World Academy of Science, Engineering and Technology
International Journal of Computer and Systems Engineering
Vol:17, No:10, 2023

Improved Computational Efficiency of Machine Learning Algorithms Based on Evaluation Metrics to Control the Spread of Coronavirus in the UK

Swathi Ganesan, Nalinda Somasiri, Rebecca Jeyavadhanam, Gayathri Karthick

Abstract**—**The COVID-19 crisis presents a substantial and critical hazard to worldwide health. Since the occurrence of the disease in late January 2020 in the UK, the number of infected people confirmed to acquire the illness has increased tremendously across the country, and the number of individuals affected is undoubtedly considerably high. The purpose of this research is to figure out a predictive machine learning (ML) archetypal that could forecast the COVID-19 cases within the UK. This study concentrates on the statistical data collected from 31st January 2020 to 31st March 2021 in the United Kingdom. Information on total COVID-19 cases registered, new cases encountered on a daily basis, total death registered, and patients' death per day due to Coronavirus is collected from World Health Organization (WHO). Data preprocessing is carried out to identify any missing values, outliers, or anomalies in the dataset. The data are split into 8:2 ratio for training and testing purposes to forecast future new COVID-19 cases. Support Vector Machine (SVM), Random Forest (RF), and linear regression (LR) algorithms are chosen to study the model performance in the prediction of new COVID-19 cases. From the evaluation metrics such as r-squared value and mean squared error, the statistical performance of the model in predicting the new COVID-19 cases is evaluated. RF outperformed the other two ML algorithms with a training accuracy of 99.47% and testing accuracy of 98.26% when n = 30. The mean square error obtained for RF is 4.05e11, which is lesser compared to the other predictive models used for this study. From the experimental analysis, RF algorithm can perform more effectively and efficiently in predicting the new COVID-19 cases, which could help the health sector to take relevant control measures for the spread of the virus.

Keywords**—**COVID-19, machine learning, supervised learning, unsupervised learning, linear regression, support vector machine, random forest.

I. Introduction

THE WHO confirmed the latest coronavirus epidemic during the January 2020 World Health International Conference in Geneva. Since January 2020, Wuhan's new coronavirus (SARS-CoV-2) has caused a global pandemic (COVID-19). COVID-19's fast spread and lack of medicines and therapies impacted worldwide life; 120 million worldwide and 4 million UK residents have COVID-19; 2.5 million people died, 126,000 in the UK [24]. Novel coronaviruses have been studied theoretically and statistically lately and since COVID-19 transmits from person to person, it is believed that ML-based digital components may help to forecast the COVID-19 epidemic and alert affected regions and help the retention of spreading it.

The main aim of this research is to critically inspect the trend of COVID-19 cases within UK, to explore various ML models which may be applied for COVID-19 prediction. A comparative performance of RF, LR, and SVM models was evaluated based on the evaluation metrics. Also, the model accuracy depends on the quality of the dataset used and the effectiveness of the proposed models to predict the cases. To improve the model performance, ensemble learning technique has been used to train the multiple models and combine their predictions to make final prediction.

II. Literature Review

ML has lately earned popularity for constructing disease forecasting design due to the difficulty and huge existence of the issue in designing epidemiological models. While ML techniques have been employed to design preceding outbreaks (e.g., Ebola, cholera, swine fever, H1N1 influenza, dengue fever, Zika, oyster norovirus), there exists a discrepancy within literary works for peer-reviewed articles on COVID-19 [1]. Many other kinds of research have centered on assessing documents that debate the use of ML to support the COVID-19 reactions. Authors in [3] recognized seven vital places where ML supervises and regulates the COVID-19 epidemic. In subsequent research, [2] expanded on such seven zones by recognizing and conducting a quick evaluation of the useable research.

