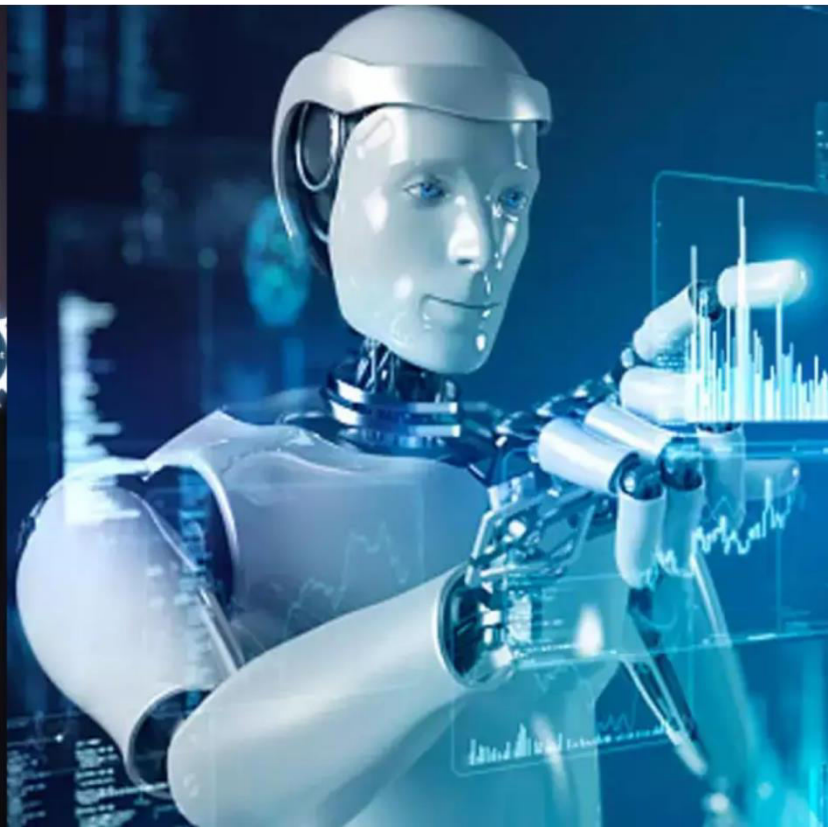
IJIRSET
International Journal of Innovative Research in Science | Engineering | Technology

ISSN(O): 2319-8753
ISSN(P): 2347-6710

# International Journal of Innovative Research in Science Engineering and Technology (IJIRSET)

*(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)*

**Impact Factor: 8.699**

**Volume 15, Issue 4, April 2026**

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 4, April 2026

|DOI: 10.15680/IJIRSET.2026.1504154|

# Enhancing Disease Diagnosis System using Machine Learning: A Review

**Dr. Sulochana B. Sonkamble , Mr. Sunit Nagsen Shinde**

Department of Computer Engineering, JSPM NTC, Pune, India

**ABSTRACT:** This paper presents an AI-driven Clinical Decision Support System (CDSS) designed to enhance disease diagnosis by leveraging patient similarity. The system integrates heterogeneous healthcare data, including electronic health records, laboratory reports, and patient demographics, to provide accurate and personalized diagnostic recommendations. Advanced machine learning and deep learning techniques, combined with Natural Language Processing (NLP), are used to extract meaningful semantic relationships between symptoms and diagnoses. The proposed model employs word embedding and similarity measures to identify patients with similar clinical profiles and predict potential diagnoses. Experimental results on real-world datasets demonstrate improved accuracy, precision, and recall compared to traditional methods. The system assists healthcare professionals in decision-making, reduces diagnostic errors, and supports early disease detection for better patient outcomes.

**KEYWORDS:** Artificial Intelligence, Clinical Decision Support, Disease Diagnosis, Patient Similarity, Machine Learning

## I. INTRODUCTION

The rapid growth of healthcare data from various sources such as Electronic Health Records (EHRs), laboratory reports, and wearable devices has created new opportunities for improving medical diagnosis. Traditional diagnostic methods rely heavily on physicians' expertise and experience, which may sometimes lead to errors, especially in complex cases involving multiple diseases. With the advancement of Artificial Intelligence (AI), intelligent systems can now assist healthcare professionals by analyzing large volumes of medical data efficiently and accurately.

Clinical Decision Support Systems (CDSS) play a crucial role in modern healthcare by providing data-driven insights and recommendations to physicians. These systems integrate patient data with medical knowledge bases to support diagnosis, treatment planning, and monitoring. However, many existing systems focus on single disease prediction and do not effectively utilize the relationships between multiple symptoms and conditions. Additionally, they often fail to leverage the full potential of heterogeneous data sources.

To address these challenges, this work proposes an AI-driven CDSS that utilizes patient similarity and semantic analysis for disease diagnosis. By applying machine learning, deep learning, and Natural Language Processing techniques, the system identifies patterns in patient data and predicts diagnoses based on similarities with historical cases. This approach improves diagnostic accuracy and provides more personalized healthcare solutions.

