Educational Administration: Theory and Practice
2024,30(3), 861-869
ISSN:2148-2403
https://kuey.net/

Research Article

Analysis Of Pattern Change In Diseases Using Machine Learning

Madhura Baiju Bhokare[1*], Prof. Dr. Sunita Dhotre[2], Assistant Prof. Sheetal S Patil[3]

[1*]M.Tech Student, Computer Engineering, Bharati Vidyapeeth (Deemed To Be University) College of Engineering Pune, (India)

[2]Associate Professor, Computer Engineering Department, Bharati Vidyapeeth (Deemed To Be University) College of Engineering Pune, (India)

[3]Assistant Professor, Computer Engineering Department, Bharati Vidyapeeth (Deemed To Be University) College of Engineering Pune, (India)

***Corresponding Author:** Madhura Baiju Bhokare
*M.Tech Student, Computer Engineering, Bharati Vidyapeeth (Deemed To Be University) College of Engineering Pune, (India)

Citation: Madhura Baiju Bhokare, Analysis Of Pattern Change In Diseases Using Machine Learning *Educational Administration: Theory And Practice*, *30*(3), 861-869, Doi: 10.53555/kuey.v30i3.1387

ARTICLE INFO

ABSTRACT

SARS-CoV-2, also known as COVID-2019, is spreading quickly and threatening our society on a global scale. Every government is employing money, infrastructure, data security equipment, survival evaluations, treatments, and other resources to make major attempts to stop the fatal disease from spreading. Nationwide, COVID-19 sepsis infections are spreading quickly. To allocate resources and classify patients in order to avert death, a reliable and accessible clinical assessment of the severity of human illness is essential. Through the analysis of genome and protein sequences, the study seeks to confirm an early stage of sickness by estimating the likelihood of further alterations depending on their length. A study involving 852 isolates of the COVID-19 Next strain application was conducted. The project included a dashboard showing the distribution and percentages of gene mutations for each base sequence in the genomic data, a transmission chart, and a visual phylogenetic analysis. Additionally, it took basic traits from genetic alterations that might help anticipate more mutations in the future. For 11191 patients, survival analysis parameters were found.

Keywords: Monitoring System, Performance evaluation, Machine learning, SARS-CoV2, Covid-19, Outbreak, Genome sequences, Time, Risk

I. INTRODUCTION

Three categories exist within machine learning: reinforcement learning, semi-supervised learning, and supervised learning. The computer is trained using tagged datasets as part of a process called supervised machine learning. The instances are correctly categorized using this method based on the category to which they belong. After analysing the data, the software will be able to forecast future events using the knowledge it has gained from past examples. Conversely, unsupervised machine learning differs from supervised machine learning in that it may acquire information on its own without the presence of appropriately classified data. Machines will be given training samples to examine during the unsupervised learning stage of machine learning, and it will be up to the computer to find any hidden patterns in the dataset. In reinforcement learning, the computer assumes the role of an agent that uses an approach based on experimentation and environmental observation to find the best appropriate behaviours. This strategy will be used repeatedly until the computer system learns how to accomplish a specific task efficiently. If the machine fails to complete a task successfully, its state will be increased; if it does complete the task, it will be forced to increase its state. As a result, this study provides a deep learning technique for classifying the COVID-19-related categorizations utilised in tweets. One can tell if someone is pleased or unhappy by looking at other people's comments. The focus is on automatically identifying texts with negative emotions within the large text. To our knowledge, previous methods have not improved the feature weights by a recurrent neural network with semantically ordered by SVM, especially when it comes to understanding feelings from tweets related to COVID-19. The main goal is to use context learning

and SVM to maximise the effectiveness of the semantic connections between individual words. The structure of the survey is described in section II by a number of current scholars. In section III, the suggested system architecture and research technique are presented. The supervised classifier training and testing process is covered in section IV, along with the algorithm specifics. Section VI presents a conclusion and a model for future work, while Section V concentrates on a thorough experimental investigation of the suggested system and numerous other systems that are now in use.

II. LITARATURE SURVEY

A technique for classifying the feelings expressed in microblog evaluations that incorporated emojis was published by Li et al. [1]. In the form of an emoji-text-incorporating bi-LSTM model that was given the name ET-BiLSTM. Emojis, as vector representations, were provided to the suggested model. Based on the findings, it is clear that ET-BiLSTM improves the overall performance of sentiment categorization. They classified people's feelings using a BERT model that they implemented. After that, they investigated how attitudes towards vaccination are dispersed around the globe by analyzing the hotspot locations and using kernel density estimation. Another research study was suggested by Balli et al. [2] during COVID-19 to classify public datasets and sentiment data manually labelled for positive and negative Turkish tweets. During the pre-processing of the dataset, the Zamburak [3] library and the Snowball library were used as separate libraries. In addition, the data were tokenized using TF-IDF before being provided to ML algorithms like CNN and RNN; however, for LSTM, the tokenizer class presented the data. This was done so that the ML algorithms could process the data. It was found that the models that were applied to Sentiment had superior performances, and the accuracy of the data when it was negatively weighted was greater than when it was positively weighted.