According to the authors of [4], the UK COVID-19 outbreak was predicted using an optimization-based approach and an updated version of SEIR model that accounts for vulnerable, susceptible, exposed, and recovered individuals. In [5], the authors reported that ML techniques can be used to predict the spread of COVID-19 both within country and across different countries. COVID-19 infections have been estimated by doctors at various levels, and various methods have been used to predict the severity of the disease, including clinical assessments, ICU grading, ML algorithms and deep learning models. Predictive models based on statistical analyses, CT scans, and symptoms have been used to anticipate the spread of COVID-19, as well as the risk of death and the potential for community transmission. However, clinicians are currently unable to accurately forecast the course of the disease.

Swathi Ganesan, Lecturer, Nalinda Somsiri, Head of the Programme, Rebecca Jeyavadhanam, Lecturer, and Gayathri Karthick, Lecturer, are with Department of Computer Science, York St John University, London, United Kingdom (e-mail: s.ganesan@yorksj.ac.uk, n.somasiri@yorksj.ac.uk, r.balasundaram@yorksj.ac.uk, g.karthick@yorksj.ac.uk).

In [6], a stochastic epidemic model was used to investigate the transmission of an illness over discrete time intervals using a binomial distribution. The National Health Commission of China has reviewed various models for predicting the spread of disease, including those described in literature. Reference [7] proposes a probabilistic pairwise comparison model to analyze the causes, epidemiology, and treatment of clinical illness, and suggests that immunization, isolation, and quarantine measures can reduce the overall sensitivity to the disease.

The authors of [8] experimented with ML models to classify COVID-19 deaths. Other studies, such as [9] and [10], utilized various ML techniques to predict vaccine design and COVID-19 forecasting, respectively. Exponential Smoothing (ES) was found to be a better performer than LR and SVM for forecasting COVID-19 [13]. Additionally, [11] and [12] conducted experiments on COVID-19 forecasting using different approaches, [14] developed an outbreak prediction system for 10 countries, while [15] performed a similar growth prediction for India. In [16], ML techniques were applied to analyze trends in disease spread and identify contributing factors, which could help experts develop strategies to combat the pandemic. In [17], a research study delved into the utilization of machine learning (ML) algorithms. The outcomes highlighted that Support Vector Machines (SVM), and Random Forest (RF) attained the highest accuracies of 77.4% and 95.4%, respectively. These findings indicate that SVM and RF show great potential in effectively comprehending data and achieving remarkable accuracy for future predictions. The researchers in [18] proposed a hybrid approach for predicting COVID-19 cases using ML adaptive neuro-fuzzy inference systems and enhanced beetle antennae search swarm intelligence metaheuristics. The researchers [19] utilized time-series data from January 22, 2020, to January 25, 2021, and implemented a Long Short-Term Memory model with 10 hidden units to predict COVID-19 confirmed and death cases.

In [20], the authors employed a range of computer vision techniques, including image classification, object detection, pattern recognition, and semantic segmentation. Their study aimed to achieve diverse levels of deep learning objectives to characterize the data. The study [21] demonstrates the potential of machine learning algorithms in analyzing and understanding sentiments and experiences related to the pandemic. This can be further extended to sentiment analysis of public opinions and emotions towards COVID-19 control measures, providing valuable insights into the effectiveness and acceptance of various strategies. Additionally, [22] emphasized that the utilization of intelligent tutoring systems can be explored to facilitate remote learning and provide personalized educational resources to students in the context of the pandemic, ensuring continuity in academic progress despite the challenging circumstances.

III. Experimental Study

A. Data Set and Pre-Processing

The data containing COVID-19 information in the United Kingdom from 31st January 2020 to 31st March 2021 are obtained from the WHO. The dataset is preprocessed according to the study of active cases in the UK. However, several noise and inconsistencies were found in the dataset. Hence, filling in missing values, reducing noise when identifying outliers, and correcting data inaccuracies were carried out as part of the cleansing procedures [23]. The dataset is first practiced using a ML model. Then, the critical function is selected based on the forecast's relevance. Using the significant function, useless characteristics were removed as feature extracting.

B. Data Visualizations

When visualizing the data, several points were identified. Distribution of the number of new cases in the UK in Fig. 1 shows that from Jan 2021, the number of daily new cases dropped significantly from more than 60,000 new cases recorded in January to less than 10,000 recorded in April.

Fig. 1 Distribution of the number of new cases in the UK

Fig. 2 illustrates that the number of death cases rose steadily from less than 50 death cases in March 2020 to more than 1200 death cases in May 2020. After that, the trend started to fall to less than 50 death cases in September 2020. The highest number

of deaths ever was recorded in Jan 2021, and then decreased tremendously to below 50, which may be due to the current vaccination plan across the UK. Moreover, Fig. 3 shows the growth of different types of cases.

Fig. 3 illustrates the pattern of confirmed cases in the UK from March 2020. However, the cases start to raise slowly from May 2020 with a sudden tremendous rise in the confirmed cases after November 2020. Fig. 3 also compares the new cases, total death and death cases during the period. Although a steady rise in confirmed cases has been noted, there is a constant pattern of COVID spread in the other cases.

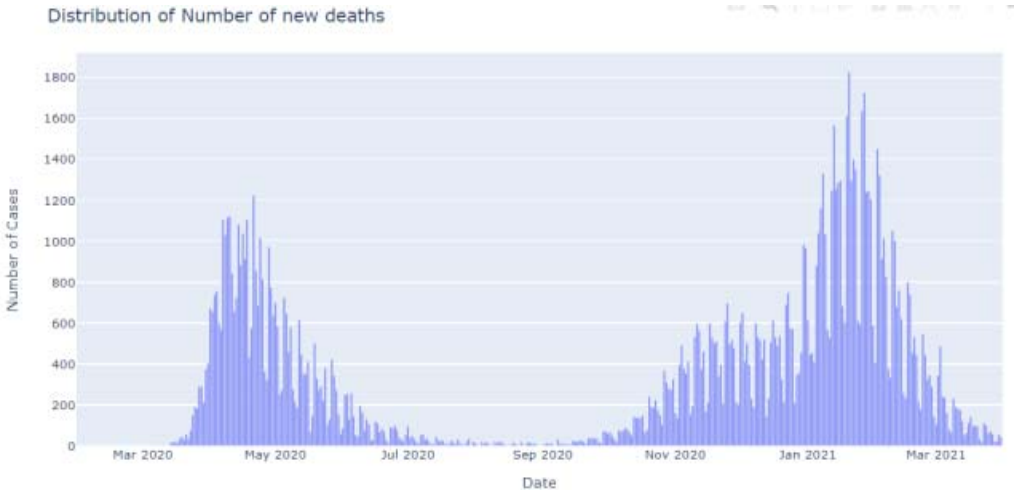

Fig. 2 Distribution of death cases

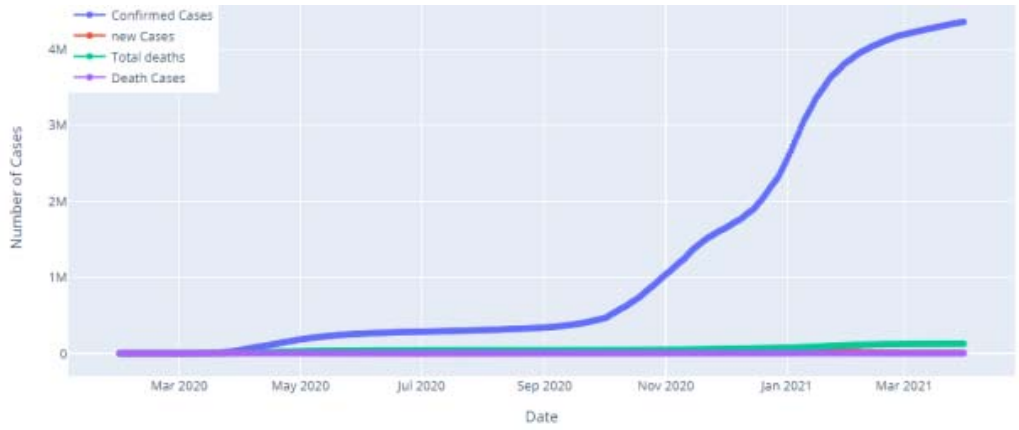

Fig. 3 Growth of different types of cases

C. Model Development

Three methods of supervised classification were used in this experiment. They are SVM, RF, and LR, and they were tested to find the best model. Impurity-reducing properties are determined by the training.

SVM classifier uses a hyper-plane to linearly separate the data using the linear kernel. The training and testing split is 8:2. Detecting COVID-19 is a high-priority issue; hence, the study aims for a hyperplane with a narrower margin of error.

Random sampling and ensemble techniques are used in Random Forest (RF) to enhance prediction accuracy. In order to achieve a more robust model, two assumptions were made. Firstly, the dataset's target variable should consist of valid values to enable the classification algorithms to generate accurate predictions instead of relying on guesswork. Secondly, it is essential for the predictions to exhibit minimal correlation, from each tree within the ensemble.

LR identifies the optimum path using one prediction. A line might link the predicted value to its predictor. While developing the model, few assumptions were made as the design must be suitable for multiple regression analysis, which requires one tool and linear connection should be effective in the model.

