International Journal of Research in Engineering and Science (IJRES)
ISSN (Online): 2320-9364, ISSN (Print): 2320-9356
www.ijres.org Volume 13 Issue 9 ǁ September 2025 ǁ PP. 100-107

A Hybrid Machine Learning Technique for Infectious Disease Prediction in Big Data Environment: A Literature Review

Kolhe Shilpa Dattatraya[1] and Ankur Khare[2]
[1]Research Scholar, Department of Computer Science and Application, Rabindranath Tagore University, Raisen, India,
[2]Assistant Professor, Department of Computer Science and Application, Rabindranath Tagore University, Raisen, India,
Corresponding Author: shilpaa3@gmail.com

Abstract
Infectious sicknesses remain a significant global health task, often resulting in large-scale outbursts that affect billions of people. The advent of big data and Machine Learning (ML) techniques proposals new avenues for cultivating infectious disease prediction, enabling quick detection and mitigation of potential epidemics. This paper presents a comprehensive literature assessment on hybrid machine learning techniques for transferable disease prediction within big data atmospheres. The review explores several hybrid models that combine numerous machine learning algorithms to increase prediction accuracy, scalability, and productivity. In particular, it examines the incorporation of supervised and unverified learning methods, deep learning agendas, and ensemble techniques. Furthermore, the paper highpoints the key tasks faced in big data atmospheres, such as high-dimensional data, real-time processing necessities, and the requirement for interpretability. The literature review goals to identify current tendencies, gaps, and future directions in the request of hybrid ML methods for infectious disease forecast, providing valuable insights for scientists and practitioners in the arena.

Date of Submission: 15-09-2025 Date of acceptance: 30-09-2025

I. Introduction

The spread of transferrable diseases has posed a dangerous public health task for centuries, affecting societies together economically and socially. Rapid development, globalization, and climate change have additional intensified the risk of transferrable disease outbreaks, making the necessity for accurate and timely expectation more pressing than forever. Traditionally, infectious disease prediction trusted on statistical models and expert-driven techniques that often struggled to preserve up with the difficulty and scale of modern outbursts. However, the rise of big data and ML tools has converted the landscape, offering sophisticated tools to examine vast and heterogeneous datasets in immediate [18].

Machine learning, mainly when applied to large-scale datasets, has exposed immense possible in disease surveillance, early outbreak discovery, and prediction of infection tendencies. Nevertheless, individual machine learning models frequently encounter limitations in handling the various nature of transferrable disease data, which may comprise clinical records, epidemiological data, and genomic orders. Hybrid machine learning methods, which combine the powers of multiple algorithms, offer a talented solution to these tasks by enhancing predictive correctness and addressing data heterogeneity [9].

This paper delivers an in-depth literature examination of hybrid machine learning techniques practical to infectious disease guess in big data atmospheres. It explores various hybrid models, counting ensemble methods, deep learning-based hybrids, and mixtures of managed and unsupervised learning. The examination also discusses the practical tasks in working with big data, such as the necessity for scalability, actual analysis, and the difficulty of high-dimensional data. By synthesizing discoveries from recent educations, this paper aims to offer a holistic opinion of the present landscape and propose instructions for future research in the arena of infectious disease forecast using hybrid ML techniques [19].

II. Literature Review

N. Sharma et al. [1] offerings a heterogeneous collaborative forecasting model pointed at improving the accuracy of disease guesses. The model combines numerous machine learning algorithms, including choice trees, support vector machines (SVM), and neural networks, to prediction the likelihood of numerous diseases. By integrating these models, the collaborative leverages the strengths of respectively algorithm to enhance prognostic performance. The proposed model was appraised using medical datasets, and its guesses were benchmarked against separate models. The authors originate that the ensemble methodology outperformed each individual model in terms of correctness and robustness. The education also highlighted the standing of combining diverse algorithms to challenge complex disease prediction difficulties. However, the model requires important computational resources, and its disposition in real-time environments could be interesting.

S. Grampurohit et al. [2] discovers the presentation of machine learning procedures, including choice trees, random forests, and provision vector machines, in predicting diseases founded on medical information. The authors developed models to prediction various diseases by examining patients' clinical data and environmental features. The study focused on the correctness and performance of dissimilar algorithms, with random forests realising the best results in rapports of prediction correctness. The authors emphasize the possible of machine learning in disease prophecy, particularly for diseases that require quick detection and interference. Despite the success, challenges continue, particularly in the quality of the contribution data, which can heavily inspiration the models' reliability. Moreover, the models lacked interpretability, which could hinder practical application in clinical settings.

