SOUTH ASIAN JOURNAL OF SOCIAL SCIENCES & HUMANITIES

ISSN: 2582-7065 (Online)

SAJSSH, VOL 2, ISSUE 3, PP. 203-221

COVID-19 Patient Health Prediction using Artificial Intelligence Boosted Random Forest Algorithm

Abdul Subhan[1], Tuba Rasheed[2], Zarwa Shah[3], Sadia Noor[4], Muhammad Aamir Khan[5] and Usman Shakoor[6]

[1,2,3,4]*Masters Scholar at Economics Department COMSATS University Islamabad.*
[5,6]*Assistant Professor at Economics Department COMSATS University Islamabad.*

Corresponding Author: Abdul Subhan, **Email:** subhan2142@gmail.com

Received: 19th January 2021 **Accepted:** 20th May 2021 **Published:** 3rd June 2021

ABSTRACT

In the current times, there is high demand for artificial intelligence (AI) techniques to be integration with real-time collection, wireless infrastructure, as well as processing in terms of end-user devices. It is now remarkable to make use of AI for detection as well as prediction of pandemics that are extremely large in nature. Coronavirus pandemic of 2019 (COVID-19) began in Wuhan, China and caused the deaths of 175,694 deaths around the world, while the number of active patients stands at 254,4792 patients around the world. In Pakistan, from January 2020 March 2021, there have been 658,132 positive cases, 603,512 recovered cases of COVID-19 with 16,208 deaths, reported by world health organization. Nonetheless, the quick and exponential increase in COVID-19 patients has made it necessary that quick and efficient predictions be made in terms of the possible outcomes with respect to the patient for the sake of suitable treatment by making use of AI techniques. A fine-tuned random forest model has been proposed by this paper, which has been given a boost by AdaBoost algorithm. The COVID-19 patient's health, geographical area, gender, and marital status are used for the prediction of severity in terms of cases as well as possible outcomes, either recovery or no recovery (i.e. death). The model is 90% accurate and has a 0.76 F1 Score on the set of data used. Analysis of data shows a positive correlation with respect to the gender of patient, and death. It also shows that most of the patients had ages between twenty years and seventy years.

KEYWORDS

COVID-19, Predictions, Boosted Random Forest Algorithm, Pakistan

DOI: http://doi.org/10.48165/sajssh.2021.2313

INTRODUCTION

Being a vast industry, needing real time gathering as well as processing with respect to medical data, the healthcare industry faces the issue of data handling that needs predictions as well as dissemination of information done in real time so that practitioners can provide medical attention on time. The main stakeholders of this industry, including physicians, hospitals, vendors, and companies based on health have made attempts at collecting, managing, reviving data so as to use it for the enhancement of medical practices as well as innovation with respect to technology. It has become a difficult task to deal with healthcare data because data has a massive volume, there are various issues related to security, incompetence related to application of wireless network application, as well as the velocity related to its increase. Therefore, for the sake of increased efficiency, more accuracy, as well as workflow, data analytics tools are required for the management of complex data by healthcare industries.

The coronavirus pandemic has caused an outbreak of respiratory illnesses around the globe that began in China's Wuhan region. Researches have demonstrated that the clinical characteristics of Covid-19 are similar to those of severe acute respiratory syndrome coronavirus 2(SARS-COV-2). The most common symptoms of COVID19 are cough and fever, meanwhile gastrointestinal symptoms are less common. Fever is more frequently found in patients who are infected with viruses such as MERS Corona Virus (2%) as well as SARS corona virus (1%) as compared to those infected with COVID-19 (2020); thus it is possible that the surveillance system misses non-febrile patients that focuses primarily on the detection of fever (Zumla A et al. 2015). The patients that are infected with COVID-19 were largely associated with the animal and seafood market in Wuhan which indicated a spread from animal to person. On the contrary, many patients have not been associated to the animal markets, demonstrating a human to human transmission. The coronavirus pandemic is considered a health emergency around the globe, spreading at a rapid pace (Pham QV et al. 2020).

COVID19 originated from Wuhan in China and has caused the deaths of 175,694 deaths around the world, while the number of active patients stands at 254,4792 patients around the world (WHO Situation Report 2020). In Pakistan, from January 2020 March 2021, there have been 658,132 positive cases, 603,512 recovered cases of COVID-19 with 16,208 deaths, reported by world health organization (see figure 1). As of March 2021, a total of 659,467 vaccine doses have been administered. (WHO 2021a). Figure 2 a, b and c illuminated the distributed

DOI: http://doi.org/10.48165/sajssh.2021.2313

percentage of COVID19 positive cases, recovered cases and deaths across provinces in Pakistan: Punjab, Sindh, Khyber-Pakhtunkhwa, and Baluchistan for the period January 2020 to March 2021. In provinces, Punjab and Sindh have more COVID19 positive and death cases, while, Baluchistan and Khyber-Pakhtunkhwa have less positive and death cases. Moreover, Figure 1c irradiated that in Sindh province COVID19 recovered cases are growing at a much faster pace than predicted.

Figure 1: **COVID19 Positive Cases, Recovered Cases and Deaths all over the Pakistan since January 2020 to March 2021**

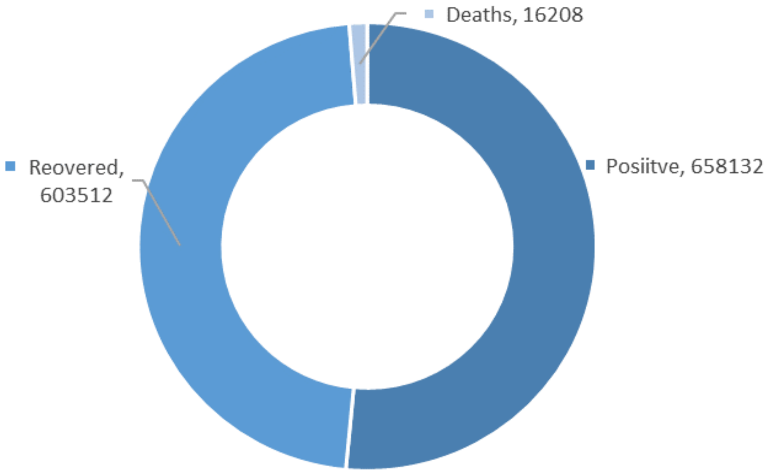
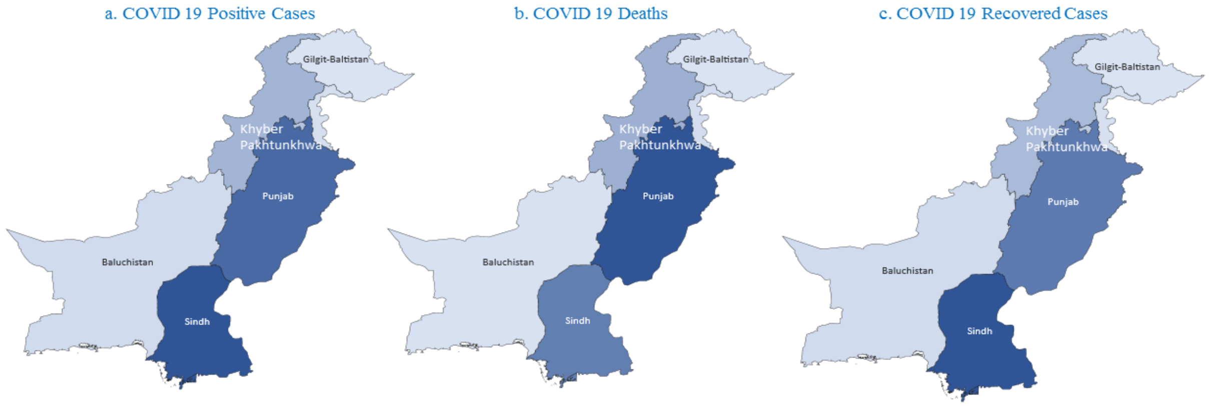


Figure 2: **COVID19 Positive Cases, Recovered Cases and Deaths across provinces in Pakistan since January 2020 to March 2021**



(Source: WHO March 2021) Author Centric Visualization

As medical facilities are under massive stress, it is important that healthcare facilities and governments focus on the identification and treatment of cases that have more probability of surviving, and thus making good use of the scarce stock of medications as well as medical resources. The 21st century has seen the advent of the breakthrough technology called artificial intelligence (AI) which has many applications in different fields including weather prediction, autonomous systems and astronomical exploration etc. (Kathiresan S et al. 2020). Some related researches have applied artificial intelligence to detect, prevent and predict so the pandemic can be fought. Researchers in Wang and Wong (2020) made use of a convolutional neural network-based model for the detection of COVID19 patients by making use of CXR images. A pretrained ImageNet was used and the model was trained on open source dataset related to Chest X-Ray images (CXR). Pal et al. (2020) on the other hand, used a LSTM model for the prediction of country specific risk related to COVID19 which depends upon trends as well as weather data of that specific country so that the likely spread of COVID19 within that country can be predicted. In Liu et al. (2020) the ML was applied by AI practitioners so that the internet activity, health organization reports, news reports and media activity can be processed so that the spread of the coronavirus outbreak can be predicted in China on providence level (Cai H, 2020). The authors in Bayes and Valdivieso (2020) used the Bayesian approach so that the number of deaths can be predicted in case of Peru about 70 days in future, making use of Chinese empirical data. In Beck et al. (2020) authors have made use of Artificial Intelligence for the identification of drugs that are commercially available and can be used for the treatment of COVID19 patients. The Bidirectional Encoder Representations have been used with respect to the Transformers (BERT) framework which lies at the center of the model. Researchers in Tang et al. (2020) used the random forest algorithm so that the severity analysis with respect to COVID-19 patients can be carried out, employing the computed tomography (CT) Scans. The authors in Khalifa et al. (2020) employed a fine-tuned model based on generative adversarial network for the detection of pneumonia using chest X-rays, that is one of COVID19's symptoms. The researchers in Sujatha et al. (2020) employed a method that may help predict how far COVID19 would stretch in India, using linear regression as well as vector autoregression and multilayer perceptron that would help give expectations related to Kaggle information on COVID-19, so that the epidemiological pattern can be anticipated with respect the COVID19 cases. Kutia et al. (2019) attempted to break down the perspectives of clients with respect to applications of E health in China as well as the framework of eHealth in Ukraine, that gave us knowledge as well as suggestions for improvement with regard to an application of eHealth called eZdorovya for benefits related to health information. Sultan et al. (2019) came up with a hybrid method for the generation and

DOI: http://doi.org/10.48165/sajssh.2021.2313

facilitation of patients that have Alzheimer’s so that their memories can be recalled. A summary based on an egocentric video used important individuals, medicines as well as objects as tools so that their method could be used. Moreover, Feng et al. (2019) has proposed emerging tacline nanonetwork based on the internet, promising an innovative range of applications related to e-health. The researchers have used transmit network based on information which uses terahertz band to go to an operator. Eventually, Iwendi et al., (2020) the researchers came up with a range of strategies with the intention of speaking to, improving, and enabling multidisciplinary as well as multi institutional ML for the exploration of healthcare informatics. A novel way regarding the deep learning structure driven by internet of health things (IoHT) used for the purpose of identifying and arranging cervical cancer in the form of pap smear pictures, using ideas related to transfer learning, was introduced by Khamparia et al. (2020). A technique used for the production of manufactured chest X-ray (CXR) pictures, making use of an Auxiliary Classifier Generative Adversarial Network (ACGAN) used model, which is given the name of CovidGAN, was suggested by Waheed et al. (2020). Sakarkar et al. (2020) has proposed a mechanized discovery & characterization model based on learning, with respect to fundus DR pictures.

The aim of this paper is to bridge the gap with respect to traditional healthcare systems, by making use of the machine learning (ML) algorithms for the simultaneous processing of travel data as well as healthcare data, together with different parameters related to patients infected with COVID19, for the prediction of probable outcome related to the patient, on the basis of symptoms, history of travel, as well as delay with regard to the reporting of a case through identification of patterns using previous data of patients. We contribute in the form of i. Processing data related to healthcare as well as travel by means of algorithms of machine learning instead of traditional healthcare systems for the identification of people infected with COVID19. ii. This research work has made a comparison of multiple algorithms to process patient data and has made an identification of boosted random forest to be the finest of these methods. After this, a grid search was executed for the fine tuning of hyper parameters with regard to boosted random forest algorithm for the improvement of performance. Through this research work, the requirement of re-comparison of existing algorithms to process patient data related to COVID19 is obliterated. Researchers will be further able to work towards the development of a solution which provides a combination of processing with regard to patient demographics, health data, as well as travel data to predict the health outcomes of COVID19 patients in a better way. The study has been organized in this way: The Methodology section

DOI: http://doi.org/10.48165/sajssh.2021.2313

dilates upon materials and methods used, as well as description of the dataset, preprocessing of data, as well as data analysis regarding the classification algorithms that have been employed. The results section focuses on the outcomes of the experiment, after which the discussion section is included. The conclusion section summarizes the outcomes, providing a conclusion as well as the scope of this current work in the future.

Table 1: ***Brief Overview of Existing Studies***

S No.	Writer(s)	State(s)	Data	Variables	Approach	Conclusion
1	Tobias et al. 2020	Italy and Spain	February 2, 2020, to March 13, 2020 Under Lockdown	COVID-19 incident & death cases	Quasi-Poisson Regression Model	After the first lockdown, incidence trends were considerably reduced in both countries. However, although the slopes have been flattened for all outcomes, the trends kept rising. During the second lockdown, implementing more restrictive measures for mobility, it has been a change in the trend slopes for both countries in daily incident cases and ICUs.
2	Chakraborty and Ghosh 2020	Canada, France, India, South Korea, and the UK	4 April 2020 to 12 May 2020	COVID-19 death cases	Integrated Moving Average Model and Wavelet-based Forecasting Model	Wavelet-based Forecasting Model outperformed in forecasting. The proposed model can be used as an early warning system to fight against the COVID-19 pandemic.
3	Sung-mok Jung et al. 2020	Outside of mainland China	24th January to 5th February 2020	COVID-19 exported cases	Exponential Growth Model	The ongoing COVID-19 epidemic is most likely to become a pandemic
4	Sujatha et al. 2020	India	January 22, 2020, to April 10, 2020	COVID-19 death & recovered cases	Linear Regression (LR), Multilayer perceptron (MLP) &Vector Autoregression Model (VAR)	It is concluded that the MLP method is giving good prediction results than that of the LR and VAR method.

DOI: http://doi.org/10.48165/sajssh.2021.2313

5	Mohammed et al. 2020	China	21 January 2020 to 18 February 2020	COVID-19 death cases	Adaptive Neuro-Fuzzy Inference System (ANFIS) Framework. Enhanced flower pollination algorithm (FPA) and Salp swarm algorithm (SSA)	The results suggest that the FPA and SSA perform better than the other parameter selection methods.
6	Hiroshi Nishiura et al. 2020	Japan	December 2019-January 2020	Pneumonia cases linked to coronavirus	Maximum likelihood method	The cumulative incidence as of January 24, 2020 of coronavirus in China was found to be 5502 cases, and number of infections may not lie in hundreds but rather in thousands
7	Leonard et al. 2020	USA	March 4 to April 4 2020	Clinical, hospital level and demographic characteristics with respect to death in COVID-19 patients	Multilevel logistic regression	It was found that 35.4 percent patients succumbed to death 28 days following admission in ICU, 37.2% got discharged while 27.4% stayed in hospital.
8	Asif Javed (2020)	Pakistan	2020	COVID-19, Global value chains, Worker's remittances	Descriptive Analysis	Coronavirus has affected human lives severely but also adversely affected economic markets around the world, putting millions of jobs at risk.
9	Rosen and Stenbeck (2020)	Sweden	2020	Expected future unemployment rates owing to COVID-19	Systematic review, Data of important statistics	Interventions made to control the pandemic as well as the shutting down of economic activities has led to an overall rise in mortality
10	IHME COVID-19 health service utilization forecasting team, Christopher JL Murray	USA & Europe Economic Area Countries	2020	COVID19 deaths, hospital capacity as well as utilisation	Mixed effects nonlinear regression framework, extended mixture model, micro-simulation model	Together with a lot of deaths due to COVID19, the pandemic will also cause great burden on the resources of health systems which is not within hospital capacity with respect to EEA and USA, for ventilator use as well as ICU care to be specific.
11	Tetsuro Kobayashi et al. 2020	Japan	2020	COVID-19 death cases	fSimple Division, Non-parametric survival analysis	Dividing total number of deaths by number of cases must be adjusted for delay from the start

						of illness to reporting. If only the confirmed cases are assessed, insights would be limited regarding how severe the case of all infected individuals is
12	Natalie M. Linton et al. 2020	Japan	2020	COVID-19 cases and COVID-19 deaths	Probability density function	Incubation period lies in the range of two days upto fourteen days within a confidence interval 95%. The period of quarantine is suggested to be fourteen days.
13	Debanjan Parbat et al.2020	India	1st March to April 30th 2020	Total deaths, total recovered and total confirmed COVID-19 patients	Vector regression model	The model has been about 97% successful in the prediction of deaths as well as patients who have recovered and total confirmed cases while being 87% accurate in the prediction of every day new cases.
14	Jianhua Wu et al. 2021	England and Wales	2014-2020	Place and cause of death during COVID-19, Excess mortality	Farrington surveillance algorithm	Carehomes accounted for a lot of deaths because of undiagnosed COVID19. Compared to this, there were lesser but significant number of excess deaths in homes as well, due to cardiac arrest and cancer which is because of the fact that hospital care is being avoided.
15	Hiroshi Nishiura et al. 2020	Japan	29th to 31st January 2020	Ascertainment rate of infection	Mean serial interval.	Death risk among those who are infected, or the infection fatality risk is between 0.3% and 0.6%. These values can be compared to the pandemic of Asian influenza that took place in 1957–1958

MATERIALS AND METHODS

For this project, the dependencies are on the given libraries and packages; Datetime, Pandas, Numpy, Scikit Learn, Scipy as well as Matplotlib. The implementation of this project has been done on the platform of Google Colab by employing the CPU runtime. For Google Colabs, the CPU specifications are model: 79, model name: Intel(R) Xeon(R) CPU @ 2.20 GHz, model name: and cache size: 56,320 KB, and CPU Family: 6. Google Drive is the storage used.

Dataset

Keeping in view the variables in the context of Pakistan, primary data was collected from 50 respondents by taking interviews from them on call. The variables included gender, marital status, patient health - weather the patient recovered or died, SOPs - whether they were followed or not followed, and whether or not symptoms such as malaise, cold, fatigue & body pain, fever and cough were observed.

