OPEN ACCESS Freely available online

Journal of Medical Diagnostic Methods

Research Article

Synergizing Machine Learning and Blockchain for Pioneering Early Disease Detection: A Focused Study on COVID-19 Diagnosis

Jayendra S. Jadhav*, Jyoti Deshmukh

Department of Computer Engineering, Rajiv Gandhi Institute of Technology, University of Mumbai, Mumbai, India

ABSTRACT

Early disease detection plays a pivotal role in modern healthcare, as it can substantially influence patient prognosis, healthcare costs, and overall public health. Machine learning algorithms serve as indispensable tools for uncovering subtle patterns, trends, and predictive factors present in complex medical data sources, such as patient records, diagnostic images, and genomic information. The integration of machine learning with Blockchain technology presents a substantial opportunity for transformative advancements in healthcare. This document examines several machine learning techniques such as LR, RF, GB, SVC, and GNB. It showcases their remarkable effectiveness in analysing symptoms for accurate disease detection, with COVID-19 serving as a primary case study. The application of cross-validation offered a sophisticated analysis of the performance capabilities, revealing that the Random Forest and Gradient-Boosting models are particularly effective, striking a vital balance in their metrics, which is vital for the reliable detection of diseases at their beginning. In addition, these models, with their significant accuracy (0.91) and precision (0.92), affirmed its status as an exceptional tool for the early identification of diseases. Ultimately, the combination of machine learning and Blockchain technologies significantly bolsters healthcare systems' ability to detect and manage diseases early, enhancing our understanding of diseases and guiding public health measures and strategies.

Keywords: Machine learning; Blockchain; Early disease detection; COVID-19

INTRODUCTION

Early disease detection plays a pivotal role in modern healthcare, as it can substantially influence patient prognosis, healthcare costs, and overall public health. Machine learning and Blockchain technology together could completely change how diseases are identified and treated in healthcare systems [1,2]. The following section delves into the importance of detecting diseases at an early stage and examines how the combined power of machine learning and Blockchain technology can significantly improve healthcare outcomes.

Context and significance of early disease detection

Detecting diseases at an early stage is essential for several reasons that significantly impact both individual health and broader public well-being. Firstly, early identification of diseases often leads to more successful treatment outcomes. By diagnosing diseases early, healthcare providers can implement effective treatment strategies, which can substantially improve patient health and enhance overall quality of life. For instance, in the case of cancer, detecting tumours at an early stage when they are smaller and localized may allow for less invasive treatments and a higher likelihood of complete remission. Secondly, early detection can alleviate the financial burden on healthcare systems. Treating diseases in advanced stages typically requires more resources, including intensive medical interventions, hospitalizations, and long-term care. By identifying and treating diseases earlier, healthcare systems can avoid the need for expensive treatments associated with late-stage diseases, ultimately reducing healthcare costs and optimizing resource allocation [3].

Additionally, the prompt identification of diseases is vital for halting the transmission of infectious conditions among populations. Prompt identification of infectious diseases enables swift intervention measures such as isolation, contact tracing, and vaccination campaigns, which are essential for containing

Correspondence to: Jayendra S. Jadhav, Department of Computer Engineering, Rajiv Gandhi Institute of Technology, University of Mumbai, Mumbai, India, E-mail: jayendra071985@gmail.com

Received: 04-May-2024, Manuscript No. JMDM-24-31182; **Editor assigned:** 07-May-2024, PreQC No. JMDM-24-31182 (PQ); **Reviewed:** 21-May-2024, QC No. JMDM-24-31182; **Revised:** 28-May-2024, Manuscript No. JMDM-24-31182 (R); **Published:** 04-Jun-2024, DOI: 10.35248/2168-9784.24.13.481

Citation: Jadhav JS, Deshmukh J (2024) Synergizing Machine Learning and Blockchain for Pioneering Early Disease Detection: A Focused Study on COVID-19 Diagnosis. J Med Diagn Meth. 13:481.

outbreaks and safeguarding public health. This becomes especially clear in situations involving highly contagious illnesses such as COVID-19, where the early identification of cases has played a crucial role in curbing the spread and reducing the strain on healthcare infrastructures and society overall [4,5].

The ability to detect health issues early, whether its infectious diseases, or chronic conditions, can indeed be transformative. Early detection not only enhances individual health outcomes but also significantly impacts healthcare systems, economies, and society as a whole. It emphasizes the importance of investing in preventive healthcare measures, screening programs, and diagnostic technologies that enable timely identification of diseases, ultimately leading to healthier populations and more resilient healthcare systems [2,6,7].

Machine learning in healthcare

Machine learning has played an instrumental part in revolutionizing the healthcare industry due to its remarkable capacity to analyse enormous data sets quickly and precisely. Within the healthcare sector, machine learning algorithms serve as indispensable tools for uncovering subtle patterns, trends, and predictive factors present in complex medical data sources such as patient records, diagnostic images, and genomic information. This analytical competence empowers healthcare professionals to make informed decisions, anticipate disease risks, and customize treatment plans to suit the individual characteristics of each patient. From disease diagnosis to treatment optimization and predictive modelling, machine learning has consistently demonstrated its value and transformative potential across various healthcare applications [8].

Utilizing machine learning methods, healthcare practitioners can derive valuable insights from extensive and varied datasets, consequently improving clinical decision processes and patient care results. For example, in the field of medical imaging, machine learning algorithms excel at identifying subtle abnormalities or anomalies in diagnostic scans, enabling early disease detection and timely interventions. Similarly, in genomic analysis, machine learning methodologies facilitate the identification of genetic markers and biomarkers associated with disease susceptibility and progression, enabling personalized risk assessment and targeted interventions [9].