In a separate piece of research, Sitaula and Shahi [4] developed a hybrid feature to represent tweets from Nepal. They used two-word representations, such as the bag-of-words (BOW), with Fast Text-based approaches and specialized methods specialized main. Convolutional neural networks with several channels were then fed the concatenated representations as input (MCNN). According to the findings, the combination of features performed better than the individual features, with an accuracy of 69.7 per cent, while the MCNN model attained 71.3 per cent accuracy in comparison to traditional methods. Using increased feature weighting and the attention mechanisms of LSTM-RNN, Singh et al. [5] performed research to categorize Twitter data linked to COVID-19. TF-IDF was used to extract tweet features. The findings of the experiments demonstrated that the suggested technique surpassed the rest of the traditional machine-learning algorithms. In a study conducted by Parimala et al. [6,] the researchers classified tweets relevant to catastrophic occurrences using LSTM in conjunction with the feature extraction approach. They suggested using an algorithm called risk assessment sentiment analysis (RASA). Compared with XGBoost and binary classifiers, the findings demonstrate that RASA attained a higher level of accuracy [6].

People regularly contributed their ideas, news, and experiences in confronting this virus via social media, regarded as a large data center throughout the epidemic. Social networking sites (SNSs) like Twitter, an important resource for identifying and monitoring various events, such as the spread of illness, have become more common. This online platform encouraged academics to analyze time and tweets, including people's sentiments and responses about various topics, including election voting, the financial market, criminal activity, and hate speech [7].

In addition, the purpose of artificial intelligence (AI) in this crisis has unquestionably helped to research the shift in human behaviors and worries associated with COVID-19 patients and fatalities during and after the pandemic. As a result, several COVID-19 surveillance models are looking for an efficient method of text processing and extracting information from COVID-19-related postings. This would result in early reports being generated, which may be critical for preventing outbreaks. The method in question is known as sentiment analysis (SA) [8], sometimes known as emotions mining [9], and it involves assigning positive, negative, or neutral connotations to a variety of texts based on their views. These phrases are given a preliminary processing using natural language processing (NLP) and then given a classification by machine learning (ML) [10]. These perspectives are highly helpful for constructing more efficient disease monitoring systems. Much research has contributed to the analysis of tweets written in English; however, these studies still need to consider the connection between syntactic and semantic information and ML approaches based on feature types.

The researchers Samuel et al. [11] suggested two machine learning methods to classify tweets based on their Sentiment: naive Bayes and logistic regression. These models separated tweets into two categories: positive and negative. In their study, the authors examine the efficacy of these models on two categories of data with varying lengths of characters. The first group has less than 77 characters, while the second category contains 120 characters per tweet. In all categories, Naive Bayes fared better than logistic regression. For shorter tweets, NB reached an accuracy of 91.43 per cent, whereas this was only 74.29 per cent for LR. NB achieved an accuracy of 57.14 per cent for lengthier tweets, while LR evaluated an accuracy of 52 per cent. A strategy for analyzing and corning the COVID-19 epidemic, based on the frequency of terms and sentiment analysis, has been utilized by utilized cheers such as Rajput et al. [17]. In their method, word-level, bi-gram, and tri-gram frequencies express word rates according to a power law distribution. As a result of this, we were able to identify three distinct kinds of tweets: negative, positive, and neutral.

An investigation that examines and depicts the global impact of COVID-19 has been presented by Muthausami et al. [18]. This investigation is based on analysis and visualization, dividing the tweets into three categories using the approach of machine learning. There is a positive class, a neutral class, and a negative class. They used a variety of classifiers, such as support vector machines (SVM), naive Bayes, random forests, decision trees, Logit Boost, and Max Entropy. The results obtained by the Logit Boost ensemble classifier were shown to be superior to those obtained by the other algorithms using the suggested approach.
Deep learning models, such as LSTM recurrent neural networks, were used in a study that was carried out by Jelodar et al. [19], and the researchers developed an algorithm to identify feelings based on these models (LSTM RNN). The subject modeler for COVID-19 was stated on social media, and natural language processing was used to create the classifier.
Aljameel et al. [20] examined a large dataset of tweets written in Arabic that were connected to COVID-19. The authors developed a model using machine learning to forecast and categorize the categories of Saudi Arabian residents to actions taken by the government and efforts to manage pandemics. [21] They used uni-gram and bi-gram TF-IDF in conjunction with SVM, naive Bayes, and KNN classifiers to improve accuracy.[22] With an accuracy of 85 per cent, the output findings demonstrated that SVM performed better than KNN and naive Bayes.
.

