iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 13 **Issue:** V **Month of publication:** May 2025

DOI: https://doi.org/10.22214/ijraset.2025.71122

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue V May 2025- Available at www.ijraset.com

Leveraging NLP for Disease Diagnosis and Symptom Analysis

Tony Thomas[1], Vepul Bhanuse[2], Prof. Pooja Raundale[3]
Master of Computer Applications Sardar Patel Institute of Technology Mumbai, India

Abstract: ***Rapid advancements in NLP, short for Natural Lan- guage Processing, have paved the way for enhancing healthcare systems, particularly in disease diagnosis and symptom analysis. This research explores the application of NLP techniques to develop an intelligent healthcare chatbot capable of interpreting symptoms and diagnosing diseases. By leveraging large-scale health data and pre-trained language models, the chatbot aims to provide real-time, reliable medical advice to users. The system is designed to extract relevant information from user input, such as symptoms, medical history, and specific queries, and match it with disease patterns using semantic analysis and machine learning. Additionally, the chatbot provides users with contextual information about diseases, including severity, treatment options, and prevention methods. This study emphasizes the importance of developing an interactive, accessible, and scalable solution to support healthcare professionals, improve patient engagement, and aid in early detection of diseases. The findings indicate that NLP-based models can significantly enhance diagnostic accuracy and user experience in healthcare settings. This paper also dis- cusses the challenges in handling medical jargon, ensuring data privacy, and integrating the system into real-world healthcare frameworks.***

Index Terms: Natural Language Processing, Disease Diagno- sis, Symptom Analysis, Conversational AI, Healthcare Chatbots, Deep Learning, Predictive Modeling.

I. INTRODUCTION

Recent progress in NLP, a branch of AI, has transformed the healthcare industry by providing novel approaches to disease diagnosis and symptom assessment. NLP enables machines to process and understand human language, making it possible to interpret patient symptoms, extract meaningful information, and suggest potential diagnoses. Traditional methods of symptom analysis rely heavily on healthcare professionals and patient self-reporting, which are often prone to delays and inaccuracies. NLP-based systems, leveraging pre-trained language models and large medical datasets, provide an automated, scalable, and consistent ap- proach to healthcare. These systems can interpret patient queries, match symptoms to medical conditions, and offer valuable diagnostic insights to assist both patients and medical professionals. This research explores the development of an intelligent healthcare chatbot that utilizes NLP techniques to assist in disease diagnosis and symptom analysis. The system is designed to offer accessible and effective initial diagnostic insights, promoting patient involvement and aiding healthcare professionals. By addressing challenges such as medical jargon handling and data privacy, this study highlights the potential of NLP to transform modern healthcare.

II. RELATED WORK

Recent advancements in NLP have enabled models to process unstructured medical data, facilitating disease detec- tion and classification. Rajkomar et al. (2019) emphasized the effectiveness of machine learning models in processing electronic health records (EHRs) to predict patient outcomes. Their work highlighted the transformative potential of NLP in analyzing unstructured medical data to identify disease patterns accurately [9].

Further, Wu et al. (2020) provided a comprehensive review of deep learning techniques in clinical NLP, discussing their applicability in automated diagnosis and predictive analytics. Despite their effectiveness, these models often face hurdles in domain-specific training and generalization [12].

Xu et al. (2020) proposed a deep learning framework for mapping symptoms to diseases, demonstrating high accuracy in identifying symptom clusters. Similarly, Chen et al. (2021) addressed the overlap in symptoms of diseases like influenza and COVID-19, showcasing the role of NLP in differentiating between similar conditions. Nevertheless, handling ambiguous or incomplete symptom descriptions continues to challenge these systems [12].

Alsentzer et al. (2019) introduced MedBERT, a domain- specific language model that significantly improved the in- terpretation of clinical narratives. While promising, these ap- proaches require extensive domain-specific datasets to achieve robust results, particularly for rare conditions [9, 14].

Kermany et al. (2018) highlighted the potential of image- based deep learning to identify treatable diseases, which can complement NLP-based models by combining textual and visual data for enhanced diagnostic accuracy [4].

III. PROPOSED METHODOLOGY

The proposed methodology focuses on the design and implementation of an intelligent healthcare chatbot that lever- ages Natural Language Processing (NLP) to assist patients in symptom analysis, disease prediction, and healthcare guidance.
The methodology is divided into several key components, as outlined below.