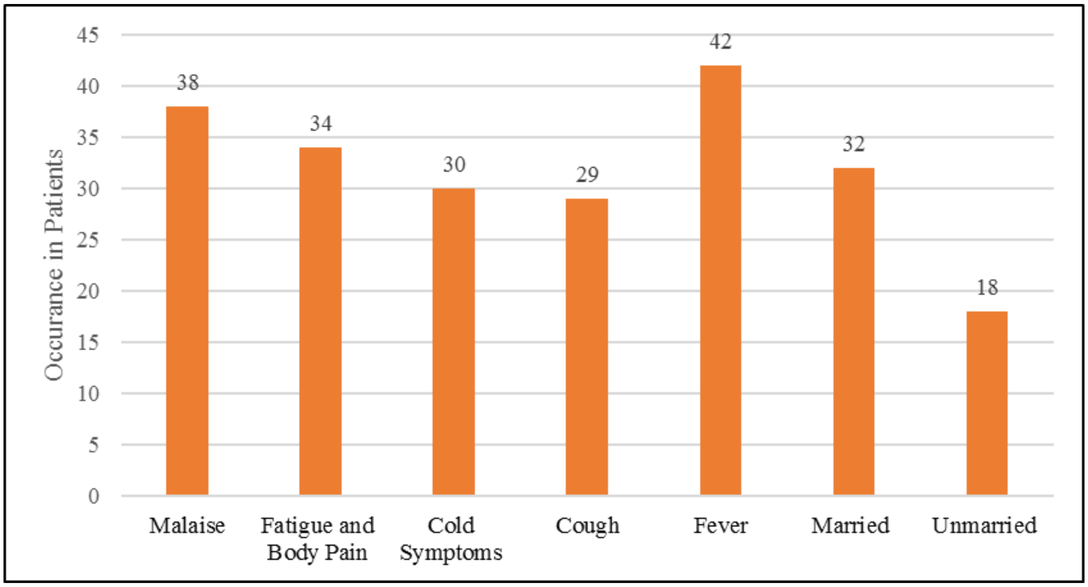
Table2: ***Data Description***

Variables		Values (for Categorical variables)	Type
Location	Urban	Yes (1), No (0)	Numeric, Categorical
	Rural	Yes (1), No (0)	Numeric, Categorical
Age		NA	Numeric
Gender	Female	Yes (1), No (0)	Numeric, Categorical
	Male	Yes (1), No (0)	Numeric, Categorical
Marital Status	Married	Yes (1), No (0)	Numeric, Categorical
	Unmarried	Yes (1), No (0)	Numeric, Categorical
Patient Health	Death	Yes (1), No (0)	Numeric, Categorical
	Recovered	Yes (1), No (0)	Numeric, Categorical
SOPS	Followed	Yes (1), No (0)	Numeric, Categorical
	Not Followed	Yes (1), No (0)	Numeric, Categorical
Symptoms	Malaise	Yes (1), No (0)	Numeric, Categorical
	Fatigue and Body Pain	Yes (1), No (0)	Numeric, Categorical
	Cold Symptoms	Yes (1), No (0)	Numeric, Categorical
	Cough	Yes (1), No (0)	Numeric, Categorical
	Fever	Yes (1), No (0)	Numeric, Categorical

DATA ANALYSIS

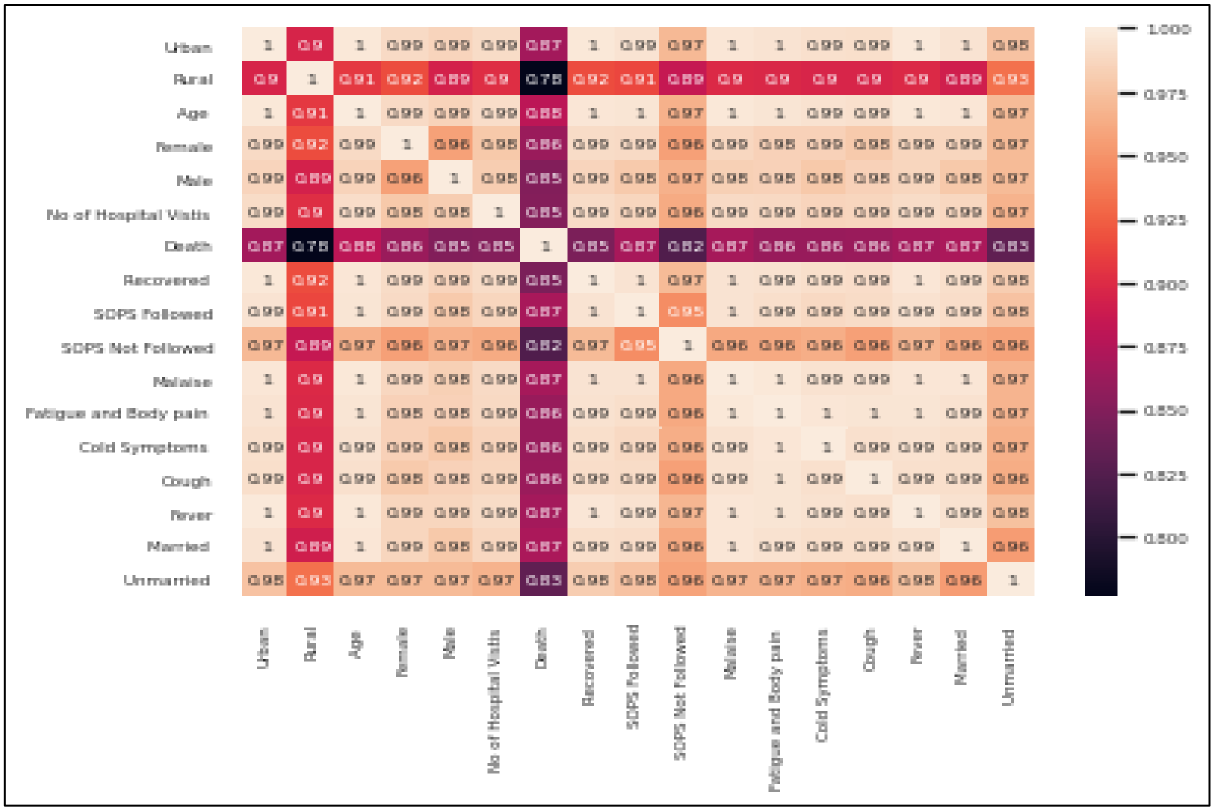
Cough, fever, cold, body pain, fatigue and malaise were found to be the symptoms that were most common among patients for whom the data was available within our set of data and displayed in the first figure.

Figure 1: **COVID 19 Symptoms in Patients**



Correlation among characteristics of the dataset gives important information regarding characteristics as well as the level of impact they have on the value targeted. The Pearson Correlation's heat map among characteristics of the set of data is displayed in the second figure.

Figure 2: **Pearson Correlation's Heat Map**



DOI: http://doi.org/10.48165/sajssh.2021.2313

The correlation of people living in urban with death is equal is lower compared to correlation of people living in rural areas. In contrast the correlation of people living in urban areas with recovery is higher compared to correlation of people living in rural areas. The correlation of age with death is lesser compared to the correlation of age with recovery. Males have a higher correlation with death due to COVID19 as compared to females, while females and males have an equal correlation with recovery. Those who visited hospitals had a lower correlation with death compared to recovery with which the correlation was higher. Those who followed SOPs had a higher correlation with recovery as compared to correlation with death and similar in case of those who did not follow Sops. Those who suffered from malaise, fatigue/body pain, cold, cough and fever had a greater correlation with recovery as compared to death. Both married and unmarried individuals had a greater correlation with recovery as compared to death.

Data Pre-processing

The set of data has consisted of columns, while the data was date, numeric type and string. The categorical variables were also included in the set of data. As the ML model needs the input data to be in the form of numbers, label encoding was done for the categorical variables. Thus, to each unique categorical value included within column, a number is assigned. Some records of patient data include missing values for ‘recov’ as well as ‘death’; columns, patient records of this type have been distinguished from the main set of data and a compilation is done into a test set of data, whereas rest of the records are compiled into a train set of data.

Evaluation Metrics

The given study aims to predict accurately the outcome of some specific patient relying upon many different factors, including demographics, travel history etc. As this is an important prediction, accuracy is crucial. So, for the evaluation of model, three evaluation metrics were considered for the study. These terms were employed in equations: TP represents true positive, TN represents true negative, FP stands for false positive, while FN represents false negative.

i. Accuracy

When the set of data includes (TP+TN) points of data, accuracy equals total correct prediction’s ratio (TP + TN + FP + FN) to total data points by classifier. Accuracy is vital in measuring the

DOI: http://doi.org/10.48165/sajssh.2021.2313

classification model's performance. Accuracy can be calculated according to what has been shown in the below equation:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}\ 0.0 < Accuracy\ < 1.0 \qquad (1)$$

ii. Precision

Precision equals ratio of True Positive (TP) samples with respect to True Positive (TP) and False Positive (FP) samples combined. Precision is an important metric for the identification of the number of patients that have been specified correctly in a set of data that is imbalanced. Precision has been calculated in the 2nd equation given below:

$$\text{Precision} = \frac{TP}{TP+FP} \qquad (2)$$

iii. Recall Score

Recall also equals ratio of True Positive (TP) samples with respect to True Positive (TP) and False Negative (FN) samples combined. Recall is an important metric for the identification of number of patients that were classified in a class set of data that was imbalanced, out of all patients that were capable of being predicted correctly. The calculation of Recall is given in the third equation in the following manner:

$$\text{Recall Score} = \frac{TP}{TP+FN} \qquad (3)$$

iv. F1 Score

The F1 Score equals the recall and precision value's harmonic mean. A perfect balance is struck between precision and recall, thus giving true evaluation with respect to the performance of the model in COVID-19 patient's classification. This is an important measure that will be used for the evaluation of model. F1 Score is calculated as displayed in the fourth equation in the following way:

$$\text{F1 Score} = 2\ X \frac{Precision\ X\ Recall}{recision\ X\ Recall} \qquad (4)$$

DOI: http://doi.org/10.48165/sajssh.2021.2313

RESULTS

Figure 3: **Evaluation Matrix for Decision Tree**

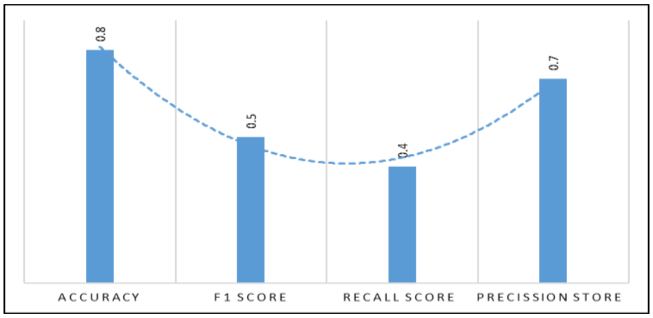


Figure 4: **Evaluation Matrix for Support Vector Machine**

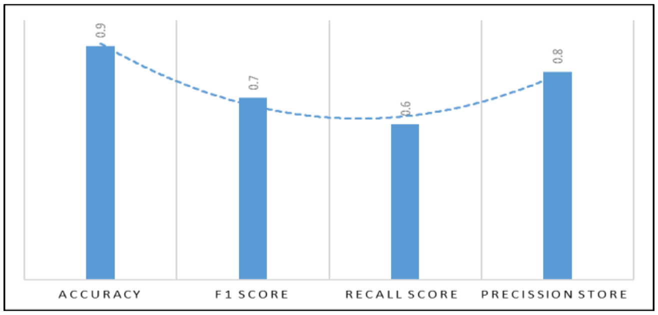


Figure 5: **Evaluation Matrix for Gaussian NB**

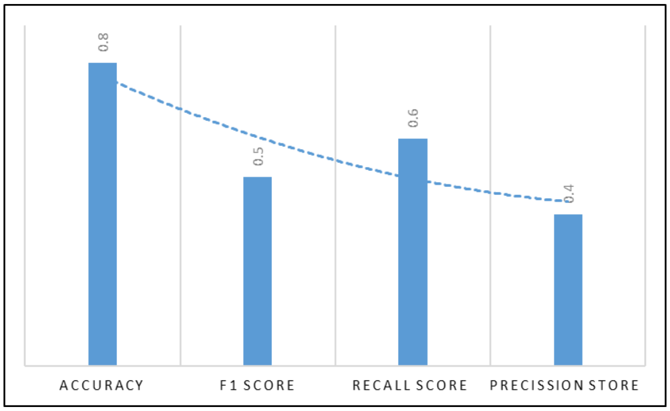
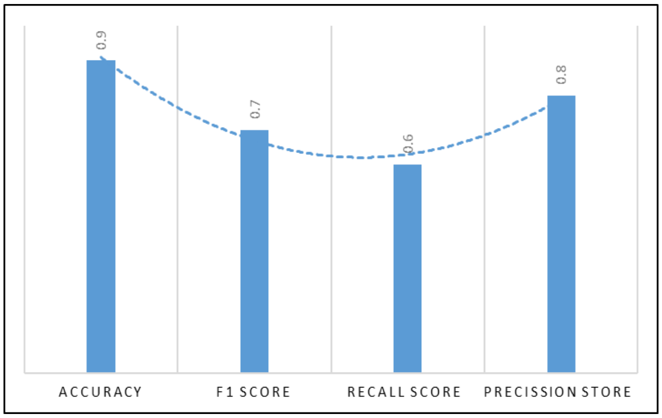


Figure 6: **Evaluation Matrix for Boosted Random Forest**



As the set of data that was employed may be an imbalanced set of data, the F1 score will be used for comparison as the primary metric. Figures three to six show performances of models for each of the mentioned models. Figure 7 demonstrates decision tree that has been constructed so that the target variable can be estimated. Decision tree has depth equal to 2, while the Gini index for each of the nodes equals s <0.5, indicating imbalance with respect to training data. As the best model in terms of performance is the Boosted Random Forest algorithm, the model will be fine-tuned for the sake of improved performance with respect to the set of data.

