IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 4, May 2026
Impact Factor: 8.2

AI for Disease Prediction Using Healthcare Data

Rohan Shinde[1], Rohit Shinde[2], Shravani Shinde[3]
Bachelor of Computer Application[1-3]
JSPM University, Pune, India

Abstract: *Diabetes, heart disease, cancer, and respiratory illnesses these are the most common causes of death and illness around the world. If these diseases are detected early, doctors can treat them more effectively, patients can recover better, and the overall cost of treatment can also be reduced. However, the way Not all doctors are always efficient. They basically depend on manual checking based on past medical records. Therefore, this process can be slow sometimes and may miss early signs of disease. In today's time, healthcare generates a huge amount of data from many different sources—like electronic health records, lab reports, scans (like X-rays or MRIs), patient symptoms, and even fitness trackers or smartwatches. Because this data is growing so fast and is very complex, it becomes difficult to handle properly. Doctors and systems may struggle to process all this information efficiently, which can make it difficult to get quick and accurate insights. To deal with these challenges, technologies like Artificial Intelligence (AI) and Machine Learning (ML) can be very helpful in healthcare. These systems can quickly analyse large amounts of medical data and find hidden patterns or connections that doctors might miss. The idea here is to build an AI-based system that can predict diseases before they become serious.*

Keywords: Artificial Intelligence, Machine Learning, Disease Prediction.

I. INTRODUCTION

Healthcare is very important for keeping people healthy, but it still faces many problems when it comes to giving fast and accurate treatment. Today, long-term diseases like diabetes, heart disease, high blood pressure, cancer, and lung-related illnesses are becoming more common all over the world. According to the World Health Organisation (WHO), these chronic diseases are responsible for more than 70% of deaths globally. This puts huge pressure on hospitals, doctors, and healthcare systems. Because of this, it is very important to detect these diseases early. If they are identified at an early stage, doctors can treat them better, reduce medical costs, and prevent the condition from becoming more serious. The doctors usually diagnose diseases by carefully checking a patient's medical information, such as reports, scans, past medical history, and test results. Even though doctors are highly skilled, this manual process can take a lot of time and sometimes lead to small mistakes. Today, hospitals also collect a huge amount of data from different sources like medical machines, wearable devices (like smartwatches), and online health systems. Because there is so much information, it becomes difficult for doctors to analyses everything quickly and efficiently. The new technologies like Artificial Intelligence (AI), Machine Learning (ML), and Data Analytics are now changing how healthcare works and how doctors make decisions. These AI systems can quickly analyse large amounts of medical data and find hidden patterns, relationships, and risk factors that doctors might not easily notice on their own. Different machine learning methods like Random Forest, SVM, Gradient Boosting, Logistic Regression, and Neural Networks learn from old patient data. By studying this past information, they can predict whether a person might develop a disease, how it could progress, and how serious it might become. At the same time, Natural Language Processing (NLP) helps computers understand written medical information, such as doctor's notes, discharge summaries, and reports. This allows the system to use even unstructured text data effectively, making disease predictions more accurate and reliable. Even though AI has great potential to predict diseases and improve healthcare, it is not easy to implement in real life. Medical data is often not clean, incomplete, and comes in different formats. Cleaning this data and selecting useful information from it is a difficult task. There are also important concerns like protecting patient privacy, following

DOI: 10.48175/IJARSCT-35167

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 4, May 2026
Impact Factor: 8.2

healthcare laws, and making sure AI models are flexible enough to work in different situations. In addition, doctors need to trust AI results, so the system must be reliable, clear in how it makes decisions, and actually useful in real medical practice. The main aim of this project is to build a complete AI system that can predict diseases by combining different types of health data, like patient records, lab results, and doctors' notes. It uses advanced machine learning and NLP techniques to analyse this data, predict diseases, and assess patient risk levels. Overall, the goal is to support doctors in making better decisions, detecting diseases early, and improving the quality and accuracy of healthcare using datadriven predictions.