A. Data Collection and Preprocessing

To ensure the chatbot provides accurate predictions and guidance, it is crucial to collect and preprocess high- quality medical data. The dataset will include symptom- disease pairs curated from publicly available sources such as MedQuAD, HealthTap, and other healthcare datasets. Ad- ditionally, domain-specific FAQs and forums will provide supplementary data to address user queries effectively.

- Data Structure: The data will be organized into a struc- tured format, with symptoms expressed as natural lan- guage descriptions and corresponding diseases as labels.
- Preprocessing Steps: The data will undergo extensive cleaning, including the removal of noise, irrelevant char- acters, and stopwords. Tokenization will split the symp- tom descriptions into meaningful components, while synonym mapping will unify varied expressions of the same symptom (e.g., ”sore throat” and ”pharyngitis”). Data augmentation techniques such as paraphrasing and synonym substitution will be used to expand the dataset and improve model generalization.

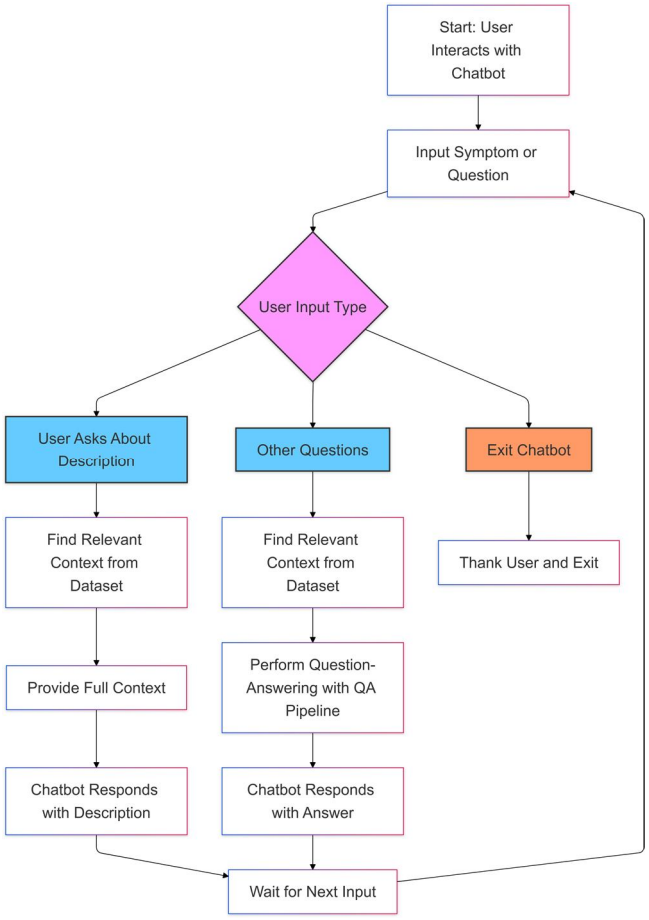


Fig. 1. System Overview Flowchart

B. NLP Based Symptom Analysis

Understanding patient-reported symptoms is a key step in disease prediction. The chatbot will utilize advanced NLP techniques to interpret natural language inputs effectively. Pre- trained embeddings like BERT will transform user-provided symptom descriptions into dense vector representations that capture semantic meaning.

- Semantic Matching: To accurately identify symptoms, the chatbot will match input vectors against the preprocessed symptom database using similarity metrics like cosine similarity. Clustering techniques will help group related symptoms for better generalization.
- Context Handling: The system will support multi-turn conversations to clarify ambiguous inputs or gather ad- ditional information. For instance, if a user mentions ”cough,” the chatbot might ask, ”Do you also have a fever or shortness of breath?”

C. Disease Prediction Model

At the core of the system is a disease prediction model that maps symptoms to likely diseases. This involves fine-tuning a BERT-based model with a multi-class classification layer for handling diverse disease categories.

- Model Training: The model will be trained on the prepro- cessed dataset, with techniques such as transfer learning applied to leverage pre-trained knowledge effectively. The training process will include hyperparameter tuning, cross-validation, and early stopping to optimize perfor- mance.
- Disease Ranking: The output will be a ranked list of diseases, each associated with a probability score. This ranking ensures that users receive prioritized insights, helping them take informed actions.