III. PROPOSED SYSTEM DETAILS

A system's behaviour, structure, and many functional viewpoints are depicted in a system architecture model. It is set up to give a broad picture of both the behavioural and structural elements. It compiles data sets gathered from many official websites, including Kaggle, NCBI, and WHO. After the data has been processed, analyse the datasets and divide them into three categories: genetic, clinical, and textual data. Following the use of machine learning algorithms for detection, prediction, diagnosis, and prognosis, accurate time-based findings for the subsequent mutation level are obtained through the application of mathematical equations and algorithms. System architectural resources for creating the machine learning system are shown in Figure 2.

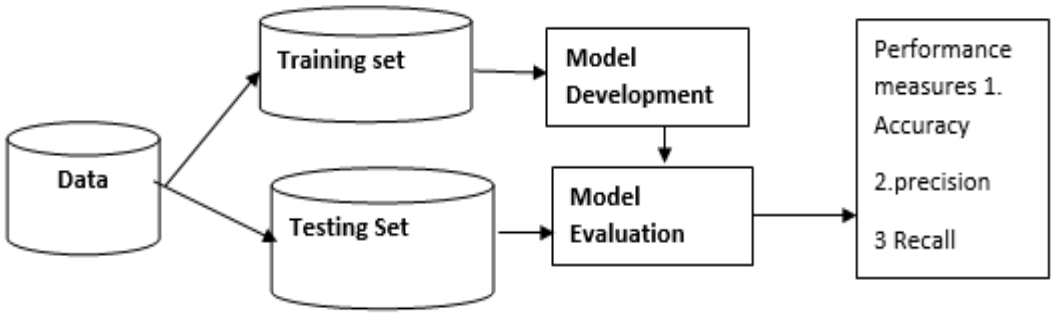

Figure 2: Proposed System architecture

Dataset Collection
The airline dataset that was used in this study was taken from multiple real-time data sources as well as Kaggle. Prior to being anticipated based on accuracy, data is first collected from aviation companies, examined, retrieved, and submitted to machine learning algorithms.

Pre-processing and normalization
It's possible that the data has several gaps and pointless details. This part is handled by data preparation, which includes a variety of pre-processing methods such data reduction, cleansing, and transformation.

Feature extraction and Selection
This process retrieves a range of features from the input data. After the features have been extracted, redundant and superfluous training features are removed by standardising the features using a feature selection threshold. Different hybrid properties are extracted from the normalised data with relational features, and training is done by choosing an optimisation approach.

Classification
The chosen features are sent as input to the training module, which generates thorough Background Knowledge for the entire system, once the module has been properly performed. We can enter the testing data into the training model after we have it in order to obtain the classification prediction. Pre-processing, vectorization,

and classification of the testing text are all part of the testing stage. Using machine learning techniques, the module testing assesses the predicted performance of the system.

Problem Statement

It is proposed that a COVID-19 outcome prediction model utilizing a Convolutional Neural Network (CNN) be constructed to effectively capture spatial patterns within medical imaging data. Alternatively, a Long Short-Term Memory (LSTM) network can be employed for time-series analysis on temporal aspects.

IV. ALGORITHM DESIGN

Input: Train Feature set {} which having values of train dataset, Test Feature set {} which having values of test dataset, Threshold T, Label L.
Output: classified all instances with weight and label.
Step 1: Read all features from Test set using below

$$\text{TestFeature} = \sum_{j=1}^{n} (\text{T[j]})$$

Step 2: Read all features from Train set using below

$$\text{TrainFeature} = \sum_{k=1}^{\text{m}} (\text{T[k]})$$

Step 3: Read all features from Train set using below
Step 4: Generate weight of both features set

$$W = (\text{TrainFeature}, \text{TestFeature})$$

Step 5: Verify Threshold
Selected Instance= result = W > T ? 1: 0;
Add each selected instance into L, when n = null
Step 6: Return L

LSTM working

Long Short-Term Memory Networks are sequential neural networks with deep learning capabilities that preserve information. It is a unique type of recurrent neural network that can solve the vanishing gradient issue that RNNs encounter. LSTM was created by Hochreiter and Schmid Huber to address issues with conventional RNNs and machine learning algorithms. Python programmers can use the Keras library to implement the LSTM Model.

A recurrent neural network (RNN) architecture called Long Short-Term Memory (LSTM) was created to solve the vanishing gradient issue with conventional RNNs. The network finds it challenging to identify long-range dependencies in sequential input because gradients, which are utilised to update the weights during training, vanishing gradient problem.

