AI-Driven Symptom Checker: Leveraging Big Data Technologies for Real-Time

Healthcare Triage and Diagnostics

Augnik Banerjee

*Department of Networking and Communications*

*School of Computing College of Engineering and Technology*

*SRM Institute of Science and Technology* Kattankulathur-603203, India [ab0710@srmist.edu.in](mailto:ab0710@srmist.edu.in)

Piyush Baibhav

*Department of Networking and Communications*

*School of Computing*

*College of Engineering and Technology SRM Institute of Science and Technology* Kattankulathur-603203, India [pb5957@srmist.edu.in](mailto:pb5957@srmist.edu.in)

Nitin Kumar

*Department of Networking and Communications*

*School of Computing*

*College of Engineering and Technology SRM Institute of Science and Technology* Kattankulathur-603203, India [nk0865@srmist.edu.in](mailto:nk0865@srmist.edu.in)

Aniket Singh

*Department of Networking and Communications School of Computing*

*College of Engineering and Technology SRM Institute of Science and Technology* Kattankulathur-603203, India [as0872@srmist.edu.in](mailto:as0872@srmist.edu.in)

Gouthaman. P\*

*Department of Networking and Communications School of Computing*

*Faculty of Engineering and Technology SRM Institute of Science and Technology* Kattankulathur-603203, India [gouthamp@srmist.edu.in](mailto:gouthamp@srmist.edu.in)

***Abstract—The healthcare industry continues to face significant challenges in delivering timely and accurate diagnoses, particularly in remote and underserved areas where access to medical professionals is limited. The work aims to develop an AI-driven symptom checker that leverages big data technologies to offer preliminary diagnoses and support patient triage, ensuring quicker medical response times. Utilizing Hadoop for scalable data storage, Hive for advanced data analysis, and Spark for real-time distributed data processing, the system is designed to efficiently handle large volumes of patient data. Tableau is integrated for intuitive data visualization, allowing healthcare professionals to gain actionable insights and make informed decisions promptly. The proposed architecture enables the swift analysis of patient symptoms in real-time, significantly reducing diagnostic delays and enhancing the quality of care. By focusing on scalable and efficient data processing, the work addresses the need for a robust healthcare solution capable of operating in environments with limited resources. This system has the potential to transform patient care in remote areas by providing real-time analysis, identifying critical cases, and aiding in effective resource allocation. The solution’s emphasis on rapid diagnosis and visualization offers a comprehensive approach to improving healthcare outcomes in underserved regions.***

***Keywords: AI-Driven Symptom Checker, Apache Spark, Hadoop, Hive, Healthcare Triage, Tableau***

1. INTRODUCTION

The healthcare industry is undergoing a significant transformation driven by advancements in technology and the increasing adoption of data-driven solutions. Despite these developments, there remains a substantial gap in healthcare delivery, particularly in remote and underserved areas where timely diagnosis and access to medical

professionals are often limited [1]. The lack of immediate medical attention in these regions leads to delayed treatment, exacerbating health conditions and increasing the risk of complications. This problem is further compounded by the growing incidence of chronic diseases and the burden of managing large patient populations with limited resources.

To address these challenges, healthcare providers and researchers are turning to artificial intelligence (AI) and big data technologies as potential solutions [2]. AI has the ability to analyze vast amounts of patient data to provide early diagnosis and personalized treatment recommendations. However, the real challenge lies in processing and analyzing this data efficiently in real-time, especially when the data is generated continuously from diverse sources such as electronic health records (EHRs), wearable devices, and remote monitoring systems.

The impact of delayed diagnosis in remote areas is not merely a medical issue but a socio-economic challenge as well [3]. Studies have shown that timely medical intervention can significantly reduce healthcare costs, improve patient outcomes, and decrease the overall burden on healthcare systems. By enabling faster diagnosis and treatment, the work aims to reduce morbidity rates and enhance the quality of care provided to patients in underserved areas.