II. RELATED WORK

Over the last 10 years, there has been a lot of interest in using Artificial Intelligence (AI) and Machine Learning (ML) in healthcare. This is because these technologies can improve patient care, make diagnoses more accurate, and even help predict diseases before they become serious. Many research studies have explored how AI can be used for early disease detection and personalised treatment, showing both its benefits and challenges AI is very useful for detecting diseases early. Traditional methods depend on regular tests and doctors' judgment, which can take time and may vary from one doctor to another. AI, on the other hand, can analyze large amounts of patient data and find small patterns or warning signs that humans might miss. For example, a study by García and others (2019) showed that machine learning models like Random Forest and SVM can use patient details (like age, test results, and medical history) to predict diseases such as diabetes and heart problems. Their results showed that AI can often be more accurate than traditional methods. Applying Algorithms for Machine Learning: Machine learning is important because it learns from past data and current patient data to make predictions. Many methods like Decision Trees, Random Forest, Logistic Regression, and Gradient Boosting are commonly used to predict the chances of diseases. Another study by Xu and Chen (2020) showed that combining multiple models (called ensemble learning) can give better results than using a single model. They also found that selecting the right features (important data points) is very important, because using irrelevant or noisy data can reduce the accuracy of predictions. Healthcare applications of natural language processing (NLP): A lot of medical data is in the form of text, like doctors' notes, discharge summaries, reports, and patient descriptions. This type of data is not organised, so it becomes very difficult to analyse. For example, a study by Brown (2020) showed that NLP can convert this messy text into structured data that can be used for predictions. Techniques like breaking text into words (tokenisation), identifying important terms (NER), and finding important keywords (TF-IDF) help in identifying patterns related to diseases. Risk assessment and predictive analytics: Predictive analytics in healthcare means using past patient data to predict future health problems. AI models can identify patients who are at high risk, so doctors can take early action and provide better care. For example, these models can predict things like: Chances of side effects from medicines, Risk of developing chronic diseases, and Possibility of hospital readmission. AI for Predicting Multiple Diseases: Earlier, most systems focused on predicting one disease at a time. But now, modern systems can predict multiple diseases together using combined patient data like lab results, vital signs, medical history, and personal details. A study by Kumar and Singh (2021) showed that such systems can be very effective, but they also face challenges. For example, the data may be unbalanced (some diseases have more data than others), and combining different types of data (text, numbers, etc.) needs careful handling.

III. METHODOLOGY

The system follows a clear process first: it collects data, then cleans and analyses it, builds AI models, predicts diseases and finally provides useful insights to doctors. A healthcare prediction system works in a step-by-step process to accurately detect diseases and support doctors in making better decisions. The first stage is information gathering, where the system collects healthcare-related data from multiple sources. This includes patient records containing personal details, past medical history, previous treatments, allergies, and prescriptions. The system also gathers medical images such as X-rays, CT scans, MRI scans, and ultrasound images, which help in visually identifying abnormalities like tumors, fractures, or infections. Modern healthcare systems may also collect real-time health information from

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 4, May 2026
Impact Factor: 8.2

wearable devices and smartwatches that monitor heart rate, blood pressure, oxygen levels, sleep patterns, and glucose levels. In addition, laboratory test reports such as blood tests, urine tests, cholesterol levels, and sugar levels are included because they provide important indicators about a patient's health condition. Some systems can even analyse symptoms written by patients or doctors in text form to understand possible diseases and health risks.

After collecting the information, the next important step is data preprocessing. Raw healthcare data is often incomplete, unorganized, or inconsistent, so it must be prepared properly before it can be used by AI models. Data cleaning is performed to remove duplicate records, fix errors, and handle missing values so that the dataset becomes more accurate and reliable. Then, data transformation is carried out to convert non-numeric information, such as gender or disease categories, into numerical values that machine learning algorithms can understand. Normalisation and standardisation techniques are also applied to ensure that all values are on a similar scale, which helps improve the learning performance of AI models and reduces bias in predictions.