For sequence prediction problems where comprehending context over long distances is critical, like speech recognition, language modelling, and machine translation, LSTM networks are a good fit. They accomplish this by adding a memory cell and a series of gates that regulate the information entering and leaving the cell.
The key components of an LSTM unit include:

Cell State (Ct): The "memory" of the LSTM unit that can store information over long periods.

Hidden State (ht): The output of the LSTM unit, which is used for making predictions or passing information to subsequent units.

Input Gate (i), Forget Gate (f), Output Gate (o): Three sigmoidal gates that regulate the flow of information into and out of the cell, allowing the LSTM to add or remove information as needed.

Cell State Update: The process by which the cell state is updated based on the input, previous cell state, and the gates' decisions.

V. CONTRIBUTION

- Developing LSTM models that accurately predict COVID-19 cases, confirmed cases, active cases, or deaths based on real-time data, helping in resource allocation and decision-making.
- Identifying the most important features or factors influencing the spread of COVID-19 and incorporating them into LSTM models to improve prediction accuracy.

- Developing LSTM models that can assess the risk of COVID-19 transmission in different regions or populations based on real-time data, helping authorities prioritize response efforts.
- Development of a novel approach to modelling for COVID-19 case prediction: we have designed a general network architecture that combines the LSTM model additively and trains the entire architecture jointly, allowing the relative weights of the interpretable predictive LSTM part to be fully determined by the data. This approach is a departure from the traditional sequential modelling approach and can contribute to the literature on sequential data prediction.

VI. RESULTS AND DISCUSSION

Data Description

The forecasting ability of the 14 examined data-based approaches is evaluated using confirmed and recovered COVID-19 data from India every day. On the 30th of January and 26th of February 2020, the WHO NCBI website has a daily record of cumulative confirmed and recovered COVID-19 cases from the first case in India and other countries. Without missing values, the data is updated automatically for delayed data on the website. The confirmed and recovered COVID-19 cases dataset consistent with the research is shown in Figures 2 a–b. India has the most confirmed cases, according to research findings.[3], [29] Given the population of each country, COVID-19 has the most influence on India. Conversely, India has seen a dramatic increase in recovered cases, demonstrating that they responded quickly and effectively to this health crisis.[28], [30]

Table 2: Summary of the used COVID-19 time-series dataset.

Country_Region	Last_Update	Lat	Long_	Confirmed	Deaths	Recovered	Active	Combined_Key	Incident_Rate
Afghanistan	21/05/2022 4:20	33.93911	67.709953	179624	7695	NaN	NaN	Afghanistan	461.422181
Albania	21/05/2022 4:20	41.15330	20.168300	275732	3497	NaN	NaN	Albania	9581.346862
Algeria	21/05/2022 4:20	28.03390	1.659600	265847	6875	NaN	NaN	Algeria	606.250118
Andorra	21/05/2022 4:20	42.50630	1.521800	42572	153	NaN	NaN	Andorra	55098.686340
Angola	21/05/2022 4:20	-11.20270	17.873900	99287	1900	NaN	NaN	Angola	302.093928