P. Dutta et al. [3] presents an AI-based feature selection procedure to forecast COVID-like diseases by using machine learning models such for example logistic regression, k-nearest neighbours (KNN) and accidental forests. The attention of the research is on recognising the most significant features from large medical datasets to streamline models without surrendering accuracy. The authors employed numerous feature selection techniques to decrease the dimensionality of the dataset, manufacture the models more effective and interpretable. They testified that the random forest model with particular features outperformed others in rapports of accuracy. However, the attribute selection process may overlook multifarious interactions between structures, and the model's performance depends seriously on the quality of the selected geographies. The research recommends further refinement in feature selection methods can enhance model dependability in real-world applications.

F. E. Ayo et al. [4] designates the expansion of a decision support system (DSS) planned to diagnose various diseases using bioinformatics methodologies. The system employs artificial neural networks (ANN) and judgment trees to investigate genomic and epidemiological data for accurate diagnosis of numerous diseases. The DSS integrates several diagnostic data sources, including genomic orders and patient health histories, to improve the correctness of disease diagnosis. The system was verified on a series of diseases, and it demonstrated high accurateness and fast processing times. The training underscores the probable of combining bioinformatics with machine learning to increase analytical tools. However, the complexity of genomic information integration presents scalability problems, especially when applied to world-wide datasets.

H. H. Thary et al. [5] recommends a structure for diagnosing infectious diseases founded on a questionnaire-driven approach combined with machine learning techniques such as conclusion trees and accidental forests. The study focuses on increasing a cost-effective and competent diagnostic tool for use in isolated areas where advanced medical services may not be accessible. The questionnaire collects patient indications and other relevant data, which is then processed through machine learning models to predict probable infectious diseases. The framework suggestions a scalable solution for nonstop diagnosis in low-resource settings. However, the support on self-reported symptoms presents potential biases and imprecisions, which could affect the model's diagnostic accurateness.

M. Mariki et al. [6] offerings a machine learning method for diagnosing malaria by integration clinical symptoms with persistent demographic features. The author's industrialized models using support vector machines (SVM), logistic reversion, and accidental forests to classify patients as moreover positive or destructive for malaria. By integrating together clinical symptoms and demographic information, the models showed improved diagnostic accurateness compared to models relying exclusively on clinical symptoms. The accidental forest model performed the greatest in terms of correctness and recall. The study highlights the standing of using comprehensive patient information to improve diagnostic accurateness, but it also notes the boundaries in generalizing the results to non-endemic zones where malaria symptoms capacity overlap with additional diseases.

Y. A. Adamu et al. [7] attentions on construction machine learning models to projection malaria outbreaks based on conservational factors such as infection, humidity, and rainfall. The author's industrialized models using decision trees, accidental forests, and SVM, with accidental forests achieving the highest correctness. The study highlights the position of environmental monitoring in forecasting malaria outbreaks and provides a agenda for integrating environmental information into disease prediction models. The model's capability to forecast outbreaks in advance can help in the opportune implementation of preventive events. However, the model's presentation is highly dependent on the superiority and availability of perfect environmental data.

J. Gao, J. Li et al. [8] offerings a time series study of the cumulative frequencies of typhoid and paratyphoid fevers in China using together Grey and Seasonal Autoregressive Integrated Moving Average

(SARIMA) models. The authors associated the two models in expressions of their forecasting accurateness for future incidences of these infections. The study originate that while the SARIMA model implemented well for short-term forecasts, the Grey model was more appropriate for long-term forecasting. The authors determined that combining together models could offer better overall forecasting accurateness. The study's methodology is useful for public health preparation but is limited to time-series information and cannot incorporate additional features such as demographic or clinical statistics.

H. Wang et al. [9] presents a Student Physical Health Information Management Model leveraging huge data analytics to progress the monitoring and assessment of students' health. The classical integrates physical fitness information collected from students and developments it using huge data techniques to recognize trends, correlations, and potential health dangers. The study emphasizes the necessity for timely and widespread health monitoring in educational organizations. Through the proposed model, health managers and instructors can quickly investigate large datasets, allowing for more knowledgeable management of student health histories. However, challenges such as information privacy and safety were noted as critical problems to address, especially when dealing with sensitive health knowledge.