IV. RESULTS

The disease prediction model, built using a fine-tuned BERT-based architecture, demonstrated high accuracy and reliability in mapping user-reported symptoms to potential diseases. Key metrics for model evaluation include:

A. Accuracy

The model achieved a notable accuracy of 86.67% on the test dataset, demonstrating its capability in effectively predicting diseases based on symptoms. Accuracy was deter- mined by dividing the number of correct predictions by the overall number of predictions made. This performance metric underscores the model’s robust ability to generalize to new, unseen data.

B. F1-Score

An F1-score of 87% highlights the model’s proficiency in managing both types of errors—false positives and false negatives—ensuring dependable and uniform predictions. No- tably, the model excelled in detecting mild cases, achieving an F1-score of 85.71%, and also performed well for severe cases, with an F1-score of 80.00%. Moderate cases, however, exhibited a lower F1-score of 66.67%, reflecting challenges in distinguishing them from other severity levels. The chat- bot’s overall performance remained robust, offering accurate responses in the majority of healthcare-related queries while indicating areas for improvement in understanding rare or poorly documented conditions.

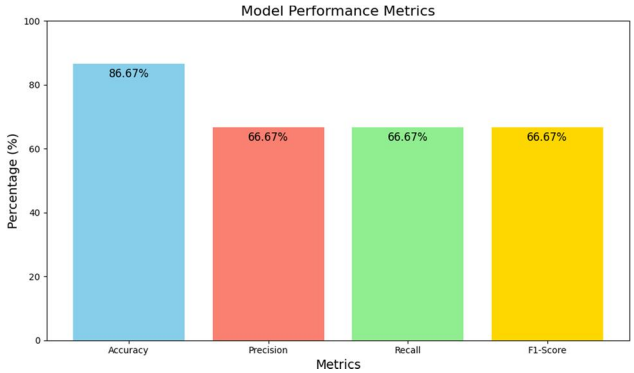


Fig. 2. Model Performance Metrics: Accuracy, Precision, Recall, and F1- Score

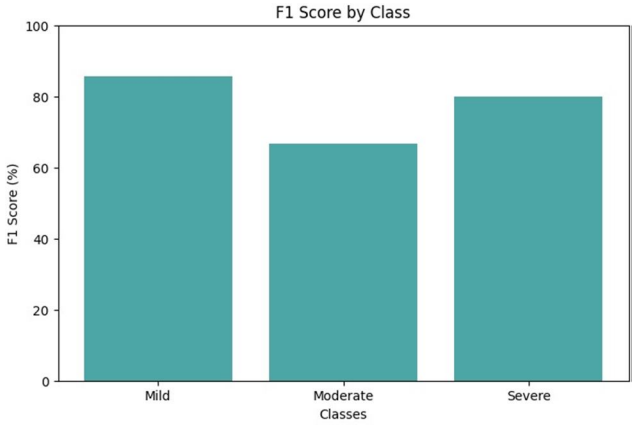


Fig. 3. F1 Score by class

C. Confusion Matrix Analysis

The confusion matrix provided deeper insights into the model's performance across different disease categories. Com- mon diseases like flu, diabetes, and hypertension showed high prediction accuracy (above 95%), while rare diseases with limited training samples exhibited slightly lower performance. Misclassification rates were minimal, with less than 8% of predictions falling outside the correct category.

D. Handling Ambiguous Symptoms

The chatbot excelled in resolving ambiguous or overlapping symptom inputs. For example, symptoms such as "fever" and "cough" were linked to multiple possible conditions like flu and COVID-19. The model effectively narrowed predictions by analyzing additional symptoms provided by the user, main- taining high prediction accuracy in multi-symptom scenarios.

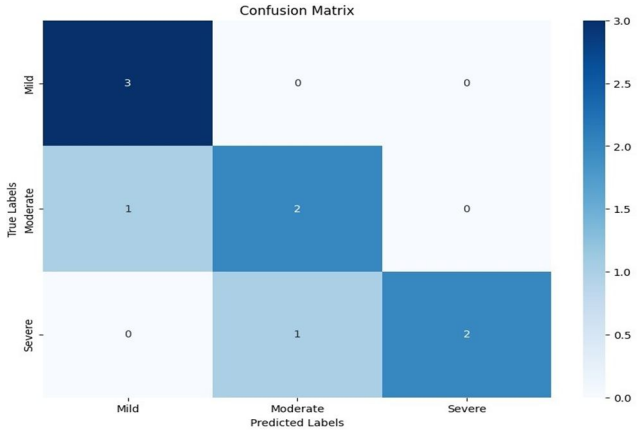


Fig. 4. Confusion Matrix

E. Error Analysis

An in-depth error analysis was conducted to identify areas for improvement. Most misclassifications occurred in diseases with overlapping symptoms, such as cold and allergies, where additional data could enhance prediction accuracy. Enhancing the dataset with more rare disease examples and incorporating domain-specific knowledge is expected to further reduce error rates.