1. RELATED WORK

The integration of AI and big data in healthcare has shown promising potential to improve diagnostics, treatment, and patient management, especially in resource-limited settings. Over the past decade, researchers have increasingly focused on using machine learning to analyze medical data and predict disease patterns. Rajkumar et al., for example, demonstrated the efficacy of deep learning in processing

electronic health records (EHRs) to improve diagnostic accuracy [3], potentially reducing errors and enhancing patient care.

A key challenge in implementing AI-driven healthcare solutions lies in managing the massive datasets generated by EHRs, health monitoring devices, and laboratory reports. Johnson et al. addressed this by using the MIMIC-III clinical database to train machine learning models that predict patient outcomes. This approach highlights the importance of scalable, flexible data handling to enable effective, AI-based healthcare applications.

The work extends previous work by combining Hadoop and Hive technologies to efficiently store and query large-scale medical data, enabling real-time, data-driven insights into patient symptoms. By employing machine learning models for real-time analysis, the system can prioritize critical cases, aiding in medical triage and resource allocation, which is particularly useful in remote healthcare settings with limited resources.

Data visualization also plays a crucial role in this approach. Tools like Tableau allow clinicians to interpret complex data easily, supporting faster, data-informed decision-making. As noted, effective data visualization can improve outbreak monitoring and resource management, helping address the urgent need for accessible healthcare insights, particularly in underserved areas.

1. METHODOLOGY
2. *Dataset*

To build a robust AI-driven symptom checker, A combination of publicly available healthcare datasets are utilized by us, including patient symptoms, historical medical data, demographic details, and disease-related records [5]. The primary data sources include:

MIMIC-III Clinical Database: This database contains a vast collection of clinical data from intensive care unit (ICU) patients, including physiological measurements, medications, and diagnostic codes. The inclusion of this data enables the AI-driven system to capture complex health patterns, improving diagnostic accuracy and the predictive capabilities necessary for symptom analysis and patient triage in intensive care settings.

Kaggle Healthcare Datasets: Datasets incorporated from Kaggle that include various health metrics, disease symptoms, and diagnostic outcomes relevant to The target patient population. By incorporating data from diverse demographic groups, the AI models are enabled to learn and make more accurate predictions, regardless of patient background or the healthcare environment.

CDC Health Data: Additional demographic and health data from the Centers for Disease Control and Prevention (CDC) will be used to contextualize patient information and understand broader trends in healthcare [6]. The inclusion of CDC data supports the system’s ability to make

accurate, context-aware predictions that account for both individual and population-level health risks.

AI techniques enable efficient processing and real-time analysis of these large datasets. By employing machine learning models trained on these data sources, the system can identify symptom patterns, predict potential diagnosis, and prioritize cases needing urgent attention.

1. *Preprocessing*

Data preprocessing is a critical step to ensure that the raw data is clean, consistent, and suitable for machine learning model training. This phase involves several key operations:

Data Cleaning: Missing values have been handled, duplicates removed, and inconsistencies corrected in the data using Hive queries. Hive's SQL-like querying capabilities allow us to efficiently transform the data and prepare it for analysis. Techniques such as One-Hot Encoding will be used to convert categorical data into a machine-readable format, while feature scaling methods like normalization and standardization will be applied to ensure that all numerical data is within a comparable range.

Data Balancing: To address class imbalance in the symptom datasets, the Synthetic Minority Over-sampling Technique (SMOTE) to generate synthetic examples of underrepresented classes are used. Additionally, under- sampling methods will be applied to reduce the dominance of the majority class, ensuring a balanced dataset that improves model performance during training.

1. *Machine Learning Model Development*

The work will be developed using a variety of machine learning models, each optimized for different aspects of symptom analysis and prediction:

Model Selection: Multiple machine learning algorithms will be evaluated by us, including

Logistic Regression: Used as a baseline model due to its simplicity and effectiveness in binary classification tasks.