IV. Results and Discussions

After splitting the cleaned dataset into 80/20 for training and test data, we have analyzed through three ML algorithms named as RF, SVM and LR to predict the new cases of COVID-19 patients. In the context of predicting COVID-19 cases, a learning curve is used to visualize the relationship between the amount of training data used to build a predictive model and the model's performance. It helps to identify the optimal amount of

training data needed to build a high performing model, as well as the point at which adding more data will no longer significantly improve the model's performance.

Fig. 4 Learning curve for RF model and SVM

Fig. 4 illustrates the training error for the RF model. It is evident that the error is quite low, indicating that the RF model has effectively customized itself to fit the training data. This is a positive indication of the model's ability to learn from the training data. However, the low training error may also suggest that the RF model has little regard for data points outside of the training set, potentially leading to overfitting.

Fig. 5 Learning curve for SVM model

In the results of our analysis, we observed that the training error for the SVM model was moderately small. This suggests that the model successfully learned from the training data, displaying a moderate level of bias towards that specific data. Furthermore, the SVM model's capability to adapt to the training data implies a moderate level of performance on this dataset.

Fig. 6 Learning curve for LR model

We observed from Fig. 6 that the training error for the LR model was high. This indicates that the model is not able to effectively fit the training data, suggesting a high level of bias with respect to that set of data. The model's inability to tailor the training data effectively may impact its performance on unseen data, and further evaluation is necessary to assess the model's generalization ability.

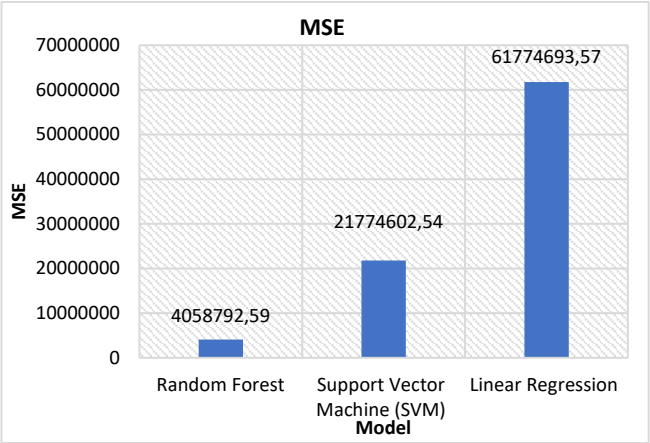

Fig. 7 MSE bar graph

Mean Squared Error (MSE) is a common metric used to evaluate the performance of predictive models. The results from the MSE showed that the RF model had the lowest MSE among the three algorithms, with a value of 4058792.59. This indicates that the RF model had the highest prediction accuracy of the three models studied. On the other hand, the LR model had the highest MSE, at 61774693.57, suggesting the lowest predictive power among the algorithms studied. These findings suggest that the RF model may be the most suitable choice for predicting COVID-19 cases in this specific dataset.

