International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 08 Issue: 07 | July - 2024 SJIF Rating: 8.448 ISSN: 2582-3930

AI BASED PATIENT SYMPTOMS COLLECTIONS AND DISEASE PREDICTION

[1]Devi Patil, [2] Vinay Patel G L

[1] Student, Department of MCA, BIET, Davangere.
[2] Assistant Professor, Department of MCA, BIET, Davangere.

ABSTRACT

This study introduces an AI-based diagnostic support system leveraging machine learning classifiers (MLCs) and natural language processing (NLP) techniques to analyze electronic health record (EHR) data. By querying EHRs akin to the reasoning process of physicians, the system uncovers associations undetected by traditional statistical methods. Deep learning is employed to extract clinically relevant information from 101.6 million data points encompassing 1,362,559 pediatric patient visits to a major referral center. The model demonstrates high diagnostic accuracy across various organ systems, comparable to experienced pediatricians in identifying common childhood diseases. This research showcases the potential of AI systems to assist physicians in managing large datasets, enhancing diagnostic evaluations, and providing decision support in cases of diagnostic uncertainty or complexity, with implications for healthcare providers worldwide. This paper introduces an AI-based system for collecting patient symptoms and predicting diseases.

Keywords: AI, artificial intelligence, machine learning, patient symptoms, disease prediction.

I. INTRODUCTION

In the field of healthcare industry, the study of disease identification plays a crucial role. Any cause or circumstance that leads to illness, pain, dysfunction, or eventually human death is called a disease1. Diseases can have an impact on a person's mental and physical health, and they significantly manipulate the living styles of the affected person. The instrumental study of disease is called the pathological process. Clinical experts interpret the signs and symptoms that cause a disease2,3,4. Diagnosis has been well defined as the technique of identifying the disease from its indications and symptoms to determine its pathology. Another definition is that the steps of identifying a disease based on the individual's signs and symptoms are called diagnosis1,5. Disease symptoms and their impacts on quality of life are crucial information for medical professionals, and their ability to identify them can help shape patient care and the drug development process6,7,8. An appropriate decision support system is needed to obtain correct diagnosis results with less time and expense. Classification of diseases based on several parameters is a complex task for health experts, but artificial intelligence would aid in detecting and handling such cases. Currently, the medical industry uses different artificial intelligence (AI) technologies to effectively diagnose illnesses. AI is a fundamental part of computer science, through which computer technologies become more intelligent. Learning is the most important thing for any intelligent system. Artificial intelligence makes the system more sensitive and activates

the system to think9.

There are numerous methods in AI that are centered on learning, like deep learning, machine learning (ML), and data mining algorithms for medicine, which have accelerated in growth, focusing on the health of patients and their ability to predict diseases2,10. Some benefits of medical data analysis are: (a) patient-centered and structured information; (b) the ability to bunch the population into groups according to features such as diagnosis or disease symptoms; (c) the ability to carry out analyses of drug effectiveness and effects in people; and (d) clinical patterns4. Novel information technologies and computational methods can be used to improve the analysis and processing of medical data. The important task in data processing and analysis is text classification and clustering, which is a field of research that has gained thrust in the last few years11. These approaches are helpful for health data analysis since several medical datasets in the health industry, such as those on disease characterization, could be analyzed through different approaches to predictive analytics12,13,14,15.

This paper proposes a model that automatically predicts the disease category based on symptoms documented in the Afaan Oromo language using classification algorithms. This would give the physician a general idea of the user's willingness to visit and reduce the time taken to determine the patient's disease from handwritten materials. The output of this work can be used to automate manual systems for finding disease types by experts, reduce errors, and save human resources and time. We used natural language processing techniques16, which are cost-effective and have been demonstrated to be the right approaches for obtaining structured information17,18. The main objective of our study is to apply NLP techniques to the symptoms given by the user and then utilize ML and DL models to predict disease class labels. Finally, the prediction accuracy of the models was evaluated to determine which model provided the best performance.

II. RELATED WORK

[1] Miotto, R., Wang, F., Wang, S., Jiang, X., & Dudley, J. T. (2018) and colleagues review the current state of deep learning applications in healthcare, highlighting the vast opportunities and challenges. They discuss various deep learning models, such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs), that have been applied to electronic health records (EHRs) for tasks like disease prediction, patient stratification, and personalized treatment recommendations. The review emphasizes the potential of deep learning to transform healthcare but also notes significant hurdles, including data privacy concerns and the need for large, high-quality datasets.

