Machine Learning Techniques for Disease Prediction

Nikhil Potnis[1*], *Dr Bhavana Tiple*[2]

[1]School of Computer Science and Engineering, Dr. Vishwanath Karad MIT World Peace University, Pune, India
[2] School of Computer Science and Engineering, Dr. Vishwanath Karad MIT World Peace University, Pune, India

Abstract. Epidemic disease prediction is a critical area of research that has garnered increasing attention in recent years, particularly in the wake of the COVID-19 pandemic. One promising avenue for predicting the spread of diseases is through the analysis of social media data, such as Twitter. Machine learning (ML) techniques can be applied to Twitter data to identify patterns and trends that may be indicative of an emerging epidemic. For example, natural language processing (NLP) techniques can be used to analyze the language used in tweets to identify keywords and phrases that are commonly associated with a particular disease. Additionally, sentiment analysis can be used to assess the overall mood of the Twitter community, which can be a useful predictor of disease outbreaks. By combining these techniques with real-world data on disease incidence and other relevant factors, it may be possible to develop highly accurate models for predicting the spread of epidemic diseases, which could have important implications for public health policy and emergency response planning.

1 Introduction

Machine learning has revolutionized many industries, including healthcare, and has the potential to transform the way we diagnose and treat diseases. The ability to predict diseases accurately can help healthcare professionals take proactive steps to prevent or treat the disease early, resulting in better health outcomes for patients.

Deep learning is a subset of machine learning that uses artificial neural networks with multiple layers to learn and extract hierarchical representations of data. This technique has revolutionized the field of artificial intelligence by enabling computers to perform complex tasks that were previously only possible for humans.

One of the most significant applications of deep learning in modern times is in the field of image and speech recognition. Deep learning algorithms can identify objects and recognize speech with a high degree of accuracy, which has led to the development of products such as voice assistants, facial recognition systems, and self-driving cars.

Deep learning has also been applied to healthcare, where it is used to predict diseases, analyze medical images, and develop personalized treatment plans. In finance, deep learning is used

* Nikhil Potnis: nikhilshilpa29@gmail.com

to analyze data and predict market trends, while in the field of natural language processing, it is used for language translation and sentiment analysis. Overall, deep learning has made significant contributions to a wide range of fields and has the potential to transform many aspects of our lives in the future.

Disease prediction is a crucial aspect of healthcare.[1] Early detection and diagnosis of diseases can significantly improve patient outcomes by allowing for timely intervention and treatment. For instance, early detection of cancer can increase the chances of successful treatment and reduce the likelihood of the cancer spreading to other parts of the body.[2]

In addition, disease prediction can help prevent the onset of diseases by identifying individuals who are at high risk. For example, predicting the risk of developing type 2 diabetes can help individuals make lifestyle changes to reduce their risk of developing the disease.[3]

2 Literature Survey

Shrikant Tiwari et al. (2023) study the usage of Machine learning algorithms in COVID-19 inquiry and other procedures. The major goals of the paper were Examining the impact of the data type and data nature, as well as obstacles in data processing for COVID-19. 2) Better grasp the importance of intelligent approaches like ML for the COVID-19 pandemic. 3) The development of improved ML algorithms and types of ML for COVID-19 prognosis. 4) Examining the effectiveness and influence of various strategies in COVID-19 pandemic. 5) To target on certain potential issues in COVID-19 diagnosis in order to motivate academics to innovate and expand their knowledge and research into additional COVID-19-affected industries.[1]

The authors proposed early warning system for COVID-19 using twitter data. They implemented a COVID-19 forecasting model through a Twitter-based linear regression model to detect early signs of the COVID-19 outbreak.[2]

Ahuja and authors used a deep learning-based model to detect COVID-19 cases on Twitter. The model was trained on a large dataset of tweets and achieved high accuracy in detecting COVID-19 cases.[3]