M. Wang et al. [10] benevolences a deep learning-based speedy warning system for transferrable diseases in hospital settings. The authors recommend a Multi-Self-Regression Deep Neural Network (MSR-DNN) that usages historical infection data to forecast potential disease outbreaks. The MSR-DNN model is organised to identify designs and correlations in hospital infection information, providing real-time warnings for possible outbreaks. By assimilating multiple layers of self-regression, the model accomplishes higher predictive correctness than traditional methods. The learning demonstrated the system's advantageousness in early disease detection, dropping the response time for hospital management to recruit preventive methods. However, the model's complexity necessitates substantial computational resources, and its application in resource-constrained hospitals could pose tasks.

M. K. Singh et al. [11] attentions on the presumption and detection of infectious diseases by a selection of machine learning algorithms, such for example decision trees, support direction machines, and neural networks. The authors manufacturing a model to spot diseases by analysing clinical and conservation factors. Their approach incorporates feature engineering to enhance the extrapolative power of the models. The unintentional forest algorithm displayed the highest accuracy in envisaging infectious diseases compared to previous models. The paper highlights the impending of machine learning in cultivating diagnostic processes but also proceedings that model interpretability and the superiority of input data endure significant challenges for real-world implementation.

D. Swain et al. [12] suggests a machine learning classifier for diagnosing Chronic Kidney Disease (CKD) with extraordinary accuracy. The authors occupied various algorithms, including logistic regression, coincidental forests, and support vector machines, to convalesce a vigorous classifier. They used a feature collection technique to improve the model's efficiency, converging on reducing completed fitting and increasing generalization. The classifier single-minded strong performance, predominantly in early-stage diagnosis of CKD, which is fundamental for timely intercession. The random forest classical was the most operative in terms of precision and recall. Although the results are encouraging, the study mentions the necessity for larger, more diverse datasets to authenticate the model additional.

Y. Liu et al. [13] presents a machine learning framework grounded on Extreme Gradient Boosting (XGBoost) to forecast the occurrence and expansion of the H9N2 avian influenza virus in positioning hen farms. The authors self-possessed data from several environmental and farm management features and applied XGBoost to pinpoint key factors contributing to sickness outbreaks. The model performed remarkably well, with high predictive correctness, and provided actionable visions for disease prevention and regulator in poultry farming. The training highlights the utility of machine learning in agricultural surroundings but appreciates that the model’s correctness may decline when applied to altered farm environments due to varying information conditions.

H. F. Ahmad et al. [14] concentrates on educating COVID-19 forecasting models depleting machine learning techniques, particularly communicable disease modelling approaches. The authors employed models such for example Long Short-Term Memory (LSTM) and Support Vector Machines (SVM) to guess the spread of COVID-19 by analysing epidemiological records. The recommended models incorporated various components such as infection rates, community mobility, and containment measures to make truthful predictions. The LSTM model presented superior performance in projecting the disease’s short-term spread equated to traditional models. The authors observed that while machine learning can suggestively enhance disease modelling, the unpredictability in COVID-19 statistics, such as underreporting, poses a task to the models’ accurateness.

M. Hussain et al. [15] reconnoitres the custom of machine learning techniques to prediction waterborne diseases, concentrating on diseases caused by contaminated water foundations. The authors industrialized a model that combines eco-friendly data, such as water characteristic and weather conditions, with clinical information to predict outbreaks of diseases comparable cholera and typhoid. The training employed algorithms such for example decision trees and accidental forests, with random forests offering the highest prediction correctness. The model helps public health officials classify high-risk areas and apparatus timely interventions. The training emphasizes the need for real-time information integration to improve the efficiency of the model in positive public health scenarios.

M. Mwamnyange et al. [16] contributions a big data analytics agenda aimed at successful the surveillance and rejoinder to childhood infectious diseases. The framework develops an improved Map Reduce algorithm to procedure large-scale health data more professionally. The system collects and analyses real-time information from several health and environmental sources to categorize potential outbreaks of childhood diseases. The anticipated framework enhances the rapidity and accuracy of disease discovery, enabling quicker reactions by healthcare professionals. The revision highlights the scalability and tractability of the framework, making it appropriate for use in together high- resource and low-resource settings. However, the authors comment the necessity for high-quality data to safeguard the framework's effectiveness.