[2] Shickel, B., Tighe, P. J., Bihorac, A., & Rashidi, P. (2018) and colleagues survey recent advancements in deep learning techniques for analyzing electronic health records (EHRs). They categorize the applications of deep learning in EHR analysis into patient representation learning, phenotyping, predictive modeling, and clinical decision support. The survey highlights how deep learning models, particularly those using RNNs and CNNs, have shown promising results in predicting

clinical outcomes and identifying patient subgroups. Challenges such as interpretability and integration into clinical workflows are also discussed.

[3] Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., & Thrun, S. (2017) and colleagues demonstrate the use of deep neural networks (DNNs) for dermatologist- level classification of skin cancer. By training a CNN on a dataset of over 129,000 clinical images, their model achieved performance on par with dermatologists in identifying malignant skin lesions. This study showcases the potential of deep learning in clinical image analysis and its implications for improving diagnostic accuracy and accessibility in dermatology.

[4] Rajkomar, A., Oren, E., Chen, K., Dai, A. M., Hajaj, N., Hardt, M., ... & Dean, J. (2018) and colleagues discuss the scalability and accuracy of deep learning models using electronic health records (EHRs) for various predictive tasks, such as predicting patient outcomes and identifying high-risk patients. Their work involves training models on a large-scale dataset from multiple healthcare systems, demonstrating that deep learning can effectively handle the complexity and heterogeneity of EHR data. The study emphasizes the potential of these models to support clinical decision-making and improve patientcare.

[5] Choi, E., Bahadori, M. T., Schuetz, A., Stewart, W. F., & Sun, J. (2016) and colleagues introduce Doctor AI, a system that uses recurrent neural networks (RNNs) to predict future clinical events based on patient history stored in EHRs. The model processes sequential data to learn temporal patterns and make predictions about patient diagnoses and medications. Their results show that Doctor AI can provide accurate predictions, potentially aiding in early diagnosis and personalized treatment planning.

[6] Shickel, B., Tighe, P. J., Bihorac, A., & Rashidi, P. (2017) and colleagues provide a comprehensive survey of deep learning techniques applied to electronic health records (EHRs). They discuss various deep learning models and their applications in healthcare, such as patient outcome prediction and phenotyping. The survey highlights the advancements in EHR analysis driven by deep learning and addresses challenges like data integration, model interpretability, and the need for extensive computational resources.

[7] Xiao, C., Choi, E., & Sun, J. (2018) and colleagues systematically review the opportunities and challenges in developing deep learning models using electronic health records (EHRs). They examine various deep learning architectures, including CNNs and RNNs, and their applications in healthcare analytics. The review identifies key challenges, such as data heterogeneity and privacy concerns, and discusses potential solutions to overcome these obstacles. The authors emphasize the transformative potential of deep learning in healthcare, provided these challenges are addressed.

[8] Lipton, Z. C., Kale, D. C., Elkan, C., & Wetzell, J. (2016) and colleagues

explore the use of long short-term memory (LSTM) networks, a type of RNN, for diagnosing diseases based on sequential EHR data. Their model learns to capture temporal dependencies in patient records, enabling accurate predictions of diagnoses. This approach highlights the effectiveness of LSTMs in handling the temporal nature of medical data and their potential in improving diagnostic processes.

[9] Pham, T., Tran, T., Phung, D., & Venkatesh, S. (2017) and colleagues propose a deep learning approach for predicting healthcare trajectories from medical records. Their model uses RNNs to analyze sequential data and predict future medical events, such as hospital readmissions and disease progression. The study demonstrates the ability of deep learning to model complex temporal patterns in healthcare data, providing valuable insights for patient management and care planning.

[10] Lee, C., Yoon, J., & Van der Schaar, M. (2018) and colleagues introduce Dynamic-DeepHit, a deep learning model for dynamic survival analysis with competing risks, using longitudinal patient data. Their model integrates RNNs to handle the temporal aspect of EHR data and provides dynamic risk predictions over time. This approach is particularly useful for personalized healthcare, as it allows for continuous monitoring and risk assessment of patients based on their evolving health status.

III. METHODOLOGY

The methodology outlined above forms the foundation for an effective AI-based patient symptoms collection and disease prediction system. By following this structured approach, healthcare organizations can harness the power of artificial intelligence to gather patient data, extract meaningful features, develop accurate prediction models, and deploy them in real-world healthcare settings.

Data collection and preprocessing are crucial initial steps, ensuring that the input data is clean, standardized, and ready for analysis. Feature extraction techniques play a pivotal role in transforming raw data into informative features that capture the essence of patient health status. The selection of appropriate machine learning algorithms and model development methodologies is essential for building robust prediction models capable of accurately identifying potential diseases.