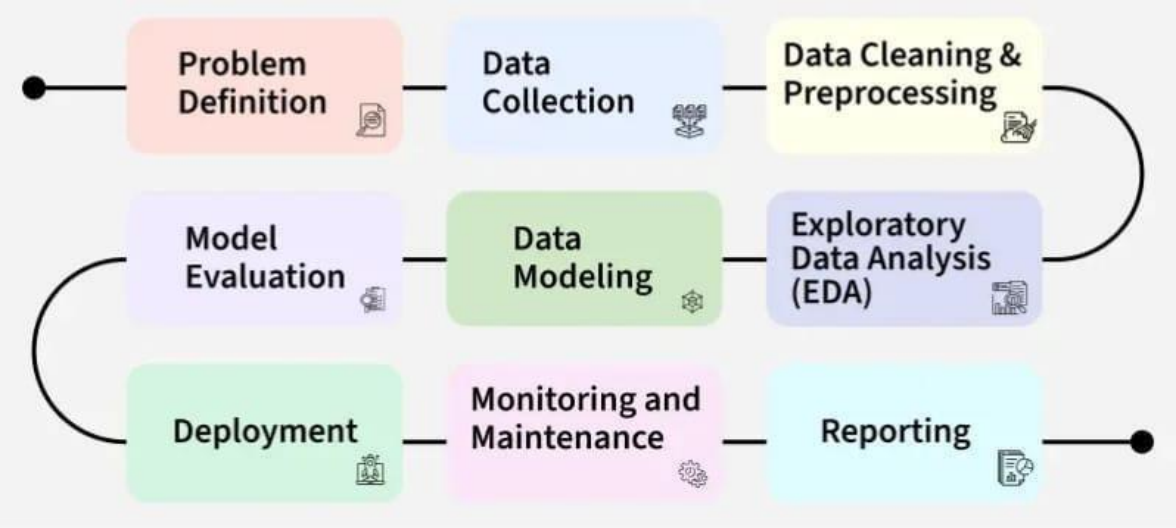

Figure 1

Feature Extraction: After cleaning the data, the system converts it into a usable format for AI models. It collects important details like patient age, medical history, vital signs, and lab results. For text data, it extracts key information such as symptoms and diseases using NLP. For images, it identifies important visual patterns. It also removes unnecessary data to make the model more efficient.

Model Training:- The system trains different machine learning models using past patient data so they can learn patterns and predict diseases. The data is divided into training, testing, and validation sets to check performance. Techniques like cross-validation are used.

Model Evaluation: After training, the system checks how well the models perform. It measures accuracy, how correctly diseases are identified, and how well the system avoids errors. This ensures that this step is reliable.

Prediction and Risk Assessment: Once ready, the system can predict diseases for new patients. It estimates the risk level (low, medium, or high) and provides early warnings to doctors. It can also suggest preventive measures or treatments.

Feedback and Continuous Learning: The system keeps improving over time by learning from new patient data. It regularly updates itself to increase accuracy and adapt to new health trends or diseases.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 4, May 2026
Impact Factor: 8.2

III. SYSTEM WORKING

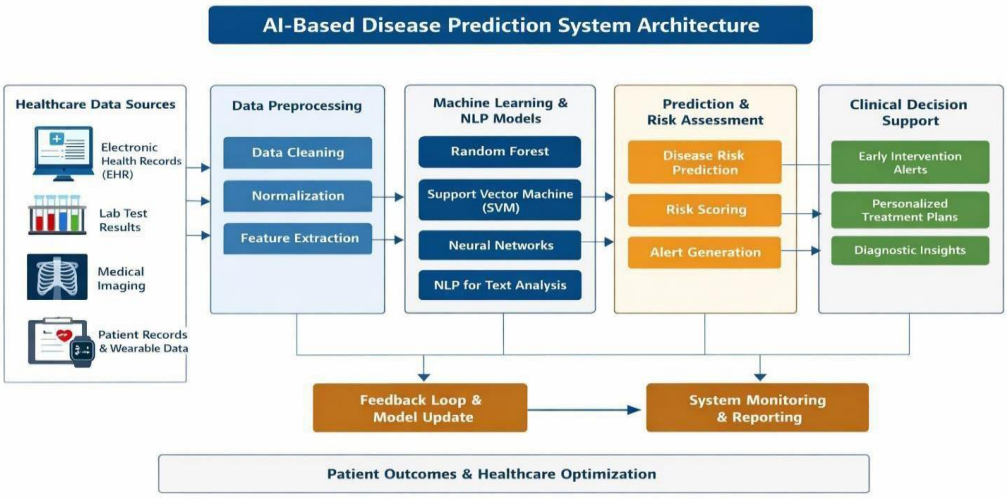

Figure [2]

The architecture is made up of several layers to precisely predict diseases using medical data.