S. Palaniappan et al. [17] recommends a machine learning-based meth000od to prediction epidemic disease dynamics and infectious danger. The authors developed models expending algorithms such as accidental forests and support vector machines to approximation the spread of infectious diseases grounded on historical information and environmental factors. The education emphasizes the importance of sympathetic disease dynamics to accomplish epidemics effectively. The results displayed that the random forest algorithm implemented best in predicting the infectious danger, providing valuable visions for public health planning. The reading underlines the importance of characteristic selection and model tuning in completing high prediction accurateness.

G. Dhiman et al. [18] presents a federated knowledge framework to precaution healthcare data while enabling documents sharing for machine learning models. The associated learning approach allows several institutions to train models on their limited data without membership sensitive patient information. The authors recommend this framework to augment the security and privacy of healthcare information in a big data atmosphere while maintaining great predictive performance. The learning demonstrates the effectiveness of the federated knowledge model in civilizing disease prediction accuracy without compromising information privacy. However, the paper similarly discusses challenges such as communiqué overhead and the necessity for standardized data formats crossways institutions.

M. Uppal et al. [19] offerings a cloud-based system considered for real-time responsibility prediction in hospital atmospheres using machine learning techniques. Hospitals rely seriously on various sensors to screen patients, equipment, and conservational conditions. Sensor failures can lead to critical problems, principally in real-time healthcare monitoring. Near address this, the authors suggest a machine learning framework that puckers sensor data on the cloud for real-time examination and liability prediction. The framework is manufactured on cloud computing arrangement to ensure scalability and efficient information processing. The machine learning model, primarily using decision trees and neural networks, can forecast potential sensor failures formerly they occur, minimizing downtime and attractive operational efficiency in hospitals. Moreover, the system emphasizes real-time nursing and data-driven decision-making to recover hospital management. Challenges related to information security, latency, and the addition of legacy methods are highlighted as areas necessitating further research.

L. B. Amusa et al. [20] behaviours a bibliometric investigation of the job of big data in infectious disease epidemiology, studying research trends and openings in the information. The authors analysed a huge volume of publications related to big data requests in tracking and monitoring infectious diseases. They acknowledged key themes such as the usage of machine learning for disease extrapolation, big data's job in improving disease surveillance, and its probable for improving public health interventions. One key discovery is the increasing reliance on real-time information collection and examination in understanding and controlling disease outbreaks. The learning also recommends a research agenda that reassures more interdisciplinary collaboration, the incorporation of diverse data foundations (including social media), and the expansion of robust ethical frameworks for management sensitive health information. The paper suggests that, despite the probable of big data, tasks such as data privacy, the digital divide, and unequal contact to knowledge in low-resource settings must be instructed to fully apprehend its benefits in epidemiology.

S. K. Yadav et al. [21] contributions a comprehensive numerical modelling approach for forecasting the spread of infectious diseases, with an attention on COVID-19. The authors used several statistical methods such as time series examination, compartmental models (SIR, SEIR), and machine learning models to prediction disease broadcasting and evaluate the influence of containment strategies. The learning highlights the critical role of epidemiological limitations, including infection rates, recapture rates, and population movement, in determining the model's correctness. The SEIR (Susceptible-Exposed-Infectious-Recovered) model achieved well in capturing the subtleties of COVID-19 transmission, exceptionally in estimating the effects of community distancing and vaccination. The authors also discovered the tasks of data uncertainty, such as underreporting or overdue reporting of events, which can skew model predictions. The learning emphasizes that while numerical modelling can significantly increase public health planning, the models necessitate constant updating and recalibration as additional data converts available.

Here is a table (Table 1) summarizing the proposed methodology, performance parameters, advantages, and limitations of the research papers:

