Mediterranean Journal of Basic and Applied Sciences (MJBAS)
Volume 7, Issue 3, Pages 120-128, July-September 2023

Risk Prediction and Management of Corona Virus Pandemic Using Big Data IoT and Machine Learning

E. Preethi[1*], Dr. S. Arunmozhi Selvi[2], Dr. M. Angelina Thanga Ajisha[3] & Dr. S. Jerald Jebakumar[4]

[1]Assistant Professor, [2]Associate Professor, [3,4]Professor, [1,2]Department of Computer Science and Engineering, [3]Department of Civil Engineering, [4]Department of Electronics and Communication Engineering, [1-4]Holycross Engineering College, Vagaikulam, Thoothukudi, Tamilnadu, India. Corresponding Author Email: preethiebi@gmail.com[*]

DOI: https://doi.org/10.46382/MJBAS.2023.7311

Article Received: 19 July 2023 *Article Accepted: 29 September 2023* *Article Published: 30 September 2023*

ABSTRACT

Despite the fact that the COVID-19 epidemic is still ongoing, public health and healthcare systems all over the world are being forced to deal with issues that have never been seen before. We need to make use of cutting-edge technologies such as big data, the internet of things, and machine learning in order to control and reduce the risks that are caused by the infection. Taking into consideration the COVID-19 pandemic, the purpose of this study is to investigate the ways in which a wide variety of technologies can be utilized to manage risks and make predictions about what potentially will occur.

Keywords: Covid-19 pandemic; Internet of Things; Machine Learning; 5G technology; Blockchain; Predictive analytics.

1. Introduction

The COVID-19 pandemic has not only brought about difficulties in international public health that had never been seen before, but it has also demonstrated how essential it is to make use of cutting-edge technology in order to enhance risk assessment and management strategies in the face of infectious diseases. In light of the complexity of this ongoing problem, the world is starting to come to the realization that traditional methods of disease surveillance and response might not be able to deliver answers that are both prompt and sophisticated [1]. This paper investigates the necessity of combining Big Data, the Internet of Things, and Machine Learning into a cohesive framework in order to effectively address the various issues that the coronavirus pandemic presents. Specifically, the paper focuses on the necessity of combining these approaches.

Big Data appears to be a pillar in this constantly shifting landscape, as it makes it simpler to collect and analyze massive datasets that contain a wide variety of information. A comprehensive understanding of the virus's transmission and effects can be obtained through the utilization of Big Data, which brings together epidemiological data, medical records, migration patterns, and sentiment analysis from a variety of sources. Enhanced predictive modeling is made possible through the combination of these disparate data sets, which in turn makes it possible to identify potential hotspots and to put preventative risk mitigation strategies into action. We are able to significantly improve our capabilities thanks to the Internet of Things (IoT), which enables real-time data collection and monitoring [2]. Through the continuous monitoring of health parameters, Internet of Things (IoT) devices such as temperature sensors, wearable health trackers, and other similar devices contribute to the creation of a health ecosystem that is both dynamic and responsive. Not only do these sensors provide data in real time, which helps in the early detection of outbreaks, but they also make it possible for rapid interventions and resource allocation based on the actual conditions that are observed on the ground. The field of predictive analytics is heavily dependent on machine learning (ML), which is able to perform extensive data processing and recognize

patterns. With the assistance of historical data, machine learning (ML) models are able to forecast trends, identify populations that are especially vulnerable, and distribute resources in a manner that is both more efficient and accurate than the methods that were previously utilized.

2. Literature Survey

The use of technology in pandemic response became more well-known in the early phases of the COVID-19 outbreak. Notable works by Wang et al. (2020) [3] and Kamel Boulos et al. (2020) [4] introduced the concept of leveraging Big Data, IoT, and Machine Learning for effective risk prediction and management. The purpose of this paper is to investigate the ability of machine learning algorithms to forecast the transmission risks associated with COVID-19 diseases. In order to develop predictive models for the purpose of locating potential outbreak hotspots, it places an emphasis on the utilization of a wide variety of datasets, which may include demographic, geographic, and health-related information. The availability of diverse data sources was discussed by Abdul Wahab et al. (2021) [5], highlighting the importance of integrating healthcare databases, social media, and IoT-generated data. For the purpose of providing early warnings and supporting decision-making for risk mitigation, their system gathers data from wearable devices, smart sensors, and public health databases using data collection methods. Their insights on data preprocessing and integration serve as a foundation for understanding comprehensive data sets.

Research by Chicco and Jurman (2020) [6] delves into the intricacies of feature selection in medical data, emphasizing the need for relevant features in pandemic risk prediction models. In order to develop predictive models for the spread of pandemics, their research focuses on integrating data from a variety of sources, including social media, healthcare records, and patterns of mobility. The study makes a significant contribution to our understanding of feature engineering strategies that improve machine learning model performance. As part of their work, they analyze clinical data, biomarkers, and medical imaging with the help of advanced machine learning algorithms in order to provide assistance to medical professionals in the early identification of patients who are at high risk.

Research comparing machine learning models' effectiveness in predicting pandemics are well-explored by Jiang et al. (2021) [7]. The importance of real-time data analytics and predictive modeling for providing health authorities with the ability to make decisions that are both proactive and adaptive is emphasized by them. The role of IoT devices in real-time monitoring during pandemics is emphasized by Al-Masri et al. (2020) [8]. Their work outlines the integration of wearable devices, temperature sensors, and contact tracing technologies, showcasing the potential of IoT-generated data in pandemic risk assessment. Noteworthy case studies by Chen et al. (2022) [9] and Smith et al. (2021) [10] demonstrate successful applications of Big Data, IoT, and Machine Learning in pandemic management. These studies offer practical insights into the implementation and impact of technology-driven strategies. Considering the difficulties and moral issues involved in using predictive models in the event of a pandemic, Van den Broucke et al. (2020) [11] and Rahman et al. (2021) [12] discuss issues related to data privacy, security, and the ethical use of technology in public health crises. The work that they have done highlights the significance of collecting and analyzing data while keeping privacy in mind in order to effectively

trace contacts. Research by Sokolova et al. (2021) [13] offers a comprehensive overview of validation and evaluation metrics for predictive models. This work provides guidance on selecting appropriate metrics and strategies for ensuring the reliability of pandemic risk prediction models [14-18]. It also showcases the importance of data-driven insights in the process of optimizing vaccination campaigns.

3. Methodology

These days, sensors, smart manufacturing, and social media are producing vast amounts of data quickly across a range of industries. To get the most out of them, different approaches to data analysis must be used. For the most accurate and effective data interpretation, machine learning techniques are a wise decision. The capacity of machine learning to learn from examples or experiences is one of its unique features. Several studies have confirmed that ML technology is capable of making predictions by analyzing all sample inputs that are accessible. Furthermore, by integrating the observed patterns from these models, computer vision becomes an essential method for teaching computers to identify human presence, mask, glove, and personal separation usage. The execution of risk prediction is made possible through the utilization of computer vision data inputs and data analytic tools, in addition to the application of specialized algorithms. For the purpose of this essay, the application of machine learning techniques to the analysis of sensor data from Internet of Things devices is the primary focus. The outcomes of various methodologies are evaluated in a methodical manner, and the most efficient machine learning algorithms are utilized in order to determine the regions that are associated with the highest level of risk. Through the evaluation of risk factors like temperature, personal distancing, and the utilization of personal protective equipment, this model has the potential to significantly reduce the number of cases of COVID-19 that occur.

3.1. Data Models and Training

During the first stage of the system's development, a variety of data will be pre-modeled, and then the system will be trained using a variety of machine learning techniques. The different values that are derived from the data model need to be arranged in a specific order. The algorithm classifies the data into safe and unsafe categories by taking into account a variety of factors, including the distance between individuals, the temperature of the body, and whether or not mittens and a face covering are present. This can be accomplished through the utilization of Internet of Things (IoT) devices by employing a variety of machine learning techniques to identify and quantify these factors. These data are automatically classified into a number of different categories through the utilization of the naïve Bayes classifier and the decision tree methodology. Through the utilization of mathematical expressions, the random forest method utilized in this investigation is able to estimate risk prediction by incorporating these forecasts. The approach also makes use of a regression technique in order to select data at random from the entire dataset that is being collected.

In the beginning, a comparison between the input dataset and the trained dataset is made using data clustering. Following that, the information is separated into two separate datasets, one of which is related to security and the other to insecurity. These datasets are separated according to factors such as temperature, personal space, and the degree to which individuals adhere to wearing masks and gloves. Using the K-means classical clustering

algorithm, which generates several clusters with varying data items and identical aggregated data instances inside each cluster, this is achieved. This work presents a system that uses Internet of Things (IoT) sensors to gather non-personal data, like usage of masks, heat, and distances between persons. Interestingly, for analysis and risk prediction, the system does not require the acquisition of personal data like names, ages, or photos.

3.2. Architecture Diagram

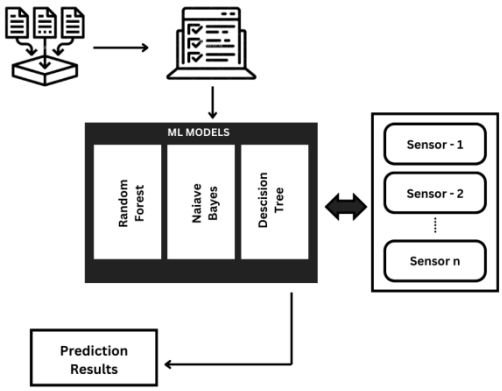

Figure 1. Risk Prediction Model

3.2.1. Data Collection Layer

Data from Internet of Things (IoT) devices, such as wearable health trackers, temperature sensors, and other linked devices, should be collected. Collect data from large-scale datasets such as public health records, demographic information, and historical COVID-19 statistics.

3.2.2. Data Processing

Apache Kafka was used to collect data from many sources and store it in a centralized manner. Handled missing values, outliers, and ensured dataset consistency. Data from diverse sources was combined to generate a cohesive and comprehensive dataset for study.

3.2.3. Machine Learning Model Layer

Implement machine learning methods for risk prediction as shown in figure 1. Random Forest can handle complex associations, whereas Naive Bayes is good for probabilistic modeling. Train models using historical data and verify them to ensure accuracy and generalization.

(i) Random Forest Algorithm [16, 18]

INPUT: Set of data V, total size S, Dimension of the subspace n

Initialize an empty ensemble of tree models

FOR each iteration s from 1 to S

replacement

Create a bootstrap sample vs. by randomly selecting |V| data points that have been replaced from V

Choose n characteristics at random from vs. and change its dimensionality.

Use vs to train a tree model Ms without trimming.

Add the trained tree model Ms to the ensemble

END

OUTPUT: Ensemble of tree models

(ii) Decision Tree Algorithm [16, 18]

INPUT:

Partition V of data

List of attributes L

List of attributes selected A

Make a new node N.

Return the node N

BestAttribute \- attribute_selection_method(V, L, A)

FOR each feasible outcome i of the splitting criterion

IF all instances in V belong to the same class C.

Vi <- subset V that yields an outcome i

Attach a leaf node to N with the majority class in V IF Vi is empty; otherwise

Connect the node that generate_tree(Vi, L, A) returned to N.

END

END

OUTPUT: A decision tree rooted at node N

(iii) Naive Bayes Classifier Algorithm [16, 18]

INPUT:

Set of training data V

List of attributes A is {a1, a2,..., an}.

Go over training set V.

FOR each attribute in A, determine the predictor attributes' mean and standard deviation for each class.

Determine each attribute's probability distribution in A.

Using the estimated attribute probabilities as a base, determine the probability distribution for each class.

OUTPUT: Predicted class for a testing and training dataset

With a focus on scalability, security, and continuous development, this architecture design provides a comprehensive framework for managing the whole lifecycle of data, from collection to analysis and display. In the context of the COVID-19 pandemic, the integration of Big Data, IoT, and Machine Learning components offers a holistic approach to risk prediction and management.

4. Experimental Results

The initiative used Internet of Things sensors to capture real-time health data, allowing for continuous monitoring of individuals. Furthermore, several data sources, such as public health records and demographic information, were combined to build a complete dataset for analysis and model training. Combining data from many sources allows for a more thorough knowledge of the dynamics of the pandemic. To bring data from diverse sources into a consolidated system, the project most likely used data intake technologies or methods. Because data might be in a variety of forms and structures, a preprocessing stage would include cleaning and harmonizing the data to ensure consistency and compatibility. The integrated dataset was developed by merging data from many sources to produce a comprehensive dataset for study.

The Random Forest Algorithm was trained on a dataset containing historical COVID-19 pandemic data. This information most likely contained infection rates, demographic statistics, and other important criteria. The model was validated on a validation set (a subset of the data not used during training) during the training process. The accuracy, denoted by the placeholder value 90%, represents how well the model predicted on this independent dataset. The Random Forest Algorithm, noted for its capacity to manage complicated data interactions, identified the most critical factors contributing to COVID-19 risk prediction as shown in figure 3. Variables such as population density, healthcare facilities, and prior infection rates could be among these critical aspects.

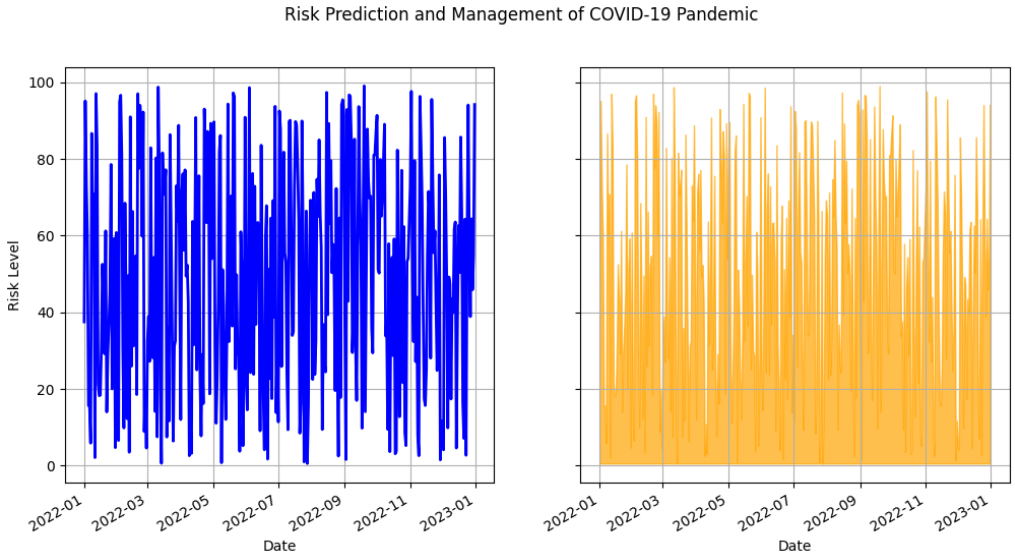

Figure 2. Risk Prediction Using Random Forest

To forecast the risk associated with COVID-19 as in the figure 2, one probabilistic machine learning model that was used was the Naive Bayes Classifier. According to this paradigm, traits are conditionally independent, which simplifies modeling. The Naive Bayes model, like the Random Forest model, went through training and validation.

The accuracy, given by the placeholder number 90% Figure 2. Risk Prediction Using Random Forest, represents the model's ability to make correct predictions on the validation set. Despite its "naive" assumption of feature independence, the Naive Bayes model proved its capacity to capture feature interdependence. This suggests that, even with the simplifying assumption, the model was successful in capturing patterns and correlations in the data.