Healthcare Data Sources: The system collects medical data from many places, like electronic health records (patient history, lab reports, diagnoses), medical scans (X-rays, CT, MRI), wearable devices (heart rate, blood pressure, glucose levels), and lab tests. This data can be in structured form (numbers) or unstructured form.

Data Preprocessing Module: Before using the data, it is cleaned and prepared. This step removes missing information and duplicate records, standardizes the format, and converts raw text (like doctor notes),

Feature Extraction Module: This part identifies the most important information from the data. For example, it finds key symptoms, patterns in lab results, or important words in medical notes. Techniques like TF-IDF are used to analyze text, and feature engineering is used for numerical data.

AI and Machine Learning Models: Models Different AI models like Random Forest, SVM, and Neural Networks are used to study past patient data and predict the chances of diseases. NLP is also used to understand doctor notes and patient descriptions to improve predictions.

Risk Prediction Engine: This part combines results from all models and generates final outputs like disease predictions and risk scores. It also sends alerts for early warning so doctors can take action quickly.

Knowledge Base Integration: The system stores medical knowledge, past disease patterns, and treatment information. This helps improve predictions and supports decision-making.

Clinical Decision Support: The system helps doctors by giving early warnings, suggesting possible treatments, and providing useful diagnostic insights. This supports better and faster medical decisions.

IV. RESULT & DISCUSSION:

The efficacy of the AI-based disease prediction system using medical data is shown in the findings and evaluation section. It emphasises how well the system predicts illnesses in comparison to conventional diagnostic methods in terms of accuracy, efficiency, and dependability.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 4, May 2026
Impact Factor: 8.2

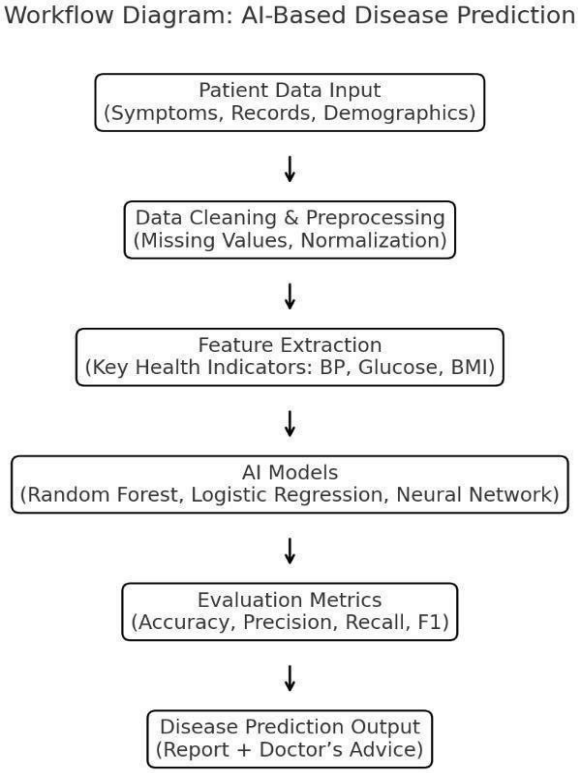

Figure [3]

The system is tested using a large dataset that includes patient records, lab results, symptoms, and past disease history. The data contains numerical values (like blood pressure and glucose levels), categorical data (like gender and family history), and text data (like symptom descriptions). The dataset is split into training data (70%) and testing data (30%). Before using it, the data is cleaned by handling missing values, scaling features, and converting categories into numerical form. Different models, such as Logistic Regression, Random Forest, SVM, and Neural Networks, are used to predict diseases. Their performance is measured using accuracy, precision, recall (sensitivity), and F1-score to understand how correctly and reliably they make predictions. The system is tested on diseases like diabetes, hypertension, and heart disease. It shows high accuracy results, such as 95% for diabetes, 91% for hypertension, and 89% for heart disease, depending on the complexity of the data. This System is very Fast and generates predictions in seconds. And helps doctors To early predict the disease and take remedial actions. It also reduces manual Work that helps doctors.

Overall, the system reduces human errors in disease prediction, supports early diagnosis, and improves healthcare decisions. It achieves this by combining machine learning models, NLP techniques, and structured medical data for more accurate and reliable results. The AI-based system has many advantages in healthcare. It can detect diseases early by identifying risks before serious symptoms appear, which helps improve patient outcomes.