Table 1: Comparison Table

Paper	**Proposed Methodology**	**Performance Parameters**	**Advantages**	**Limitations**
N. Sharma et al. [1], 2021	Heterogeneous cooperative model combining decision trees, SVM, and neural systems	Prediction correctness, robustness	Improves accurateness by leveraging multiple procedures' strengths	Necessitates significant computational resources
S. Grampurohit et al. [2], 2020	Machine learning algorithms (choice trees, random forests, SVM) for infection prediction	Correctness, precision, recall	Effective for forecasting multiple diseases	Restricted interpretability and reliance on quality information
P. Dutta et al. [3], 2021	AI-based feature variety techniques for COVID-like disease forecasts using logistic reversion, KNN, random forests	Prediction accurateness, feature importance	Reduces model difficulty, focuses on key structures	May overlook multifaceted interactions between structures
F. E. Ayo et al. [4], 2020	Decision support system (DSS) using ANN and conclusion trees for multi-target disease analysis	Diagnostic correctness, response time	Assimilates genomic and epidemiological data for correct diagnosis	Scalability may be imperfect in huge datasets
H. H. Thary et al. [5], 2021	Questionnaire-based agenda integrating machine learning models (decision trees, unplanned forests)	Accurateness of diagnosis based on symptom information	Cost-effective for remote areas, quick conclusion	Dependent on truthful patient-reported data
M. Mariki et al. [6], 2022	Combining scientific symptoms with patient structures using SVM, logistic regression, and unplanned forests	Diagnostic accurateness, recall, precision	Improved predictive correctness by combining symptoms and demographic information	May not streamline well in non-endemic sections
Y. A. Adamu et al. [7], 2021	Machine learning model for malaria expectation based on environmental influences (temperature, humidity, rainfall)	Accuracy, memory, precision	Helps in precautionary measures by forecasting malaria incidences	Requires high-quality conservational data for optimum performance
J. Gao et al. [8], 2020	Time sequence analysis using Grey and SARIMA models for forecasting typhoid and paratyphoid frequencies	Model fitting metrics, forecast correctness	Operative in analysing cumulative incidence information	Restricted to time series data, tests in handling missing data
H. Wang et al. [9], 2021	Big data-based student health data administration model	Prediction accurateness, efficiency in data dispensation	Scalable for huge datasets, supports real-time nursing	Limited submission in non-academic settings
M. Wang et al. [10], 2022	Multi-self-regression deep neural network for speedy warning of infectious diseases	Prediction correctness, response time	Effective for speedy detection, high performance in huge hospital datasets	Computationally expensive, incomplete interpretability
M. K. Singh et al. [11], 2023	Machine learning models (random forests, SVM) for communicable disease prediction and detection	Correctness, recall, precision	Capable of identifying multiple diseases	Limited interpretability, complex to data superiority
D. Swain et al. [12], 2023	Chronic kidney infection classifier using machine learning techniques	Classification correctness, AUC-ROC	Healthy model for chronic disease organization	May not oversimplify well across different populaces
Y. Liu et al. [13], 2023	XGBoost agenda for predicting H9N2 disease in placing hen farms	Model correctness, precision, recall	Highly correct for predicting disease outbreaks in hatchling	Engrossed on poultry industry, limited request to other diseases
H. F. Ahmad et al. [14], 2021	Infectious disease displaying using improved machine learning models for COVID-19 forecasting	Forecast correctness, RMSE	Better-quality model performance for pandemic predicting	Sensitive to sound and data contradictions
M. Hussain et al. [15], 2023	Machine learning for waterborne disease prediction	Prediction correctness, precision, recall	Effectual for predicting positive cases in actual	Imperfect by available training information quality
M. Mwamnyange et al. [16], 2021	Improved MapReduce algorithm for childhood disease investigation in big data atmospheres	Model performance, competence, response time	Scalable for huge datasets, suitable for public health checking	Difficulty in algorithm implementation

Paper	Proposed Methodology	Performance Parameters	Advantages	Limitations
S. Palaniappan et al. [17], 2022	Machine learning algorithms for widespread disease dynamics forecast	Accuracy, precision, memory	Suitable for epidemic threat predictions	Requires high-quality input information for reliable outcomes
G. Dhiman et al. [18], 2022	Federated learning method to protect healthcare information in big data atmospheres	Privacy protection, model performance	Enhances information security and privacy, ascendable	Requires a multifaceted infrastructure for application
M. Uppal et al. [19], 2022	Cloud-based machine learning for actual hospital sensor information monitoring	Fault detection accurateness, response time	Real-time watching, scalable in hospital atmospheres	Vulnerable to cloud-based security threats
L. B. Amusa et al. [20], 2023	Bibliometric examination for big data and infective disease epidemiology research	Citation scrutiny, research trends	Provides a all-inclusive research agenda	Does not contain practical implementation of prototypes
S. K. Yadav et al. [21], 2021	Statistical demonstrating for infectious disease forecast, focusing on COVID-19	Model correctness, RMSE, MAE	Effective for COVID-19 banquet modelling	Limited generalizability separate COVID-19

This table 1 presents an organized view of the methodology, performance metrics, advantages, and limitations of each paper.