DISCUSSION: BOOSTED RANDOM FOREST CLASSIFICATION

Boosted Random Forest represents an algorithm consisting of two parts; including the boosting algorithm: Random Forest classifier algorithm and the AdaBoost algorithm (27), that are made

DOI: http://doi.org/10.48165/sajssh.2021.2313

up of many decision trees. Models are built by decision trees that are quite like an actual tree. The data is divided into small subsets by the algorithm, while also adding branches with respect to the tree at the same time. The result of this is a tree which consists of decision nodes as well as leaf nodes. There are two or more than two branches of a decision node, which represent value for every characteristic for example: age, symptom1 and so on that has been tested. Leaf nodes have the resultant value with respect to prospective condition of the patient which is the target value. Multiple classifier decision trees that are ensemble of classifiers, removes risk of failure with respect to one single decision tree so that the target value can be predicted correctly. Therefore, the result obtained from multiple trees is averaged by the random forest so that the final result can be provided. Equation 5 expresses margin function with respect to random forest. Equation 6 shows generalization error, while equation 7 shows confidence in prediction. h_1 (x) , h_2 (x) , . . . , h_k (x) represent ensemble of classifiers that are the decision trees & training data has been taken from X and Y vectors.

Margin function may be expressed in the following manner:

$$\text{mg (X, Y) = a}v_k\text{I (}h_k\text{ (X) = Y)} - \text{ma}x_j \neq \text{Ya}v_k\text{I (}h_k\text{ (X) = j) (5)}$$

Here, I(.) represents indicator function. Following is the generalization error:

$$PE^{*} = P_{X,Y}\ (mg(X,Y) < 0)$$

On the X, Y space, probability is expressed. h_k (X) = h(X, Θ_k) in random forests, thus number of classifiers or decision trees rises, for the entire sequences with respect to trees. The probability PE^{*} and Equation (7) show convergence, from tree structure as well as Strong Law of Large Numbers.

$$P_{X,Y}\ \text{(P }\Theta\text{ (h (X, }\Theta\text{) = Y)}-\text{ ma}x_j \neq YP_{\Theta}\ \text{(h(X, }\Theta\text{) = j) < 0) (7)}$$

Boosting algorithm AdaBoost (28) is applied and it gives corrective mechanism so that the model may be improved following each prediction of the state of patient. In the end, the decision is based on summation of every base model. This is one of ML’s most efficient techniques.

Corrective mechanism may be given in the following manner:

Equation (8). Given $(x_1,\ y_1)$, . . . , $(x_m,\ y_m)$. Here, x_i ∈ X, y_i ∈

$$\text{Y} = \{-1 + 1\}\ \text{ For, t =1, . . . , T. Initialize } D_1 \text{ (i)=} \frac{1}{m}$$

Following the training of a weak learner, which in this case is a random forest, by making use of distribution D_t

DOI: http://doi.org/10.48165/sajssh.2021.2313

Getting hypothesis, h_t : X → {−1, +1},

Alongside error $e_t = P_{r_{i\sim D_t}}[h_t(x_i) \neq y_i]$

Following the choice of α $=\frac{1}{2}\ln\ln\left(\frac{1-e_t}{e_t}\right)$

Update: $D_{t+1}(\mathrm{i}) = \frac{D_t(i)}{Z_t}$ x $\{e^{-\alpha t}\ if\ h_t(x_i) = y_i\ e^{-\alpha t} if h_t(x_i) \neq y_i$

where, normalization factor is Z_t. Final hypothesis is obtained in the following manner:

$$H(x) = Sign\left(\sum_{t=1}^{T}\alpha_t h_t\ (x)\right)$$

The dependent variable is this case is state of patient (recovered/dead) whereas the explanatory variables are gender, age, marital status, whether or not SOPS were followed, or none of the symptoms (1–6). The boosted random forest has been used as it has accurate classification performance even in case of sets of data that are imbalanced (25, 29).