Dashboards

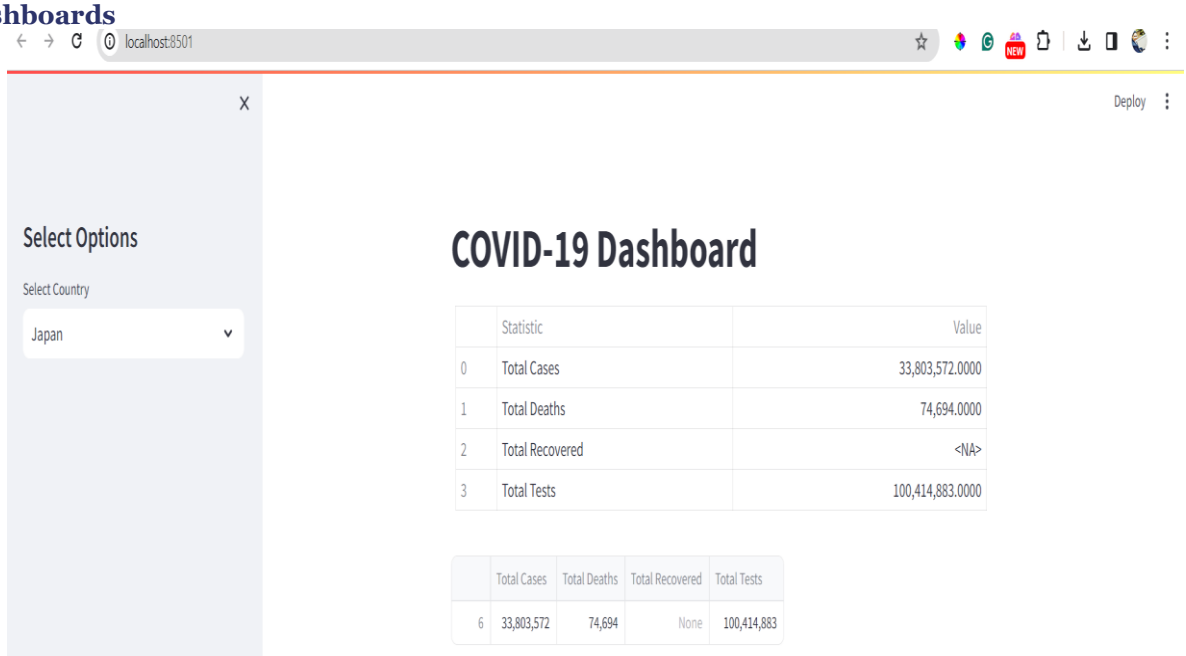

Figure 2: Covid19 Dashboard

To compare COVID-19 statistics for Japan, you can use various metrics such as total cases, total deaths, cases per capita, and testing rates. Obtain the total number of confirmed cases and deaths for Japan. You can find this data from sources like the World Health Organization (WHO), the Japanese Ministry of Health, Labour and Welfare, or other reliable sources.

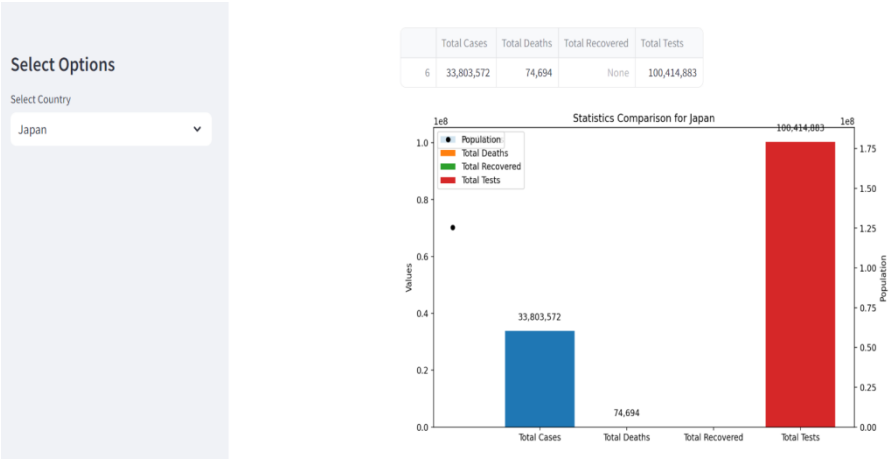

Figure 3: Statistic Comparison for Japan

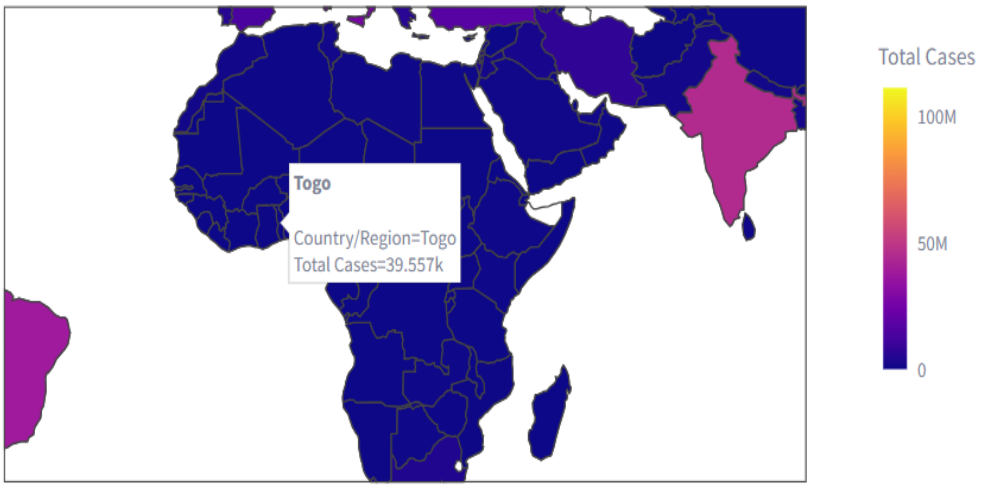

Figure 4: Country/Region Test Cases show

The evolution of COVID-19 worldwide cases has been dynamic and has varied over time.

