2021 5th International Conference on Computer, Communication and Signal Processing (ICCCSP) | 978-1-6654-3277-1/20/$31.00 ©2021 IEEE | DOI: 10.1109/ICCCSP52374.2021.9465494

Symptoms based Early Clinical Diagnosis of COVID-19 Cases using Hybrid and Ensemble Machine Learning Techniques

C Koushik
School of Computer Science and Engineering
Vellore Institute Of Technology, Chennai
Chennai, India
c.koushik2018@vitstudent.ac.in

Ritwika Bhattacharjee
School of Computer Science and Engineering
Vellore Institute Of Technology, Chennai
Chennai, India
ritwika.bhattacharjee2018@vitstudent.ac.in

C Sweetlin Hemalatha
School of Computer Science and Engineering
Vellore Institute Of Technology, Chennai
Chennai, India
sweetlinh@gmail.com

***Abstract*— This paper aims to develop a classification system to distinguish COVID-19 positive and negative cases based on common symptoms and could be used as a first-level screening tool for early detection of mild cases. Accordingly, existing classification models such as Logistic Regression, Gradient Boosting (GB), Random Forest (RF) and K-Nearest Neighbours have been tried on the COVID-19 symptoms dataset to identify the best performing model. Although traditional machine learning models provide promising results in terms of accuracy, precision and recall, this paper analyses the possibilities of improvement in classification results through ensemble and hybrid approaches. It is observed from the results that K-mode clustering followed by classification-based hybrid modelling resulted in improved classification accuracy in the clusters leading to an average accuracy of 87.17% and 87.24% with GBand RF respectively. Finally, the MaxVoting ensemble model, comprising GB and RF algorithms further boosted the accuracy to almost 90%.**

Keywords*— *Machine Learning, Classification, Clustering, Ensemble Models, Hybrid Models, COVID-19, Healthcare.

I. INTRODUCTION

Nowadays, predictive analytics [1] provides an alert about the events that might occur in the future to clinicians, doctors and other healthcare domain experts. It enables them to take appropriate measures to prevent them before hand or in the worst case prepare measures to tackle the situation in the event of their occurrence. In healthcare domain, it helps to make clinical decisions for individual patients such as detecting the presence of COVID-19 from their symptoms [2]. Also, time-series prediction models [3] are used to determine the growth of COVID-19.

Cough, fever and pneumonia are the observed to be the most common clinical manifestations of COVID-19 [4]. Over 570, 000 confirmed cases and more than 26,000 deaths have been reported in 199 countries and regions as reported on Mar 28, 2020 by the World Health Organization [5].

This situation around the world has created a lot of havoc and managing such a situation requires real time collection and processing of medical data [6]. Hence, the need of the hour is data analytics tools and algorithms which can very conveniently manage and work on large complex data for building predictive models [7].

Most of the existing literatures focus on deep learning models for CT scan images [8], [9], [10], and chest X-ray images [11] which could predominantly detect the infection only after 5 days or more. Nevertheless, our work focuses on early detection of COVID-19 cases based on symptoms using Machine Learning (ML) algorithms to facilitate the first level screening of infected patient for self-isolation.

Though the traditional ML models for classification discovers the relationship between the predictor variables and the target variable, they fail to capture the structural characteristics and similarities in the sample space which could provide better classification results.

The main contribution of this paper is two folds:

1. Building hybrid classification model for identifying COVID-19 cases.
2. Building Ensemble model to predict the positive cases of COVID-19.

II. MATERIAL

The dataset used in this work is maintained by Israel Ministry of Health (Israeli government) [12] and publicly available online in Kaggle. It has 112345 records with 102233 records of negative cases and 10112 records of positive cases. The structured dataset in '.csv' format contains the common symptoms such as cough, fever, sore throat, shortness of breath, headache and other fields like age above 60 or not, gender, test indication (Contact with infected person, for Abroad Travel and otherwise).