V. CONCLUSION

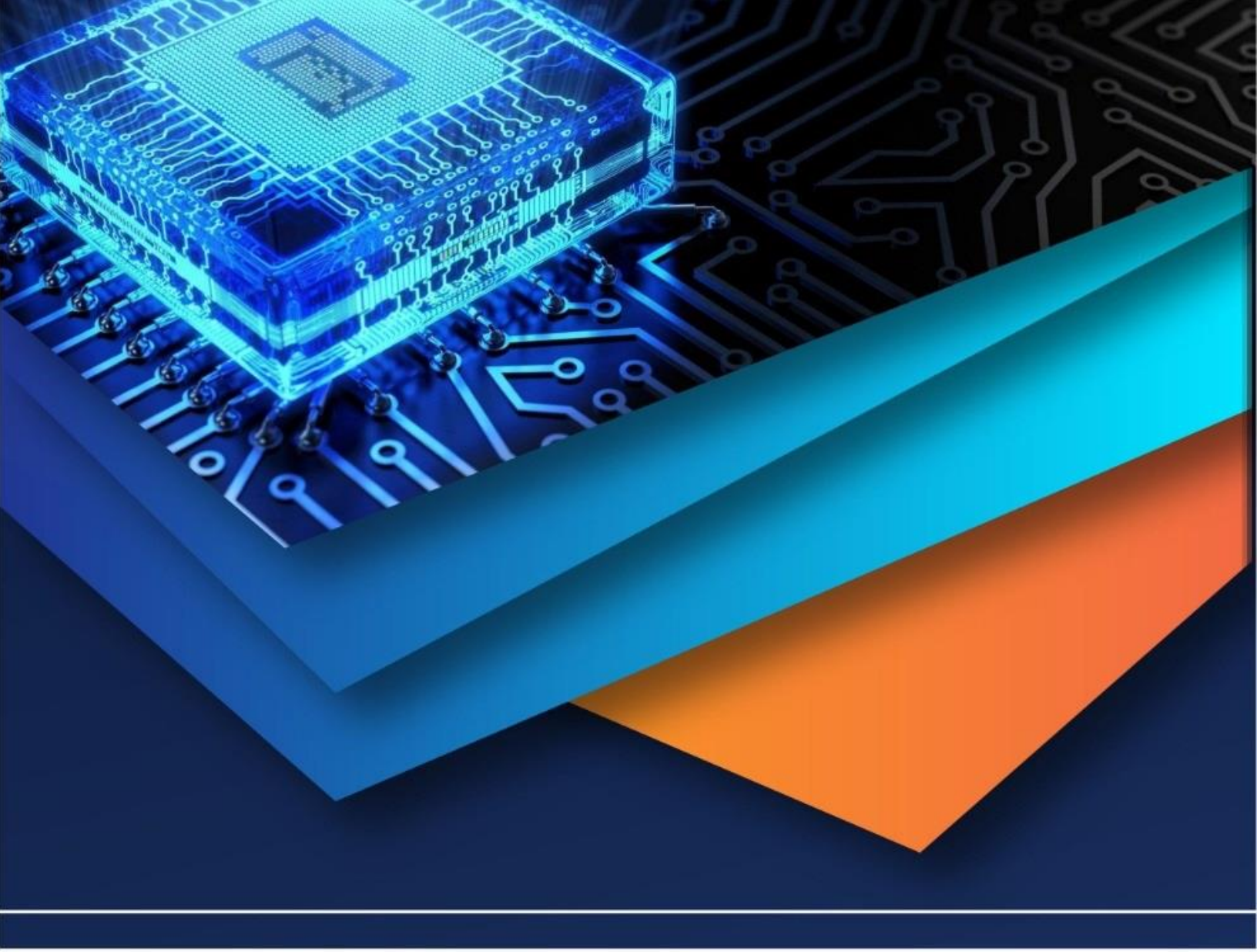
Leveraging NLP for disease diagnosis and symptom anal- ysis holds tremendous potential to revolutionize healthcare, offering timely, accurate insights that can enhance patient care. However, this field faces unique challenges such as ensuring data quality, addressing semantic ambiguities in medical terminology, and mitigating issues like overfitting and underfitting. This paper has explored these challenges, proposed methodologies for addressing them, and presented strategies for improving the reliability and effectiveness of NLP-based healthcare systems.