According to the results presented in Fig. 8, the RF model had the highest prediction accuracy, at 98.26%, followed by the SVM model with an accuracy of 90.69%. The LR model had the lowest accuracy, at 73.59%. These results suggest that the

RF model is the most effective at predicting future outbreaks of COVID-19 in the UK based on the data analyzed in this study.

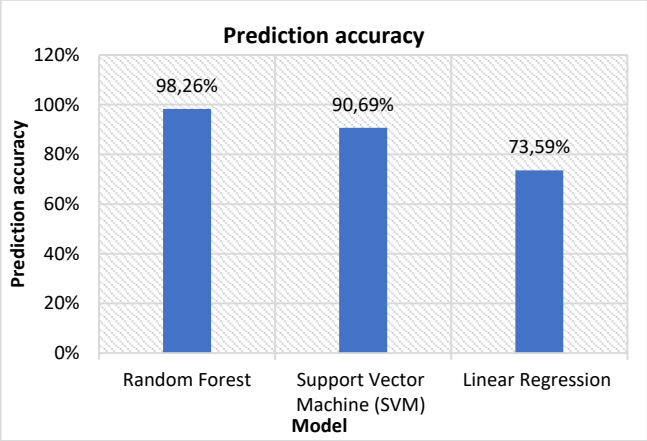

Fig. 8 Prediction accuracy of all models

TABLE I
TRAINING ACCURACY, PREDICTION ACCURACY AND MSE

Model	Training accuracy	Prediction accuracy	MSE
RF	99.47%	98.26%	4.05e11
SVM	91.2%	90.69%	2.17e13
LR	89.23%	73.59%	6.17e13

Table I exemplifies the accuracy of the model after training and testing. The RF model performed well during model training with an accuracy of 99.47% and a testing accuracy of 98.26%. On the other hand, the SVM and LR models achieved lower training accuracies of 91.2% and 89.23% respectively, with corresponding testing accuracies of 90.69% and 73.59%. From the evaluation metrics, RF has proven its robustness and accuracy especially when the data is noisy. RF is prone to less overfitting which is evaluated in this study from the MSE value.

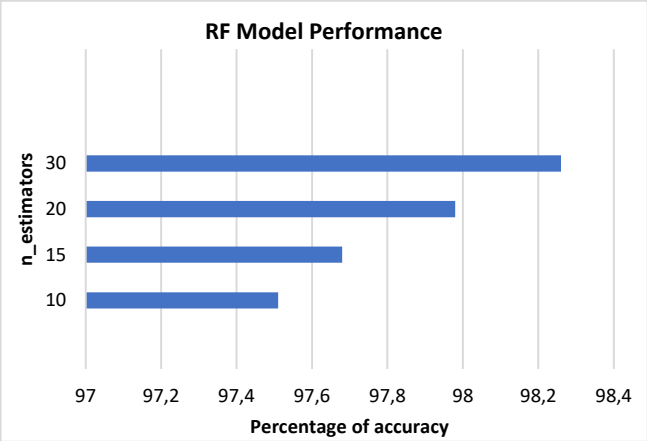

Fig. 9 RF model accuracy for n number of estimators

Ensemble learning methods have been used to train multiple models and combine their performance for final prediction. As a result, complex patterns in the data are captured and lead to accurate predictions. Fig. 9 illustrates the accuracy of the model with n number of estimators, which is determined by the number of decision trees. As the number of decision trees is increased, the model performance is more accurate. In this study, for n = 30, the RF model has attained a highest percentage of accuracy. At some point, adding more decision trees will deteriorate the model performance. So, it is important to tune the model with right number of trees that gives accurate model performance.