1. Research Gaps and Possible Solutions

Here are specific research gaps identified from the research papers (presented in Table 1):

1. **Ensemble Model Optimization**: The heterogeneous ensemble forecasting model presented by N. Sharma et al. [1] necessitates further exploration into optimizing the mixture of algorithms. While the model realizes a high level of correctness, its performance under rehabilitated disease scenarios and optimization approaches for real-time applications remains a opening.
2. **Handling Imbalanced Datasets:** S. Grampurohit et al. [2] attention on disease prediction using machine learning processes but do not methodically address the issue of excessive datasets, particularly for rare diseases. Methods to handle skewed datasets without compromising accurateness require further learning.
3. **Model Generalization Across Diseases**: P. Dutta et al. [3] industrialized a feature-selection-based AI technique for forecasting COVID-like diseases. However, the model's generalization proficiency across additional infectious diseases has not been fully verified, suggesting a gap in generating universally applicable models.
4. **Interpretability in Multi-Target Models:** F. E. Ayo et al. [4] characteristic the importance of multi-target disease conclusion using bioinformatics. A major gap remains in refining the comprehensibility of machine learning models for multi-target calculations to ensure improved clinical adoption.
5. **Data Collection Bias:** H. H. Thary et al. [5] contemporary a framework for diagnosing transferrable diseases using machine learning, however they do not address probable biases in data collection that may pretend the accuracy of calculations. Research into bias mitigation strategies is wanted.
6. **Limited Use of Non-Clinical Data:** M. Mariki et al. [6] attention on combining clinical symptoms and patient structures for malaria analysis. However, non-clinical factors like environmental and social factors, which also donate to disease spread, are generally ignored.
7. **Temporal Dynamics in Predictive Models**: Y. A. Adamu et al. [7] industrialized a malaria prediction model using machine learning, however a gap remains in assimilating temporal dynamics to version for seasonal or long-term tendencies in disease outbreaks.
8. **Hybrid Model Evaluation:** J. Gao et al. [8] used together Grey and SARIMA models for time sequences analysis of typhoid and paratyphoid fevers. However, merging these models with machine learning techniques to recover forecasting correctness has not been discovered extensively.
9. **Scalability in Big Data:** H. Wang et al. [9] advanced a student physical health data management model in a big data atmosphere, but scalability problems, particularly in real-time health watching scenarios, remain underexplored.
10. **Data Security in Cloud-Based Healthcare**: M. Uppal et al. [19] highpoint fault prediction in real-time checking in a hospital environment. However, safeguarding the security and privacy of sensor information in cloud-based systems, particularly in the context of patient health documents, is unmoving a significant gap.

Here are potential explanations (solutions) for the acknowledged research gaps in the research papers (presented in Table 1):

1. Ensemble Model Optimization:

Appliance adaptive ensemble learning techniques related stacking, blending, or dynamic collection, which dynamically adjust the amalgamation of models based on the input information characteristics. Meta-learning can likewise be used to automatically handpicked the best combination of procedures for different disease scenarios, informative real-time performance.

2. Handling Imbalanced Datasets:

Use information augmentation techniques, to stability imbalanced datasets. Additionally, cost-sensitive knowledge and ensemble methods with prejudiced classes can be employed to handle the unevenness.

3. Model Generalization Across Diseases:

Progress a transfer learning agenda where models trained on one disease container be fine-tuned on extra diseases with limited documents. Using multi-task learning can also recover generalization by sharing structures across multiple related diseases. This wanted allow models to adapt improved to new infectious diseases.

4. Interpretability in Multi-Target Models:

Participate explainable AI (XAI) methods to progress the interpretability of multi-target models. These methods can transport clear insights into how the model spreads its forecasts, increasing its clinical acceptance.

5. Data Collection Bias:

Appliance bias detection algorithms and frequently audit datasets to recognize and mitigate biases. Collect information from diverse sources and safeguard the inclusion of demographic variables to decrease bias. Additionally, expending fairness-aware machine learning algorithms can help improve the effects of bias in guesses.