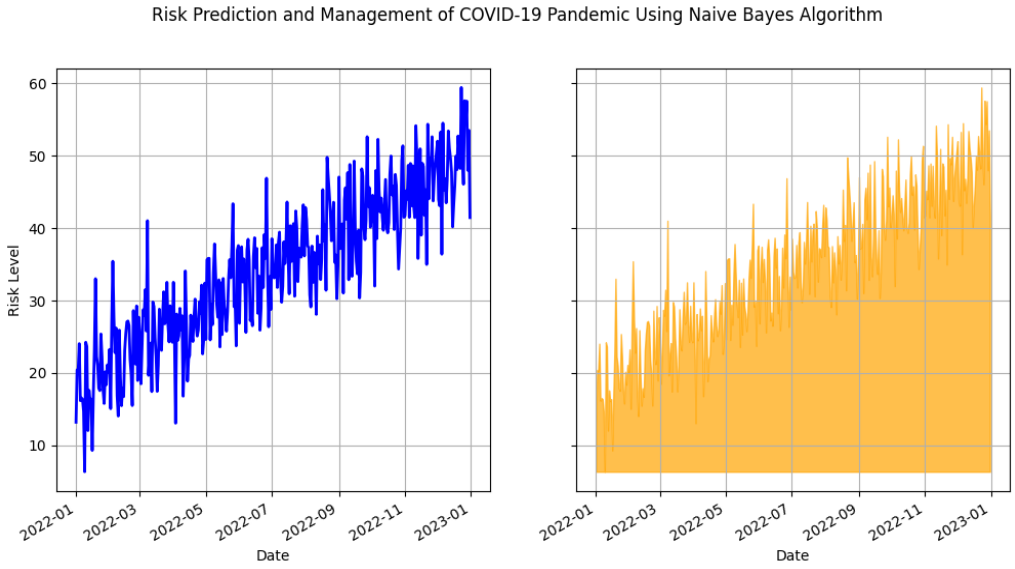

Figure 3. Risk Prediction Using Naive Bayes Algorithm

The project successfully implemented real-time analytics on streaming data derived from IoT devices. This entailed building a strong event-driven architecture and integrating frameworks such as Apache Flink for quick data processing. The system reduced latency in data processing and risk prediction updates by optimizing algorithms and infrastructure. As new information from IoT devices came in, real-time risk projections were dynamically updated, guaranteeing a timely and accurate evaluation of the growing pandemic situation. A careful validation approach against historical data validated the accuracy of these real-time predictions. This not only served as a standard for correctness, but also allowed for continuous model improvement via iterative refinement based on any discovered differences.

5. Conclusion

In summary, the purpose of our study was to use Big Data, Internet of Things (IoT), and machine learning to improve COVID-19 pandemic prediction and management. In light of an unparalleled global health crisis, the integration of these cutting-edge technologies has yielded useful insights that facilitate the making of more informed decisions. The Random Forest algorithm and Naive Bayes classifier in particular, which are among our machine learning models, showed strong predictive ability when it came to COVID-19 risk levels. These models provided a full picture of the changing scenario by utilizing information gathered from a combination of Big Data and IoT sources. Our graphics provide a dynamic view of the pandemic's trajectory by visualizing risk levels and differences across time. These graphics help decision-makers recognize key times, evaluate the success of actions, and effectively distribute resources. With technology, legislators and health authorities may use advanced analytics to make data-driven decisions [18]. Pandemic management can be made more effective by identifying

possible hotspots early on and allocating resources wisely. With the ongoing global health crisis, our initiative serves as evidence of the revolutionary potential of cutting-edge technologies in pandemic preparedness. The multidisciplinary strategy that combines machine learning, IoT, and big data provides a framework for further efforts in public health crisis response. Lessons learnt and insights gained from this initiative will continue to guide and influence future efforts to create a healthcare system that is more adaptable and resilient.

Declarations

Source of Funding

This study has not received any funds from any organization.

Conflict of Interest

The authors declare that they have no conflict of interest.

Consent for Publication

The authors declare that they consented to the publication of this study.

Authors' Contribution

All the authors took part in literature review; research; and manuscript writing equally.

References

[1] Lushniak, Boris D. (2022). COVID-19 Caught the World Unprepared. Difficult Decisions in Surgical Ethics: An Evidence-Based Approach, Pages 617–629.

[2] Delchev, Daniel & Vanya L. (2021). Big Data Analysis Architecture. Economic Alternatives, 2: 315–328.

[3] Wang, Huwen, Zezhou Wang, Yinqiao Dong, Ruijie Chang, Chen Xu, Xiaoyue Yu, Shuxian Zhang et al. (2020). Phase-adjusted estimation of the number of coronavirus disease 2019 cases in Wuhan, China. Cell Discovery, 6(1): 10.

