Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 2, 2021, Pages. 3525 - 3531
Received 20 January 2021; Accepted 08 February 2021.

COVID Health Prediction Using Hybrid Additive Model

Donepudi Babitha[1*], M R Narasingarao[2]
[1, 2]Koneru Lakshmaiah Education Foundation, India
*donepudi.babitha@gmail.com

ABSTRACT

The immense application and broad usage of Artificial Intelligence has incorporated it into all real-world domains. Now, with the increase in severity of globalpandemics, the usage of AI in healthcare has been increased. The Corona virus disease 2019, originated in China,has caused devastating effect on public-health and finance worldwide. It even caused in rise in number of deaths in the infected patients. Hence, quick prediction using AI algorithms and treatment of infection is required to evade increase in casualties. This paper proposed a hybrid Additive system, a mixture of the swarm intelligence for the selection of attributes and the prognosis employing the base model Additive Tree. It is assessed by applying several metrics like accuracy, sensitivity, count of decision nodes and features, specificity. It is then contrasted to other conventional approaches like Support Vector Machine, Hybrid Random Forest, and the base model as well. The projected prototype attained highest accuracy in detecting the health outcome of patient with 95.42% accuracy with just 8 features and 9 split nodes resulting in clear visual explanation for the medical practitioner.

Keywords
Random Forest + AdaBoost; Support Vector Machine + SGD; Artificial Intelligence; Additive Tree (ADT); Particle Swarm Optimization (PSO).

Introduction

In health care applications, real-time analysis of data performs a key part for medical examiners to take critical clinical decisions. However, these forecasts have evolved to be immensely complicated with surge in the amount of such health data. To overcome this complexity, ML based analytical tools are adapted to provide efficient and accurate results. COVID-19 is a string of Corona virus family which causes respiration related issues in the people affected. It was originated from Wuhan in China, which consequently outspread all through the globe. The clinical characteristics of this illness are indeed very identical to those of SARS-CoV. Cough, Fever are the major symptoms noticed in the people infected by this virus. However, few people are infected without having any such symptoms. The initial spread of this virus started from animal-to-person in Wuhan market, which later spread its transmission from person-to-person. This triggered to the declaration of this disease as a pandemic as its incidence has escalated at an unprecedent pace. It caused a lot of stress on medical practitioners and government to identify the patients and treat them.

Artificial Intelligence (AI) has been widely used in several applications like finance, predicting climate and astronographical explorations. There are several notable works experimented by scientists by applying AI in predicting and detecting COVID-19. However, this paper aims to apply ML algorithms to handle the medical data and travel history simultaneously to predict the health condition of the patient.

Literature Review

Siraj A (1) contributed by finding sites with S-Nitrosylation by applying deep networks, which utilized the sequences of proteins as inputs. An advanced analysis was done by Gambhir et.al (2) using ML algorithms to find the trends in dissemination of COVID in India. Abdelsalam M and

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 2, 2021, Pages. 3525 - 3531
Received 20 January 2021; Accepted 08 February 2021.

Zahran A (3) formulated a computerized determination of Diabetic Retinopathy by employing SVM to categorize angiography snapshots. A sampling-based system to capture the data of stroke-patients, which is categorized by using Random Forest, was framed by Wang M et.al (4) to classify if a human will have cardiac attack.

An auxiliary model CovidGAN has been suggested by Waheed et.al (5) to generate manufactured pictures of chest X-rays. Equity prices are prognosticated unerringly by enforcing upgraded PSO along with LSTM by Yi Ji et.al (6). In research done by X Wang et.al (7) on CT scans, a supervised framework was developed to identify COVID and the localization of lesion was discovered by employing deep learning models. Pal et al. (8) exploited LSTM framework to detect the risk specific to country and the diffusion of disease in that country. To prescribe insurance products, Y. Guo et.al (9) enforced a Random Forest approach and differentiated with various other core ML models.

To reveal fake information, multiple AI based frameworks which used integrated approaches like Term frequency are executed by Jiang T et.al (10). In survey executed by A. U. Haq et al. (11), Breast Cancer was unerringly distinguished by enforcing SVM categorizer which employed the Relief algorithm for feature assertion. A refined and composite model to anticipate the disease with clear explication has been formulated by Luna J et.al (12). To prognosticate health of COVID victim, Iwendi C et al. (13) applied boosted ML algorithm Random Forest to give precise output.