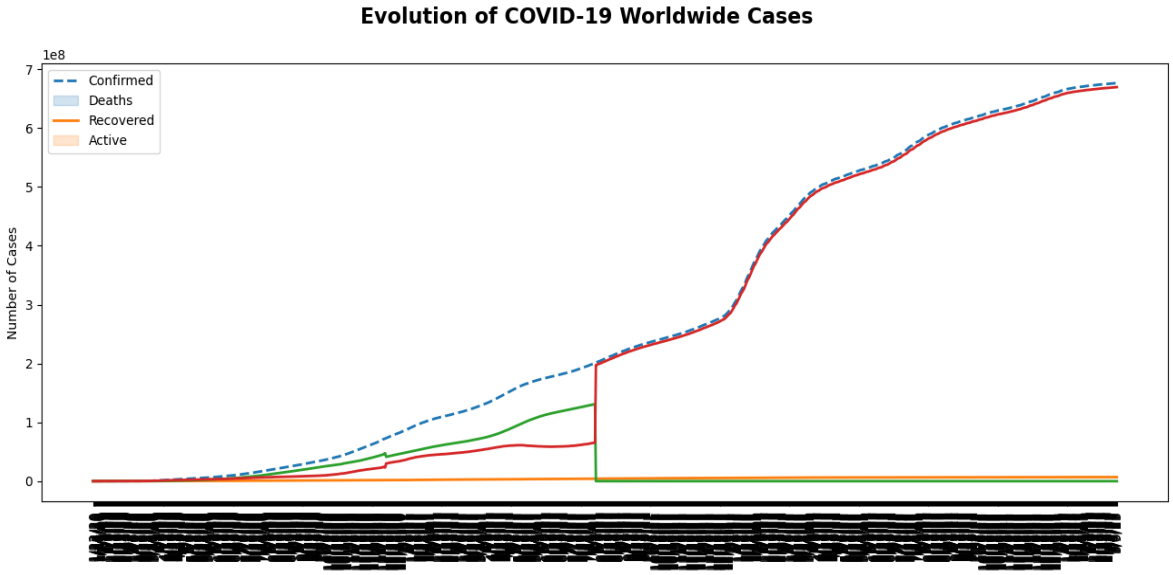

Figure 5: Evolution of Covid-19 Worldwide Cases

Choose a machine learning model suitable for time series forecasting, such as LSTM (Long Short-Term Memory) or Prophet. Use the trained model to predict for the next seven days based on the most recent data available.

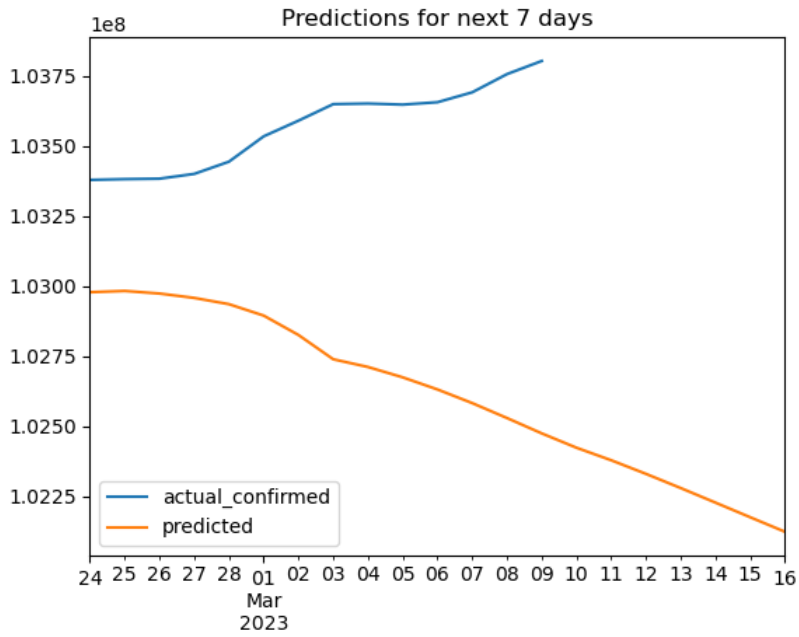

Figure 6 Predictions for next 7 days

There is a measure of how well the model performs on the training data. It indicates how well the model fits the training data. This measures how well the model performs on data it has yet to be trained on, typically a separate validation dataset.