Moreover, machine learning algorithms play a pivotal role in optimizing treatment strategies by analysing patient data to identify the most effective treatment regimens and predict treatment responses. By utilizing historical patient data and clinical outcomes, machine learning models can guide the selection of personalized treatment options tailored to the unique characteristics and medical history of each patient [10]. Additionally, machine learning-based predictive modelling techniques enable healthcare providers to forecast disease trajectories, anticipate potential complications, and intervene proactively to mitigate adverse outcomes [11].

Incorporating machine learning methods into healthcare processes marks a transformative movement towards a more personalized and data-centric approach in medicine. Utilizing the predictive power of these algorithms, medical professionals can discover novel insights, refine clinical operations, and, ultimately, elevate the standard of patient care throughout the healthcare journey [10].

Blockchain in medical sector

Blockchain offers a secure and transparent means of managing healthcare data. It establishes an immutable, decentralized ledger where sensitive patient information, including medical histories and treatment records, can be securely stored and shared among authorized stakeholders. Blockchain ensures data integrity, confidentiality, and accessibility, with patients retaining control over their own health data. Moreover, it provides a comprehensive audit trail, bolstering trust and accountability in healthcare systems [1].

The synergy ML and blockchain

The merging of machine learning and Blockchain technology offers significant potential for innovative improvements in the healthcare sector. Merging the analytic prowess of machine learning with the robust security features of Blockchain allows for significant improvements within healthcare infrastructures. Machine learning thrives on delving into the vast amounts of data preserved on the Blockchain, extracting critical insights that can bolster the early identification of diseases, evaluate patient risk profiles, and refine therapeutic approaches. Concurrently, Blockchain provides a secure, distributed, and tamper-proof repository for sensitive health information, ensuring its confidentiality and integrity during exchange and storage. Patients can securely share their health data with healthcare providers and researchers, confident in the knowledge that the Blockchain's immutable records guarantee the integrity and authenticity of their information. This synergy between machine learning and Blockchain inculcates confidence in the healthcare ecosystem, reinforcing trust and facilitating the seamless exchange of information for improved patient care and outcomes [12,13].

In summary, early disease detection is a pivotal aspect of healthcare, and the integration of machine learning with Blockchain systems offers a promising approach to enhance detection while ensuring data security and transparency. This combination has the potential to reshape healthcare delivery by making it more personalized, efficient, and patient-centred, ultimately benefiting both individuals and public health at large.

A literature survey on the identification of unknown diseases at an early stage using machine learning and Blockchain can be organized around key thematic areas. These include the transformative role of machine learning in healthcare, the integration of Blockchain for enhanced data security and management, and the combined potential of these technologies in pioneering early disease detection.

Machine learning in early disease detection

The Reddy, et al., and Potnis and Tiple both underscore the transformative role of machine learning within healthcare [2,3]. These studies delve into a variety of algorithms, evaluating their capacity to identify diseases at their nascent stages, thereby potentially improving patient outcomes significantly. Reddy, et al., present an in-depth analysis of machine learning's advancements in healthcare, focusing on early disease detection [2]. They examine the progress, challenges, and future directions of machine learning applications, highlighting the technology's pivotal role in enhancing diagnostic processes and healthcare

delivery. Gopisetti, et al., and Potnis N and Tiple B discuss the development of systems utilizing machine learning to predict multiple diseases. These papers demonstrate the adaptability of machine learning models in processing extensive datasets to diagnose a wide array of conditions, underscoring the potential for more comprehensive and inclusive healthcare solutions [3,14].

Santangelo, et al., has explored the application of machine learning in predicting infectious diseases. Their detailed analysis emphasizes the critical role that data-driven models play in anticipating disease outbreaks, essential for timely and effective intervention strategies. The review points out how machine learning algorithms can significantly improve public health responses by enabling precise and early detection of infectious diseases [15]. Alanazi R [8], study focuses on the early identification and treatment of chronic diseases through the use of machine learning techniques. The study shows how early detection of chronic diseases by predictive modeling may result in far superior treatment regimens and outcomes. The research conducted by Alanazi adds to the increasing corpus of evidence supporting the use of machine learning in the treatment of chronic illnesses in order to enhance patient satisfaction and increase the efficacy of medical care [8,16].

These papers collectively highlight the critical impact of machine learning in revolutionizing healthcare through early disease detection. They point to a future where healthcare is not only reactive but also proactive, with technologies like machine learning paving the way for earlier interventions, personalized treatment plans, and overall improved patient care.

Blockchain in healthcare

Bangare, et al., highlight the pivotal importance of Blockchain technology in securing patient data across healthcare systems. By facilitating secure and transparent data exchanges, Blockchain technology aids collaborative efforts in disease detection and management. The study showcases how Blockchain can be effectively employed to safeguard sensitive health information, ensuring that patient data is protected against unauthorized access and breaches, thereby bolstering trust in digital healthcare services [17]. Chen, et al., propose an innovative framework that combines the strengths of Blockchain technology with machine learning to enhance patient privacy during the diagnostic process. The framework addresses prevalent concerns regarding data privacy in healthcare by ensuring that patient information remains confidential while still benefiting from the predictive power of machine learning algorithms. This approach mitigates the challenges associated with data sharing and collaboration among healthcare providers, making it a pivotal development in the pursuit of secure and efficient disease diagnosis systems [4].