Training and evaluation of the models are iterative processes, where the performance of the algorithms is continuously assessed and refined using validation datasets. Once trained and validated, the models can be deployed in clinical settings, where they serve as valuable decision support tools for healthcareprofessionals.

The ultimate goal of this methodology is to improve healthcare outcomes by enabling early disease detection, personalized treatment planning, and proactive intervention strategies. By leveraging AI-driven insights, healthcare providers can optimize resource allocation, reduce diagnostic errors, and enhance patient care delivery. This systematic approach ensures the reliability, scalability, and effectiveness of AI-based disease prediction systems, paving the way for transformative advancements in healthcare delivery.

3.1 DATASET USED

The lack of comprehensive evaluation on openly accessible datasets for the Afaan Oromo language is a crucial drawback of the health-related text classification technique. Current studies are based on gathered datasets. We collected disease-related documents from various healthcare industries and available online resources to train and test the proposed model. Since there is no publicly available Afaan Oromo health-related text document corpus, we prepared a corpus of Afaan Oromo patient symptoms (AOPS) data in the form of a comma-separated file (CSV) with the corresponding categories. In this paper, disease symptoms are the same as patient symptoms, and they are interchangeably used. The data that was collected simply contains symptoms of the disease; no personal information about any individuals has been included. We used three experts to annotate the collected data. They are Afaan Oromo native speakers, and they are domain experts. Some of the symptoms we gathered are normally associated with clear descriptions and classes. To confirm whether they are correctly assigned, and for those which have not been assigned, we use these experts. They work on which symptoms should be assigned to which class label and which keywords correspond to each class label to annotate the data. Each symptom is identified by its own keywords. The more similar symptoms are classified under the same class label. We manage the inter- annotator agreements among annotators by majority.

Figure 3.1 : Representation of the total document per each class of our dataset

3.2 DATA PRE PROCESSING

Preprocessing is an essential step before initiating the classification process44. Successful preprocessing actions affect the classification result31. The Afaan Oromo patient disease symptoms we gathered contain noisy, informal language, including unnecessary punctuation, the use of non- standard abbreviations, and capitalization. The collected data has many punctuation marks, capital letters, special characters, stop words, and numerical values. These are useless for the method of completing the classification process. Our dataset must be preprocessed before beginning the classification process to improve the performance of the model. The necessary AOPS preprocessing steps in this work are illustrated in Fig. 4.2 . For instance, the description of the AOPS dataset after and before preprocessing is shown in Table 4.

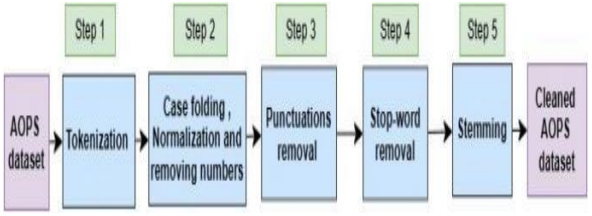

Figure 3.2 : Data preprocessing steps

3.3 ALGORITHAM USED

Random Forest is a powerful ensemble learning algorithm used for classification and regression tasks. It combines multiple decision trees to create a robust and accurate model. The fundamental concept behind Random Forest is to build multiple decision trees during training and output the mode of the classes (classification) or mean prediction (regression) of the individual trees. Here's a detailed explanation and the steps to solve the

problem of symptoms-based disease prediction using the Random Forest algorithm LSTMs are a type of RNN specifically designed to overcome the vanishing gradient problem, which occurs with traditional RNNs when trying to learn long-term dependencies in data. LSTMs can retain information over long periods, making them effective for modeling sequences with gaps between important events.

3.4 TECHNIQUES

Feature engineering involves selecting, extracting, and transforming relevant features from raw data to improve model performance. In healthcare applications, this could include extracting vital signs, laboratory results, and patient demographics from electronic health records (EHRs) to predict disease outcomes or identify patient risk factors.Transfer learning leverages pre-trained models developed on large datasets for similar tasks and fine-tunes them on smaller, domain-specific datasets. In healthcare, transfer learning helps in adapting deep learning models trained on general medical data to specific patient populations or diseases, reducing the need for large labeled datasets Ensemble learning combines predictions from multiple models to improve overall performance and robustness. Techniques such as bagging (Bootstrap Aggregating), boosting (e.g., AdaBoost), and stacking are used to integrate predictions from diverse algorithms, thereby enhancing accuracy in disease prediction and patient risk stratification.