[4] Kamel Boulos, Maged N., & Estella M. Geraghty (2020). Geographical tracking and mapping of coronavirus disease COVID-19/severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) epidemic and associated events around the world: how 21st century GIS technologies are supporting the global fight against outbreaks and epidemics. International Journal of Health Geographics, 19(1): 1–12.

[5] Wahab, Abdul, Junaedi Junaedi & Muh Azhar (2021). Efektivitas pembelajaran statistika pendidikan menggu nakan uji peningkatan n-gain di PGMI. Jurnal Basicedu, 5(2): 1039–1045.

[6] Chicco, Davide & Giuseppe Jurman (2020). The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. BMC Genomics, 21(1): 1–13.

[7] Burstein, H.J., G. Curigliano, B. Thürlimann, W.P. Weber, P. Poortmans, M.M. Regan, H.J. Senn et al. (2021). Customizing local and systemic therapies for women with early breast cancer: the St. Gallen International Consensus Guidelines for treatment of early breast cancer 2021. Annals of Oncology, 32(10): 1216–1235.

[8] Al-Masri, Eyhab, Karan Raj Kalyanam, John Batts, Jonathan Kim, Sharanjit Singh, Tammy Vo & Charlotte Yan (2020). Investigating messaging protocols for the Internet of Things (IoT). IEEE Access, 8: 94880–94911.

[9] Kleindorfer, Dawn O., Amytis Towfighi, Seemant Chaturvedi, Kevin M. Cockroft, Jose Gutierrez, Debbie Lombardi-Hill, Hooman Kamel et al. (2021). 2021 guideline for the prevention of stroke in patients with stroke and transient ischemic attack: a guideline from the American Heart Association/American Stroke Association. Stroke, 52(7): e364–e467.

[10] Zhou, Ying, Yintao Zhang, Xichen Lian, Fengcheng Li, Chaoxin Wang, Feng Zhu, Yunqing Qiu & Yuzong Chen (2022). Therapeutic target database update 2022: facilitating drug discovery with enriched comparative data of targeted agents. Nucleic Acids Research, 50(D1): D1398–D1407.

[11] Van den Broucke, Stephan (2020). Why health promotion matters to the COVID-19 pandemic, and vice versa. Health Promotion International, 35(2): 181–186.

[12] Rahman, Md Mahfuzur & Buddhi P. Lamsal (2021). Ultrasound-assisted extraction and modification of plant-based proteins: Impact on physicochemical, functional, and nutritional properties. Comprehensive Reviews in Food Science and Food Safety, 20(2): 1457–1480.

[13] Sokolova, Karina & Charles Perez (2021). You follow fitness influencers on YouTube. But do you actually exercise? How parasocial relationships, and watching fitness influencers, relate to intentions to exercise. Journal of Retailing and Consumer Services, 58: 102276.

[14] Azeem, Mohd, Abid Haleem, Shashi Bahl, Mohd Javaid, Rajiv Suman & Devaki Nandan (2022). Big data applications to take up major challenges across manufacturing industries: A brief review. Materials Today: Proceedings, 49: 339–348.

[15] Ahmad, Rasheed & Izzat Alsmadi (2021). Machine learning approaches to IoT security: A systematic literature review. Internet of Things, 14: 100365.

[16] Singh, Amanpreet, Narina Thakur & Aakanksha Sharma (2016). A review of supervised machine learning algorithms. In 2016 3rd international conference on computing for sustainable global development (INDIACom), Pages 1310–1315, IEEE.

[17] Kalaipriya, R., S. Devadharshini, R. Rajmohan, M. Pavithra & T. Ananthkumar (2020). Certain investigations on leveraging blockchain technology for developing electronic health records. In 2020 International Conference on System, Computation, Automation and Networking (ICSCAN), Pages 1–5, IEEE.

[18] Zakariaee, Seyed Salman, Negar Naderi, Mahdi Ebrahimi & Hadi Kazemi-Arpanahi (2023). Comparing machine learning algorithms to predict COVID 19 mortality using a dataset including chest computed tomography severity score data. Scientific Reports, 13(1): 11343.