These papers contribute valuable insights into the ongoing efforts to improve healthcare services through technology. They highlight the importance of data security and privacy in the context of utilizing advanced computational methods like machine learning for disease diagnosis, illustrating the potential of Blockchain as a solution to some of the most pressing challenges in digital healthcare.

Integrated machine learning and blockchain approaches

Kayikci, et al., and Bangare, et al., research both look at how Blockchain and ML can be applied in the healthcare domain [12,17]. They highlight how this synergy not only bolsters the predictive accuracy of early disease detection systems but also ensures the security and integrity of healthcare data. This integrated approach presents a forward-looking model for healthcare systems, combining the predictive power of machine learning with the secure and decentralized nature of Blockchain [17,4]. Potnis N and Tiple B [3], provide an in-depth analysis of cutting-edge machine learning techniques and their applicability in the realm of disease prediction. Their research contributes to the expanding body of knowledge on how advanced algorithms can enhance diagnostic processes, offering insights into the development of more sophisticated and accurate healthcare diagnostic tools [3]. Alzubaidi, et al., concentrate on using deep learning for identifying COVID-19, illustrating the critical role that machine learning technologies play in addressing worldwide health crises. Their research highlights the significance of technological agility in public health emergencies by providing an example of the critical role that machine learning plays in quickly recognizing and handling pandemic events [18].

Collectively, these investigations offer a detailed overview of both the present and future uses of machine learning and Blockchain in healthcare. They explore how these technologies can improve diagnostic precision, protect confidential patient information, and enable quick responses to emerging global health challenges. Overall, the literature suggests a promising horizon where machine learning and Blockchain collectively enhance the capabilities of healthcare systems in the early detection and management of unknown diseases. These technologies not only offer advanced diagnostic tools but also ensure the secure and ethical use of medical data, driving forward the paradigm of proactive and personalized healthcare.

MATERIALS AND METHODS

Detecting unfamiliar diseases in their early stages poses a considerable hurdle within the healthcare sector. Access to meticulously organized and effectively managed datasets plays a pivotal role in conducting research and experimentation in this field. To address this need, we have chosen to employ the "COVID-19 Symptoms Dataset" compiled by Takbir [19], for our investigative pursuits.

This dataset provides a wealth of valuable insights into the symptoms linked with COVID-19, thereby facilitating thorough research analysis in this domain. Notably, given that COVID-19 itself was once an unidentified disease, the availability of a comprehensive dataset focusing on its symptoms proves exceptionally beneficial for our research undertakings. The dataset consists of data detailing typical symptoms observed in COVID-19 patients, including fever, cough, fatigue, shortness of breath, etc. Each record in the dataset likely corresponds to a specific case or patient, with various attributes that reflect whether certain symptoms are present or absent.

The following (Figure 1) clearly delineates how different parameters, including COVID-19 symptoms, demographic attributes, and key factors, interrelate, offering insights into various potential scenarios (Figures 1 and 2). Some potential analyses that can be performed using the dataset.

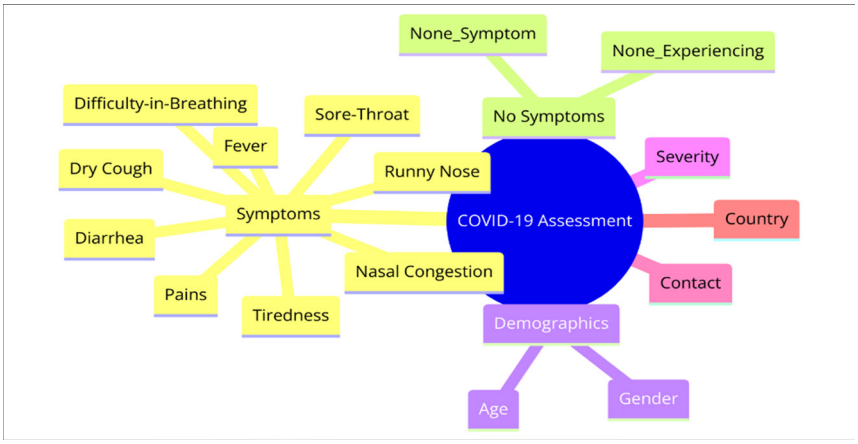

Figure 1: COVID-19 assessment parameter list.

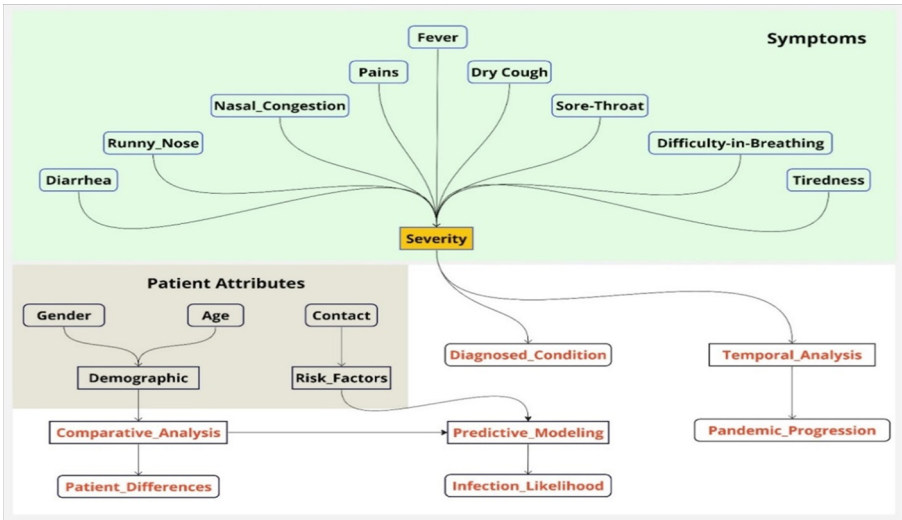

Figure 2: Dataset analysis based on parameters.