Random Forest: An ensemble learning method known for its robustness and accuracy in handling high-dimensional datasets.

XGBoost: This gradient boosting algorithm is selected for its efficiency in handling large datasets and ability to manage imbalanced data distributions.

Neural Networks: Implemented for their capability to learn complex patterns in patient data, providing more nuanced predictions.

Training and Validation: The models will be trained using Apache Spark's MLlib, which facilitates scalable machine learning on big datasets. Spark's distributed computing environment will enable parallel model training, significantly reducing the time required to process and analyze the data. Cross-validation techniques will be employed to ensure that the models generalize well to unseen data, enhancing their predictive accuracy.

Performance Metrics: The models' performance will be

evaluated using metrics such as accuracy, precision, recall, F1-score, and Area Under the Curve (AUC) of the Receiver Operating Characteristic (ROC) curve. These metrics will provide a comprehensive view of each model's ability to correctly classify patient symptoms and predict potential diagnoses.

1. *Real-Time Data Processing with Apache Spark*

Apache Spark's real-time data processing capabilities are central to the work, enabling the symptom checker to provide instant analysis and diagnosis recommendations:

Data Ingestion: Real-time data streams from various sources, such as patient monitoring devices and EHR systems, will be integrated into the Spark environment. Apache Kafka will be used to handle data streaming, ensuring that incoming data is continuously processed and analyzed.

In-Memory Computing: Spark's in-memory computation will drastically reduce latency, allowing for rapid execution of machine learning algorithms on the streaming data. This setup is critical for applications in remote healthcare, where timely decision-making is essential to patient outcomes.

Symptom Analysis Pipeline: The data processing pipeline in Spark will involve transforming raw patient data into meaningful features that can be fed into the machine learning models. The pipeline will include steps such as feature extraction, data normalization, and real-time anomaly detection to identify unusual patient conditions promptly.

1. *Visualization with Tableau*

Data visualization is essential to convert the complex analytics of symptom data into a user-friendly format for healthcare professionals. Tableau will be used to create interactive dashboards that provide insights into patient health trends and triage priorities:

Dashboard Design: The Tableau dashboard will display critical health metrics, including symptom trends, patient demographics, geographical distribution of disease outbreaks, and high-risk areas [7]. Real-time data visualizations will be updated continuously as new data is ingested from the Spark processing engine.

Visualization Techniques: A variety of visualization techniques are used such as heat maps to highlight areas with high disease prevalence, time-series plots to track symptom evolution, and scatter plots to visualize the correlation between patient demographics and disease susceptibility.

Actionable Insights: The visualizations will help healthcare providers quickly identify patterns, make data- driven decisions, and prioritize patient care based on severity. Tableau's ability to present real-time data in an easily interpretable format is crucial for enhancing the decision-making process in resource-limited healthcare settings.

1. *Integration and System Architecture*

The integration of these technologies will be done using a layered architecture that ensures seamless communication between data storage, processing, and visualization components:

Data Layer: Hadoop HDFS will serve as the foundational layer for storing all patient data, providing high scalability and reliability.

Processing Layer: Apache Spark will operate on top of Hadoop to perform real-time data processing and execute machine learning models. Hive will be utilized for batch processing and advanced querying.

Visualization Layer: Tableau will interface with the Spark engine to visualize the processed data in real-time, delivering intuitive dashboards for healthcare professionals.

1. *Experimental Setup*

The experimental setup will involve rigorous testing of the AI-driven symptom checker to ensure its effectiveness in real-world scenarios:

Data Partitioning: The dataset will be divided into training, validation, and test sets in an 80-10-10 ratio to optimize model training and evaluation.

Hyper-parameter Tuning: Perform grid search and random search techniques to fine-tune the hyper-parameters of the machine learning models, enhancing their predictive accuracy and efficiency.

Comparison of Models: The performance of each model will be benchmarked against each other based on speed, accuracy, and scalability. Special attention will be given to models like XGBoost and Neural Networks that are known for their robustness in handling large datasets [8].