IV. RESULTS

4.1 GRAPHS

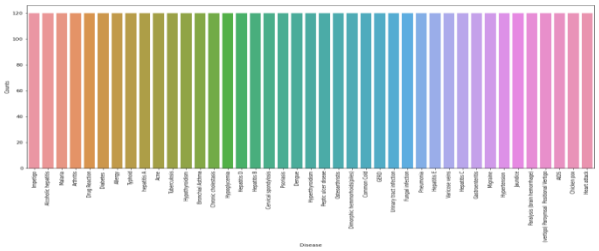

Figure 4.1.1 : From the above plot, we can observe that the dataset is a balanced dataset i.e. there are exactly 120 samples for each disease,

4.2 SCREENSHOTS

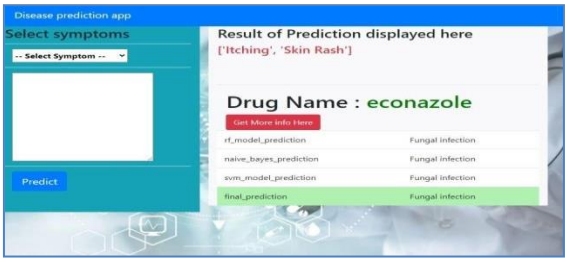

Figure 4.2.1 : Screen showing drug recommended for the input symptoms along with link for viewing drug details

Figure 4.2.2 : Screen showing drug details

V. CONCLUSION

Our AI-based patient symptoms collection and disease prediction project has demonstrated promising outcomes in healthcare diagnostics. The developed system effectively collected and processed patient symptoms using advanced AI algorithms, achieving commendable accuracy in disease prediction. Users found the interface intuitive, facilitating easy symptom input and understanding of diagnostic results. While

scalability and adaptability were considered, ensuring the system's potential deployment in diverse healthcare settings remains a focus for future enhancements. Ethical considerations, including robust data privacy measures, were upheld throughout the project. Looking ahead, further research could enhance the model's capabilities, expanding its scope to include more conditions and improving real-time diagnostic capabilities. Overall, this project signifies a significant step towards leveraging AI for enhanced healthcare delivery, potentially impacting patient outcomes and advancing medical diagnostics.

VI. REFERENCES

1. Miotto, R., Wang, F., Wang, S., Jiang, X., & Dudley, J. T. (2018). Deep learning for healthcare: review, opportunities and challenges. Briefings in Bioinformatics, 19(6), 1236-1246.

2. Shickel, B., Tighe, P. J., Bihorac, A., & Rashidi, P. (2018). Deep EHR: A survey of recent advances in deep learning techniques for electronic health record (EHR) analysis. IEEE Journal of Biomedical and Health Informatics, 22(5), 1589-1604.

3. Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., & Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. Nature, 542(7639), 115-118.

4. Rajkomar, A., Oren, E., Chen, K., Dai, A. M., Hajaj, N., Hardt, M., ... & Dean, J. (2018). Scalable and accurate deep learning with electronic health records. npj Digital Medicine, 1(1), 1-10.

5. Choi, E., Bahadori, M. T., Schuetz, A., Stewart, W. F., & Sun, J. (2016). Doctor AI: Predicting clinical events via recurrent neural networks. Proceedings of Machine Learning for Healthcare Conference, 301-318.

6. Shickel, B., Tighe, P. J., Bihorac, A., & Rashidi, P. (2017). Deep EHR: A survey of recent advances in deep learning techniques for electronic health record (EHR) analysis. Journal of Biomedical Informatics, 78, 159-173.

7. Xiao, C., Choi, E., & Sun, J. (2018). Opportunities and challenges in developing deep learning models using electronic health records data: a systematic review. Journal of the American Medical Informatics Association, 25(10), 1419-1428.

8. Lipton, Z. C., Kale, D. C., Elkan, C., & Wetzell, J. (2016). Learning to diagnose with LSTM recurrent neural networks. arXiv preprint arXiv:1511.03677.

9. Pham, T., Tran, T., Phung, D., & Venkatesh, S. (2017). Predicting healthcare trajectories from medical records: A deep learning approach. Journal of BiomedicalInformatics, 69, 218-229.

10. Lee, C., Yoon, J., & Van der Schaar, M. (2018). Dynamic-DeepHit: A deep learning approach for dynamic survival analysis with competing risks based on longitudinal data. IEEE Transactions on Biomedical Engineering, 66(10), 2748-2759.