Descriptive Analysis

Summarizing the frequency and distribution of different symptoms among COVID-19 patients.

Predictive Modelling

Creating models for estimating, using symptom profiles, the probability of contracting COVID-19.

Comparative Analysis

Contrasting symptom profiles between different demographic groups, geographic regions, or severity levels of COVID-19 cases.

Temporal Analysis

Examining how symptom prevalence and severity evolve over time during the course of the pandemic.

These analyses can provide valuable insights into the patterns, trends, and characteristics of COVID-19 symptoms, contributing to our understanding of the disease and informing public health interventions and strategies.

The Figure 3 outlines a structured approach to processing and analysing health data to potentially identify an unknown disease.

Input

The analysis focused on three primary input parameters:

Patient symptoms: The study utilized a range of primary symptoms from the dataset, such as fever, dry cough, and sore throat, tiredness, body aches, diarrhea, and the severity of these symptoms.

Geographic region: Geographical area (Postal code) where the patients are reported, providing a geographical lens to the investigation.

Date: The specific dates associated with the patient data, which is essential for temporal analysis (Figure 3).

Patient data collection

Electronic Health Records (HER), encompass the compilation of patient information, which is gathered *via* a secure system underpinned by Blockchain technology. This method ensures the capture of key data such as the patient's location and the date details, enhancing the reliability and traceability of the records. In the context of research, the dataset pertaining to symptoms of COVID-19, which includes 316,800 individual entries from a diverse patient population, is employed [19].

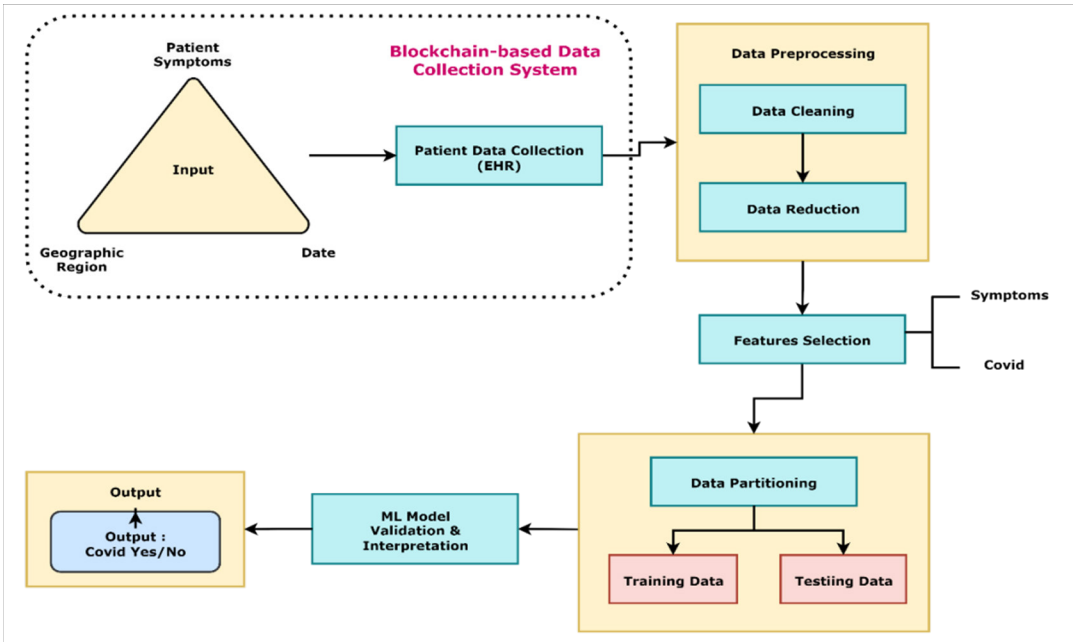

Figure 3: Methodology of Block chain-based data collection system.

Data cleaning

The dataset contains 316,800 entries, among which approximately 0.03% exhibit null values. Notably, all records are encoded in binary format (0 or 1), underscoring the significance of even the smallest value for analytical accuracy. To preserve data integrity and ensure uniformity throughout, meticulous cleaning procedures have been employed to rectify missing values, thereby enhancing the dataset's consistency for subsequent analyses.

Data reduction

The dataset comprises records from several locations (Country=10), each exhibiting a different degree of data completeness. Among these, 'China' stands out with a total of 31,860 entries, showcasing the lowest incidence of missing fields. Further analysis, particularly through standard deviation metrics, reveals China's data as the most consistent across the board. This consistency, coupled with the comprehensive nature of its dataset, makes China an ideal choice for the focal point of our experimental analysis. Upon reviewing the dataset (with 27 parameters), 14 parameters (Symptoms, Severity levels and Country) crucial for analysis were identified, aligning with the research objectives. To determine the severity of the COVID-19, four criteria are taken into consideration: Severity_Mild, Severity_Moderate, and Severity_Severe and Severity_None. These variables would enable us to determine the patient's degree of symptoms. Following this, a refined dataset was created, encompassing only the chosen parameters.

