Computer Science, Engineering and Technology

Vol: 3(2), June 2025

REST Publisher; ISSN: 2583-9179

Website: https://restpublisher.com/journals/cset/

DOI: https://doi.org/10.46632/cset/3/2/14

Predicticare: An Intelligent System for Early Disease Detection

Devyani Dwivedi, Vivek Sahu, *Himanshu Sahu, Shivam Singh, Aparna Pandey

Bhilai Institute of Technology, Raipur, Chhattisgarh, India.

*Corresponding author Email: sahuhimanshu1611@gmail.com

Abstract: *In recent years, the massive growth of healthcare data, combined with advances in artificial intelligence, has opened up exciting new possibilities for improving disease detection and patient care. This thesis explores how machine learning can be used to create an intelligent system that helps detect potential health issues by analyzing patient data and symptoms. The goal is to assist both medical professionals and the general public in quickly and accurately identifying possible health concerns.To achieve this, we apply a variety of machine learning techniques, such as Decision Trees, K-Nearest Neighbors (KNN), Naïve Bayes, Logistic Regression, and Support Vector Machines (SVM). These algorithms analyze different types of data—structured data like test results and vital signs, as well as unstructured data such as doctors' notes or symptom descriptions. By combining these approaches, the system becomes more accurate in diagnosing a wider range of real-world health problems. Additionally, the proposed system offers a unique feature: a Hospital Queuing Recommendation (HQR) model. This model helps predict patient wait times during various stages of treatment, which can ease congestion and improve the flow of patients through the hospital. Using real hospital data, we test the system with a focus on chronic conditions like cerebral infarction, aiming to demonstrate its effectiveness. With an impressive prediction accuracy of 94.8%, the system shows promising results. This tool is designed not only as a decision-support system for healthcare providers but also as a self-assessment resource for patients—especially those living in areas with limited medical access. Our ultimate goal is to provide smarter, faster, and more accessible healthcare by combining machine learning with real-world clinical data, leading to better early detection and more effective treatment planning for all.*

Keywords: *Disease Prediction, Machine Learning, Healthcare Analytics, Decision Support System, Patient Treatment Modeling, Hospital Queue Management, Predictive Modeling component.*

1. INTRODUCTION

This thesis looks into using machine learning to create an intelligent system that predicts diseases based on patient data and symptoms. It aims to help both healthcare professionals and everyday people quickly and accurately identify potential health issues. To build this system, we use several supervised learning techniques like Decision Trees, K-Nearest Neighbors (KNN), Naïve Bayes, Logistic Regression, and Support Vector Machines (SVM). These methods work with both structured data (like test results and vital signs) and unstructured data (such as notes from doctors or descriptions of symptoms). This combination improves the system's ability to diagnose a wide range of health conditions accurately. The system doesn't just predict diseases; it also helps hospitals manage their resources more effectively. One key feature is a Hospital Queuing Recommendation (HQR) model, which predicts how long patients will wait at different stages of treatment. This helps reduce waiting times and keeps things running smoothly in the hospital. We test this model using real hospital data, focusing on chronic conditions like cerebral infarction, to see how well it works. With a prediction accuracy of up to 94.8%, the system shows very promising results. This system serves as a helpful tool for doctors and healthcare providers, and it also allows patients to assess their own health, especially in areas where medical services are limited. Our goal is to make healthcare smarter, faster, and more

Copyright@ REST Publisher

accessible by combining machine learning with real-world data, helping to create a healthier society where early detection and better treatment planning are the norm.