6. Limited Use of Non-Clinical Data:

Integrate environmental, social, and demographic documents into predictive models. This could comprise using external datasets alongside scientific features to advance the overall accuracy and real-world applicability of expectations. Feature engineering after these additional data foundations can significantly enhance model performance.

7. Temporal Dynamics in Predictive Models:

Integrate time-series analysis techniques or hybrid models merging statistical techniques with machine learning techniques. These models can imprisonment temporal dependences and seasonal trends, making them appropriate for predicting the periodicity of transferrable disease outbreaks.

8. Hybrid Model Evaluation:

Improve a hybrid forecasting framework that syndicates machine learning with arithmetical models. Implement model calculation metrics and prediction intervals to measure the effectiveness of the hybrid method and fine-tune the model grounded on real-world performance.

9. Scalability in Big Data:

Exploit distributed computing frameworks such for example Apache Spark or Hadoop for ascendable processing of huge health data in real-time atmospheres. Implement cloud-native architectures with containerization to guarantee that the system can measure automatically based on the load, assembly real-time health inspecting more feasible.

10. Data Security in Cloud-Based Healthcare:

Employment privacy-preserving machine learning techniques like discrepancy privacy, homomorphic encryption and amalgamated learning to secure patient documents in cloud-based systems. This would safeguard that information remains secure while being administered in real-time across distributed systems, addressing secrecy concerns effectively.

By covering these solutions, the acknowledged research gaps can be effectively addressed, important to advancements in disease forecast models and improved health outcomes.

III. Conclusion

Transferrable disease prediction remains a critical zone of research, particularly as global health systems remain to face the risk of large-scale outbursts. This literature review has inspected the use of hybrid machine learning methods in big data atmospheres, demonstrating their probable to enhance the accuracy, scalability, and effectiveness of predictive models. By assimilating multiple machine learning algorithms—such for example supervised and unsupervised knowledge, deep learning frameworks, and collaborative techniques—researchers have been talented to address complex tasks associated with high-dimensional data and real-time allowance. Despite the advances accentuated, significant gaps remain, principally in terms of interpretability, computational effectiveness, and the ability to simplify across diverse datasets. As hybrid ML methods continue to progress, there is a growing necessity for further research that efforts on overcoming these tasks, improving the robustness of models, and attractive their practical applicability cutting-edge real-world settings. This analysis serves as a foundation for upcoming work, emphasizing the standing of interdisciplinary collaboration in developing operative tools for infectious disease forecast in the age of big data.

References

[1]. N. Sharma, J. Dev, M. Mangla, V. M. Wadhwa, S. N. Mohanty and D. Kakkar, "A Heterogeneous Ensemble Forecasting Model for Disease Predictions", *New Gen Computing*, Vol. 39 (3), pp. 701-715, 2021.

[2]. S. Grampurohit and C. Sagarnal, "Disease Predictions using Machine Learning Algorithm", *International Conferences for Emerging Technology (INCET)*, Vol. 4 (3), pp. 1-7, 2020.

[3]. P. Dutta, S. Paul, A. J. Obaid, S. Pal and K. Mukhopadhyay, "Feature Selections based Artificial Intelligence Technique for the Predictions of COVID like Diseases", *2nd International Conference on Physics and Applied Sciences (ICPAS), Journal of Physics: Conference Series*, Vol. 1963 (1), pp. 1-11, 2021.

[4]. F. E. Ayo, J.B. Awotunde, R.O. Ogundokun, S.O. Folorunso and A.O. Adekunle, "A Decision Support System for Multi-Target Disease Diagnosis: A Bioinformatics Approach", *Heliyon, cell press*, Vol. 6, pp. 1-14, 2020.

[5]. H. H. Thary and K. Azidan, "A Framework Questionnaire for Diagnosing Infectious Disease using Machine Learning Techniques", INTCSET 2020, *IOP conference series: Materials science and Engineering, IOP Publishing*, Vol. 1094, pp. 1-10, 2021.

[6]. M. Mariki, E. Mkoba and N. Mduma, "Combining Clinical Symptoms and Patient Features for Malaria Diagnosis: Machine Learning Approach", *Applied Artificial Intelligence, Taylor & Francis,* Vol. 36 (1), pp. 1-25, 2022.

[7]. Y. A. Adamu and J. Singh, "Malaria Prediction Model using Machine Learning Algorithms", *Turkish journal of computer and mathematics Education*, Vol. 12 (10), pp. 7488-7496, 2021.