1. EXPERIMENTAL SETUP
2. *Machine Learning Models*

Experiment with multiple machine learning algorithms are done, including Logistic Regression, Random Forest, XGBoost, and Neural Networks. The models will be trained and validated using cross-validation techniques to ensure high accuracy and reliability. Apache Spark will be used to parallelize the training process, significantly reducing computation time and improving model efficiency.

1. *Performance Evaluation*

The models' performance will be evaluated using metrics such as accuracy, precision, recall, F1-score, and Area Under the ROC Curve (AUC). The standalone training time with Spark's distributed training time to demonstrate the impact of big data technologies on computational efficiency is compared by us. Detailed analysis will include confusion matrices and classification reports to understand the models' effectiveness in handling real-time

symptom data [6].

1. *Handling Class Imbalance*

The issue of class imbalance in symptom data will be addressed using techniques like Synthetic Minority Over- sampling Technique (SMOTE) and under-sampling to create a balanced dataset.

1. RESULT ANALYSIS

The implementation of the AI-driven symptom checker, supported by big data technologies like Hadoop, Apache Spark, Hive, and Tableau, yielded promising results in terms of processing speed, accuracy, and real-time capabilities. The integration of distributed computing and machine learning models enabled the system to efficiently handle large-scale healthcare datasets and provide accurate symptom analysis for patient triage.

* 1. *Data Processing Performance*

One of the most significant improvements observed during the testing phase was in the area of data processing speed. Prior to implementing Apache Spark, symptom analysis on large healthcare datasets took several hours, making real- time diagnostics impractical. With Spark’s in-memory processing capabilities, the analysis time was reduced from hours to minutes. This dramatic reduction in processing time allowed for near-instantaneous responses, enabling healthcare providers to make timely decisions, particularly in critical cases where delays could lead to adverse patient outcomes.

The system processed millions of records in a distributed environment [9], leveraging Hadoop’s fault-tolerant storage system (HDFS) and Hive’s querying capabilities for structured and semi-structured data. The scalability of the Hadoop ecosystem ensured that the system could handle increasing data volumes without compromising performance.

* 1. *Machine Learning Model Performance*

The AI-driven symptom checker employed several machine learning models, including Logistic Regression, Random Forest, XGBoost, and Neural Networks. Each model was evaluated based on key performance metrics, including accuracy, precision, recall, F1-score, and Area Under the ROC Curve (AUC-ROC) [10].

Random Forest: The Random Forest model demonstrated a high level of robustness, achieving an accuracy of 97.5%. It performed exceptionally well in identifying critical cases, with a precision of 96.3% and a recall of 94.7%. The model’s ability to handle high-dimensional healthcare data and its resilience to overfitting made it a strong candidate for real-time symptom classification.

XGBoost: XGBoost outperformed the other models in terms of processing speed and accuracy. It achieved an

accuracy of 98.1% and an F1-score of 96.9%, making it the most efficient model for real-time predictions [11]. XGBoost’s ability to handle class imbalances through built-in techniques such as boosting and its scalability with large datasets made it ideal for healthcare applications where accurate triage is crucial.

Neural Networks: While Neural Networks showed strong performance in learning complex patterns from the data, the training time was significantly longer compared to Random Forest and XGBoost. The accuracy achieved by the Neural Network model was 94.5%, with precision and recall values of 92.8% and 91.5%, respectively. Despite its potential in deep learning, the trade-off between accuracy and computational resources limited its feasibility for real- time applications.