The authors used a machine learning approach to predict the spread of infectious diseases on social media. The authors used a dataset of tweets related to the Zika virus and developed a prediction model using a combination of text analysis and network analysis.[4]

The authors researched used Twitter datasets to explore user sentiment from the COVID-19 perspective. They focused on exploiting machine learning (ML), and deep learning (DL) approaches to classify user sentiments regarding COVID-19. [5]

The authors proposed a machine learning-based approach to predict the Zika virus outbreak using Twitter data. The authors used a combination of text analysis and network analysis to develop their prediction model.[6]

The authors proposed a real-time prediction model for infectious disease outbreaks using social media data and machine learning. The authors used Twitter data to predict the spread of influenza in real-time and achieved high accuracy in their predictions [7]

In [8] the authors applied a mathematical epidemic model (MEM), statistical model, and recurrent neural network (RNN) variants to forecast the cumulative confirmed cases. They proposed a reproducible framework for RNN variants that addressed the stochastic nature of RNN variants leveraging z-score outlier detection.

The authors collect the data from Twitter in the Arabic language related to the spread of influenza using many Arabic keywords. Then, they applied several machine learning algorithms. They also found the correlation between the collected tweets and the reports collected from the World Health Organization website.[9]

The authors aimed to depict whether it is possible to predict infectious disease outbreaks early, by using machine learning. This study was carried out following the guidelines of the Cochrane Collaboration and the meta-analysis of observational studies in epidemiology and the preferred reporting items for systematic reviews and meta-analyses.[10]
The main purpose of this research is to illustrate the use of ML, DL, and mathematical models that can be helpful for the researchers to generate valuable solutions for higher authorities and the healthcare industry to reduce the impact of this epidemic. [11]
The authors presented an Artificial Intelligence (AI)-based meta-analysis to predict the trend of epidemic Covid-19 over the world. The powerful machine learning algorithms namely Naïve Bayes, Support Vector Machine (SVM) and Linear Regression were applied on real time-series dataset, which holds the global record of confirmed, recovered, deaths and active cases of Covid-19 outbreak. Statistical analysis has also been conducted to present various facts regarding Covid-19. [12]
The authors attempted to incorporate the public discourse in the design of forecasting models particularly targeted for the steep-hill region of an ongoing wave. They proposed a sentiment-involved topic-based latent variables search methodology for designing forecasting models from publicly available Twitter conversations. [13]
The authors conducted a mathematical and numerical analyses based on closed-loop decisions for COVID-19. The Susceptible, Infectious, and Recovered (SIR) model was analyzed using a machine learning model to estimate the optimal constant parameters, which are the recovery and infection rates of the coupled nonlinear differential equations that govern the epidemic model [14]
The authors studied and investigated the transmission pathways of SARS-CoV-2 in the environment and provides current updates on the surveillance of viral outbreaks using WBE, viral air sampling, and AI. They framework based on an ensemble of ML and DL algorithms to provide a beneficial supportive tool for decision makers [15]

3 Methodology

Machine learning algorithms used for disease prediction are often considered as "black boxes" due to their lack of interpretability. Understanding the reasons behind the predictions made by these models is crucial for healthcare professionals to make informed decisions. There is a need to develop more interpretable models to improve their adoption in clinical practice.[4]
Also training the model on one dataset that is from one data source like Twitter and then using some other datasets like from YouTube or LinkedIn to verify the output of the model, check whether it can predict the data accurately[5]

3.1 Material

The dataset used is generated using Tweepy library. It is a set of tweets from twitter which are based on diseases. The authors used hashtags like covid19 , zika, malaria, to generate the tweets.
Another dataset used for the analysis was YouTube comments dataset from Kaggle. The dataset contains comments on YouTube videos, and are then labelled as either zero for negative, one for neutral and two for positive.

3.2 Material

The figure below indicates the methodology used by the authors to gather the results for this research.