Features selection

COVID-19 parameters and analysis

Symptoms: From the entire set of parameters, nine specific features—namely "Fever," "Tiredness," "Dry-Cough," "Difficulty-in-Breathing," "Sore-Throat," "Pains," "Nasal-Congestion," "Runny-Nose," and "Diarrhea"— were chosen for subsequent processing. These selected features are referred to as "symptoms" features, as they represent a combination of symptoms potentially indicative of COVID-19 in individuals [6].

COVID-19: The parameters "Severity_Mild," "Severity_Moderate," "Severity_Severe," and "Severity_None" are utilized for identifying the presence or absence of COVID.

The deadly COVID-19 pandemic first emerged in Wuhan, within China's Hubei Province, and swiftly expanded across the world. Initially, a significant concentration of COVID-19 cases was observed in Wuhan, indicating that early detection of the pandemic's onset could have mitigated numerous adverse outcomes.

Therefore, in our research, postal codes and dates were prioritized as key variables. By developing a system capable of monitoring disease outbreaks based on specific geographic locations and timelines, it might be possible to identify the initial stages of unknown diseases, potentially preventing widespread transmission. The fundamental aim of this research is to understand the development and dissemination of an unidentified disease, especially with regards to COVID-19, by analysing data related to symptoms, geographic information, and time-based patterns. To address this, the dataset has been augmented with two new columns, 'postal code' and 'day,' which are populated with postal codes and dates, respectively, assigned at random.

Data partitioning

The dataset exhibits a slight imbalance, with class 0 comprising 51.53% and class 1 representing 38.54%. While this disparity isn't significant, it could potentially lead to a bias favouring the majority class (class 0) during predictive analysis. To evaluate the model's accuracy, we employ two distinct data partitioning strategies: one without cross-validation and the other utilizing cross-validation.

For the approach without cross-validation: We allocate 80% of the data for training purposes, enabling the model to learn patterns from the dataset. The remaining data is set aside for testing purposes, enabling us to evaluate the accuracy and effectiveness of the model in making predictions.

For the cross-validation approach: Employed the k-fold method, dividing the dataset into subsets of equal size. Each fold is used as a validation set one time while the other folds serve as the training set. This cycle is repeated five times, ensuring that each fold acts as the validation set exactly once. By calculating the average performance across all folds, we achieve a more stable and reliable assessment of the model's effectiveness. Utilizing both approaches provides comprehensive insights into the model's predictive capabilities, allowing us to make informed decisions regarding its deployment.

Validation and interpretation

The dataset is organized into binary categories (0, 1) and shows a slight imbalance. While this imbalance is not significant, it could still lead to a bias favouring the more common class during predictive analysis. To evaluate the model's accuracy, we adopt two distinct strategies for data partitioning: one method excludes cross-validation, while the other includes it. The model employs testing data to predict whether an individual has COVID-19 ('COVID Yes') or not ('COVID No'), based on their symptoms and severity scores.

These predictions are then assessed by comparing them to the actual outcomes in the testing data. Understanding the model's results involves determining the likelihood of COVID-19 presence. A prediction of 'COVID Yes' indicates that the individual is likely to have COVID-19, while 'COVID No' suggests the opposite. These insights assist healthcare professionals in making informed decisions, in conjunction with other tests and clinical assessments.

Output

The final output of the model is a binary result indicating 'COVID Yes/No', signifying whether the patient is predicted to have COVID-19.

RESULTS AND DISCUSSION

Models comparison and discussion

The performance of each model is assessed using four key metrics: Accuracy, Precision, Recall, and F1-Score. The dataset analysed shows that class 0 accounts for 51.53% and class 1 for 38.54%, presenting a slight class imbalance. This imbalance could potentially lead to a bias towards the dominant class (class 0) in predictive analysis. To accurately gauge the model's effectiveness, we utilize two different data partitioning methods: One method without cross-validation and another with the integration of cross-validation. Specifically, we apply the k-fold cross-validation technique to ensure a thorough evaluation of the model's performance across various data segments (Table 1), (Figure 4).

Table 1: Comparison of ML models without cross validation.

Model	Accuracy	Precision	Recall	F1-Score
Logistic regression	0.75	0.75	1	0.86
Random forest	0.749	0.75	1	0.86
Gradient boosting	0.75	0.75	1	0.86
Support vector classification	0.75	0.75	1	0.86
Gaussian naive bayes	0.73	0.75	0.8	0.86

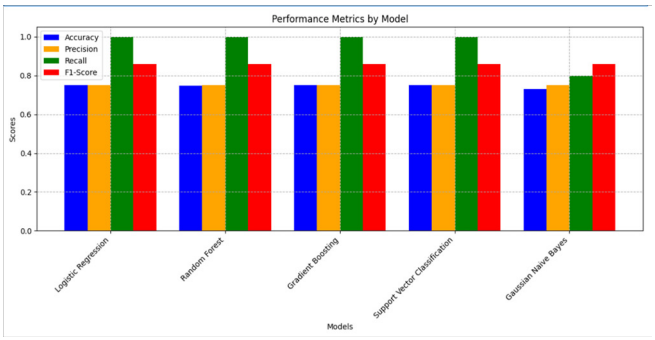

Figure 4: Performance Metrics of ML Models without cross validation. **Note:** (■): Accuracy; (■): Precision; (■): Recall; (■): F1-Score.

Model performance without performing cross validation

The performance of various machine learning models, evaluated without employing cross-validation, is presented in Table 1 and Figure 4, using both tabular and graphical methods. In this study, Logistic Regression, Random Forest, Gradient Boosting, and Support Vector Classification all perform similarly in terms of accuracy, precision, recall, and F1-Score.

This consistency suggests that, even without cross-validation, these models perform successfully in recognizing true positives (evidenced by high Recall) and maintain a decent balance of Precision and F1-Score. On the other hand, the Gaussian Naive Bayes model shows a slight decline in Accuracy (0.73) and Recall (0.8), although it matches the other models in Precision and F1-Score. The diminished Recall score suggests a lesser ability to capture all positive cases.

Model performance incorporating cross validation

The analysis using cross-validation, as demonstrated in Table 2 and Figure 5, presents the enhanced statistical performance of various machine learning models both visually and numerically. The application of k-fold cross-validation leads to marked improvements in all models' performance, underscoring the method's effectiveness in boosting the reliability and applicability of models across varied data segments. Notably, Random Forest and Gradient Boosting exhibit outstanding performance, with Gradient Boosting marginally leading.

This advancement is considerable, as shown by Gradient Boosting's impressive metrics: 0.91 in Accuracy, 0.92 in Precision, and 0.90 in Recall, confirming its high efficacy in correctly identifying true positives while minimizing false positives. Both Logistic Regression and Support Vector Classification also display notable enhancements in their metrics, demonstrating their robustness and effectiveness in disease prediction under cross-validation. Gaussian Naive Bayes, on the other hand, shows a significant increase in Precision (0.88), but a decrease in Recall to 0.79, indicating a balance between accurately detecting true positives and limiting false positives (Table 2), (Figure 5).

Table 2: Comparison of machine learning models with cross validation.

Model	Accuracy	Precision	Recall	F1-Score
Logistic regression	0.86	0.88	0.84	0.86
Random forest	0.9	0.91	0.88	0.9
Gradient boosting	0.91	0.92	0.9	0.91
Support vector classification	0.86	0.87	0.85	0.86
Gaussian naive bayes	0.84	0.88	0.79	0.83

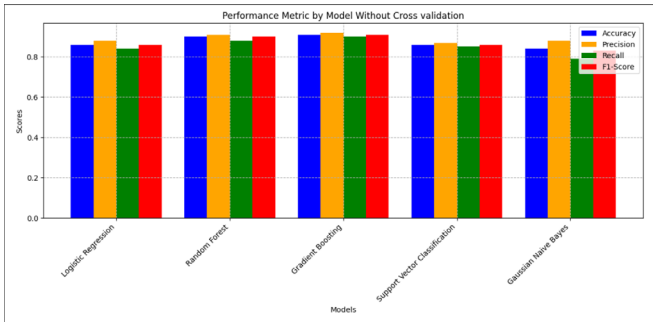

Figure 5: Performance metrics of Machine learning models with cross validation. **Note**: ( ): Accuracy; ( ): Precision; ( ): Recall; ( ): F1-Score.

Models summary and insights

When determining the usefulness of above machine learning models for the early diagnosis of diseases with symptoms similar to COVID-19, the metrics of accuracy, precision, recall, and F1-score are pivotal. Cross-validation significantly enhances the understanding of model performance by simulating a more rigorous evaluation process, reflecting potential real-world application scenarios more accurately. The analysis reveals:

- Gradient Boosting and Random Forest as potentially the best choices for tasks requiring high accuracy and a balanced approach to Precision and Recall, crucial for sensitive applications like early disease detection.
- Logistic Regression and Support Vector Classification offer robust alternatives with their balanced performance, particularly useful when model stability across diverse data samples is desired.
- Gaussian Naive Bayes presents a viable option where computational efficiency and model simplicity are prioritized, despite some trade-offs in Recall (Figure 6). Evaluating the ROC curves depicted in Figure 6 for ML models reveals their relative ability to distinguish accurately between classes in a dataset that could mirror the identification of early-stage diseases. Both the Gradient Boosting and Random Forest models demonstrate exemplary proficiency, as evidenced by their AUC scores of 0.92, underscoring their potential for highly accurate predictive tasks. Meanwhile, Logistic Regression and Support Vector Classification show commendable efficacy, each with an AUC of 0.91, suggesting their dependability for uniform detection across diverse datasets. On the other hand, Gaussian Naive Bayes, with a slightly lower AUC of 0.89, provides a quick and straightforward option, albeit with a modest compromise in detection sensitivity. The collective assessment of these models sheds light on their respective capabilities, offering strategic insights for selecting an optimal model for the vital role of early disease detection where precise, timely identification is key.

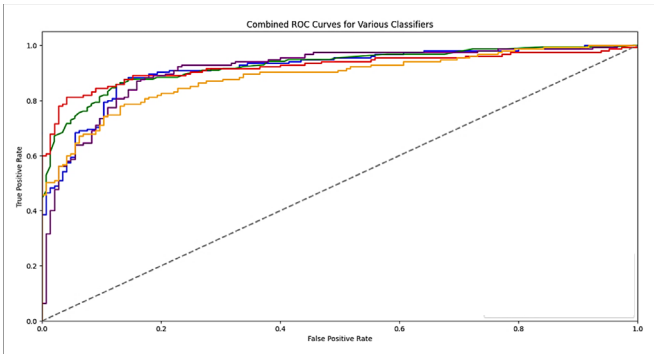

Figure 6: ROC curve of combine ML models. **Note**: ( ): Logistic Regression (AUC=0.91); ( ): Random Forest (AUC=0.92); ( ): Gradient Boosting (AUC=0.92); ( ): SVC (AUC=0.91); ( ): Gaussian Naïve Bayes (AUC=0.89); ( ): Chance.

Symptoms analysis based on area/postal code

Figures 7 and 8 show a set of area and line charts depicting the proportion of COVID-19 symptoms reported in various postal code locations. Each chart refers to a certain postal code region and displays the incidence of symptoms such as fever, cough, and sore throat, among others, culminating in a total number of COVID-19 cases. In the first image, distinct charts for each location display a trend line across symptoms, indicating how frequent each symptom is in that area. This could help identify hotspots with higher incidences of particular symptoms (Figures 7 and 8).

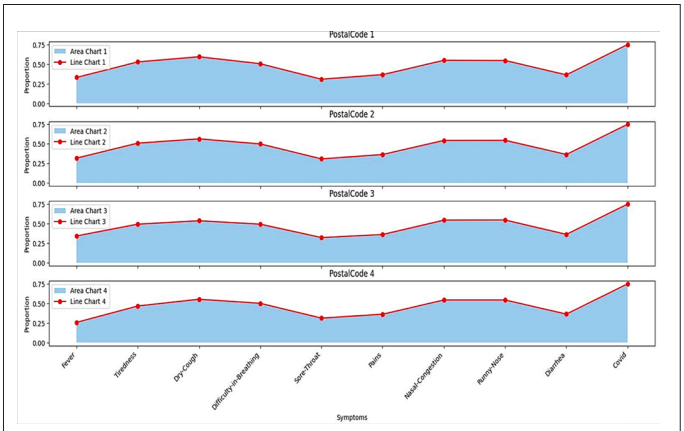

Figure 7: Area chart for different postal codes. **Note**: ( ): Area chart; ( ): Line chart.

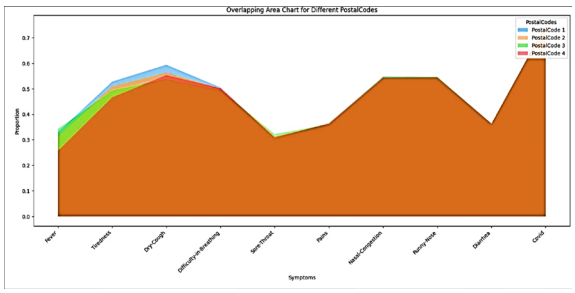

Figure 8: Overlapping area chart for different postal codes. **Note**: (): Postal code 1; (): Postal code 2; (): Postal code 3; (): Postal code 4.

In the composite visual representation of the second figure, the amalgamation of data allows for a succinct contrast of symptom reports by region. Peaks within these plots serve as markers for increased symptom reports, potentially indicating zones of intensified COVID-19 transmission. The method of analysis is pivotal for health authorities to detect and concentrate efforts on regions exhibiting a rise in symptoms like fever, cough, and sore throat. Using such graphical data facilitates the prudent distribution of medical interventions as well as the proactive identification of growing hotspots, hence improving response to possible epidemics.

Analysis based on observed symptoms on date

Figures 9 and 10 shows area and line charts that monitor the proportion of COVID-19 symptoms over four distinct dates. Each figure corresponds to a certain day and depicts the variation in symptom reporting over time (Figures 9 and 10).

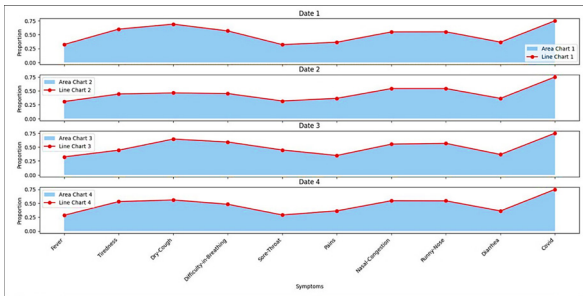

Figure 9: Area chart for different days. **Note**: (): Area chart; (): Line chart.

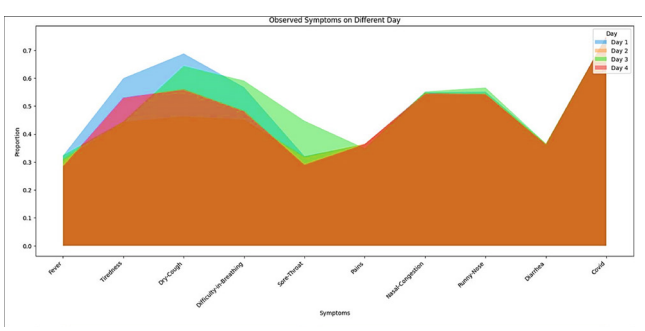

Figure 10: Overlapping area chart for different days. **Note**: (): Day 1; ():Day 2; ():Day 3; (): Day 4.

Figure 9 provides a date-by-date breakdown, highlighting fluctuations in specific symptoms, which may indicate changes in the prevalence or nature of the disease over time. For example, one may see if symptoms like fever or dry cough are becoming more prevalent on date 1 and date 3. The second figure merges all four dates into a single overlapping region chart, allowing for a visual comparison of symptoms between the dates. This layered approach can identify trends, such as an increase or reduction in certain symptoms over time, which could indicate COVID-19's spread or regression within a community. These charts are useful for tracking the evolution of the disease, since an increase in symptom reporting might suggest an epidemic, whilst a decrease could indicate successful containment efforts. Such visual data may be critical for early detection and reaction to the epidemic.