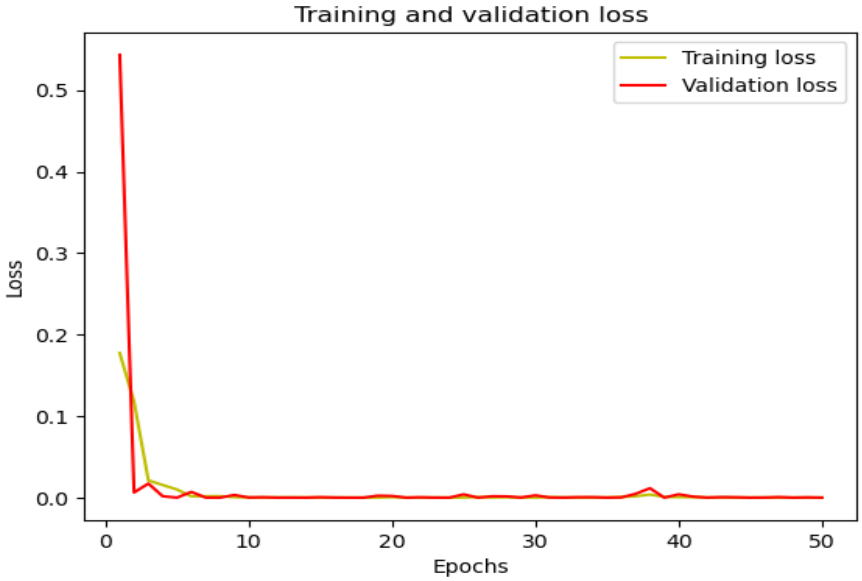

Figure 7: Training and Validation Loss

Table 3: Total infected and death cases of most populated countries

Countries	Population	Total cases	Total deaths	Mortality rate (%)
US	330769230	4495014	152070	3.38%
India	1378364845	1638870	35747	2.18%
China	1438656995	87489	4659	5.32%
Pakistan	220367900	278305	5951	2.14%
Indonesia	273173742	106336	5058	4.76 %
Proposed System (Japan Countries)	100414883	33803572	74694	0.221%

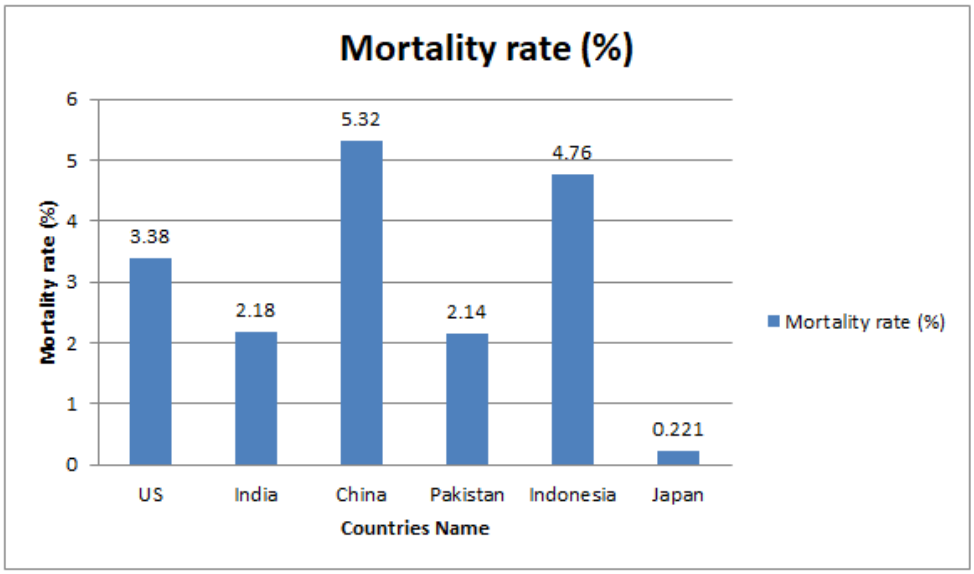

Figure 8: Mortality rate

VII. CONCLUSIONS

The study's conclusions emphasise the development of a cutting-edge machine learning technique that tackles a number of issues specific to COVID-19, such as faster installation, shorter training times, and models that are hardware agnostic. The study shows how unsupervised learning may be used to predict disease outbreaks and analyse the intervals between them, resulting in accurate prognostic predictions for patients that occur well in advance of clinical findings. Additionally, this work evaluates the efficacy of the proposed method on lung area-specific radiography by utilising the WHO, NCBI, and COVIDGR databases. In computational biology and medicine, the use of mathematics and the machine learning paradigm is highly valued. The search for antiviral compounds has yielded several successful results. It takes a virus to make a machine-learning model necessary. An integral part of the research process is the analysis and study of viral protein structures. The goal of this research is to identify data sets exhibiting fast change that could be harmful in the near future by analysing mutations throughout a range of time periods, from one month to one year. A vital step in the scientific method is the analysis and study of viral protein structures. The goal of this study is to employ data analysis to discover greater rates of harmful changes in data sets that could happen in the next weeks by examining mutations over different time periods, ranging from one month to one year.

Future Work

Future applications can utilize real-time data and generated datasets with deep learning techniques. They were deploying deep learning algorithms on real-time IoT data. Conducting experiments on a broader spectrum of health profiles can enhance the precision of the evaluation.