Methods

The software prerequisites applied in our proposal include SciPy, Pandas, Jupyter Notebook and Matplotlib. Other software requirements include R language and rtemis. The core components of CPU include Intel CORE i5 8^{th} gen 1.60 GHz @3.4 GHz and 6MB cache. The dataset 'Novel Corona Virus 2019 Dataset' (NCVD) (14) is obtained from Kaggle. It was annexed from multiple sources like John Hopkins University and WHO.

Methodology

By applying statistical techniques, an interesting insight discovered was that cold, malaise, fever, cough, body pain is deemed to be the quite classic signs. The birthplace of the person is associated significantly to recovery. Additionally, the symptoms are further positively interlinked. These insights provide crucial information about feature selection and greatly impact the target feature.

The dataset contains mixed data of various data types like Date, Numerical, Categorical and String. Label Encoding has been implemented and features are reshaped into numerical values. Several pre-processing techniques like restoring null values with "NA" are enacted. The records with no information about 'recov' and 'death' columns have been included in the test dataset. The general intent of this examination is to precisely estimate the medical situation of victim based on several components like eugenics and travel history. The dataset is sliced into two assortments in the ratio 80:20.

The principal model is constructed first and then differentiated with conventional ML algorithms like boosted RF (RF + AdaBoost), SVM+SGD. Finally, the three techniques are weighed against our proposed hybrid Additive Model (PSO-ADT). The swarm based PSO (15) algorithm is implemented on the input records to obtain the optimal characteristics which are transferred to

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 2, 2021, Pages. 3525 - 3531
Received 20 January 2021; Accepted 08 February 2021.

the core model, thereby yielding the hybrid Additive tree. The multi-objective function implemented is to achieve a trade-off between accuracy (C) and features count.

$$f(A) = \gamma(1 - C) + (1 - \gamma)(1 - Tf/Td)$$

Where the total attributes are denoted by A and the mutation factor (γ) tunes the optimal features (Td) out of total set (Tf)with high accuracy (C). The particles size is 30 with number of iterations taken as 100. Learning rate has been fine-tuned using GridSearchCV.

Data Analysis

The evaluation metrics applied on each prototype are Sensitivity, Accuracy, Specificity, count of split nodes and dimensions of features. The formulaic representation of the evaluators include as below:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Where the values in numerator include the precisely detected values and the denominator states the total predictions in dataset.

$$Sensitivity = \frac{True\ Positives}{True\ Positives + False\ Negatives}$$

$$Specificity = \frac{True\ Negatives}{True\ Negatives + False\ Positives}$$

The model which yields better results in all metrics can be regarded as better approach in terms of performance. The visual explanation can be used as other metric to decide the best approach.

Results

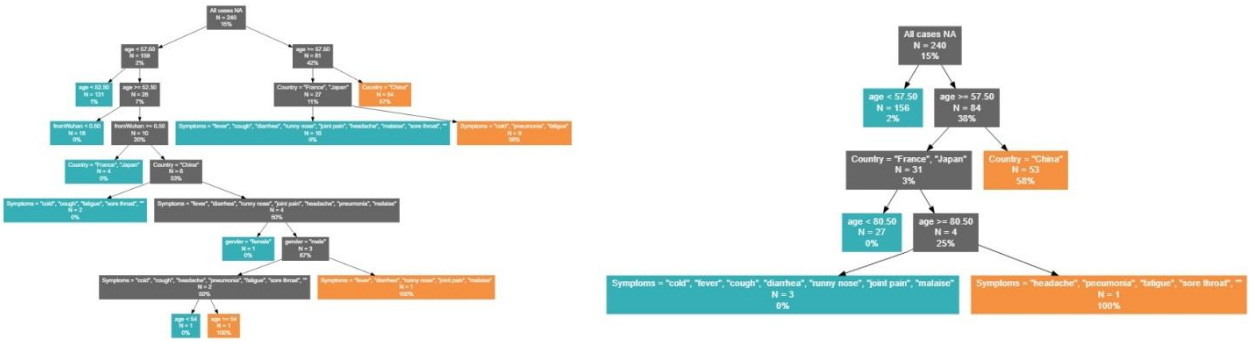

Figure 1. Visual Contrast between standard and recommended models

From figure 1, the vital motives of our research can be clearly interpreted. The comparison done based on count of decision nodes can be clearly observed. With a decrease in count of nodes from 21 to 9 improves the interpretability. It further reduces the tree complexity and time complexity in deriving classification rules.

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 2, 2021, Pages. 3525 - 3531
Received 20 January 2021; Accepted 08 February 2021.