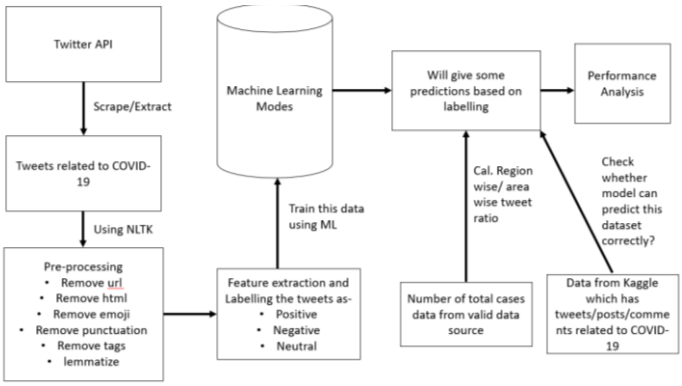

Fig. 1. Visual display of methodology used

The following steps were taken to perform the necessary implementation by the authors.

- Using twitter API, the required tweets were extracted.
- The hashtags used for extracting tweets included covid19, malaria, zika, cholera.
- To pre-process the tweets, NLTK (Natural Language Toolkit) library was used. For pre-processing following steps were used -
  - Removing URL
  - Removing HTML tags
  - Removing punctuation
  - Removing any unnecessary tags
  - Lemmatizing
- On pre-processing, the necessary features were selected and then the text was labelled. It was labelled as either – positive, negative, or neutral.
- Machine Learning models like Decision Tree, Naïve Bayes, Support Vector Machine, were applied.
- The accuracy achieved from these models is recorded for comparison purpose.
- Another dataset based on similar terms, was trained using the above-mentioned models and its accuracy was also recorded.
- Accuracy from both the datasets for Machine Learning models was compared.

4 Results and Discussion

As the unstructured data is converted into a supervised learning process, it is important to see the distribution and the counts of the different classes in the dataset. Let us now see how the distribution of the different classes belonging to the tweets is:

- Positive tweets –

Out of the whole set of 1000 tweets, 142 were identified as positive tweets. Which means only 14.2% of the whole set of tweets were positive.

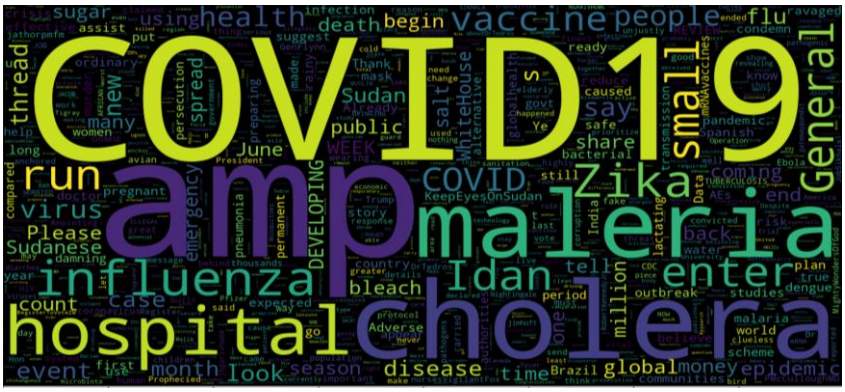

Fig. 2. Positive Word Cloud

- Negative tweets –

Out of the whole set of 1000 tweets, 19 tweets were identified as negative tweets. Which means only 1.9% of the whole set of tweets were negative.