V. CONCLUSION AND FUTURE WORK

In conclusion, the primary goal of forecasting the model in the context of COVID-19 is to aid in decision-making in the medical field. It is important to carefully define the targeted demographic and use a representative dataset to accurately evaluate the performance of a forecasting model. The results of this study showed that the RF model had the highest prediction accuracy of 98.26%, followed by the SVM and LR model. The lesser training error shows RF model is generalized well for the dataset used. The MSE for RF is much lesser than SVM and LR which enhanced the level of RF model performance for this study.

To improve our analysis for future research, we will consider enhancing the data collection process based on the effect of factors such as temperature, humidity, terrain on the spread of COVID-19 in different cities within UK and prepare ML models. We plan to investigate the use of a diagnostic model based on deep learning techniques to predict COVID-19 using chest X-ray images.

REFERENCES

[1] Cheng Fu-Yuan and Joshi, Himanshu and Tandon, Pranai and Freeman, Robert and Reich, David L and Mazumdar, Madhu and Kohli-Seth, Roopa and Levin, Matthew A and Timsina, Prem and Kia, Arash, "Using machine learning to predict ICU transfer in hospitalized COVID-19 patients," Journal of clinical medicine, vol. 9, p. 1668, 2020.

[2] Monteiro, Ana Carolina Borges and Fran, Reinaldo Padilha and Arthur, Rangel and Iano, Yuzo, "An Overview of Medical Internet of Things, Artificial Intelligence, and Cloud Computing Employed in Health Care from a Modern Panorama," The Fusion of Internet of Things, Artificial Intelligence, and Cloud Computing in Health Care, pp. 3--23, 2021.

[3] Hasan, Najmul, "A methodological approach for predicting COVID-19 epidemic using EEMD-ANN hybrid model," Internet of Things, vol. 11, p. 100228, 2020.

[4] Tuli, Shreshth and Tuli, Shikhar and Tuli, Rakesh and Gill, Sukhpal Singh, "Predicting the growth and trend of COVID-19 pandemic using machine learning and cloud computing," Internet of Things, vol. 11, p. 100222, 2020.

[5] Singh, Prabhdeep and Kaur, Rajbir, "An integrated fog and Artificial Intelligence smart health framework to predict and prevent COVID-19," Global transitions, vol. 2, pp. 283--292, 2020.

[6] Chakraborty, Chinmay and Abougreen, Arij, "Intelligent internet of things and advanced machine learning techniques for COVID-19," EAI Endorsed Transactions on Pervasive Health and Technology, vol. 7, 2021.

[7] Hota, HS and Handa, Richa and Shrivas, AK, "COVID-19 pandemic in India: forecasting using machine learning techniques," in Data Science for COVID-19, Elsevier, 2021, pp. 503--525.

[8] Ahmad, Amir and Garhwal, Sunita and Ray, Santosh Kumar and Kumar, Gagan and Malebary, Sharaf Jameel and Barukab, Omar Mohammed, "The number of confirmed cases of covid-19 by using machine learning: Methods and challenges," Archives of Computational Methods in Engineering, vol. 28, pp. 2645--2653, 2021.

[9] Kushwaha, Shashi and Bahl, Shashi and Bagha, Ashok Kumar and Parmar, Kulwinder Singh and Javaid, Mohd and Haleem, Abid and Singh, Ravi Pratap, "Significant applications of machine learning for COVID-19 pandemic," Journal of Industrial Integration and Management, vol. 5, pp. 453--479, 2020.