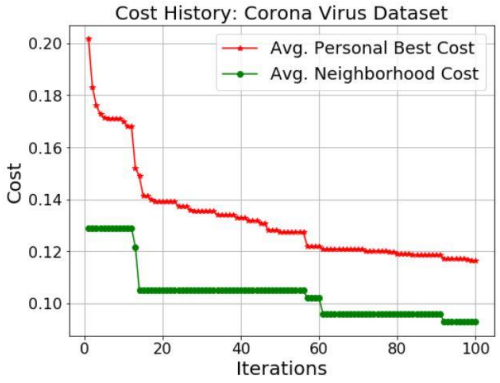

Figure 2. Cost History of PSO in our proposed model

The average values of personal costs and neighbourhood costs are traced for each single loop in the image 2. The framework is run for 100 echoes to acquire the optimal values. The optimal costs are procured progressively upon 60 iterations. The loss values and cost values are utilized to recognize the best features which yield high classification accuracy.

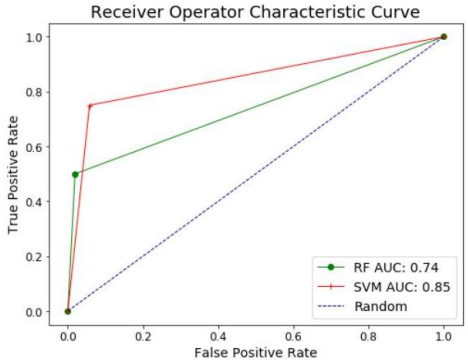

Figure 3. ROC Curve RF Vs SVM

As both the models, RF and SVM, are comparatively equal in predicting the accurate results, we evaluate them using ROC curves as shown in Figure. The curve of SVM has an elevated AUC of 0.85, thus indicating that it is further precise than the former model with AUC of 0.74.

Discussions

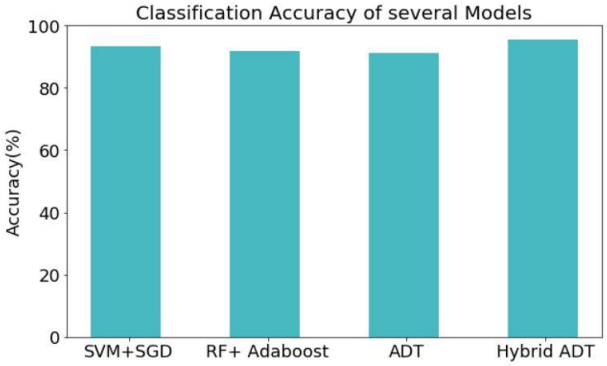

Figure 4. Comparison of accuracy of the applied models

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 2, 2021, Pages. 3525 - 3531
Received 20 January 2021; Accepted 08 February 2021.

The chart depicts the effectiveness of the recommended prototype with other standard ML models applied in terms of accuracy. The former one is highly accurate with 95.42 % compared to other models with ADT (91.25 %), SVM+SGD (93.3 %), RF+ADA (91.66%) respectively.

Table 1. Model Evaluation of applied algorithms

Model	Accuracy (%)	Sensitivity (%)	Specificity (%)	No. of features used	No. of decision Nodes
RF +AdaBoost	91.66	91	92	21	---
SVM + SGD	93.3	94	93	21	---
ADT	91.25	88	92	21	21
PSO-ADT	95.42	97.3	95	8	9

The table differentiates the results of application of several evaluation metrics on all applied models. In terms of statistical terms, the methods are looked at as far as mathematical outcomes acquired. It can be depicted clearly that the proposed framework is extremely precise with 95.42%, 97.3 % sensitivity, 95% specificity, with just 8 features and 9 split nodes.

Conclusion

In this paper, our recommended model PSO-ADT can yield accurate predictions with high interpretability. Algorithms like ADT, RF+ Adaboost, SVM+SGD and PSO-ADT are implemented on the Novel Corona Virus dataset and evaluated using several metrics. The original and recommended approaches are contrasted based on visual explanation, the size of features used and the split nodes in the tree. With a tradeoff between two factors, our recommended classifier can be applied to any kind of real-time application like medicine, healthcare.

Limitations and Future Studies

Theresearch can be further improved by implementing several sophisticated swarm-based algorithms. It can be applied to datasets which contains data of multi-class labels. This model can also be tested by applying on datasets of other domains like finance, education etc.

Acknowledgement

I thank my professor M R Narasinga Rao for providing extreme guidance throughout my research and providing his great mentorship.

References

[1] Siraj, A., Chantsalnyam, T., Tayara, H.,& Chong,K. T. (2021). RecSNO: Prediction of Protein S-Nitrosylation Sites Using a Recurrent Neural Network. *IEEE Access*. 9: 6674-6682. doi: 10.1109/ACCESS.2021.3049142.