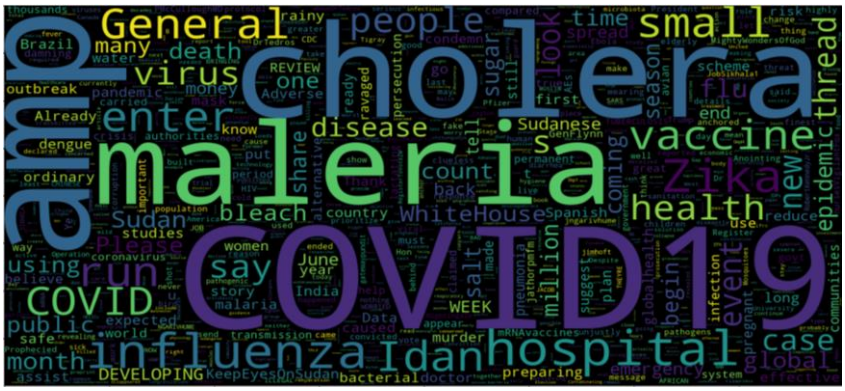

Fig. 3. Negative Word Cloud

- Neutral tweets –

Out of the whole set of 1000 tweets, 839 tweets were identified as neutral tweets. Which means only 84% of the whole set of tweets were neutral.

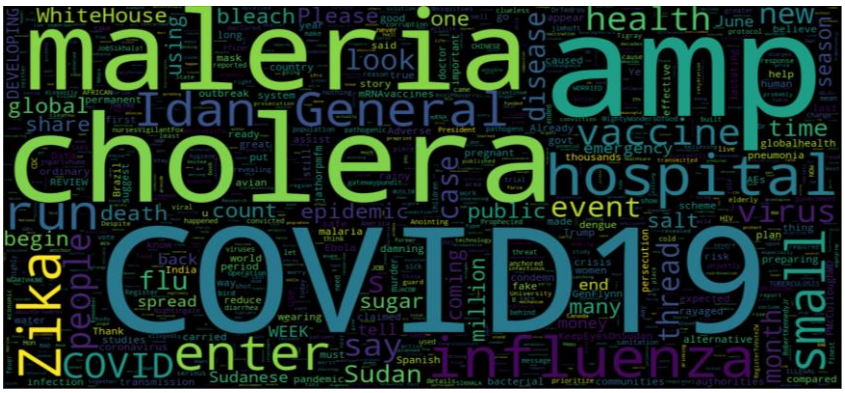

Fig. 4. Neutral Word Cloud

4.1 Machine Learning on Disease Tweets Dataset

Once we have cleaned, pre-processed, and labelled the tweets, they are now ready to be trained and validated against the various Machine Learning.
We used the scikit-learn's TFidfTransformer to arrive at the features to be input to the different classifier. The Machine Learning models used by the authors are –

- Support Vector Machine
- Naïve Bayes
- Logistic Regression
- Random Forest
- Decision Tree

- Support Vector Machine

Here we have used Linear Support Vector Machine model for training and validating the labels made on the tweets.

	precision	recall	f1-score	support
negative	1.00	0.75	0.86	4
neutral	0.93	1.00	0.96	172
positive	1.00	0.50	0.67	24
accuracy			0.94	200
macro avg	0.98	0.75	0.83	200
weighted avg	0.94	0.94	0.93	200

	True_Positive	True_Negative	False_Positive	False_Negative
Negative label	3	196	0	1
Positive label	172	15	13	0
Neutral label	12	176	0	12

Fig. 5. Performance Matrix for SVC Model

As indicated in the figure above, the model had an accuracy of 94%. Talking about the confusion matrix, we tried to find the true positive, true negative, false positive and false negative based on each label.

In the case of a negative label, the classification model correctly identified 3 true positives (cases where it correctly predicted the negative class), and 196 true negatives (cases where it correctly predicted the absence of the negative class). It made no false positive predictions (incorrectly predicting the negative class) but had 1 false negative (incorrectly predicting the positive class).

For the positive label, the model correctly identified 172 true positives and 15 true negatives. It made 16 false positive predictions but had no false negatives.

Regarding the neutral label, the model correctly identified 12 true positives and 176 true negatives. It made no false positive predictions but had 12 false negatives.