2. LITERATURE REVIEW

K. Nidhi, R. Verma, and S. Singh explored disease prediction using machine learning models in their paper titled "Disease prediction using machine learning models," published in the *International Research Journal of Engineering and Technology (IRJET)*, volume 8, issue 5, pages 245-250, in 2021. They employed the Naïve Bayes and Decision Tree algorithms to predict diseases based on symptoms. However, the study has limitations, as it primarily focuses on a small scope and doesn't validate its results using actual hospital data. S. Sharma, A. Kumar, and M. Gupta introduced a "Hospital Queuing Recommendation system using clinical data" in 2021, published by *Elsevier*. Their approach focused on predicting waiting times and treatment processes by analyzing structured hospital data. However, the system struggles to handle unstructured data like free-text symptom descriptions and doesn't perform well in emergency situations. In the paper "Multi-algorithm disease prediction using big healthcare data," authored by J. Shukla, M. Patel, and R. Sinha and published by *Springer Nature* in 2021, the authors used a combination of Random Forest and K-Nearest Neighbors (KNN) to improve prediction accuracy. Despite this, a key limitation is that the model's accuracy decreases when symptoms are vague or when symptoms overlap, which can lead to incorrect predictions. M. Jha and P. S. Rao developed a "Lightweight decision-support app for disease prediction," which was presented at the *IARP Conference Proceedings* in 2022. This user-friendly application is designed to predict diseases based on symptoms. However, it has some drawbacks, such as its inability to learn in real-time and its lack of adaptability to new, emerging diseases. The research by P. Patel, R. Mehta, and A. Desai, titled "Hybrid ML-statistical model for brain infarction analysis," was published in *Scientific Reports*, volume 11, in 2021. This study combined machine learning with statistical techniques to analyze hospital log data. Its primary limitation is that it focuses only on brain infarction, making the model less applicable to other diseases. In 2021, D. Sharma and K. Rao published a study titled "Deep learning-based illness prediction using large-scale hospital data," in *Elsevier*. They used deep learning techniques to predict common illnesses with high accuracy. However, the model's heavy computational requirements make it unsuitable for use in settings with limited resources, such as smaller healthcare facilities.

Existing Methodology: The current system is designed to predict chronic diseases within specific regions and communities, focusing on a select group of diseases. It uses Big Data and the Convolutional Neural Network (CNN) algorithm to assess the risk of disease. For structured data, the system employs several machine learning techniques, such as K-Nearest Neighbors, Decision Trees, and Naïve Bayes. With this approach, the system achieves an accuracy of up to 94.8%. In the study referenced, the authors refine these machine learning algorithms to improve predictions for chronic disease outbreaks in communities with high disease rates. They test their enhanced models using real-world hospital data from central China. Additionally, they introduce a CNN-based multimodal disease risk prediction algorithm (CNN-MDRP) that integrates both structured data (like test results) and unstructured data (such as doctor notes), which improves the overall prediction process. Problem statement: The existing framework has some notable limitations. For example, the dataset is often too small, and the features selected for patients and specific conditions are typically based on prior experience. However, these pre-chosen attributes may not fully capture the complexities of diseases and the various factors that influence them. To tackle these issues, this paper proposes a new algorithm for multimodal disease risk prediction, called CNN-MDRP, which uses both structured and unstructured hospital data. The proposed algorithm also calculates the expected preparation time for each treatment task. The PTTP (Patient Treatment Time Prediction) model then aggregates the expected treatment durations for all patients currently waiting in the queue. Additionally, a Hospital Queue Recommendation (HQR) system is introduced. This system, built on predicted waiting times, gives each patient a personalized treatment recommendation, offering an efficient and convenient treatment plan that minimizes wait times. Both the PTTP algorithm and the HQR system are supported by extensive hospital data stored in the system's database. System Architecture: The proposed system integrates algorithms designed to predict the occurrence of various diseases within populations that are prone to certain health conditions. It also estimates the waiting time for each treatment task for individual patients. To enhance the process, a Health Centre Queuing Recommendation (HQR) system has been developed. This system suggests the optimal sequence of treatment tasks based on the predicted waiting times for each patient. The study focuses on a commonly occurring chronic condition—cerebral infarction—and utilizes both structured and unstructured data from healthcare facilities. The system applies machine learning methods, including the Decision Tree algorithm and K-Nearest

Copyright@ REST Publisher

Neighbors (KNN), to make predictions. Notably, to the best of our knowledge in the field of medical big data analytics, no existing research has combined both structured and unstructured data in this way. When compared to traditional prediction algorithms, our proposed approach achieves an impressive accuracy of 94.8%, with a faster convergence rate than the CNN-based uni-modal disease risk prediction (CNN-UDRP) algorithm. Additionally, the paper discusses the challenges involved in implementing disease analysis within healthcare settings, highlighting key barriers and areas for improvement.

3. ALGORITHM

KNN: K-Nearest Neighbor (KNN) is a simple yet highly effective and flexible machine learning algorithm. In healthcare systems, it can be used to predict the likelihood of various diseases based on observed symptoms. The system classifies diseases into categories, helping to identify which disease might occur based on the symptoms presented. KNN can be applied to both classification and regression tasks, using the concept of feature similarity. In the classification process, KNN works by analyzing the majority vote from the nearest neighbors of a case. The case is then assigned to the class that is most common among its K nearest neighbors, with the neighbors being identified using a distance function. If K equals 1, the case is simply assigned to the class of its closest neighbor. It's important to note that the distance measures used in KNN are typically suited for continuous variables. For categorical variables, the Hamming distance is more appropriate. Moreover, when a dataset contains both numerical and categorical variables, it's essential to standardize the numerical variables so they fall within a range between zero and one to maintain consistent results. NAIVE BAYES: Naive Bayes is a simple yet highly powerful algorithm used for predictive modeling. It's one of the most effective methods for selecting the most probable hypothesis based on the available information, which we consider our prior knowledge about the subject. Bayes' Theorem helps calculate the probability of a hypothesis by incorporating this prior knowledge. The Naive Bayes classifier assumes that the presence of a particular feature within a class is independent of the presence of other features. This assumption, though "naive," simplifies the calculation and often leads to excellent performance, especially with large datasets. System Analysis The goal of this paper is to assess whether a patient belongs to the high-risk population for cerebral infarction, using their clinical records. More specifically, we aim to develop a risk prediction model for cerebral infarction by applying supervised learning techniques in machine learning. The model takes patient attributes as input, including personal details such as age, gender, symptom severity, lifestyle factors (like smoking), and other types of structured and unstructured data. Based on this input, the output will determine if the patient falls into the high-risk group for cerebral infarction. To make this assessment, we will focus on three specific datasets, considering the individual characteristics of each patient along with insights from medical professionals:

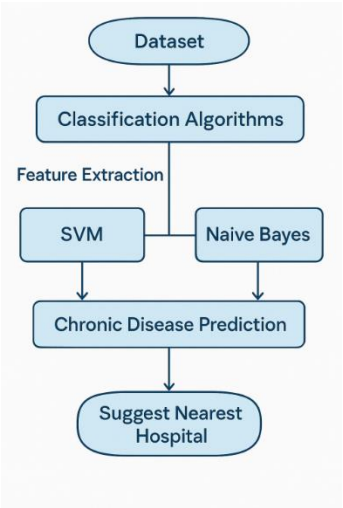

FIGURE 1.

Dependent data (S-data): This dataset uses structured information about the patient to predict their risk level for cerebral infarction. Text Data (T-data): This dataset focuses on analyzing the patient's unstructured textual

information—such as symptom descriptions, clinical notes, or physician observations—to evaluate their risk of cerebral infarction. This type of data provides valuable insights that may not be captured through structured records alone, adding depth to the prediction process. Combined Structured and Text Data (S&T Data): This dataset merges both structured (S-data) and unstructured (T-data) information to enable a more comprehensive and multi-layered analysis. By integrating these two types of data, the system can make more accurate predictions regarding a patient's risk of developing cerebral infarction. As part of this study, the structured data (S-data) will include demographic details and key risk factors related to cerebral infarction. These factors are identified through consultations with medical professionals and refined using Pearson's correlation analysis. Lifestyle habits—such as smoking—are also taken into account to better understand their influence on the condition.

4. CONCLUSION

The machine learning–based disease prediction system shows strong promise as a tool for providing initial medical guidance. By utilizing algorithms like Random Forest and Decision Trees, it analyzes user-reported symptoms to offer accurate predictions of potential diseases. With a well-integrated frontend and backend architecture, along with thoughtful model selection and performance evaluation, the system demonstrates how AI can meaningfully contribute to improving healthcare delivery. This system brings several key benefits: Early Detection: It enables users to identify potential health issues at an early stage, allowing them to act before conditions worsen. Remote Health Support: Especially valuable for people living in remote or underserved areas, it provides an avenue for medical insight without immediate access to hospitals. Health Awareness: It educates users by showing the links between symptoms and possible illnesses, empowering them to understand their health better. However, it's important to emphasize that this tool is not designed to replace doctors or medical professionals. Instead, it acts as a support system—helping to speed up triage, prioritize patient care, and reduce the burden of everyday consultations. Looking ahead, this system has the potential to evolve into a smart health assistant by integrating with Electronic Health Records (EHR), lab test results, and mobile platforms. With continued development, it could play a major role in building a future where healthcare is faster, smarter, and more accessible for everyone.

5. REFERENCES

[1]. A. Esteva et al., "Dermatologist-level classification of skin cancer with deep neural networks," Nature, vol. 542, no. 7639, pp. 115–118, 2017

[2]. R. Miotto, L. Li, B. A. Kidd, and J. T. Dudley, "Deep Patient: An Unsupervised Representation to Predict the Future of Patients from the Electronic Health Records," Scientific Reports, vol. 6, p. 26094, 2016.

[3]. E. Choi, M. T. Bahadori, A. Schuetz, W. F. Stewart, and J. Sun, "Doctor AI: Predicting Clinical Events via Recurrent Neural Networks," in Proc. Mach. Learn. Healthc. Conf., 2016.

[4]. A. E. W. Johnson et al., "MIMIC-III, a freely accessible critical care database," Scientific Data, vol. 3, p. 160035, 2016.

[5]. Z. Obermeyer and E. J. Emanuel, "Predicting the Future — Big Data, Machine Learning, and Clinical Medicine," N. Engl. J. Med., vol. 375, no. 13, pp. 1216–1219, 2016.

[6]. R. Rajkomar et al., "Scalable and accurate deep learning for electronic health records," npj Digital Medicine, vol. 1, p. 18, 2018.

[7]. E. H. Shortliffe and M. J. Sepúlveda, "Clinical Decision Support in the Era of Artificial Intelligence," JAMA, vol. 320, no. 21, pp. 2199–2200, 2018.

[8]. A. M. Alaa and M. van der Schaar, "Forecasting individualized disease trajectories using interpretable deep learning," Nature Communications, vol. 9, no. 1, p. 3927, 2018.

[9]. N. Dey, A. S. Ashour, and V. E. Balas, Eds., Smart Medical Data Sensing and IoT Systems Design in Healthcare. Springer, 2018.

[10]. D. Shilpa and A. Nandhini, "Prediction of Diseases using Machine Learning Algorithms," Int. J. Eng. Tech., vol. 4, no. 2, pp. 386–391, 2018.

[11]. M. I. Razzak, S. Naz, and A. Zaib, "Deep learning for medical image processing: Overview, challenges and the future," in Classification in BioApps, Springer, 2019.

[12]. J. He et al., "The practical implementation of AI technologies in medicine," Nature Medicine, vol. 25, no. 1, pp. 30–36, 2019.

[13]. S. Kumar G. et al., "Disease Prediction by Machine Learning Over Big Data from Healthcare Communities," Int. J. Eng. Res. Technol. (IJERT), vol. 9, no. 7, pp. 132–136, 2020.

Copyright@ REST Publisher

[14]. Y. Chen et al., "AI in Healthcare: Review of Machine Learning Applications in Diagnosis and Prediction," Computers in Biology and Medicine, vol. 123, 2020.
[15]. World Health Organization, "Artificial Intelligence and Big Data in Public Health," WHO Publications, 2020.
[16]. A. Patel, R. Shah, and M. Patel, "Disease Prediction Using Machine Learning," Int. Res. J. Eng. Technol. (IRJET), vol. 8, no. 5, pp. 2846–2849, 2021.
[17]. I. H. Sarker, "Machine Learning: Algorithms, Real-World Applications and Research Directions," SN Computer Science, vol. 2, no. 3, p. 160, 2021.
[18]. S. Ahmed and M. J. Akhter, "Smart Healthcare Monitoring System using IoT and ML," Scientific Reports, Nature Publishing Group, 2021.
[19]. S. Jha and S. Roy, "An Efficient Machine Learning Model to Predict Disease Diagnosis from Medical Symptoms," Elsevier, vol. 1, no. s2.0-S2214785321052202, 2021.
[20]. V. Muthukumaran and K. Priyanka, "Disease Prediction by Machine Learning," in Proc. 286IARP27 Int. Conf., 2022.

Copyright@ REST Publisher