[10] Ong, Edison and Wong, Mei U and Huffman, Anthony and He, Yongqun, "COVID-19 coronavirus vaccine design using reverse vaccinology and machine learning," Frontiers in immunology, vol. 11, p. 1581, 2020.
[11] Rustam, Furqan and Reshi, Aijaz Ahmad and Mehmood, Arif and Ullah, Saleem and On, Byung-Won and Aslam, Waqar and Choi, Gyu Sang, "COVID-19 future forecasting using supervised machine learning models," IEEE access, vol. 8, pp. 101489--101499, 2020
[12] Zeroual, Abdelhafid and Harrou, Fouzi and Dairi, Abdelkader and Sun, Ying, "Deep learning methods for forecasting COVID-19 time-Series data: A Comparative study," Chaos, Solitons \& Fractals, vol. 140, p. 110121, 2020.
[13] Hota, HS and Handa, Richa and Shrivas, AK, "COVID-19 pandemic in India: forecasting using machine learning techniques," in Data Science for COVID-19, Elsevier, 2021, pp. 503—525.
[14] Khakharia, Aman and Shah, Vruddhi and Jain, Sankalp and Shah, Jash and Tiwari, Amanshu and Daphal, Prathamesh and Warang, Mahesh and Mehendale, Ninad, "Outbreak prediction of COVID-19 for dense and populated countries using machine learning," Annals of Data Science, vol. 8, pp. 1--19, 2021.
[15] Saha, Aindrila and Mishra, Vartika and Rath, Santanu Kumar, "Prediction of growth in COVID-19 Cases in India based on Machine Learning Techniques," 2022 International Conference on Innovative Trends in Information Technology (ICITIIT), IEEE, 2022, pp. 1--6.
[16] R. S. M. L. Patibandla, B. T. Rao, and V. L. Narayana, "11 - Prediction of COVID-19 using machine learning techniques," *ScienceDirect*, Jan. 01, 2022. https://www.sciencedirect.com/science/article/pii/B978012824145500007 1 (accessed Jan. 07, 2023).
[17] N. Leelawat *et al.*, "Twitter data sentiment analysis of tourism in Thailand during the COVID-19 pandemic using machine learning," *Heliyon*, vol. 8, no. 10, p. e10894, Oct. 2022, doi: 10.1016/j.heliyon.2022.e10894.
[18] M. Zivkovic *et al.*, "COVID-19 cases prediction by using hybrid machine learning and beetle antennae search approach," *Sustainable Cities and Society*, vol. 66, p. 102669, Mar. 2021, doi: 10.1016/j.scs.2020.102669.
[19] Vellido, A. The importance of interpretability and visualization in machine learning for applications in medicine and health care. Neural Computing & Applications 32, 18069–18083 (2020). https://doi.org/10.1007/s00521-019-04051-w
[20] Ganesan, S., Somasiri, N. and Colombage, C., 2023, January. Deep Learning Approaches for Accurate Sentiment Analysis of Online Consumer Feedback. IEEE Proceedings.
[21] Pokhrel, A.S., Somasiri, N., Jeyavadhana, C.R. and Ganesan, S., 2022, December. Web Data Scraping Technology using TF-IDF to Enhance the Big Data Quality on Sentiment Analysis. In ICDSBDA 2022: XVI. International Conference on Data Science and Big Data Analytics. (pp. 1-8). https://waset. org/.
[22] Ganesan, Swathi, Nalinda Somasiri, and Sangita Pokhrel. "The Role of Artificial Intelligence in Education." IEEE Proceedings, 2023.
[23] K. K. A. Ghany, H. M. Zawbaa, and H. M. Sabri, "COVID-19 prediction using LSTM algorithm: GCC case study," *Informatics in Medicine Unlocked*, vol. 23, p. 100566, 2021, doi: 10.1016/j.imu.2021.100566.
[24] WorldOMeter (2023). *Coronavirus toll update: Cases & deaths by country*. (online) Worldometer. Available at: https://www.worldometers.info/coronavirus/.