### Motivation

Accurate and early disease diagnosis is critical for effective treatment and patient survival. However, traditional diagnostic approaches face limitations such as dependency on physician expertise, inability to process large-scale data, and lack of integration of diverse medical information. Additionally, patients often suffer from multiple conditions simultaneously, making diagnosis more complex.

The motivation behind this work is to develop an intelligent system that can analyze heterogeneous healthcare data, identify similarities among patients, and provide accurate diagnostic predictions. By leveraging AI techniques, the system aims to reduce diagnostic errors, improve efficiency, and support healthcare professionals in making informed decisions.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 4, April 2026

|DOI: 10.15680/IJIRSET.2026.1504154|

## II. LITERATURE SURVEY

Sanjeev Arora [1], The paper also gives a theoretical explanation of the success of the above unsupervised method using a latent variable generative model for sentences, which is a simple extension of the model in Arora et al. (TACL'16) with new "smoothing" terms that allow for words occurring out of context, as well as high probabilities for words like and, not in all contexts.

Houping Xiao [2], In this paper, we propose a novel framework to capture the trajectory of latent states for patients from behavioral data while exploiting their demographic dierences and similarities to other patients. We conduct a hypothesis test to illustrate the importance of the demographic data in diabetes management, and validate that each behavioral feature follows an exponential or a Gaussian distribution.

Koen Bruynseels [3], This perspective redefines the concept of 'normality' or 'health,' as a set of patterns that are regular for a particular individual, against the backdrop of patterns observed in the population. This perspective also will impact what is considered therapy and what is enhancement, as can be illustrated with the cases of the 'asymptomatic ill' and life extension via anti-aging medicine.

Matteo Pagliardini [4], The recent tremendous success of unsuper vised word embeddings in a multitude of applications raises the obvious question if simi lar methods could be derived to improve em beddings (i.e. semantic representations) of word sequences as well. We present a simple but efficient unsupervised objective to train distributed representations of sentences. Our method outperforms the state-of-the-art unsupervised models on most benchmark tasks, highlighting the robustness of the produced general-purpose sentence embeddings.

Fenglong Ma [5], In this paper, we propose a novel and e,ective un supervised model Si•er to automatically discover drug side-e,ects. Si•er enhances the estimation on drug side-e,ects by learning from various online platforms and measuring platform-level and user-level quality simultaneously. In this way, Si•er demonstrates be‹er performance compared with existing approaches in terms of correctly identifying drug side-e,ects. Experimental results on €ve real-world datasets show that Si•er can signi€cantly improve the performance of identifying side-e,ects compared with the state-of the-art approaches.

GanjarAlfian [6], In this study, we propose a personalized healthcare monitoring system by utilizing a BLE-based sensor device, real-time data processing, and machine learning-based algorithms to help diabetic patients to better self-manage their chronic condition. BLEs were used to gather users' vital signs data such as blood pressure, heart rate, weight, and blood glucose (BG) from sensor nodes to smartphones, while real-time data processing was utilized to manage the large amount of continuously generated sensor data.

MuhammadFazalIjaz [7], The CCPM first removes outliers by using outlier detection methods such as density-based spatial clustering of applications with noise (DBSCAN) and isolation forest (iForest) and by increasing the number of cases in the dataset in a balanced way, for example, through synthetic minority over-sampling technique (SMOTE) and SMOTE with Tomek link (SMOTETomek). Finally, it employs random forest (RF) as a classifier.

Parvathaneni Naga Srinivasu [8], Deep learning models are efficient in learning the features that assist in understanding complex patterns precisely. This study proposed a computerized process of classifying skin disease through deep learning based MobileNet V2 and Long Short Term Memory (LSTM). The MobileNet V2 model proved to be efficient with a better accuracy that can work on lightweight computational devices. The proposed model is efficient in maintaining stateful information for precise predictions.

**Proposed System**

The proposed system is an AI-driven Clinical Decision Support System that integrates multiple healthcare data sources to provide accurate disease diagnosis. It collects patient data such as symptoms, medical history, laboratory results, and demographic information from Electronic Health Records and other sources.

The system applies preprocessing techniques to clean and normalize the data, followed by feature extraction using NLP and word embedding methods. A machine learning model is then trained to learn semantic relationships between

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 4, April 2026

|DOI: 10.15680/IJIRSET.2026.1504154|

symptoms and diagnoses. The system calculates similarity scores between patients using cosine similarity and identifies the most similar cases. Based on this similarity, it predicts the most probable diagnosis for a new patient.

This approach enables personalized diagnosis, improves accuracy, and supports physicians in clinical decision-making.

## III. SYSTEM ARCHITECTURE

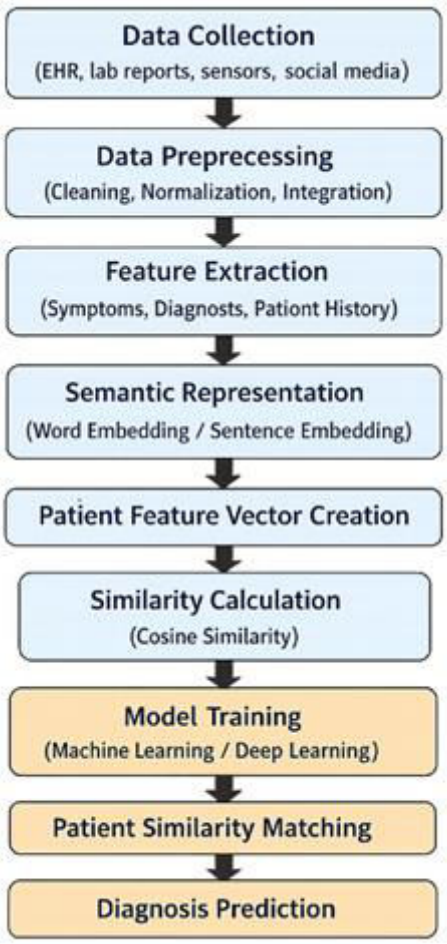



**Figure 1.System Architecture**

## IV. CONCLUSION

The proposed AI-driven Clinical Decision Support System effectively enhances disease diagnosis by utilizing patient similarity and semantic analysis. By integrating heterogeneous healthcare data and applying machine learning and NLP techniques, the system provides accurate and meaningful diagnostic predictions. Experimental results demonstrate improved performance in terms of precision and recall. The system reduces dependency on manual diagnosis, supports healthcare professionals, and enables early detection of diseases, ultimately improving patient outcomes.

## V. FUTURE SCOPE

The system can be further improved by incorporating real-time data from wearable devices and IoT-based health monitoring systems. Advanced deep learning models such as transformers can be used to enhance prediction accuracy. Integration with cloud computing can enable large-scale deployment and real-time analysis. Additionally, the system

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 4, April 2026

|DOI: 10.15680/IJIRSET.2026.1504154|

can be extended to support treatment recommendation, disease progression prediction, and personalized healthcare planning. Expanding the dataset and including multi-modal data such as medical images can further improve system performance and reliability.

## REFERENCES

1. P. N. Srinivasu, J. G. Sivasai, M. F. Ijaz, A. K. Bhoi, W. Kim, and J. J. Kang, ‘‘Classification of skin disease using deep learning neural networks with MobileNet V2 and LSTM,’’ Sensors, vol. 21, no. 8, p. 2852, Apr. 2021. [Online]. Available: https://www.mdpi.com/1424-8220/21/8/2852
2. M. F. Ijaz, M. Attique, and Y. Son, ‘‘Data-driven cervical cancer prediction model with outlier detection and over-sampling methods,’’ Sensors, vol. 20, no. 10, p. 2809, May 2020. [Online]. Available: https://www.mdpi.com/1424-8220/20/10/2809
3. G. Alfian, M. Syafrudin, M. Ijaz, M. Syaekhoni, N. Fitriyani, and J. Rhee, ‘‘A personalized healthcare monitoring system for diabetic patients by utilizing BLE-based sensors and real-time data processing,’’ Sensors, vol. 18, no. 7, p. 2183, Jul. 2018.
4. K. Bruynseels, F. S. di Sio, and J. van den Hoven, ‘‘Digital twins in health care: Ethical implications of an emerging engineering paradigm,’’ Frontiers Genet., vol. 9, p. 31, Feb. 2018. [Online]. Available: https://www.frontiersin.org/article/10.3389/fgene.2018.00031
5. M. Pagliardini, P. Gupta, and M. Jaggi, ‘‘Unsupervised learning of sentence embeddings using compositional N-gram features,’’ in Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics: Hum. Lang. Technol., vol. 1, 2018, pp. 528–540.
6. S. Arora, Y. Liang, and T. Ma, ‘‘A simple but tough-to-beat baseline for sentence embeddings,’’ in Proc. Int. Conf. Learn. Represent. (ICLR), 2017, pp. 1–14.
7. F. Ma, C. Meng, H. Xiao, Q. Li, J. Gao, L. Su, and A. Zhang, ‘‘Unsupervised discovery of drug side-effects from heterogeneous data sources,’’ in Proc. 23rd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, New York, NY, USA, Aug. 2017, pp. 967–976, doi: 10.1145/3097983.3098129
8. H. Xiao, L. Vu, and D. Turaga, ‘‘Learning temporal state of diabetes patients via combining behavioral and demographic data,’’ in Proc. 23rd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, Aug. 2017, pp. 2081–2089.

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi cross ref

निस्केयर NISCAIR

# INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

## IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

www.ijirset.com

Scan to save the contact details