[2] Gambhir, E., Jain, R., Gupta, A.,& Tomer, U. (2020). Regression Analysis of COVID-19 using Machine Learning Algorithms. *International Conference on Smart Electronics and Communication, Trichy, India*. (pp. 65-71). doi: 10.1109/ICOSEC49089.2020.9215356.

[3] Abdelsalam, M. M.,& Zahran, M. A. (2021). A Novel Approach of DiabeticRetinopathy Early Detection Based on Multifractal Geometry Analysis for OCTAMacular Images Using Support Vector Machine. *IEEE Access*. 9:22844-22858. doi: 10.1109/ACCESS.2021.3054743.

[4] Wang, M.,Yao, X., & Chen, Y. (2021). An Imbalanced-Data Processing Algorithm forthe Prediction of Heart Attack in Stroke Patients. *IEEE Access*. 9:25394-25404. doi: 10.1109/ACCESS.2021.3057693.

[5] Waheed, A., Goyal, M., Gupta, D., Khanna, A., Al-Turjman, F.,&Pinheiro, P. R. (2020). Covidgan: data augmentationusing auxiliary classifier gan for improved covid-19 detection. *IEEE Access*. 8:91916–91923. doi: 10.1109/ACCESS.2020. 2994762.

[6] Ji, Y., Liew, A. W. -C.,& Yang, L. (2021). A Novel Improved Particle Swarm Optimization with Long-ShortTerm Memory Hybrid Model for Stock Indices Forecast.*IEEE Access*.9: 23660-23671. doi: 10.1109/ACCESS.2021.3056713.

[7] Wang, X., Deng, X., Fu, Q., Zhou, Q., Feng, J., Ma, H., Liu, W., & Zheng, C. (2020). A Weakly-Supervised Framework for COVID-19 Classification and Lesion Localizationfrom Chest CT.*IEEE Transactions on Medical Imaging*.39(8): 2615-2625. doi: 10.1109/TMI.2020.2995965.

[8] Guo, Y., Zhou, Y., Hu, X.,& Cheng, W. (2019). Research on Recommendation of Insurance Products Based onRandom Forest.*International Conference on Machine Learning, Big Data and Business Intelligence*.(pp. 308-311). doi: 10.1109/MLBDBI48998.2019.00069.

[9] Liu, D., Clemente, L., Poirier, C., Ding, X., Chinazzi, M., Davis, J. T., Vespignani, A. & Santillana, M. (2020). A machine learning methodology forreal-time forecasting of the 2019–2020 COVID-19 outbreak using Internet searches, news alerts, and estimates from mechanistic models. arXiv. 2004.04019. Available online at: https://arxiv.org/abs/2004.04019.

[10] Jiang, T., Li, J. P., Haq, A. U., Saboor, A. & Ali, A. (2021). A Novel Stacking Approach for Accurate Detectionof Fake News. *IEEE Access*.9:22626-22639. doi: 10.1109/ACCESS.2021.3056079.

[11] Haq, A. U., Li, J. P., Saboor, A., Khan, J., Wali, S., Ahmad, S., Ali, A., Ahmad, G. (2021). Detection of Breast Cancer Through Clinical Data Using Supervised and UnsupervisedFeature Selection Techniques. *IEEE Access*. 9:22090-22105. doi: 10.1109/ACCESS.2021.3055806.

[12] Luna, J., Gennatas, E. D., Ungar, L. H., Eaton, E., Diffenderfer, E. S. (2019). Building more accurate decision trees with the additivetree.*Proceedings of the National Academy of Sciences.* 116(40): 19887-19893.

[13] Iwendi, C., Bashir, A. K, Peshkar, A., Sujatha, R., Chatterjee, J. M., Pasupuleti, S., Mishra, R., Pillai, S.,& Jo, O.(2020). COVID-19 Patient Health Prediction Using Boosted Random Forest Algorithm. Front. Public Health 8:357. doi: 10.3389/fpubh.2020.00357.

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 2, 2021, Pages. 3525 - 3531
Received 20 January 2021; Accepted 08 February 2021.

[14] Novel Corona Virus 2019 Dataset. (2020). Available online at:https://www.kaggle.com/sudalairajkumar/novel-corona-virus-2019-dataset/.

[15] Hamed, E., Anushya, A., Alzoubi, R., Vincy, B. S. A.,& Balawneh, D. A. (2017). An analysis of particle swarmoptimization for feature selection on medical data. *International Conference on Energy, Communication, Data Analytics and Soft Computing*.(pp.227-231).