- Naïve Bayes

We used Bernoulli Naïve Bayes model as the second model for training and validation of the tweet's prediction.

	precision	recall	f1-score	support
0.0	0.00	0.00	0.00	4
1.0	0.50	0.03	0.05	37
2.0	0.80	0.99	0.89	159
accuracy			0.80	200
macro avg	0.43	0.34	0.31	200
weighted avg	0.73	0.80	0.71	200

	True_Positive	True_Negative	False_Positive	False_Negative
Negative label	0	196	0	4
Positive label	1	162	1	36
Neutral label	158	1	40	1

Fig. 6. Performance Matrix for Naïve Bayes Model

The model had performance accuracy of 80%. For each label, we calculated the confusion matrix.

For the negative label, TP (True Positives) is 0, indicating that the model correctly classified zero instances as negative. TN (True Negatives) is 196, meaning the model correctly identified 196 instances as negative. FP (False Positives) is 0, indicating that the model incorrectly classified zero instances as negative when, they were positive. FN (False Negatives) is 4, meaning the model mistakenly classified four instances as negative when, they were positive.

For the positive label, TP is 1, indicating that the model correctly classified one instance as positive. TN is 162, meaning the model correctly identified 162 instances as negative. FP is 1, indicating that the model incorrectly classified one instance as positive when it was negative. FN is 36, meaning the model mistakenly classified 36 instances as negative when, they were positive.

For the neutral label, TP is 158, indicating that the model correctly classified 158 instances as neutral. TN is 1, meaning the model correctly identified one instance as negative. FP is 40, indicating that the model incorrectly classified 40 instances as neutral when they were positive or negative. FN is 1, meaning the model mistakenly classified one instance as negative when it was neutral.

- Logistic Regression

Another model used was Logistic Regression. It had performance accuracy of 94% as seen from below figure.

	precision	recall	f1-score	support
negative	1.00	0.75	0.86	4
neutral	0.93	1.00	0.96	172
positive	1.00	0.50	0.67	24
accuracy			0.94	200
macro avg	0.98	0.75	0.83	200
weighted avg	0.94	0.94	0.93	200

	True_Positive	True_Negative	False_Positive	False_Negative
Negative label	3	196	0	1
Positive label	172	15	13	0
Neutral label	12	176	0	12

Fig. 7. Performance Matrix for Logistic Regression Model

For the negative label, the model correctly predicted 3 positive cases and 196 negative cases. It generated no false positive cases but failed to identify 12 cases that were negative, resulting in false negatives.

For the positive label, the model correctly predicted 172 positive cases and 15 negative cases. However, it generated 13 false positive cases, indicating the presence of the positive label when it was not there, and did not identify any false negative cases.

For the neutral label, the model correctly predicted 12 positive cases and 176 negative cases. It generated no false positive cases but failed to identify 12 cases that were neutral, resulting in false negatives.

- Random Forest

The performance of random forest algorithm can be visualized from below figure.

	precision	recall	f1-score	support
negative	1.00	0.75	0.86	4
neutral	0.92	1.00	0.96	172
positive	1.00	0.46	0.63	24
accuracy			0.93	200
macro avg	0.97	0.74	0.82	200
weighted avg	0.94	0.93	0.92	200

	True_Positive	True_Negative	False_Positive	False_Negative
Negative label	3	196	0	1
Positive label	172	14	14	0
Neutral label	11	176	0	13

Fig. 8. Performance Matrix for Random Forest Model

Its accuracy came out as 93%. For each label, we calculated the confusion matrix.

In the case of negative labels, out of a total of 200 instances, the model correctly identified 3 true positives (correctly predicted negative cases), 196 true negatives (correctly predicted non-negative cases) and had no false positives (incorrectly predicted negative cases). However, there was one false negative (incorrectly predicted non-negative case).

For positive labels, out of a total of 200 instances, the model correctly identified 172 true positives (correctly predicted positive cases), 14 true negatives (correctly predicted non-positive cases) but had 14 false positives (incorrectly predicted positive cases). There were no false negatives (incorrectly predicted non-positive cases).

Regarding neutral labels, out of a total of 200 instances, the model correctly identified 11 true positives (correctly predicted neutral cases), 176 true negatives (correctly predicted non-neutral cases) and had no false positives. However, there were 13 false negatives (incorrectly predicted non-neutral cases).

- Decision Tree

	precision	recall	f1-score	support
negative	0.75	0.75	0.75	4
neutral	0.95	0.92	0.93	172
positive	0.57	0.71	0.63	24
accuracy			0.89	200
macro avg	0.76	0.79	0.77	200
weighted avg	0.90	0.89	0.89	200

	True_Positive	True_Negative	False_Positive	False_Negative
Negative label	3	195	1	1
Positive label	158	20	8	14
Neutral label	17	163	13	7

Fig. 9. Performance Matrix for Decision Tree Model

The Decision tree model had accuracy of 89%. For the negative label, out of 200 instances, the model correctly predicted 3 as true negatives (TN), meaning it correctly identified them as negative. It incorrectly predicted 1 instance as a false positive (FP), indicating a false alarm. Moreover, the model failed to identify 1 instance as a false negative (FN), meaning it missed detecting a negative case.

For the positive label, out of 209 instances, the model correctly predicted 158 as true positives (TP), accurately identifying them as positive. However, it incorrectly predicted 8 instances as false positives (FP), indicating it incorrectly labelled them as positive. Additionally, the model failed to identify 14 instances as false negatives (FN), missing them as positive cases.

Regarding the neutral label, out of 200 instances, the model correctly predicted 17 as true positives (TP), correctly identifying them as neutral. It correctly classified 163 instances as true negatives (TN), accurately recognizing them as non-neutral. However, it incorrectly predicted 13 instances as false positives (FP) and failed to identify 7 instances as false negatives (FN) within the neutral label.

Fig. 10. Accuracy of all Machine Learning Models

In the context of machine learning, accuracy is a metric that measures the overall correctness of a classification model's predictions. It represents the percentage of correctly classified instances compared to the total number of instances in the dataset. Based on the provided accuracies, the decision tree model achieved an accuracy of 89%, indicating that it correctly classified 89% of the instances. The Bernoulli Naive Bayes model achieved an accuracy of 80%, implying that it accurately predicted 80% of the instances. The random forest model

achieved an accuracy of 92%, suggesting that it correctly classified 92% of the instances. The linear support vector machine model and the logistic regression model both achieved an accuracy of 94%, indicating that they both accurately classified 94% of the instances. Higher accuracy values generally indicate better performance, but it's important to consider other evaluation metrics and domain-specific requirements when choosing the most suitable model.

4.2 Results on YouTube Comments dataset

The dataset used is YouTube comments data. The dataset is pre-processed using the same procedure as done for the twitter dataset. The following results were achieved when Machine Learning models were trained and tested on the dataset.

- Support Vector Machine

The model had an overall accuracy of 62%. We can say that it was able to predict the true negatives i.e., predicted the absence of the negative labels correctly.

- Naïve Bayes

The model had an overall accuracy of 62%. Its performance was like the SVM model. Logistic Regression

The model had an overall accuracy of 62%. Its performance was like the SVM model.

- Random Forest

The model had an overall accuracy of 62%. Its performance was like the SVM model.

- Decision Tree

The model had an overall accuracy of 62%. Its performance was like the SVM model. Overall, the performance of all the Machine Learning models has been similar in terms of accuracy. But for other parameters like precision, recall they are varying.

5 Conclusion

From the results, it is clearly visible that Machine Learning models had higher accuracy than Deep Learning models. Amongst the Machine Learning models, Logistic Regression had the highest accuracy. It means that the model was clearly able to predict the tweets either as positive or negative or neutral. For the other dataset i.e., YouTube comments dataset, all models had the same accuracy of 62%.

Talking about the Deep Learning models, they had comparatively low accuracy for both the datasets. CNN and LSTM both had similar accuracy of 18 and 21 respectively for the two datasets. Hence it can be said that we can't use Deep Learning models to predict the disease or create a model using them. But in case of Machine Learning models, we can surely use them to create a predictive model.