[8]. J. Gao, J. Li and M. Wang, "Time Series Analysis of Cumulative Incidences of Typhoid and Paratyphoid Fevers in China Using both Grey and SARIMA Models", *PLOS ONE,* Vol. 15 (10), pp. 1-14, 2020.

[9]. H. Wang, N. Wang, M. Li, S. Mi and Y. Shi, "Student Physical Health Information Management Model under Big Data Environment", Hindawi Scientific Programming, Vol. 2021, pp. 1-10, 2021.

[10]. M. Wang, C. Lee, W> Wang, Y. Yang and C. Yang, "Early Warning of Infectious Diseases in Hospitals based on Multi-Self-Regression Deep Neural Network", Hindawi Journal of Healthcare Engineering, Vol. 2022, pp. 1-13, 2022.

[11]. M. K. Singh, K. P. Singh and D. Kumar, "Prediction and Detection of Infectious Disease through Machine Learning", European Chemical Bulletin, Vol. 12 (1), pp. 4433-4446, 2023.

[12]. D. Swain, U. Mehta, A. Bhatt, H. Patel, K. Patel, D. Mehta, B. Acharya, V. C. Gerogiannis, A. Kanavos and S. Manika, "A Robust Chronic Kidney Disease Classifier using Machine Learning", Electronics, MDPI, Vol. 12 (212), pp. 1-13, 2023.

[13]. Y. Liu, Y. Zhuang, L. Yu, Q. Li, C. Zhao, R. Meng, J. Zhu and X. Guo, "A Machine Learning Framework based on Extreme Gradient Boosting to predict the Occurrence and Development of Infectious Diseases in Laying Hen Farms, Taking H9N2 as an Example", Animals, MDPI, Vol. 13 (1494), pp.1-15, 2023.

[14]. H. F. Ahmad, H. Khaloofi, Z. Azhar, A. Algosaibi and J. Hussain, "An Improved COVID-19 Forecasting by Infectious Disease Modelling using Machine Learning", Applied Sciences, MDPI, Vol. 11 (11426), pp. 1-38, 2021.

[15]. M. Hussain, M. A. Cifci, T. Sehar, S. Nabi, O. Cheikhrouhou, H. Maqsood, M. Ibrahim an F. Mohammad, "Machine Learning based efficient Prediction of Positive Cases of Waterborne Disease", BMC Medical Informatics and Decision Making, Vol. 23 (11), pp. 1-16, 2023.

[16]. M. Mwamnyange, E. Luhanga and S. R. Thodge, "Big Data Analytics Framework for Childhood Infectious Disease Surveillance and Response System using Modified MapReduce Algorithm", International Journal of Advanced Computer Science and Applications (IJACSA), Vol. 12 (3), pp. 1-13, 2021.

[17]. S. Palaniappan, R. V, B. David and P. N. S, "Prediction of Epidemic Disease Dynamics on The Infectious Risk using Machine Learning Algorithms", SN Computer Science, Vol. 3 (47), pp. 1-3, 2022.

[18]. G. Dhiman, S. Juneja, H. Mahafez, I. E. Bayoumy, L. K. Sharma, M. Hadizadeh, M. A. Islam, W. Viriyasitavat and U. Khandaker, "Federated Learning Approach to Protect Healthcare Data over Big Data Scenario", Sustainability, MDPI,Vol. 14 (2500), pp. 1-14, 2022.

[19]. M. Uppal, D. Gupta, S. Juneja, A. Sulaiman, K. Rajab, A. Rajeb, M. A. Elmagzoub and A. Shaikh, "Cloud based Fault Prediction for Real Time Monitoring of Sensor Data in Hospital Environment using Machine Learning", Sustainability, MDPI, Vol. 14 (11667), pp. 1-19, 2022.

[20]. L. B. Amusa, H. Twinomurinzi, E. Phalane and R. N. P. Mafuya, "Big Data and Infectious Disease Epidemiology: Bibliometric Analysis and Research Agenda", Interactive Journal of Medical Research, Vol. 12, pp. 1-16, 2023.

[21]. S. K. Yadav and Y. Akhter, "Statistical Modelling for the Prediction of Infectious Disease Dissemination with Special Reference to COVID-19 Spread", Frontiers in Public Health, Vol. 9, pp. 1-27, 2021.