Genetic Algorithm (GA)

GA is a class of Evolutionary Algorithms which is stochastic, parallel search heuristics inspired by the biological model of evolution.[7] It is very robust and its successes has enabled it to be applied to various problems. A GA uses a population of candidate solutions. The individuals within the population compete with each other based on the fitness function. The population undergoes evolution by selecting parents from one generation and genetic operators are applied to form new generation of individuals. These genetic operators, generally, are crossover and mutation. Crossover is the equivalent of sexual reproduction in the nature while mutation introduces diversity in the gene pool.

GA is used to deduce optimal feature set from the entire dataset. In this paper 5 categories are used to incorporate GA: GA + Decision Tree, GA + Linear SVM, GA + RBF SVM, GA + Random Forest and GA + ANN.

4 Result and Discussions

The first model that was used to perform classification task on this dataset was Decision tree. For decision tree, maximum depth allowed was 2. Various depths were tried, but for depths more than 3, the accuracy was coming to be 100 percent. The model was starting to overfit the dataset. The main reason behind this is the small number of data samples. At the depth of 2, the model gave an accuracy of 73.63 percent (Fig. 4) on test set and 75.34 percent on training set (Fig. 5) as can be seen from the confusion matrix. The application of GA improved the accuracy level from 73 percent on test set to 100 percent (Fig. 6). The optimal feature set obtained for GA + Decision Tree was: 'Temp 8am-High', 'Temp 12pm-Slight', 'Temp 12pm-Mod', 'Temp 12pm-High', 'Temp 4pm-Slight', 'Temp 4pm-Mod', 'Temp 4pm-High', 'Temp 8pm-Mod', 'BP Systolic-High', 'BP Diastolic-High', 'Nausea-Slight', 'Nausea-Med', 'Nausea-High', 'Abdominal Pain-Yes'.

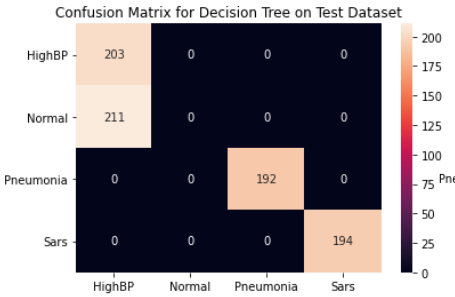

Fig. 4. Confusion matrix for Decision Tree on test dataset

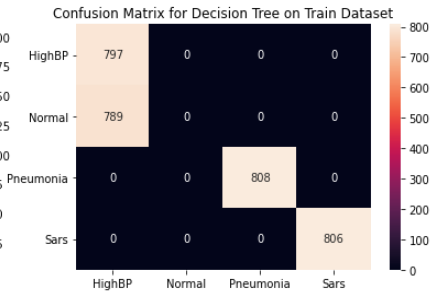

Fig. 5. Confusion matrix for Decision Tree on training dataset

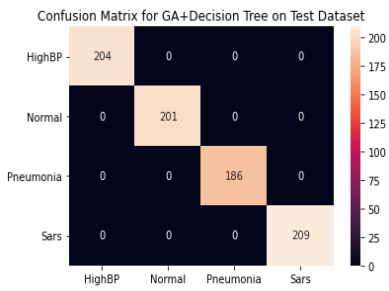

Fig. 6. Confusion matrix for GA + Decision Tree on test set

The second model used for classification was Linear SVM. The accuracy of classification was similar to Decision Tree without the application of GA, i.e, 73.63 percent on test set (Fig. 7). The accuracy for training set came to be 75.34 percent (Fig. 8). For linear SVM, the value of regularization parameter, C, was kept at 0.0003. The strength of the regularization is inversely proportional to C. Here, the model is severely regularized by keeping the value of C very low because otherwise the model was giving 100 percent accuracy similar to Decision Tree. After the application of GA, the accuracy was 74.88 percent on the test set (Fig. 9). Only a small increase in accuracy was observed, but this is expected owing to the amount of regularization that was done. Optimal features identified were: 'Temp 8am-Slight', 'Temp 12pm-Slight', 'Temp 12pm-High', 'Temp 4pm-High', 'Temp 8pm-Slight', 'BP Systolic-Slight', 'BP Systolic-Med', 'BP Systolic-High', 'BP Diastolic-Slight', 'Nausea-Slight', 'Nausea-Med', 'Abdominal Pain-No', 'Abdominal Pain-Yes'.