***Figure 7*: Decision Tree**

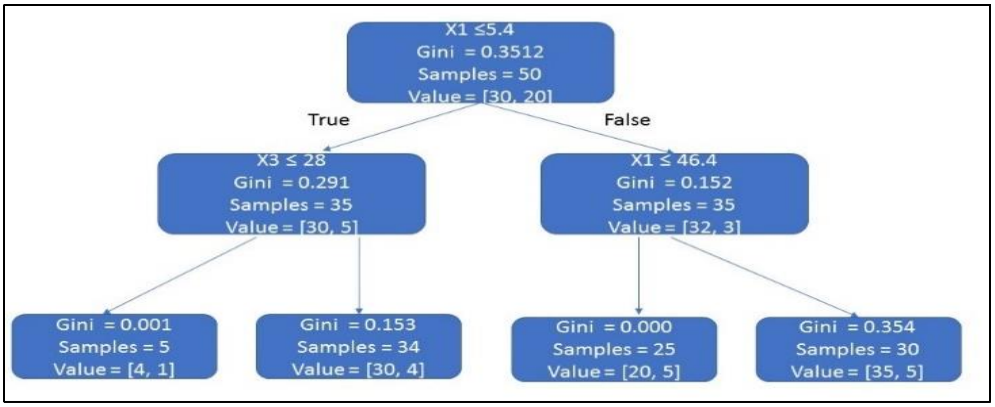


***Figure 8:* Decision Tree 1**

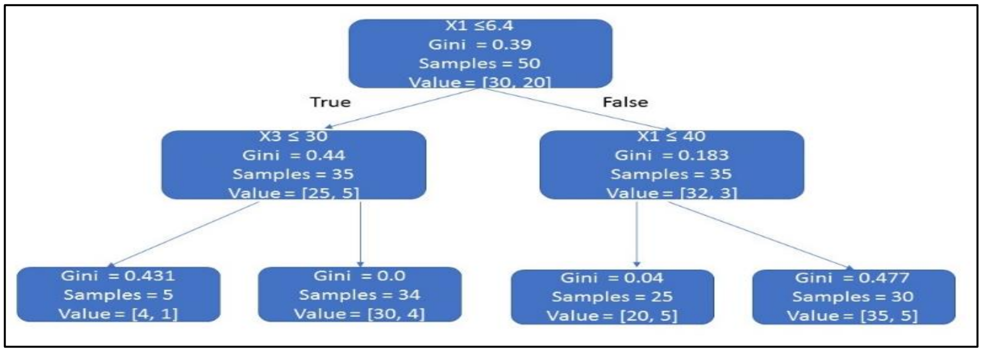


Figure 9: **Decision Tree 10**

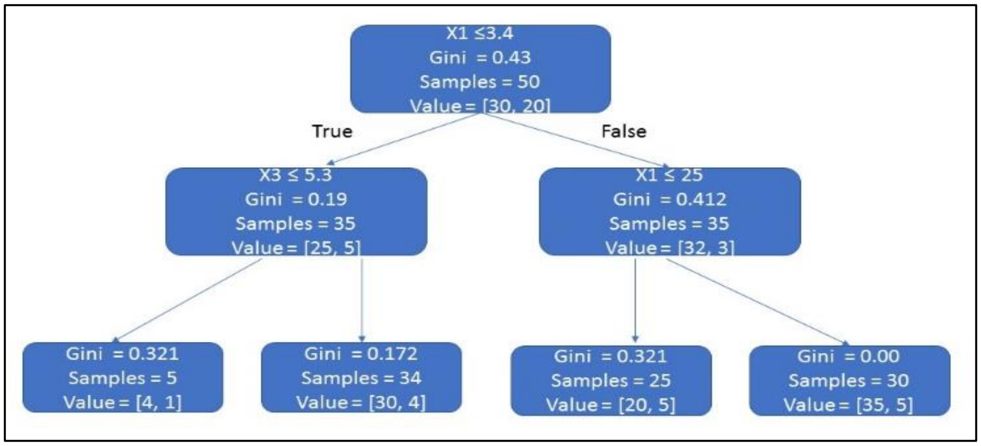


Figure 10: **Decision Tree 25**

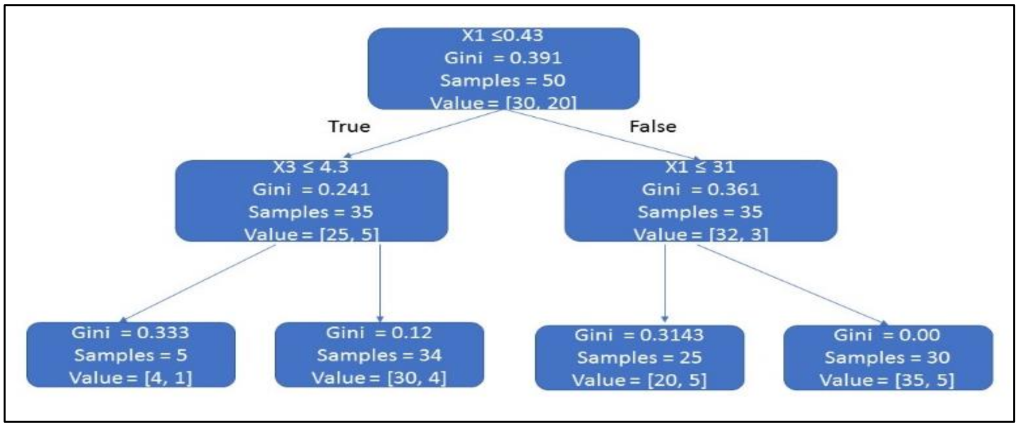
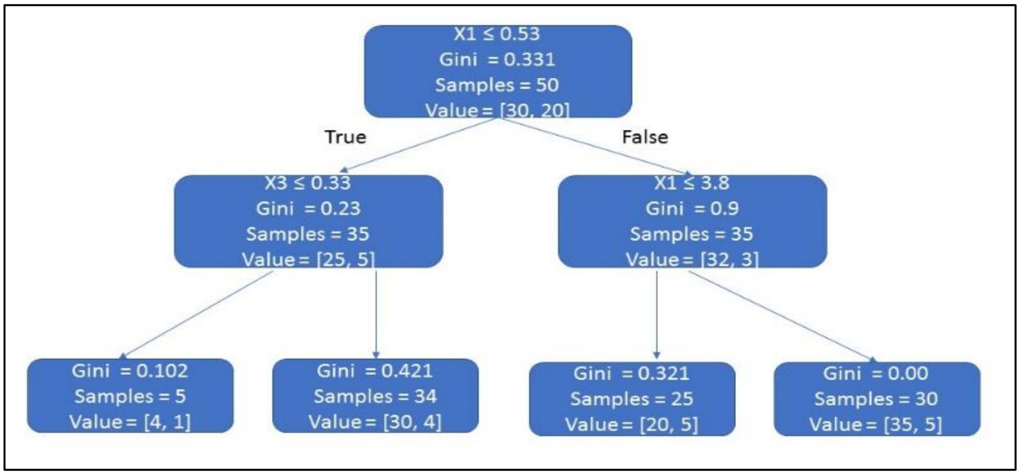


Figure 11: **Decision Tree 100**



Decision trees that have been shown in he figures 8, 9, 10 and 11 equals 2 in terms of depth. Additionally, Gini index with respect to every leaf node of each tree is that depth of trees has been decreased to 2 and number of decision trees (estimators) have been increased to 50 within random forest. Thus, high variance is prevented in the model and accurate predictions are provided.

CONCLUSION AND FUTURE WORK

It is very important to apply Artificial Intelligence so that patient data can be processed for the efficacy of strategies related to treatment. This research work brought forward a model which provides implementation of Random Forest algorithm that has been boosted through AdaBoost algorithm that has an F1 score equal to 0.76 with respect to the set of data for COVID-19 patients. It has been discovered that accurate predictions are provided by the Boosted Random Forest algorithm even with respect to sets of data that are imbalanced. The data employed in the analysis for this study has shown that urban areas had higher death rates compared to death rate in Pakistan's rural areas. Also, death rates were higher among male patients in comparison with female patients. Most of the patients that were affected were between 20 years and 70 years of age. Work done in the future will be directed towards the creation of a pipeline which provides a combination of CXR scanning computer vision models as well as healthcare and demographic models that deal with data processing. Later, integration among models and applications will be done so that growth with respect mobile healthcare can be supported. Thus, a step can be taken towards a diagnostic system which is semi-autonomous and can supply screening as well as detection for regions affected by COVID-19 at a rapid pace, so that we can be well prepared for outbreaks in the future.

REFERENCES

Bai, Y., Yao, L., Wei, T., Tian, F., Jin, D. Y., Chen, L., & Wang, M. (2020). Presumed asymptomatic carrier transmission of COVID-19. *Jama*, *323*(14), 1406-1407.

Bayes, C., & Valdivieso, L. (2020). Modelling death rates due to COVID-19: A Bayesian approach. *arXiv preprint arXiv:2004.02386*.

Beck, B. R., Shin, B., Choi, Y., Park, S., & Kang, K. (2020). Predicting commercially available antiviral drugs that may act on the novel coronavirus (SARS-CoV-2) through a drug-target interaction deep learning model. *Computational and structural biotechnology journal*, *18*, 784-790.

Cai, H. (2020). Sex difference and smoking predisposition in patients with COVID-19. *The Lancet Respiratory Medicine*, *8*(4), e20.

Chen, L., Zhou, M., Dong, X., Qu, J., & Gong, F. Y. (2001). han Y. *Yang F, Zhang tJ*.

Feng, L., Ali, A., Iqbal, M., Bashir, A. K., Hussain, S. A., & Pack, S. (2019). Optimal haptic communications over nanonetworks for E-health systems. *IEEE Transactions on Industrial Informatics*, *15*(5), 3016-3027.

Freund, Y., Schapire, R., & Abe, N. (1999). A short introduction to boosting. *Journal-Japanese Society For Artificial Intelligence*, *14*(771-780), 1612.

Iwendi, C., Bashir, A. K., Peshkar, A., Sujatha, R., Chatterjee, J. M., Pasupuleti, S., .& Jo, O. (2020). COVID-19 patient health prediction using boosted random forest algorithm. Frontiers in public health, 8, 357.