Currently a very few reliable open symptoms datasets are available due to confidentiality of patient's data. Hence, we chose this dataset which contains symptoms data in the initial period of March-April where the growth of cases was observed to be similar in Israel and India (approximately 500 daily average in Israel and similar trends in India towards end of March 2020), which justifies our assumption of possible similarity in patients. The dataset is considered reliable as it is based on the data maintained officially by Ministry of Health, Israel Government. Moreover, the total number of cases in the dataset matches the monthly statistics of cases as recorded by the Israel Government.

978-1-6654-3277-1/21/$31.00 ©2021 IEEE

III. Proposed Solution

The aim of this work is to study the effect of traditional classification machine leaning models and further develop efficient models through ensemble and hybrid machine learning techniques to improve the performance of the overall system. Figure 1 gives a general idea about the workflow of the proposed work.

A. Data Preprocessing

In the preprocessing phase, a few categorical variables are encoded into numerical values for classification. For example, in the column age_above_60, "No" and "Yes" values are encoded as 0 and 1, in the gender column, "Female" and "Male" values are encoded as 0 and 1 respectively. In the test_indication column, "Contact with Confirmed" is encodedas 1,"Abroad" is encoded as 2 and 3, otherwise.

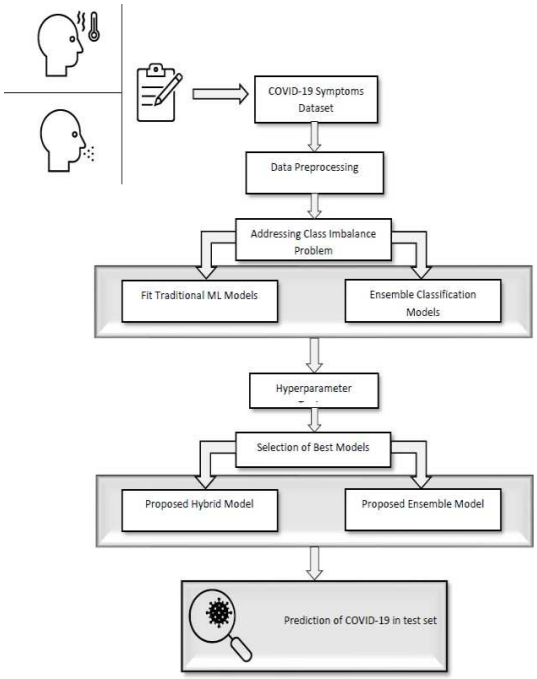

Fig. 1. Proposed Model Framework

B. Addressing Class Imbalance Problem

One of the major issues with the dataset is the Class Imbalance problem caused due to a larger number of negative cases (around 90%) and a smaller number of positive cases (around 10%). In order to avoid biased model training and prediction, we used the over-sampling method because of the reason that, more the data, better the reliability.

Oversampling is implemented by adding copies of instances of the under-represented class i.e., the positive corona results in our case. This way the number of positive and negative samples are balanced in the modified dataset. Figure 2 shows the graphical distribution of the number of cases after oversampling of the original dataset such that the number of negative cases (about 99k) is now comparable to the number of positive cases (about 107k). Figure 3 presents the distribution of symptoms in terms of the percentage of each symptom present in the patients after resampling. It is evident from the figure that cough is the most common symptom observed in more than 39% of all the patients, followed by fever observed in almost 32% patients and headache is present in almost 13.5% patients.

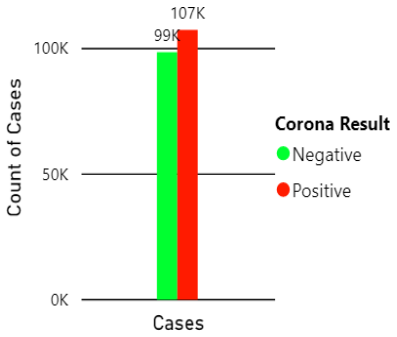

Fig. 2. Distribution of Cases after resampling

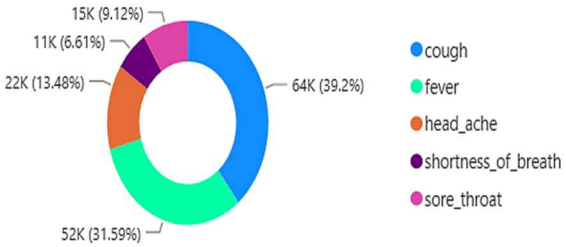

Fig. 3. Distribution of Symptoms after resampling

C. Fitting Traditional ML Models for prediction

After cleaning and resampling the dataset, the next process is to split the data into training and testing sets. Although there are several existing methods to split the data, we decided to adopt sklearn's train_test_split() to split the data into two random partitions in the ratio of 0.75:0.25 because its time- efficient when dealing with such a huge dataset. After splitting the data, since we are treating a binary classification problem,we decided to adopt the following specific classification techniques such as Logistic Regression, K-Nearest Neighbours and Decision Trees to predict the target binary variable corona_result in the testing set.

D. Fitting Ensemble Models for prediction

Ensemble learning is a process in which more than one machine learning models such as classifiers are strategically trained and generated in order to solve a particular computational intelligence problem [13]. The primary objective behind ensemble learning algorithms is to improve the performance of the classification, regression/prediction and approximation models or to reduce the possibility of selecting a poor model due to some unfortunate reasons or assumptions. In our paper, we primarily focus on applications of ensemble learning techniques such as Random Forest and Gradient Boosting.

E. Hyperparameter Tuning

When selecting the models for fitting, it is important to select the appropriate combination of parameters and suitablevalues of those parameters that will lead to better performancein terms of various factors of consideration. To tune the parameters, we have used RandomizedSearchCV() from sklearn.model_selection package. This technique

assesses models for a given vector of hyperparameters using cross- validation.

F. Selection of Best Models

In the case of covid-19 case prediction, Type II errors (false negative cases) are more dangerous because it will neglect the actual positive cases by predicting them as negative and hence, increase the chance of spreading. We choose the best model by comparing Type I (false positive cases) and Type II errors of the models considering other metrics such as accuracy, precision and recall.

G. Proposed Hybrid Model for prediction

The proposed hybrid ML model combines supervised (classification) and unsupervised learning (clustering) to further improve and explore the prediction modelling system [14]. This approach divides the original dataset into clusters using K-Modes clustering [15] and then builds top two classification models (Random Forest and Gradient Boosting) resulted in base prediction, on each of the clusters. The final accuracy is then calculated by taking the mean of accuracy values in all clusters. There is no possibility of collusion in deciding the result as the clustering and classification model are applied in sequence. The output from the clustering model acts as input for classification model which gives the final result. Algorithm 1 depicts the proposed hybrid model of clustering followed by classification.

Algorithm 1: Hybrid Model Approach

Input: Dataset X = {X1, X2, …….,Xn} having all the feature vectors X_i of n observations

Output: Accuracy of the model obtained after Hybrid Classification, i.e. *totalAccuracy*

1. **procedure** *hybridClassification(X)*
2. **Select** a suitable number of clusters k using elbow method
3. **Select** k random points from the dataset X and storethem as the initial set of centroids
   $V = 4. \{v1,v2,……..,vk\}$
4. **Initialise** Set of k empty lists S = $\{S1,S2,……..,Sk\}$ such that S_i contains all the datapoints in that clusterhaving centroid v_i
5. **while** true **do**
6. **for** each $x_i \in X$ **do**
7. **Initialise** *chosenClusterCentroid,, minimumDistance* = 0
8. **for** each $v_i \in V$ **do**
9. **Calculate** *dissimilarity*=Matching Dissimilarity between x_i and v_i
10. **if** *dissimilarity* < *minimumDissimilarity*
11. **Update** *minimumDissimilarity = dissimilarity*
12. **Update** *chosenClusterCentroid = vi*
13. **end for**
14. **Assign** datapoint x_i to cluster having centre = *chosenClusterCentroid*, i.e, $S_{chosenClusterCentroid}$
15. **end for**
16. **for** each of the k clusters $v_i \in V$ **do**
    Recalculate new cluster centre v_i by taking mode of all datapoints in current cluster
17. **end for**
18. Recalculate the dissimilarity between each datapoint and new cluster centers
19. **if** no datapoint is reassigned **then**
20. **Break**
21. **end if**
22. **end while**
23. **initialize** *Accuracy* = []
24. **for** each cluster $Si \in S$ **do**
25. append to *Accuracy* list the value of accuracy obtained after applying gradientBoostingClassification(*Si*)
26. **end for**
27. **calculate** *totalAccuracy* = mean(*Accuracy*)
28. **return** *totalAccuracy*
29. **end procedure**

H. Proposed Ensemble Model for prediction

MaxVoting Ensemble model is used in the proposed design that combines the top two base ensemble classifiers, Random Forest and Gradient Boosting. For each data point, these base classifiers models are combined to make predictions based on majority voting. This can also be presumed as effectively taking mode of all the predictions. Any chance of collision among the individual models within the ensemble is overcome by majority voting which considers the maximum reported class by all base models as the final output, thus taking into account the collective decision rather than being biased towards any particular classifier. This method is depicted in Algorithm 2.

Algorithm 2: Ensemble Model Approach

Input:
Training dataset D with labels representing C classes Learning algorithm of a classifier model L
Y labels of the training set Number of learning algorithms N

Output: Accuracy of the model obtained after Ensemble VotingClassifier, i.e. *aggregateVote*

1. **procedure** *EnsembleModel(D)*
2. Do i=1 to N:
3. **Call** algorithm associated with the dataset, Di and build the model classifier Li
4. **Compare** Yi with Ci generated from the model Li and update *vote*
5. *aggregateVote* ← **mode(***vote***)**
6. *aggregateVote* to the ensemble
7. **end procedure**

IV. RESULTS

This section presents the distribution of COVID-19 cases through graphical representation and discusses the performance of traditional models vs. proposed hybrid and ensemble models.

A. Visualizing the distribution of cases with respect to the different categorical variables

Figure 4 shows the graphical distribution of COVID-19 cases with respect to the age group (60 above and below) of the patients present in the dataset.

Table I shows the distribution of cases based on test indications of Abroad, Contact with confirmed and other reasons in males and females.

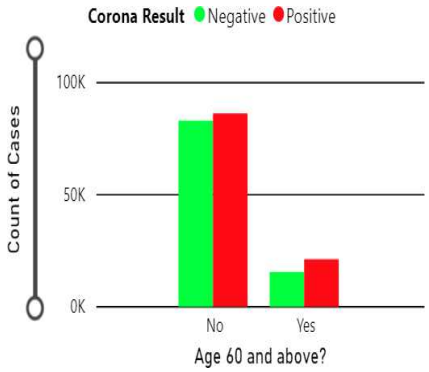

Fig. 4. Distribution of cases with respect to age groups

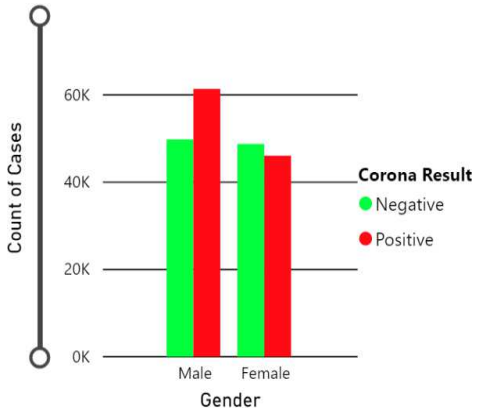

Fig. 5. Distribution of cases across gender

TABLE I

DISTRIBUTION OF CASES FOR DIFFERENT TEST INDICATIONS ACROSS GENDER

Gender	Test Indication	Positive Cases	Negative Cases
Female	Abroad	6108	4057
Male	Abroad	10164	4474
Female	Other	17256	44178
Female	Contact with confirmed	22716	546
Male	Contact with Confirmed	24096	691
Male	Other	27132	44640
Total		**107472**	**98586**

B. Performance Comparison of ML Classification Models

Table II shows the performance comparison of classification models such as Logistic Regression, KNN Classification, Entropy Based Decision Tree Classifier, Random Forest Classifier and Gradient Boosting Classifier. The models' performance was evaluated based on Accuracy, Precision and Recall scores as well as Type I and Type II errors using the confusion matrix as discussed in sub-section F in section III.

It is observed from the results that the first two models, i.e., Logistic Regression and KNN Classification have a remarkably high value of Type II errors as compared to the other models and hence we outright reject them as they are unreliable and dangerous. Also, the Random Forest classifier performs better in terms of accuracy, precision and recall than the traditional classifiers. However, the Gradient Boosting classifier shows lesser Type II error compared to Random Forest classifier. Hence, those two models are considered for further analysis using the proposed hybrid and ensemble modelling techniques.

TABLE II

PERFORMANCE COMPARISON OF CLASSIFICATION MODELS

Model Name	Type I Error	Type II Error	Accuracy	Precision	Recall
Logistic Regression	28.62%	71.38%	84.70%	0.85	0.85
KNN Classification	71.38%	86.46%	82.19%	0.83	0.82
Gradient Boosting Classifier	33.82%	66.18%	86.18%	0.86	0.86
Random Forest Classifier	31.59%	68.41%	86.32%	0.87	0.86
Decision Tree	40.95%	59.05%	83.23%	0.83	0.83

C. Proposed Hybrid Method Approach

K-modes clustering is performed followed by Gradient Boosting Classification on each cluster to effectively predict traces of COVID-19 in people. The number of clusters obtained using Elbow method is depicted in Figure 6 and it is found to be 3

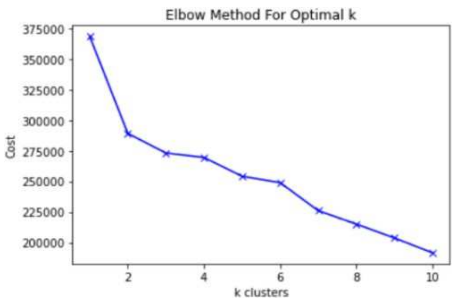

Fig. 6. Elbow Method for finding optimal k

Table III shows the accuracy obtained after training each of the clusters and the overall classification accuracy which is

the average accuracy of all the three clusters. The same approach was carried out for Random Forest Classifier too. Having obtained scores of 87.17 and 87.24, we can verify and confirm that the hybrid model approach outperforms traditional ML classification models in this specific case. The combination with the Gradient Boosting Classifier not only increases overall accuracy but also reduces Type 2 errors significantly to 46% as compared to the best case of 59% in Decision Tree Classifier, thus proving superior. This is important since Type 2 errors are extremely risky because it will neglect the actual positive cases by predicting them as negative and hence increase the chance of spreading. The overall results of the hybrid model using Gradient Boosting Classifier have been summarised in Table IV.

TABLE III

ACCURACY SCORES ON EACH CLUSTER AND THE OVERALL SCORE

	Gradient Boosting	Random Forest
Cluster	**Accuracy**	**Accuracy**
0	89.82%	90.03%
1	87.85%	87.84%
2	83.84%	83.85%
Average Accuracy	87.17%	87.24%

TABLE IV

HYBRID MODEL RESULTS USING GRADIENT BOOSTING CLASSIFIER

Accuracy	Type I Errors	Type II Errors
87.17%	43.06%	56.94%

The overall results of the hybrid model using Random Forest Classifier have been summarised in Table V.

TABLE V

HYBRID MODEL RESULTS USING RANDOM FOREST CLASSIFIER

Accuracy	Type I Errors	Type II Errors
87.24%	43.81%	56.19%

In both these cases, we could observe an increase in overall scores and reduction in Type I and Type II Errors. However, hybrid model with Random Forest works slightly better than the former giving better average accuracy and lesser Type II error.

Both the ensemble models are considered efficient in giving more reliable decisions as compared to traditional classification models as they combine the results of various individual models trained on the diverse subsets of the original model, thus reducing the variance of the overall model through diversification even if the individual models may have been deeply overtrained. Ensemble methods minimize the disadvantage and errors of the single models and improve the performance to provide the best prediction possible. [16]

$$Var(f_{com}) = Var\left(\sum_j \frac{1}{L} d_j\right) = \frac{1}{L^2} Var\left(\sum_j d_j\right) = \frac{1}{L^2} L.Var(d_j) = \frac{1}{L} Var(d_j) \quad (1)$$

As given by Eq. (1) the variance of the overall model (f_{com}) is guaranteed to be lower that an individual model d_j by a factor of total number of models L.

In the study of bagging models, Bühlmann and Yu [17] have established that the hard decisions create diversity and instability rather bagging is observed to smooth those decisions, thus resulting in much smaller variance and mean squared error compared to all single models.

In the work on Bagging predictors [18], the author has described the method for generating several versions of a same predictor by taking bootstrapped replicates of the original training data for getting an ensembled classifier through majority voting or predictor by averaging. Later tests on real-time datasets showed that bagging gave considerable increase in accuracy.

D. Proposed Ensemble Model Approach

Table VI shows the accuracy scores of the individual models used in the ensemble algorithm as well as the accuracy obtained by combination of the models under MaxVoting Ensemble learning approach. It is observed that after combining both Random Forest Classifier and Gradient Boosting Classifier using the Voting Classifier, the accuracy of this ensemble model was found to be 89.82%.

TABLE VI

PERFORMANCE COMPARISON OF ENSEMBLE MODELS VS ENSEMBLE VOTING CLASSIFIER

Models	**Accuracy**
Gradient Boosting Classifier	86.16%
Random Forest Classifier	86.32%
Ensemble VotingClassifier	**89.82%**

Based on the results, we can notice that the hybrid model performs better than the individual traditional models but overall, the weakness of a hybrid ML approach where the cases may not be equally distributed across the clusters, can be eliminated and the overall system can be enhanced by merging more techniques through ensemble learning algorithms.

V. CONCLUSION

In this paper, hybrid and ensemble models are experimented to aid the healthcare units in identifying the COVID-19 cases among patients simply based on symptoms before going for any physical tests. Thus, it provides them with a prior prediction of the situation which might be helpful in preparing precautions and measures beforehand and not causing havoc due to unhandled cases in the future. The analysis began by fitting various traditional classification models such as Logistic Regression based classifier, KNN classifier, Entropy based decision tree classifier, Random Forest classifier and Gradient Boosting classifier to the data ofsymptoms. The summarised results show that Gradient Boosting and Random Forest classifier emerges as the best choice for further analysis in Hybrid and Ensemble models owing to their high accuracy and acceptable type II errors. Furthermore, clustering followed by classification hybrid modelling outperformed all the traditional models with an improved accuracy of 87%. Finally, the MaxVoting ensemble

model, comprising Random Forest and Gradient Boosting algorithms further boosted the accuracy to almost 90% and emerged as the best approach for our current data. In future, it is planned to study the choice of optimal hyper parameters tuned for different ML algorithms based on bio-inspired and nature-inspired algorithms.

REFERENCES

[1] Alharthi, H. (2018). Healthcare predictive analytics: An overview with a focus on Saudi Arabia. *Journal of infection and public health*, *11*(6), 749-756.

[2] Huang, C., Wang, Y., Li, X., Ren, L., Zhao, J., Hu, Y., ... & Cao, B. (2020). Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. *The lancet*, *395*(10223), 497-506.

[3] Salgotra, R., Gandomi, M., & Gandomi, A. H. (2020). Time series analysis and forecast of the COVID-19 pandemic in India using genetic programming. *Chaos, Solitons & Fractals*, *138*, 109945.

[4] Zhu, N., Zhang, D., Wang, W., Li, X., Yang, B., Song, J., ... & Tan, W. (2020). A novel coronavirus from patients with pneumonia in China, 2019. New England journal of medicine.

[5] World Health Organization. (2020). Novel Coronavirus (2019-nCoV): situation report, 3.

[6] Rossman, H., Keshet, A., Shilo, S., Gavrieli, A., Bauman, T., Cohen, O., ... & Segal, E. (2020). A framework for identifying regional outbreak and spread of COVID-19 from one-minute population-wide surveys. *Nature Medicine*, *26*(5), 634-638.

[7] Van Calster, B., Wynants, L., Timmerman, D., Steyerberg, E. W., & Collins, G. S. (2019). Predictive analytics in health care: how can we know it works?. *Journal of the American Medical Informatics Association*, *26*(12), 1651-1654.

[8] Ardakani, A. A., Kanafi, A. R., Acharya, U. R., Khadem, N., & Mohammadi, A. (2020). Application of deep learning technique to manage COVID-19 in routine clinical practice using CT images: Results of 10 convolutional neural networks. *Computers in Biology and Medicine*, *121*, 103795.

[9] Pathak, Y., Shukla, P. K., Tiwari, A., Stalin, S., & Singh, S. (2020). Deep transfer learning based classification model for COVID-19 disease. *Irbm*.

[10] Silva, P., Luz, E., Silva, G., Moreira, G., Silva, R., Lucio, D., & Menotti, D. (2020). COVID-19 detection in CT images with deep learning: A voting-based scheme and cross-datasets analysis. *Informatics in Medicine Unlocked*, *20*, 100427.

[11] Minaee, S., Kafieh, R., Sonka, M., Yazdani, S., & Soufi, G. J. (2020). Deep-covid: Predicting covid-19 from chest x-ray images using deep transfer learning. *Medical image analysis*, *65*, 101794.

[12] Shir Avr – Samples to COVID-19 in Israel Dataset - https://www.kaggle.com/shiravr /israel-covid19-dataset

[13] https://en.wikipedia.org/wiki/Ensemble_learning

[14] Xiao, J., Tian, Y., Xie, L., Jiang, X., & Huang, J. (2019). A hybrid classification framework based on clustering. *IEEE Transactions on Industrial Informatics*, *16*(4), 2177-2188.

[15] K. Lakshmi, N. Karthikeyani Visalakshi , S. Shanthi and S. Parvathavarthin. Clustering Categorical Data using K-Modes Based on Cuckoo Search Optimization Algorithm. Ictact journal on Soft Computing, October 2017, Volume: 08, Issue: 01

[16] Benjamin Fredrick David. H, A. Suruliandi. Performance Evaluation of Ensemble Classifiers on Benchmark Datasets. International Conference on Recent Trends in Multi-Disciplinary Research (ICRTMDR-18), A.P.C. Mahalaxmi College for Women, Tuticorin, December 2018.

[17] Bühlmann, P., & Yu, B. (2002). Analyzing bagging. *The Annals of Statistics*, *30*(4), 927-961.

[18] Bühlmann Peter, Bin Yu. Analyzing bagging. The Annals of Statistics 30.4 (August 2002): 927-961.

[19] Leo Breiman. Bagging predictors. Machine learning 24, 123-140 (August 1996).