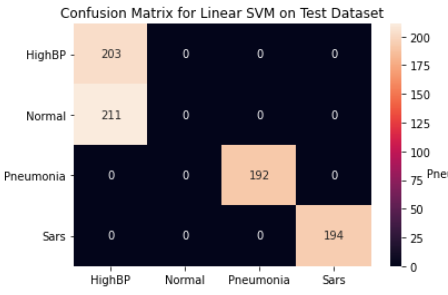

Fig. 7. Confusion matrix for Linear SVM on test dataset

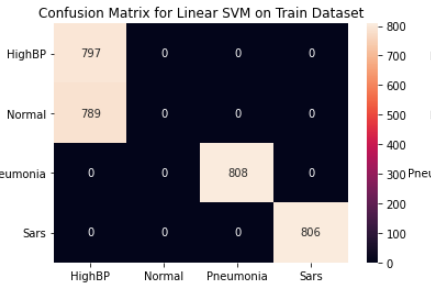

Fig. 8. Confusion matrix for Linear SVM on training dataset

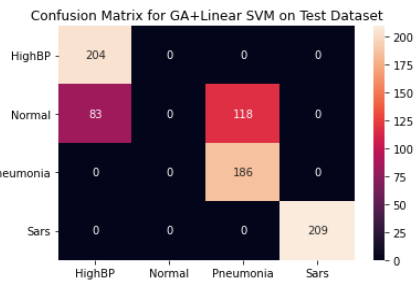

Fig. 9. Confusion matrix for GA + Linear SVM on test set

For the purpose of classification, the third model that was used was RBF SVM. The accuracy for RBF SVM for test set came to be 97.63 percent (Fig. 10), while for training set the same was achieved to be 98.13 percent (Fig. 11). The value for regularization parameter, C, was kept to be 0.0045 because of the same reasons as explained above (to avoid overfitting). With the application of GA, the accuracy rose to 100 percent (Fig. 12). The optimal features identified by the GA model were: 'Temp 12pm-Slight', 'Temp 4pm-Mod', 'Temp 8pm-Mod', 'Temp 8pm-High', 'BP Systolic-Slight', 'BP Systolic-Med', 'BP Systolic-High', 'BP Diastolic-Med', 'BP Diastolic-High', 'Nausea-Slight', 'Nausea-Med', 'Nausea-High', 'Abdominal Pain-No', 'Abdominal Pain-Yes'.

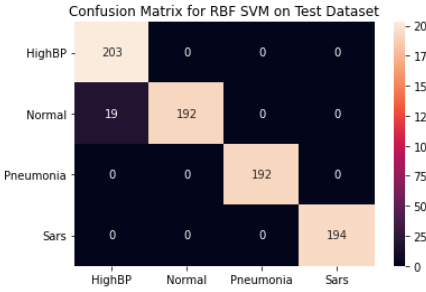

Fig. 10. Confusion matrix for RBF SVM on test dataset

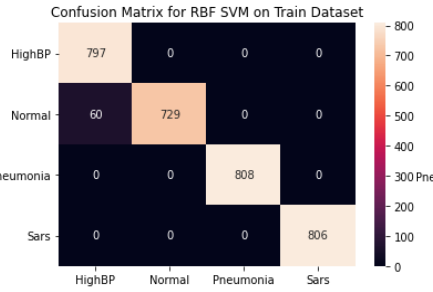

Fig. 11. Confusion matrix for RBF SVM on training dataset

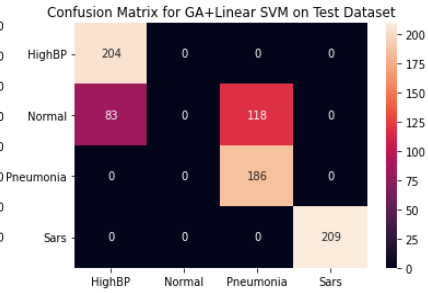

Fig. 12. Confusion matrix for GA + RBF SVM on test set

Next model used for this task of classification was Random Forest. The value of hyperparameters, maximum depth and number of estimators, were set to be 2 and 3 respectively. These values were obtained after a lot of iterations of execution of the model and where the model didn't seem to overfit the samples. The accuracy of Random Forest was 100 percent for test dataset (Fig. 13) and training dataset (Fig. 14). The application of GA also resulted into the same accuracy (Fig. 15). The optimal features identified by GA were: 'Temp 8am-High', 'Temp 12pm-Slight', 'Temp 12pm-High', 'Temp 4pm-Mod', 'Temp 8pm-Slight', 'Temp 8pm-Mod', 'Temp 8pm-High', 'BP Systolic-Med', 'BP Diastolic-Slight', 'BP Diastolic-Med', 'Nausea-High', 'Abdominal Pain-No'.

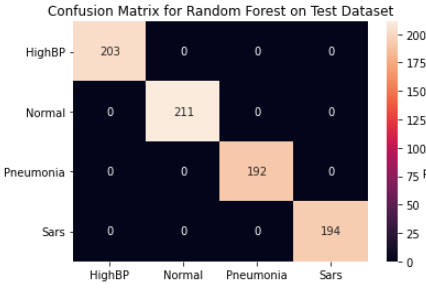

Fig. 13. Confusion matrix for Random Forest on test dataset

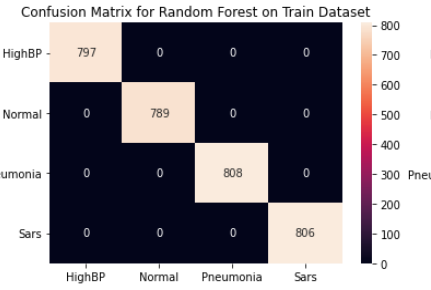

Fig. 14. Confusion matrix for Random Forest on training dataset

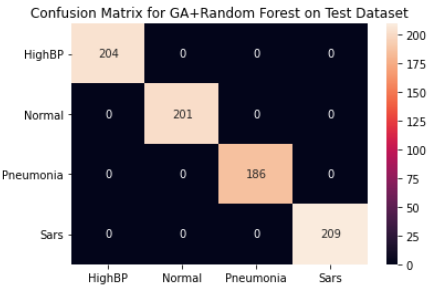

Fig. 15. Confusion matrix for GA + Random Forest on test set

The final model used for classification was ANN. For ANN, the number of hidden layers that were chosen was 2 and number of neurons 80 and 100 for 1st hidden layer and 2nd hidden layer respectively. Again, to avoid overfitting these, values were tested to a great range. But as the model was seeming to overfit the samples, these values were reduced and brought to this point where it seemed that the model was generalizing well. To further prevent overfitting, dropout of 20 percent was included in the model itself. The activation function used was a sigmoid function. The ANN model used Cross Entropy Loss for backpropagation and Adam as the Optimizer. With all these parameters, the accuracy on the test set came out to be 99.88 percent (Fig. 16) while that for training set was 99.81 percent (Fig. 17). The accuracy after the application of GA to ANN was 99.88 percent (Fig. 18). This didn't change much. But this accuracy was obtained from the optimal features and not the entire dataset. The optimal features identified in this case were: 'Temp 8am-High', 'Temp 12pm-High', 'Temp 4pm-Mod', 'Temp 4pm-High', 'Temp 8pm-Mod', 'BP Systolic-Slight', 'BP Systolic-Med', 'BP Diastolic-Slight', 'BP Diastolic-Med', 'BP Diastolic-High', 'Nausea-Slight', 'Abdominal Pain-Yes'.

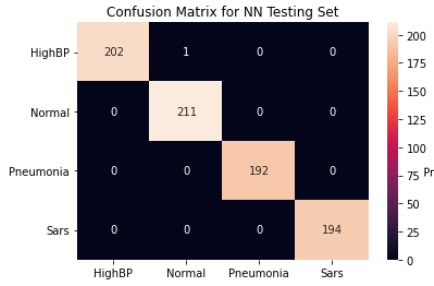

Fig. 16. Confusion matrix for ANN on test dataset

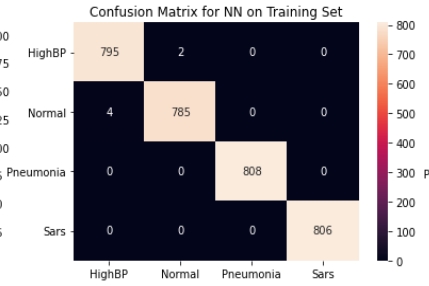

Fig. 17. Confusion matrix for ANN on training dataset

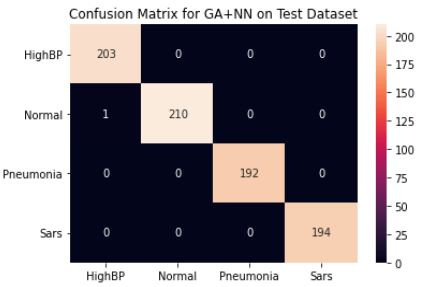

Fig. 18. Confusion matrix for GA + ANN on test set

For GA + ANN, various values of threshold were used to analyse at what value, does the model attain maximum accuracy. Table 1 summarizes the results (accuracy) obtained as a result of varying threshold value.

Threshold (θ)	Accuracy on Training set	Accuracy on test set
0.1	59.88	58.75
0.15	80.19	79.5
0.2	91.6	89.75
0.25	95.94	95.34
0.3	97.56	98
0.35	99.16	98.75
0.4	99.44	**99.75**
0.45	**99.47**	99.5
0.5	99.38	99.25
0.55	98.41	98.63
0.6	96.47	96
0.65	93.59	93.5
0.7	88.81	88.38
0.75	78.13	78.5
0.8	63.47	63
0.85	43.63	42.62
0.9	29.03	29.125

Table 1. Analyzing threshold values to attain maximum accuracy

As can be seen from Table 1, the accuracy is the highest when the threshold value is 0.45 for the training set where the accuracy comes to be 99.47 percent, and a value of 0.4 for test dataset where the accuracy comes to be 99.75 percent. Any of these two values would present great results.

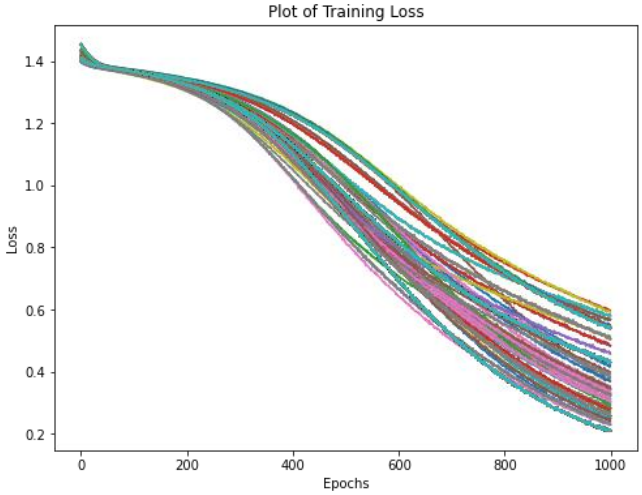

Fig. 19. Plot for Total Loss for different generations

Fig. 19 shows the plot of training loss for different generations. The starting loss was almost same for all the generations, and they all decreased with the number of epochs.

The accuracy for all the methods including with and without the application of GA over test set and training set can be visualized in Table 2.

Models	Accuracy before application of GA	Accuracy after application of GA
Decision Tree	**73.63**	**100**
Linear SVM	73.63	74.88
RBF SVM	97.63	100
Random Forest	100	100
ANN	99.88	99.88

Table 2. Comparison of Accuracy between methods before and after applying GA

For all the models, when GA is being integrated, the population size was kept as 10, and the maximum number of iterations, was kept at 1000. The fitness function was defined on the basis of accuracy. The fitness function is trying to maximise the accuracy of classification.

According to Sharma and Gedeon (2013)[2], the best result was obtained from the combination of GA + SVM where recognition rate was 0.89. But with respect to this paper, as can be seen from Table 2, the most remarkable improvement was in the case of Decision Tree where the accuracy improved from 73.63 percent to a straight 100 percent. All other models also show some improvements. Only in the case of ANN, it doesn't seem to improve. It might be because of the fact that for this model, not the most optimal features were chosen as a result of variation in population set. Also, as ANN works on the principle of allotting weights, it generally tends to identify the important features and hence, probably, GA didn't have much effect on its accuracy.

GA is an indeterministic model wherein there are many variables and various hyperparameters. So, it is not important that same results would be obtained each time GA is applied to the above-mentioned models. In this paper, after performing several iterations, the best result has been showcased.

5 Conclusion and Future Work

This paper tries to classify the symptoms of a patient (as given in the dataset) into four classes or labels: High Blood Pressure, Pneumonia, SARS/COVID and Normal conditions. This classification can help in preventing people with SARS/COVID like symptom from moving in public or getting into hospital where they have a chance of infecting others and hence building up the chain. The objective of this paper was to analyse the effect of including GA in the simple classification models and gauge its results. GA was used to identify the optimal features from the dataset, which helped in preventing overfitting while improving generalization.

The integration of GA with the models led to improvement in classification and hence in the accuracy. Although the improvements were of varying degree, improvements were still there. The largest difference was obtained in the case of Decision Tree, where initially the accuracy was 73.63 percent before GA was applied. After integrating with GA, the model's accuracy shot upto 100 percent. This indicates that a huge increase in accuracy is possible with the application of GA. This is not just limited to supervised learning techniques, instead GA when used with an unsupervised learning, the results were similar, in the paper's case it was Fuzzy c-means Clustering. The accuracy rose up by 2 percent and better clustering was achieved. In case of GA + ANN, various values of threshold were tested to obtain the highest accuracy. The highest accuracy obtained was at the threshold values of 0.4 and 0.45.

As this dataset was too small, in future an effort can be made to obtain more data samples. Also, as with the growing cases of COVID, the number of variants is also growing, the dataset could also include the symptoms in these variants as well. Having been trained on a good sample set, the transfer learning model can be used to identify the type of variant and help in fighting this COVID pandemic. During the application of GA, even the effect of elitism can be analysed on the accuracy of the overall model. As elitism passes on fit individuals to the next generation, the accuracy might increase for the model.

Reference

[1] Singh, M., 2021. Comparison of Neural Network and Machine Learning Techniques to detect SARS-COV. ABCs2021 (4th ANU Bio-inspired Computing Conference).

[2] Sharma, N. & Gedeon, T., 2013, Hybrid Genetic Algorithms for Stress Recognition in Reading. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 117-128.

[3] Wang, F., Franco-Penya, H., Kelleher, J., Pugh, J. and Ross, R., 2017. An Analysis of the Application of Simplified Silhouette to the Evaluation of k-means Clustering Validity. Machine Learning and Data Mining in Pattern Recognition, pp.291-305.

[4] Rousseeuw, P., 1987. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics, 20, pp.53-65.

[5] DeepAI. 2021. Decision Tree. [online] Available at: <https://deepai.org/machine-learning-glossary-and-terms/decision-tree> [Accessed 30 May 2021].

[6] DeepAI. 2021. Random Forests. [online] Available at: <https://deepai.org/machine-learning-glossary-and-terms/random-forest> [Accessed 30 May 2021].

[7] Nowé, A., 2011. Genetic Algorithms. Encyclopedia of Astrobiology, pp.635-639.

[8] Mendis, B. S., Gedeon, T. D., & Koczy, L. T. (2005). Investigation of aggregation in fuzzy signatures, in Proceedings, 3rd International Conference on Computational Intelligence, Robotics and Autonomous Systems, Singapore. (used for dataset)