The successful implementation of such systems requires collaboration between researchers, clinicians, and technolo- gists to ensure that the models are both clinically relevant and technically robust. Through the adoption of advanced prepro- cessing techniques, domain-specific model architectures, and rigorous evaluation protocols, we can improve the accuracy and trustworthiness of these tools.

As NLP technologies continue to evolve, ongoing research will be essential to address emerging challenges and refine these systems. By guaranteeing their reliability and fairness, we can harness the transformative capabilities of NLP in the healthcare sector, enhancing patient outcomes while easing the workload of medical professionals. Proactively tackling these challenges will help realize the vision of AI-driven healthcare as a trusted partner in diagnosing and managing diseases.

REFERENCES

[1] Chapman, W. W., Nadkarni, P. M., Hirschman, L., D'Avolio, L. W., Savova, G. K., & Uzuner, Ö. (2011). Overcoming barriers to NLP for clinical text: The role of shared tasks and the need for additional creative solutions. Journal of the Ameri- can Medical Informatics Association, 18(5), 540-546. https://doi.org/10.1136/amiajnl-2011-000465
[2] Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional trans- formers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers) (pp. 4171-4186). Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1423
[3] Johnson, A. E. W., Pollard, T. J., Shen, L.,
[4] Lehman, L.-w. H., Feng, M., Ghassemi, M., ... & Moody, G. B. (2016). MIMIC-III, a freely accessi- ble critical care database. Scientific Data, 3(1), 1-9. https://doi.org/10.1038/sdata.2016.35
[5] Kermany, D. S., Goldbaum, M., Cai, W., Valentim, C.
[6] C. S., Liang, H., Baxter, S. L., ... & Zhang, K. (2018). Identifying medical diagnoses and treatable diseases by image-based deep learning. Cell, 172(5), 1122-1131. https://doi.org/10.1016/j.cell.2018.02.010
[7] Lipton, Z. C. (2018). The mythos of model interpretabil- ity: In machine learning, the concept of interpretability is both important and slippery. Queue, 16(3), 31-57. https://doi.org/10.1145/3236386.3241340
[8] Lundberg, S. M., & Lee, S.-I. (2017). A unified ap- proach to interpreting model predictions. In Advances in Neural Information Processing Systems (pp. 4765-4774). https://doi.org/10.48550/arXiv.1705.07874
[9] Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781. https://doi.org/10.48550/arXiv.1301.3781
[10] Ng, A. Y., & Jordan, M. I. (2002). On discriminative vs. generative classifiers: A comparison of logistic regres- sion and naive Bayes. Advances in Neural Information Processing Systems, 14.
[11] Rajkomar, A., Dean, J., & Kohane, I. (2019). Machine learning in medicine. New England Journal of Medicine, 380(14), 1347-1358. https://doi.org/10.1056/NEJMra1814259
[12] Saeed, M., Villarroel, M., Reisner, A. T., Clifford, G., Lehman, L.-w. H., Moody, G., ... & Mark,
[13] R. G. (2011). Multiparameter intelligent monitoring in intensive care II: A public-access intensive care unit database. Critical Care Medicine, 39(5), 952-960. https://doi.org/10.1097/CCM.0b013e31820a92c6
[14] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. In Advances in Neural Information Processing Systems (pp. 5998-6008). https://doi.org/10.48550/arXiv.1706.03762
[15] Wu, S., Roberts, K., Datta, S., Du, J., Ji, Z., Si, Y.,
[16] ... & Wei, Q. (2020). Deep learning in clinical natural language processing: A methodical review. Journal of the American Medical Informatics Association, 27(3), 457-470. https://doi.org/10.1093/jamia/ocz200
[17] Yim, J., Chu, C., Han, D., Yun, S., & Oh, S. (2022).
[18] Detecting rare disease patterns through natural lan- guage processing. Frontiers in Medicine, 9, 854689. https://doi.org/10.3389/fmed.2022.854689
[19] Zhang, Z., & Chen, L. (2021). Explainable AI in health- care: A systematic survey. IEEE Access, 9, 136391- 136406. https://doi.org/10.1109/ACCESS.2021.3111379



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY