International Journal of Innovative Research in Engineering and Management (IJIREM)
ISSN (Online): 2350-0557, Volume 13, Issue 3, June 2026
DOI: https://doi.org/10.55524/ijirem.2026.13.3.1
Article IJIR3208, Pages 1-6
www.ijirem.org

AI-Based Medical Diagnosis Using Machine Learning and Natural Language Processing for Symptoms Prediction

Mahak[1], G. Hariharan[2], Sakshi Kathuria[3], and Jyoti Chaudhary[4]

[1, 2] MCA Scholar, Department of Computer Application, Amity Institute of Information and Technology, Amity University, Gurugram, Haryana, India
[3, 4] Assistant Professor, Department of Computer Application, Amity School of Engineering and Technology, Amity University, Gurugram, Haryana, India

Correspondence should be addressed to Mahak; avika7527@gmail.com

Received: 26 April 2026; Revised: 11 May 2026; Accepted: 24 May 2026

ABSTRACT- Given the increasing importance of fast, accurate and cost-effective medical diagnoses, it is now crucial to develop smart systems that can assist in diagnosing patients. This paper is a design and development concept to develop a tool that can be used by patients to diagnose their medical condition and offer early diagnostic advice to the patient. The system is set in such a way that to pass in information that is not strictly structured as the one we normally speak in and here, the symptoms begin to be described in a manner that is easily understandable and realistic in the real world.

The NLP module uses steps such as text normalization, tokenization and lemmatization (word normalization), as well as entity recognition, in order to extract key features, which are important to the medical context. These characteristics are assessed over the machine learning to answer the potential diseases. The system also establishes the severity of the symptoms and classifies cases into various degrees basing on urgency. The system is not aimed to replace the professional medical consultation but is a supportive tool to help in diagnosing at the early stages and may be enhanced further with the improved models and real time data availability.

KEYWORDS- Artificial Intelligence, Medical Diagnosis, Natural Language Processing, Machine Learning, Symptom Analysis, Risk Assessment, Predictive Modeling.

I. INTRODUCTION

Artificial Intelligence (AI) has continued to be a significant facilitator in the contemporary healthcare framework, and its primary application to boost diagnostic effectiveness and accessibility. Traditional medical diagnosis is still possible where diagnosis relies on clinical expertise, physical examination and laboratory testing that still results in delays in the treatment of medical patients especially in a setting with limited resources. As digital health technologies rapidly grow, there is a high demand to create intelligent systems capable of helping to diagnose in the initial stage and lessen the load on healthcare professionals [1] [6] [12]. With the emergence of Natural Language Processing (NLP) and the concomitant rise of the Machine Learning (ML) together, a new line of understanding how to organize the process of creating an automated diagnostic system has unravelled themselves. To learn the patterns on historical medical data machine learning is used, and to process the unstructured input of patients into text written in a natural language, NLP is used. This would particularly be helpful because the description of symptoms as controlled by the patient is done not in any organized way, but in an informal one [2] [5] [9]. Similar applications and approaches have also emerged in medical research that are based on using machine learning models on complex patterns of data to avert neurodegenerative disorders E.g. Alzheimer's disease [15]. In addition, NLP within it and machine learning have been effectively used to process large scale textual information and draw meaningful insights of unstructured information [16].

The low access to healthcare services, the increasing delays in the other traditional procedures of diagnosis, increased patient load, and the importance of early detection of the disease necessitate such systems. The need to develop systems that can be efficient as well as be available in real life conditions stems out of these issues [15] [16]. The aim of the current project is to develop a smart diagnostic system, which is going to analyze the symptoms entered by a user, the unstructured input is going to be interpreted with the help of the NLP, the prediction of the potential disease with the help of the machine learning models and return the confidence score is expected, the reason of this is to predict the potential diseases with the help of the machine learning models and returning confidence scores.

II. LITERATURE REVIEW

It has been observed that over the last couple of years, the adoption of Artificial Intelligence (AI) in healthcare has grown dramatically thanks to the presence of extensive medical datasets and enhanced computational processes. Researchers have discussed various methods to improve diagnostic accuracy, minimize human error and aid in decision-making. These solutions are based on the move towards more data-driven and adaptable solutions with a further division into rule-based systems, machine learning and Natural Language Processing (NLP)-based approaches [18] [20].

A. Conventional Diagnostic Systems.

The previous diagnostic systems largely employed the rule-based methodology, wherein rule and decision tree based diagnostic systems were utilized; to diagnose diseases based on symptoms. They came up with these systems through expert understanding making them dependable to well-established medical conditions. They were also rather easy to implement, and did not need large datasets.

Yet these systems were not very flexible and could be easily adapted to the changes in the diseases or the arrangement of the symptoms. They also discerned problems in dealing with unpredictability as well as incomplete information, which are natural in real-world medical circumstances. The more the number of rules, the more complicated would be the maintenance and update of the system. These limitations caused the gradual replacement of rule-based systems with more sophisticated ones that are able to learn by looking at the data and adjust to the new environment [7][8].

B. Machine Learning based methods.

In order to improve on the existing diagnostic systems, the machine learning solutions have been applied to enable the models to learn the patterns on their own using the medical data, and without the necessity to rely on predefined rules. Innovations in clinical machine learning in the recent past have demonstrated additional how these models could enhance the diagnosis, prognosis and treatment outcomes of healthcare applications [20]. These strategies enable systems to process both structured and unstructured medical data, such as descriptions of symptoms, patient history, and electronic health records.

Among the primary strengths of machine learning is the fact that as more data is exposed the job will be optimized and its performance will improve accordingly. The model, through training is able to learn relationships between symptoms and the illness which helps in making more accurate predictions. More complex relationships between more than two variables, can also be tackled by machine learning algorithms, which increases the realism of the diagnostic results [9][10].

Besides that, the models are scalable and can be suitably altered to various fields of medicine without a significant overhaul. They further give probabilistic results, e.g. confidence scores, allowing users to read into the probability of various conditions. Although these are the advantages, machine learning models heavily rely on data quality. Bad or biased data might affect performance, and not all models can be interpreted, which can decrease confidence in the medical use case [13][14].

C. Natural Language Processing and its role in healthcare.

Natural Language Processing (NLP) has a significant role in healthcare, as it allows any system to comprehend and process human language. Much of medical data such as descriptions of various patients and also clinical notes are available in the form of unstructured text. This unstructured data is converted to structured data which can be analyzed through machine learning algorithms [2][5].

Features of NLP enable users to specify symptoms using the natural language, which enhances ease of use and access. Some of these methods include tokenization, lemmatization, and named entity recognition which serve to extract meaningful medical information out of text. This enhances the communication between the users and diagnostic systems and makes them more user-friendly and intuitive.

Even more elaborate NLP models exist that help in ambiguity, as well as, understanding the context that is significant in medical diagnosis where medical symptoms may have many interpretations. By combining both NLP and machine learning, systems can be made to process end-to-end and become familiar with user input and provide predictions to the user. However, the same problems, such as controlling the language used in the field and maintaining absolute accuracy in all circumstances remain. These methods are still kept up to date into making more efficient healthcare applications [16].

D. Research Gap

Despite the process of formulating the concept of a real-world implementation, a variety of challenges linked to the concept of AI-based diagnostic systems development have taken place. One of the issues is that it is not always easy to infer the uncertain descriptions of the symptoms as patients usually prove to describe the symptoms informally and differently. This may result in wrong feature extraction and less accuracy of prediction.

The second constraint is that it relies on quality of data and incomplete or biased data sets are detrimental to the model performance. There are also concerns that existing systems have problems in reduction of the false prediction and handling the uncertainty which is a vital criterion in diagnosis. Also, most systems fail to completely take into account the contextual and patient-related factors like ages, lifestyle and medical history.

Another problem is the consequence of the lack of interpretability as complex machine learning models are more likely to be black boxes and, therefore, it becomes difficult to comprehend how predictions are made. Despite these improvements, there are still issues of reality variability and the stable truth of prediction in the current systems.

To remedy these issues, this project postulates the application of an integrated framework that integrates the NLP and machine learning on how to overcome these problems and offer a reliable decision support. This is to enhance the usability, flexibility and overall performance of AI-based diagnostic systems [13][14]. Unlike many existing healthcare diagnostic systems that focus on individual tasks such as disease prediction or symptom identification separately, the proposed work combines Natural Language Processing, Machine Learning, risk assessment, and preliminary decision support within a single framework.

The system is designed to process symptoms described in natural language, generate confidence-based predictions, and evaluate severity levels to support early-stage diagnosis. In addition, features such as guided symptom input and integrated analysis improve accessibility and usability for users without technical or medical knowledge. This combination of multiple functionalities within one platform provides a more practical and user-oriented approach compared to several existing systems that offer limited prediction capability without integrated support mechanisms.

III. METHODOLOGY

The proposed medical diagnosis system using AI is a structured system, which uses Natural Language Processing (NLP), machine learning (ML), and rule-based techniques to convert User input into meaningful diagnostic output. The system is programmed to process unstructured descriptions of the symptoms, make efficient predictions and risks evaluation.

A. System Workflow and Processing Pipeline

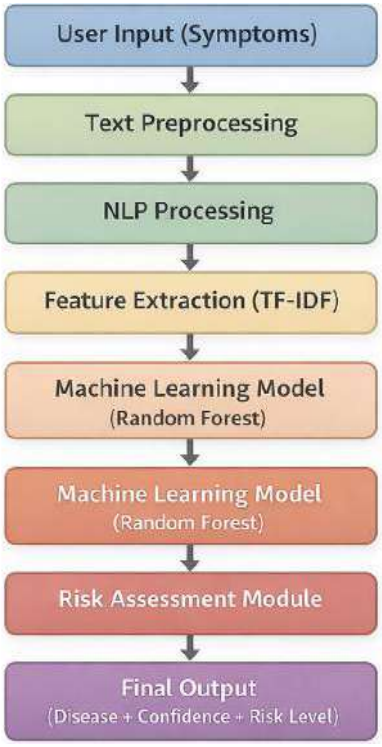

Figure 1: System Workflow of AI-Based Medical Diagnosis System

Like in Figure 1, the system adheres to a structured workflow that receives user input which is then processed through various phases until diagnostic output is generated. The system starts with user input, whereby, symptoms are entered in the natural language. The input is pre-processed to eliminate noise and equalize the text format. Upon which, NLP techniques are used to derive valuable features of the input.

These characteristics are brought into numerical values via methods like TF-IDF they are then processed through a machine learning model. The system has a Random Forest classifier that examines the patterns among symptoms and diseases and produces predictions [3].

Lastly, a risk assessment module will help assess the severity of the condition and classify it into various levels. The system will provide the predicted diseases and confidence scores and risk levels.

B. Natural Language Processing Module

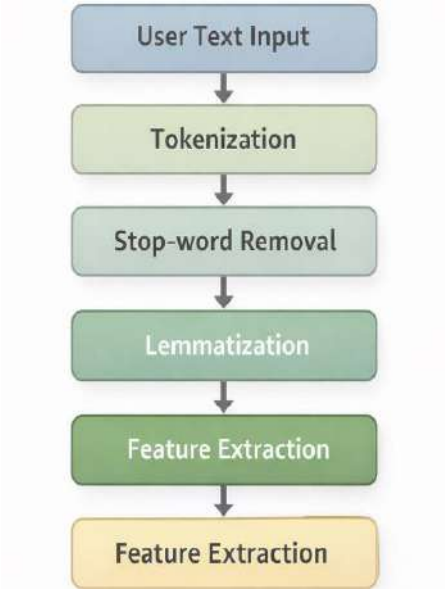

Figure 2: NLP Processing Pipeline

The **NLP** module is shown to process text in various stages such as tokenization, remove stop-words and lemmatization into lemmas, as shown in Figure 2.
The NLP module takes the user input and transforms it with the use of several steps which include tokenization, stop-word, and lemmatization. These are used to clean and normalize the text, so as to further analyze it.
Text processing tools such as spaCy and NLTK are employed, as well as phrase matching algorithms that are used to find relevant medical terms. The processes of feature extraction, such as TF-IDF, transform the processed text into a numerical form to be used by a machine learning approach [5].

C. Machine Learning and Prediction.

Processing characteristics are then fed to a machine learning model to predict disease. The system mainly relies on the Random Forest classification system which enhances features of the system by integrating a number of decision trees. Likewise, other AID-based methods have also been considered to predict disease symptoms to enhance diagnostic efficiency [19]. Training of the model is done with labeled datasets that consist of symptoms and diseases associated with them. The model produces a number of potential diagnoses and confidence scores, as medical diagnosis is probabilistic.

D. Risk assessment and Decision support.

It is a system that evaluates the severity of the symptoms and combinations used on the system to determine the risk level. There is also a segmentation of cases into the low, medium and high-risk groups.
Such demographic aspects as age and prevailing conditions are also taken into account to increase accuracy. The system offers minimum suggestions, depending on the level of risk, to assist in making decisions.

E. Optimization and System Handling

In order to make sure that the system is highly efficient in its performance, the system has in it caching and efficient search mechanisms which are meant to ensure that the system is highly efficient in its performance.
Fuzzy matching (and fallback) methods are employed to cope with discrepancies in medical terms whereas fallback mechanisms are used to ensure system reliability in situations where medical terms differ. The personalization and precision are additionally improved with context-aware processing.

IV. RESULTS AND ANALYSIS

The given AI-based medical diagnosis system was tested on the basis of its capabilities to properly identify diseases to the user who provides his or her symptom descriptions. Standard evaluation measures, such as accuracy, precision, recall, and F1-score, were used in the evaluation of the system performance. To mitigate the risk of not meeting the demands of the required criteria of reliability, accuracy and applicability in the real world, clinical evaluation of AI-based diagnostic systems is necessary [17]. This is because these measurements help to determine the efficiency and reliability of the model in real life applications [1].
To test the system representative inputs of symptoms that represent common symptom combinations were used to test the system. The outcomes suggest that the model can be used to make credible predictions with a minimum response time. Also, the incorporation of NLP and machine learning algorithms is sure to make the system correct when it comes to handling unstructured input data.

A. Performance Evaluation

The performance of the machine learning model is summarized in the table 1:

Table 1: Performance Evaluation Metrics of the Proposed AI-Based Medical Diagnosis System

Metric	Value
Accuracy	76%
Precision	75%
Recall	72%
F1-Score	73%

B. Analysis of Results

- The 76% accuracy of the model shows that the model is reasonably helpful in the prediction of diseases, given the input symptoms.
- The precision score indicates that the majority of the findings that are predicted are applicable and sound.
- The value at recall indicates that the system is capable of detecting a remarkable percentage of the actual cases but it can be improved further.
- F1-score is a score that also represents a balanced performance in terms of both precision and recall.

On the whole, the findings show that the system can effectively offer a preliminary diagnosis, and may be considered as a supportive decision-making tool.

C. Graphical Representation

A bar graph to compare accuracy, precision, recall and F1-score can be used to visualize the performance of the model. As it is evident in the graph, the highest metric was accuracy with the recall slightly lower, which leaves room to improve by ensuring the identification of all possible cases.

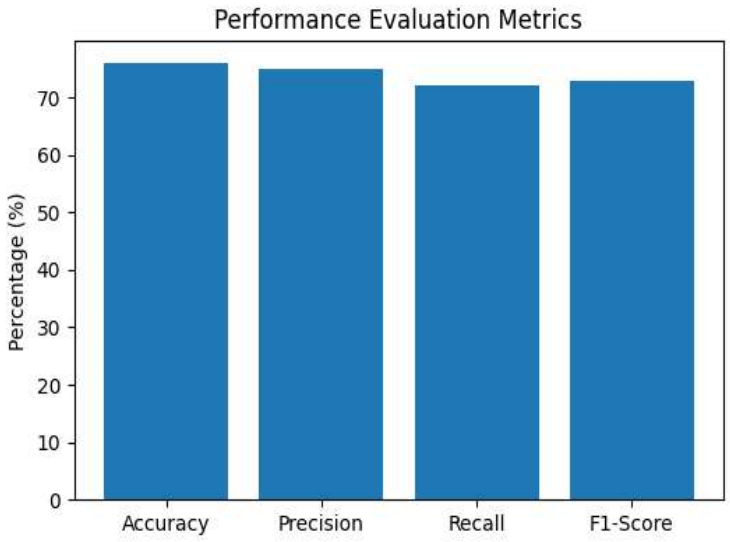

Figure 3: Graphical Representation of Performance Metrics of the Proposed AI-Based Medical Diagnosis System

In the above Figure 3 illustrates the comparative performance of the proposed AI-based diagnostic system using standard evaluation metrics. The model achieved the highest performance in accuracy (76%), while recall showed comparatively lower values, indicating scope for further improvement in identifying all positive cases.

V. CONCLUSION

This paper introduced the design and development of an AI-based medical diagnosis system which combines the natural language processing and machine learning models to predict diseases by evaluating the symptoms presented by users. It is also effective in terms of processing unstructured input, feature extraction, and generating a list of possible diagnoses, and confidence scores. The addition of a risk assessment module adds further functionality to the system because of case prioritization according to their risk level and a rudimentary decision support system. The outcomes of the experiment prove that the model has a good level of performance in a realistic accuracy range, which made it optimal in initial diagnostics. Even though the system is not an alternative to professional medical consultation there is still good evidence that the system can be used as a reliable support tool to help in early-stage diagnosis as well as enhancing healthcare accessibility. Overall, the proposed solution highlights the potential of AI-driven solutions in the modern healthcare systems [12] [14].

In addition, the system may also be equipped with some advanced features like adding some state-of-the-art models of deep learning to further enhance its prediction as well as contextual representation capabilities [15]. The system may also be supplemented with real-time clinical data and electronic health records that will help in ensuring that the system is more reliable. This can be further improved in future by incorporation of voice-based symptom entry, multiple language implementation and implementation of mobile application to further improve access. Furthermore, it is possible to enhance transparency and user trust in the system by means of incorporating explainable AI techniques [16].

CONFLICTS OF INTEREST

The authors declare that they have no Conflicts of Interest.

REFERENCES

[1] Mitchell, T. M., *Machine Learning*. New York, NY, USA: McGraw-Hill, 1997. Available from: http://www.cs.cmu.edu/~tom/mlbook.html

[2] Jurafsky, D. and Martin, J. H., *Speech and Language Processing*, 3rd ed. (draft), 2023. Available from: https://web.stanford.edu/~jurafsky/slp3/

[3] Breiman, L., "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001. Available from: https://doi.org/10.1023/A:1010933404324

[4] Devlin, J., Chang, M. W., Lee, K., and Toutanova, K., "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. 2019 Conf. North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, 2019, pp. 4171–4186. Available from: https://aclanthology.org/N19-1423/

[5] Bird, S., Klein, E., and Loper, E., *Natural Language Processing with Python*. Sebastopol, CA, USA: O'Reilly Media, 2009. Available from: https://www.nltk.org/book/

[6] World Health Organization, *Global Strategy on Digital Health 2020–2025*. Geneva, Switzerland: World Health Organization, 2021. Available from: https://www.who.int/publications/i/item/9789240020924

[7] Shortliffe, E. H., *Computer-Based Medical Consultations: MYCIN*. New York, NY, USA: Elsevier, 1976. Available from: https://www.sciencedirect.com/book/9780444001795/computer-based-medical-consultations-mycin

[8] Szolovits, P., *Artificial Intelligence in Medicine*. Boulder, CO, USA: Westview Press, 1982. Available from: https://books.google.com/

[9] Goodfellow, I., Bengio, Y., and Courville, A., *Deep Learning*. Cambridge, MA, USA: MIT Press, 2016. Available from: https://www.deeplearningbook.org/

[10] Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., and Thrun, S., "Dermatologist-level classification of skin cancer with deep neural networks," *Nature*, vol. 542, no. 7639, pp. 115–118, 2017. Available from: https://doi.org/10.1038/nature21056

[11] Miotto, R., Li, L., Kidd, B. A., and Dudley, J. T., "Deep patient: An unsupervised representation to predict the future of patients from electronic health records," *Scientific Reports*,

vol. 6, Art. no. 26094, 2016. Available from: https://doi.org/10.1038/srep26094

[12] Topol, E., *Deep Medicine: How Artificial Intelligence Can Make Healthcare Human Again*. New York, NY, USA: Basic Books, 2019. Available from: https://www.basicbooks.com/titles/eric-topol/deep-medicine/9781541644632/

[13] Rajkomar, A., Dean, J., and Kohane, I., "Machine learning in medicine," *New England Journal of Medicine*, vol. 380, no. 14, pp. 1347–1358, 2019. Available from: https://doi.org/10.1056/NEJMra1814259

[14] Beam, A. L. and Kohane, I. S., "Big data and machine learning in health care," *JAMA*, vol. 319, no. 13, pp. 1317–1318, 2018. Available from: https://doi.org/10.1001/jama.2017.18391

[15] Kaur, M., Arora, A., Kathuria, S., Arshad, M. W., and Singh, S. P., "Machine Learning for Alzheimer's Disease Detection and Categorization in Brain Images," *International Journal of Intelligent Systems and Applications in Engineering*, vol. 12, no. 21s, pp. 630–636, 2024. Available from: https://ijisae.org/index.php/IJISAE/article/view/5459

[16] AL Mahadin, G. *et al*., "Deep learning framework for identifying and synthesizing news headlines on social media," 2025. *(Complete publication details and source URL not provided.)*

[17] Park, S. H., Han, K., Jang, H. Y., Park, J. E., Lee, J. G., Kim, D. W., and Choi, J., "Methods for clinical evaluation of artificial intelligence algorithms for medical diagnosis," *Radiology*, vol. 306, no. 1, pp. 20–31, 2023. Available from: https://doi.org/10.1148/radiol.220182

[18] Kaur, S., Singla, J., Nkenyereye, L., Jha, S., Prashar, D., Joshi, G. P., and Islam, S. R., "Medical diagnostic systems using artificial intelligence (AI) algorithms: Principles and perspectives," *IEEE Access*, vol. 8, pp. 228049–228069, 2020. Available from: https://doi.org/10.1109/ACCESS.2020.3042273

[19] Budhraja, H., Gupta, M., Bhardwaj, N., and Agarwal, L., "AI-driven symptom-based disease prediction for efficient healthcare diagnosis," in *Proc. 2025 Int. Conf. Modeling, Simulation & Intelligent Computing (MoSICom)*, 2025, pp. 1–6. Available from: https://doi.org/10.1109/MoSICom67153.2025.11398223

[20] Swanson, K., Wu, E., Zhang, A., Alizadeh, A. A., and Zou, J., "From patterns to patients: Advances in clinical machine learning for cancer diagnosis, prognosis, and treatment," *Cell*, vol. 186, no. 8, pp. 1772–1791, 2023. Available from: https://doi.org/10.1016/j.cell.2023.01.035