Jain, V., & Chatterjee, J. M. (2020). *Machine Learning with Health Care Perspective*. Springer International Publishing.

Khalifa, N. E. M., Taha, M. H. N., Hassanien, A. E., & Elghamrawy, S. (2020). Detection of coronavirus (covid-19) associated pneumonia based on generative adversarial networks and a fine-tuned deep transfer learning model using chest x-ray dataset. *arXiv preprint arXiv:2004.01184*.

Khamparia, A., Gupta, D., de Albuquerque, V. H. C., Sangaiah, A. K., & Jhaveri, R. H. (2020). Internet of health things-driven deep learning system for detection and classification of cervical cells using transfer learning. *The Journal of Supercomputing*, 1-19.

Khalilia, M., Chakraborty, S., & Popescu, M. (2011). Predicting disease risks from highly imbalanced data using random forest. *BMC medical informatics and decision making*, *11*(1), 1-13.

Kobayashi, T., Jung, S. M., Linton, N. M., Kinoshita, R., Hayashi, K., Miyama, T., ... & Nishiura, H. (2020). Communicating the risk of death from novel coronavirus disease (COVID-19).

Kutia, S., Chauhdary, S. H., Iwendi, C., Liu, L., Yong, W., & Bashir, A. K. (2019). Socio-Technological factors affecting user's adoption of eHealth functionalities: A case study of China and Ukraine eHealth systems. *IEEE Access*, *7*, 90777-90788.

Liu, D., Clemente, L., Poirier, C., Ding, X., Chinazzi, M., Davis, J. T., ... & Santillana, M. (2020). A machine learning methodology for real-time forecasting of the 2019-2020 COVID-19 outbreak using Internet searches, news alerts, and estimates from mechanistic models. *arXiv preprint arXiv:2004.04019*.

Linton, N. M., Kobayashi, T., Yang, Y., Hayashi, K., Akhmetzhanov, A. R., Jung, S. M., ... & Nishiura, H. (2020). Incubation period and other epidemiological characteristics of 2019 novel coronavirus infections with right truncation: a statistical analysis of publicly available case data. *Journal of clinical medicine*, *9*(2), 538.

Leonard, S., Atwood, C. W., Walsh, B. K., DeBellis, R. J., Dungan, G. C., Strasser, W., & Whittle, J. S. (2020). Preliminary findings on control of dispersion of aerosols and droplets during high-velocity nasal insufflation therapy using a simple surgical mask: implications for the high-flow nasal cannula. Chest, 158(3), 1046-1049.

Nishiura, H., Kobayashi, T., Yang, Y., Hayashi, K., Miyama, T., Kinoshita, R., ... & Akhmetzhanov, A. R. (2020). The rate of under ascertainment of novel coronavirus (2019-nCoV) infection: estimation using Japanese passengers data on evacuation flights.

Pal, R., Sekh, A. A., Kar, S., & Prasad, D. K. (2020). Neural network based country wise risk prediction of COVID-19. *Applied Sciences*, *10*(18), 6448.

Parbat, D., & Chakraborty, M. (2020). A python based support vector regression model for prediction of COVID19 cases in India. *Chaos, Solitons & Fractals*, *138*, 109942.

Pham, Q., & Nguyen, D. C. (2020). T. Huynh-The, W. Hwang, and PN Pathirana,". *Artificial intelligence (AI) and big data for coronavirus (COVID-19) pandemic: A survey on the state-of-the-arts,''IEEE Access*, *8*, 130820-130839.

Pillai, S. K., Raghuwanshi, M. M., & Gaikwad, M. (2020). Hyperparameter tuning and optimization in machine learning for species identification system. In *Proceedings of International Conference on IoT Inclusive Life (ICIIL 2019), NITTTR Chandigarh, India* (pp. 235-241). Springer, Singapore.

Sakarkar, G., Pillai, S., Rao, C. V., Peshkar, A., & Malewar, S. (2020). Comparative study of ambient air quality prediction system using machine learning to predict air quality in smart city. In *Proceedings of International Conference on IoT Inclusive Life (ICIIL 2019), NITTTR Chandigarh, India* (pp. 175-182). Springer, Singapore.

Shankar, K., Sait, A. R. W., Gupta, D., Lakshmanaprabu, S. K., Khanna, A., & Pandey, H. M. (2020). Automated detection and classification of fundus diabetic retinopathy images using synergic deep learning model. *Pattern Recognition Letters*, *133*, 210-216.

Sujatha, R., & Chatterjee, J. (2020). A machine learning methodology for forecasting of the COVID-19 cases in India.

Sultan, S., Javed, A., Irtaza, A., Dawood, H., Dawood, H., & Bashir, A. K. (2019). A hybrid egocentric video summarization method to improve the healthcare for Alzheimer patients. *Journal of Ambient Intelligence and Humanized Computing*, *10*(10), 4197-4206.

Tang, Z., Zhao, W., Xie, X., Zhong, Z., Shi, F., Liu, J., & Shen, D. (2020). Severity assessment of coronavirus disease 2019 (COVID-19) using quantitative features from chest CT images. *arXiv preprint arXiv:2003.11988*.

Tan, Z., Zhang, J., He, Y., Zhang, Y., Xiong, G., & Liu, Y. (2020). Short-term Load Forecasting based on Integration of SVR and stacking. *IEEE Access*.

Wang, L., Lin, Z. Q., & Wong, A. (2020). Covid-net: A tailored deep convolutional neural network design for detection of covid-19 cases from chest x-ray images. *Scientific Reports*, *10*(1), 1-12.

Wu, J., Mafham, M., Mamas, M. A., Rashid, M., Kontopantelis, E., Deanfield, J. E., & Gale, C. P. (2021, April). Place and underlying cause of death during the COVID-19 pandemic: retrospective cohort study of 3.5 million deaths in England and Wales, 2014 to 2020. In *Mayo Clinic Proceedings* (Vol. 96, No. 4, pp. 952-963). Elsevier.

WHO, (2020). Situation Report-94 Coronavirus disease 2019 (COVID-19) 2020. World Health Organization, Geneva.

Yang, X., Yu, Y., Xu, J., Shu, H., Liu, H., Wu, Y., & Shang, Y. (2020). Clinical course and outcomes of critically ill patients with SARS-CoV-2 pneumonia in Wuhan, China: a single-centered, retrospective, observational study. *The Lancet Respiratory Medicine*, *8*(5), 475-481.

Zumla, A., & Hui, D. S. (2015). to C, Stanley Perlman P. *Middle East Respiratory Syndrome HHS Public Access. Lancet*, *386*(9997), 995-1007.

DOI: http://doi.org/10.48165/sajssh.2021.2313