REFERENCES

1. Li, X.; Zhang, J.; Du, Y.; Zhu, J.; Fan, Y.; Chen, X. A Novel Deep Learning-based Sentiment Analysis Method Enhanced with Emojis in Microblog Social Networks. Enterp. Inf. Syst. 2022, 1–22.
2. N. V. A. Ravikumar, R. S. S. Nuvvula, P. P. Kumar, N. H. Haroon, U. D. Butkar and A. Siddiqui, "Integration of Electric Vehicles, Renewable Energy Sources, and IoT for Sustainable Transportation and Energy Management: A Comprehensive Review and Future Prospects," 2023 12th International Conference on Renewable Energy Research and Applications (ICRERA), Oshawa, ON, Canada, 2023, pp. 505-511, doi: 10.1109/ICRERA59003.2023.10269421.
3. Zemberek, NLP Tools for Turkish. Available online: https://github.com/ahmetaa/zemberek-nlp (accessed on 20 September 2021).
4. Uamakant, B., 2017. A Formation of Cloud Data Sharing With Integrity and User Revocation. International Journal Of Engineering And Computer Science, 6(5), p.12.
5. Butkar, U. (2014). A Fuzzy Filtering Rule Based Median Filter For Artifacts Reduction of Compressed Images.
6. Butkar, M. U. D., & Waghmare, M. J. (2023). Hybrid Serial-Parallel Linkage Based six degrees of freedom Advanced robotic manipulator. Computer Integrated Manufacturing Systems, 29(2), 70-82.
7. Kabir, M.; Madria, S. CoronaVis: A real-time COVID-19 tweets data analyzer and data repository. arXiv 2020, arXiv:2004.13932.
8. Taboada, M. Sentiment analysis: An overview from linguistics. Annu. Rev. Linguist. 2016, 2, 325–347.

9. Butkar, U. (2016). Review On-Efficient Data Transfer for Mobile devices By Using Ad-Hoc Network. International Journal of Engineering and Computer Science, 5(3).
10. Sailunaz, K.; Alhajj, R. Emotion and sentiment analysis from Twitter text. J. Comput. Sci. 2019, 36, 101003.
11. A. K. Bhaga, G. Sudhamsu, S. Sharma, I. S. Abdulrahman, R. Nittala and U. D. Butkar, "Internet Traffic Dynamics in Wireless Sensor Networks," 2023 3rd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE), Greater Noida, India, 2023, pp. 1081-1087, doi: 10.1109/ICACITE57410.2023.10182866.
12. Liu, R.; Shi, Y.; Ji, C.; Jia, M. A survey of sentiment analysis based on transfer learning. IEEE Access 2019, 7, 85401–85412.
13. Tyagi, P.; Tripathi, R. A review towards the sentiment analysis techniques for the analysis of twitter data. In Proceedings of the 2nd International Conference on Advanced Computing and Software Engineering (ICACSE), Sultanpur, India, 8–9 February 2019.
14. Saura, J.R.; Palacios-Marqués, D.; Ribeiro-Soriano, D. Exploring the boundaries of open innovation: Evidence from social media mining. Technovation 2022, 102447.
15. Mackey, T.; Purushothaman, V.; Li, J.; Shah, N.; Nali, M.; Bardier, C.; Liang, B.; Cai, M.; Cuomo, R. Machine learning to detect self-reporting of symptoms, testing access, and recovery associated with COVID-19 on Twitter: Retrospective big data infoveillance study. JMIR Public Health Surveill. 2020, 6, e19509.
16. Wan, S.; Yi, Q.; Fan, S.; Lv, J.; Zhang, X.; Guo, L.; Lang, C.; Xiao, Q.; Xiao, K.; Yi, Z.; et al. Relationships among lymphocyte subsets, cytokines, and the pulmonary inflammation index in coronavirus (COVID-19) infected patients. Br. J. Haematol. 2020, 189, 428–437.
17. Rajput, N.K.; Grover, B.A.; Rathi, V.K. Word frequency and sentiment analysis of twitter messages during coronavirus pandemic. arXiv 2020, arXiv:2004.03925.
18. Muthusami, R.; Bharathi, A.; Saritha, K. COVID-19 outbreak: Tweet based analysis and visualization towards the influence of coronavirus in the world. Gedrag Organ. Rev. 2020, 33, 8–9.
19. Jelodar, H.; Wang, Y.; Orji, R.; Huang, S. Deep sentiment classification and topic discovery on novel coronavirus or COVID-19 online discussions: Nlp using lstm recurrent neural network approach. IEEE J. Biomed. Health Inform. 2020, 24, 2733–2742.
20. Aljameel, S.S.; Alabbad, D.A.; Alzahrani, N.A.; Alqarni, S.M.; Alamoudi, F.A.; Babili, L.M.; Aljaafary, S.K.; Alshamrani, F.M. A sentiment analysis approach to predict an individual's awareness of the precautionary procedures to prevent COVID-19 outbreaks in Saudi Arabia. Int. J. Environ. Res. Public Health 2021, 18, 218.