Though there are other parameters that affect the prediction of disease like climate, area, temperature but we can still motivate ourselves to try and use Machine Learning algorithms for prediction. We tried to check through a limited but the best fitting Machine Learning algorithm. More algorithms can be used to try and check their accuracy on the datasets used by the authors.

Here the size of the dataset considered was less but more larger datasets with more annotations can also be referred for more better accuracy and results.

References

1. Tiwari, S., Chanak, P., & Singh, S. K. (2023). A review of the machine learning algorithms for covid-19 case analysis. IEEE Transactions on Artificial Intelligence, 4(1), 44–59
2. Zhang, Y., Chen, K., Weng, Y., Chen, Z., Zhang, J., & Hubbard, R. (2022). An intelligent early warning system of analyzing Twitter data using machine learning on COVID-19 surveillance in the US. Expert Systems with Applications, 198(116882), 116882
3. Ahuja C, “Deep learning-based detection of COVID-19 cases on Twitter”(2021) . New Gener. Comput,189–212.
4. J. Wang ,"A machine learning approach for predicting the spread of infectious diseases on social media",2020.
5. Yeasmin, N., Mahbub, N. I., Baowaly, M. K., Singh, B. C., Alom, Z., Aung, Z., & Azim, M. A. (2022). Analysis and prediction of user sentiment on COVID-19 pandemic using tweets. Big Data and Cognitive Computing, 6(2), 65.
6. S. S. Hasan, "Twitter-based prediction of Zika virus outbreak using machine learning algorithms", 2020.
7. E. G. Althouse, "Real-time prediction of infectious disease outbreaks using social media data and machine learning", 2019.
8. Masum, M., Masud, M. A., Adnan, M. I., Shahriar, H., & Kim, S. (2022). Comparative study of a mathematical epidemic model, statistical modeling, and deep learning for COVID-19 forecasting and management. Socio-Economic Planning Sciences, 80(101249), 101249.
9. Bani Baker, Q., Shatnawi, F., & Rawashdeh, S. (2022). Forecasting epidemic diseases with Arabic Twitter data and WHO reports using machine learning techniques. Bulletin of Electrical Engineering and Informatics, 11(2), 738–749.
10. Santangelo, O. E., Gentile, V., Pizzo, S., Giordano, D., & Cedrone, F. (2023). Machine learning and prediction of infectious diseases: A systematic review. Machine Learning and Knowledge Extraction, 5(1), 175–198.
11. Saleem, F., Al-Ghamdi, A. S. A.-M., Alassafi, M. O., & AlGhamdi, S. A. (2022). Machine Learning, Deep Learning, and mathematical models to analyze forecasting and epidemiology of COVID-19: A Systematic Literature Review. International Journal of Environmental Research and Public Health, 19(9), 5099.
12. Tiwari, D., Bhati, B. S., Al-Turjman, F., & Nagpal, B. (2022). Pandemic coronavirus disease (Covid-19): World effects analysis and prediction using machine-learning techniques. Expert Systems, 39(3), e12714.
13. Lamsal, R., Harwood, A., & Read, M. R. (2022). Twitter conversations predict the daily confirmed COVID-19 cases. Applied Soft Computing, 129(109603), 109603.
14. Narayan, K., Rathore, H., & Znidi, F. (2022). Using epidemic modeling, machine learning and control feedback strategy for policy management of COVID-19. IEEE Access: Practical Innovations, Open Solutions, 10, 98244–98258.
15. Abdeldayem, O. M., Dabbish, A. M., Habashy, M. M., Mostafa, M. K., Elhefnawy, M., Amin, L., Al-Sakkari, E. G., Ragab, A., & Rene, E. R. (2022). Viral outbreaks detection and surveillance using wastewater-based epidemiology, viral air sampling, and machine learning techniques: A comprehensive review and outlook. The Science of the Total Environment, 803(149834), 149834.