Analysis on combining symptoms, postal code and date

The chart arranges a series of area graphs by date and postal code to track the prevalence of COVID-19 symptoms as shown in Figure 11. Each row represents a different date, while each column aligns with a specific postal code. The area under the curve in each graph quantifies the frequency of symptoms reported in that region on that date. By cross-referencing temporal and geographical data, public health officials can discern patterns, identify potential disease hotspots, and predict the spread of the illness, which is vital for deploying resources and implementing containment measures.

Figure 11: Combine visualization of symptoms with postal code and date. **Note**: (): Area chart; (): Line chart.

CONCLUSION

The ability to identify unknown diseases is a crucial task that holds the potential to prevent significant losses worldwide. Early detection, especially of viral diseases like COVID-19, can play a pivotal role in mitigating the adverse effects they may have on global health, economies, and societies at large. By recognizing and addressing such diseases in their embryonic stages, we can prevent many potential losses, thereby safeguarding the well-being of communities and ensuring the continuity of daily life without the disruption of widespread health crises.

This research thoroughly investigates the promising roles of

machine learning and Blockchain in enhancing healthcare, particularly in early disease detection. It focuses on various machine learning algorithms-Logistic Regression, Random Forest, Gradient Boosting, Support Vector Classification, and Gaussian Naive Bayes-demonstrating their impressive capabilities in symptom analysis for precise disease identification, using COVID-19 as a key example. The application of cross-validation offered a sophisticated analysis of the performance capabilities, revealing that the Random Forest and Gradient Boosting models are particularly effective, striking a vital balance in their metrics, which is vital for the reliable detection of diseases at their beginning. The Gradient Boosting algorithm, with its significant accuracy (0.91) and precision (0.92), affirmed its status as an exceptional tool for the early identification of diseases. This finding underscores its readiness for clinical adoption, where its application could lead to more timely and precise disease management.

FUTURE WORK

Subsequent research avenues include broadening the scope of diseases for which machine learning and Blockchain technologies can be applied, incorporating these technologies into healthcare infrastructures to improve patient care and data sharing, and conducting real-world trials to assess their efficacy in diverse settings. Moreover, tackling ethical and regulatory challenges, along with promoting interdisciplinary collaboration, becomes essential to align these technologies with the complex demands of healthcare delivery. Such endeavours could pave the way for more personalized, efficient, and proactive healthcare systems, thereby improving health outcomes and strengthening public health infrastructure.

REFERENCES

1. Jadhav JS, Deshmukh J. A review study of the blockchain-based healthcare supply chain. Soc Sci Humanit Open. 2022;6(1):100328.
2. Reddy KP, Satish M, Prakash A, Babu SM, Kumar PP, Devi BS. Machine learning revolution in early disease detection for healthcare: Advancements, challenges, and future prospects. Int Conf Inform Sci and Technology. 2023:638-643.
3. Potnis N, Tiple B. Machine learning techniques for disease prediction. ITM Web Conf. 2023; 57:01004.
4. Chen X, Wang X, Yang K. Asynchronous blockchain-based privacy-preserving training framework for disease diagnosis. Int Conf Big Data. 2019;5469-5473.
5. Kumar N, Das NN, Gupta D, Gupta K, Bindra J. Efficient automated disease diagnosis using machine learning models. J Healthc Eng. 2021.
6. Aswathy AL, Anand HS, Chandra SV. COVID-19 severity detection using machine learning techniques from CT-images. Evol Intell. 2023;16(4):1423-1431.
7. Demilie WB. Plant disease detection and classification techniques: A comparative study of the performances. J Big Data. 2024;11(1):5.
8. Alanazi R. Identification and prediction of chronic diseases using machine learning approach. J Healthc Eng. 2022.
9. Shahid AH, Khattak WA. Improving patient Care with machine learning: A game-changer for healthcare. Appl Res Artif Intell Cloud Comput. 2022;5(1):150-163.
10. Grampurohit S, Sagarnal C. Disease prediction using machine learning algorithms. Int Conf Emerg. Technol. 2020;1-7.
11. Uddin S, Khan A, Hossain ME, Moni MA. Comparing different supervised machine learning algorithms for disease prediction. BMC Med Inform Decis Mak. 2019;19(1):1-6.
12. Kayikci S, Khoshgoftaar TM. Blockchain meets machine learning: A survey. J Big Data. 2024;11(1):9.
13. Hasanova H, Tufail M, Baek UJ, Park JT, Kim MS. A novel blockchain-enabled heart disease prediction mechanism using machine learning. Comput Electr Eng. 2022;101:108086.
14. Gopisetti LD, Kummera SK, Pattamsetti SR, Kuna S, Parsi N, Kodali HP. Multiple disease prediction system using machine learning and streamlit. Int Conf Smart Syst. Technol. 2023:923-931.
15. Santangelo OE, Gentile V, Pizzo S, Giordano D, Cedrone F. Machine learning and prediction of infectious diseases: A systematic review. Mach learn knowl extr. 2023;5(1):175-198.
16. Richens JG, Lee CM, Johri S. Improving the accuracy of medical diagnosis with causal machine learning. Nat Commun. 2020;11(1):3923.
17. Bangare S, Verma M, Siddiqui ZA, Shankar SA, Kumar P. A Blockchain and machine learning-based smart healthcare system. Int Adv Res Sci Commun Technol. 2023:56-66.
18. Alzubaidi M, Zubaydi HD, Bin-Salem AA, Abd-Alrazaq AA, Ahmed A, Househ M. Role of deep learning in early detection of COVID-19: Scoping review. Comput Methods Programs Biomed Update. 2021;1:100025.
19. Takbir A. Covid19 symptoms dataset. 2022.