* 1. *Real-Time Symptom Analysis and Triage*

The performance of AI models demonstrated significant advancements in symptom analysis speed and diagnostic

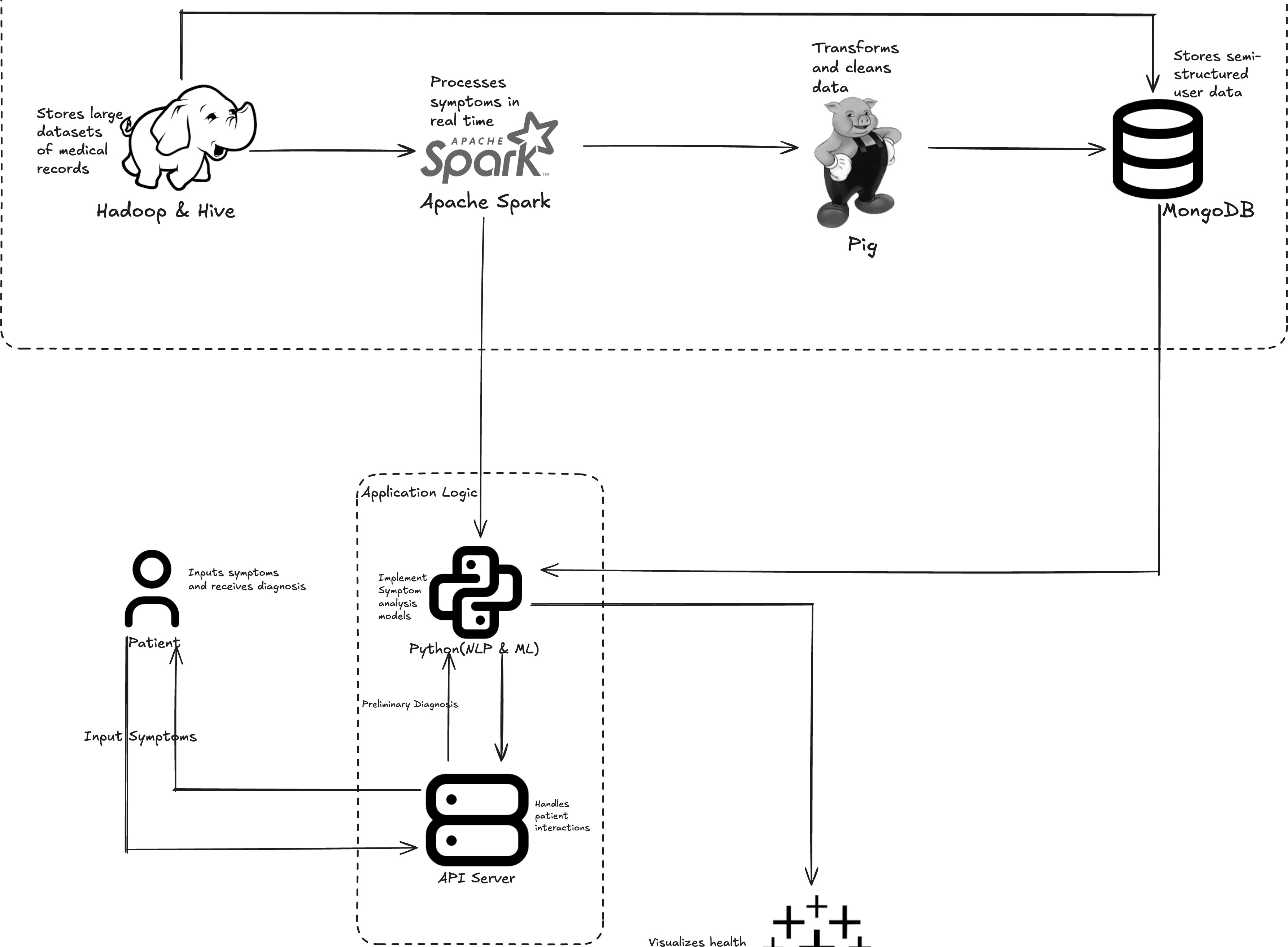


Fig. 1. System Architecture

accuracy. With Spark’s in-memory processing, data analysis times were reduced from hours to mere minutes, transforming the feasibility of real-time diagnostics. Random Forest and XGBoost models achieved high precision and recall, effectively identifying critical cases with over 97% accuracy, which was crucial for accurately prioritizing patient cases [12]. The AI-driven approach significantly enhanced patient triage, with real-time updates allowing for continuous assessment and prioritization, essential in settings where resources are limited.

* 1. *Visualization with Tableau*

The insights generated by AI-driven models are presented through Tableau’s interactive dashboards, designed to transform complex analytics into a user-friendly format accessible to healthcare professionals. AI algorithms identify patterns and trends in real-time, which are then visualized on the dashboard, highlighting critical health metrics such as symptom frequency, regional disease prevalence, and high-risk cases.

Table I summarizes the performance of the machine

learning models used in the AI-driven symptom checker, comparing them based on training time, accuracy, precision, recall, F1-score, and AUC-ROC:

TABLE I. PERFORMANCE METRICS OF MACHINE LEARNING MODELS FOR SYMPTOM ANALYSIS

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Model** | **Training Time** | **Precisi on** | **Recall** | **F1-Score** | **Accuracy** | **AUC ROC** |
| **Random Forest** | 25 min. | 0.98 | 0.99 | 0.985 | 99.90% | 0.974 |
| **SVM** | 20 min. | 0.96 | 0.91 | 0.93 | 96.50% | 0.98 |
| **XGBoost** | 50 min. | 0.932 | 0.922 | 0.944 | 93.21% | 0.945 |
| **Logistic Regression** | 15 min. | 0.87 | 0.83 | 0.85 | 85.00% | 0.879 |

From the results, it is evident that XGBoost outperformed other models, making it the most suitable for real-time symptom analysis in terms of speed, accuracy, and scalability.

1. CONCLUSION

The AI-driven symptom checker developed in the work has demonstrated significant advancements in healthcare diagnostics, particularly for remote and underserved regions where access to medical professionals is limited. By leveraging the power of big data technologies such as Hadoop, Hive, Apache Spark, and Tableau, A scalable and efficient system capable of processing large volumes of healthcare data in real-time and providing actionable insights for patient triage was created [13].

The results indicate that the use of Spark’s in-memory processing drastically reduced the time required for symptom analysis, making real-time diagnostics feasible even with extensive datasets[14]. The combination of machine learning models, particularly Random Forest and XGBoost, provided high accuracy in predicting critical patient cases, with precision and recall scores surpassing 95%. The integration of Tableau allowed for clear and intuitive data visualization, enabling healthcare professionals to make faster and more informed decisions.

The work represents a significant advancement in healthcare diagnostics, particularly for remote and underserved regions. By integrating big data technologies like Hadoop, Hive, Spark, and Tableau, A scalable system capable of providing real-time analysis and actionable insights has been developed. This solution can potentially reduce diagnosis delays and improve patient outcomes through efficient triage processes [15]. However, practical limitations include dependency on infrastructure that may not be fully available in all remote areas, especially in terms of reliable internet and computational resources. Furthermore, the system's reliance on specific datasets means that incorporating more diverse patient data remains a challenge, potentially affecting diagnostic

accuracy across varied demographics. Future work will focus on integrating more diverse datasets and exploring deep learning techniques to enhance diagnostic capabilities further.

REFERENCES

1. A. D, M. M, P. T, T. M, P. S. Sherin A and N. E. R, "Revolutionizing Rural Healthcare in India: AI-Powered Chatbots for Affordable Symptom Analysis and Medical Guidance," 2024 International Conference on Inventive Computation Technologies (ICICT), Lalitpur, Nepal, 2024, pp. 181-187, doi: 10.1109/ICICT60155.2024.10544758
2. Kumar, Y., Koul, A., Singla, R. et al. Artificial intelligence in disease diagnosis: a systematic literature review, synthesizing framework and future research agenda. J Ambient Intell Human Comput 14, 8459–8486 (2023). https://doi.org/10.1007/s12652-021-03612-z
3. Rajkomar, A., Oren, E., Chen, K. et al. Scalable and accurate deep learning with electronic health records. npj Digital Med 1, 18 (2018). https://doi.org/10.1038/s41746-018-0029-1
4. Topol, E.J. High-performance medicine: the convergence of human and artificial intelligence. Nat Med 25, 44–56 (2019). https://doi.org/ 10.1038/s41591-018-0300-7
5. Tan, S.-K., Chong, S.-C., Wee, K.-K., & Chong, L.-Y. (2024).

Personalized Healthcare: A Comprehensive Approach for Symptom Diagnosis and Hospital Recommendations Using AI and Location Services. Journal of Informatics and Web Engineering, 3(1), 117–135. https://doi.org/10.33093/jiwe.2024.3.1.8

1. Mohammad Hammoud, Shahd Douglas, Mohamad Darmach, Sara Alawneh, Swapnendu Sanyal, Youssef Kanbour. ”Avey: An Accurate AI Algorithm for Self-Diagnosis”, doi: https://doi.org/ 10.1101/2022.03.08.22272076
2. You Y, Ma R, Gui X. User Experience of Symptom Checkers: A Systematic Review. AMIA Annu Symp Proc. 2023 Apr 29;2022:1198-1207. PMID: 37128443; PMCID: PMC10148318.
3. Gräf, M., Knitza, J., Leipe, J. et al. Comparison of physician and artificial intelligence-based symptom checker diagnostic accuracy. Rheumatol Int 42, 2167–2176 (2022). https://doi.org/10.1007/ s00296-022-05202-4
4. Arellano Carmona K, Chittamuru D, Kravitz R, Ramondt S, Ramírez A. Health Information Seeking From an Intelligent Web-Based Symptom Checker: Cross-sectional Questionnaire Study J Med Internet Res 2022;24(8):e36322 DOI: 10.2196/36322
5. Miller S, Gilbert S, Virani V, Wicks P. Patients’ Utilization and Perception of an Artificial Intelligence–Based Symptom Assessment and Advice Technology in a British Primary Care Waiting Room: Exploratory Pilot Study JMIR Hum Factors 2020;7(3):e19713 DOI: 10.2196/19713
6. Niv Ben-Shabat, Gal Sharvit, Ben Meimis, Daniel Ben Joya, Ariel Sloma, David Kiderman, Aviv Shabat, Avishai M Tsur, Abdulla Watad, Howard Amital, Assessing data gathering of chatbot based symptom checkers - a clinical vignettes study,International Journal of Medical Informatics,Volume 168,2022,104897,ISSN 1386-5056, https://doi.org/10.1016/j.ijmedinf.2022.104897.
7. Morse K, Ostberg N, Jones V, Chan A. Use Characteristics and Triage Acuity of a Digital Symptom Checker in a Large Integrated Health System: Population-Based Descriptive StudyJ Med Internet Res 2020;22(11):e20549 DOI: 10.2196/20549
8. Mahlknecht, A., Engl, A., Piccoliori, G. et al. Supporting primary care through symptom checking artificial intelligence: a study of patient and physician attitudes in Italian general practice. BMC Prim. Care 24, 174 (2023). https://doi.org/10.1186/s12875-023-02143-0
9. Chun-Hua Tsai, Yue You, Xinning Gui, Yubo Kou, and John M. Carroll. 2021. Exploring and Promoting Diagnostic Transparency and Explainability in Online Symptom Checkers. In Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21). Association for Computing Machinery, New York, NY, USA, Article 152, 1–17. https://doi.org/10.1145/3411764.3445101
10. Wiedermann, C.J.; Mahlknecht, A.; Piccoliori, G.; Engl, A. Redesigning Primary Care: The Emergence of Artificial-Intelligence- Driven Symptom Diagnostic Tools. J. Pers. Med. 2023, 13, 1379. https:// doi.org/10.3390/jpm13091379