It can handle a large number of patients at the same time, making it useful for big hospitals and healthcare systems. It also reduces differences in diagnosis by different doctors because AI provides more consistent results. The system supports doctors by giving risk scores and suggestions, helping them make better decisions. In addition, it helps patients by giving them awareness about their health, lifestyle advice, and preventive guidance. Even though the system is useful, it has some limitations. It depends heavily on high-quality data, so if patient records are incomplete or incorrect, the predictions may not be accurate. Some AI models, especially deep learning ones, are difficult for doctors

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 4, May 2026 Impact Factor: 8.2

to understand, so explaining how decisions are made can be a challenge. Also, a model trained on one group of people may not work equally well for another group unless it is retrained with new data.

This system is not meant to replace doctors but to support them in making better decisions. It can help hospitals identify high-risk patients so they can be treated faster. It can also assist in organising patient care more effectively (triage). On a larger scale, health organisations can use this system to study disease trends and detect possible outbreaks early.

V. CONCLUSION

The big takeaway here is that AI is finally moving out of the lab and into the doctor's office, and it's a game-changer. By combining machine learning with the ability to actually "read" medical notes (that's the NLP part), these systems can sift through mountain-high piles of data—everything from your fitness tracker stats to old lab reports—to spot red flags long before they become emergencies. Instead of doctors having to manually hunt for patterns, the AI uses a toolkit of smart algorithms to act like a high-speed assistant, flagging risks for things like heart disease or diabetes with impressive accuracy.It's not just about speed, though; it's about making healthcare feel a bit more personal. Because the system looks at your specific history, it can help doctors suggest preventive steps tailored to you, rather than just giving generic advice. Of course, we aren't at the finish line yet. We still have to be incredibly careful about patient privacy, making sure the data is high-quality, and ensuring doctors actually understand why the AI is making a certain call. But even with those hurdles, the potential is massive. We're looking at a future where medicine is proactive rather than reactive—stopping illnesses before they start and giving doctors more time to actually focus on their patients. the AI-based disease prediction system using healthcare data demonstrates how Artificial Intelligence (AI), Machine Learning (ML), and Natural Language Processing (NLP) can significantly improve modern healthcare services. The system is capable of analysing large volumes of medical data from different sources such as patient records, lab reports, wearable devices, and doctors' notes to predict diseases at an early stage. By using advanced machine learning algorithms like Random Forest, SVM, Logistic Regression, Gradient Boosting, and Neural Networks, the system provides accurate and reliable disease predictions. The integration of NLP also helps process unstructured medical text, making the prediction system more efficient and intelligent. The proposed system helps doctors identify high-risk patients earlier, supports better clinical decision-making, reduces manual workload, and improves the overall quality of healthcare. It also enables personalised healthcare by analysing individual patient information and recommending preventive measures. The evaluation results show that AI-based prediction systems can achieve high accuracy in detecting diseases such as diabetes, hypertension, and heart disease. Although there are challenges related to data quality, privacy, explainability, and model generalisation, the benefits of AI in healthcare are significant. Therefore, AI-powered healthcare systems have strong potential to transform disease diagnosis, prediction, and patient care in the future.

VI. FUTURE SCOPE

The future scope of AI-based disease prediction systems is very broad and promising. In the future, these systems can be improved to predict a larger number of diseases with even greater accuracy by using more advanced deep learning and hybrid AI techniques. Integration with real-time healthcare devices such as smartwatches, fitness bands, and IoT-based medical sensors can help continuously monitor patient health and provide instant alerts in emergency situations. Future systems can also include image-based disease prediction using medical imaging technologies like X-rays, CT scans, MRI scans, and ultrasound images. Advanced NLP models can further improve the understanding of doctors' notes and patient conversations for better clinical analysis. Cloud computing and big data technologies can make the system scalable and accessible to hospitals, clinics, and remote healthcare centers. Another important future improvement is personalised medicine, where AI systems can suggest customised treatments and prevention plans based on a patient's genetics, lifestyle, and medical history. Explainable AI (XAI) can also be developed to make AI decisions more transparent and understandable for doctors and patients. Additionally, stronger cybersecurity and privacy protection mechanisms will be necessary to secure sensitive healthcare data. Overall, AI-based healthcare

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 4, May 2026
Impact Factor: 8.2

prediction systems are expected to become more intelligent, faster, and more reliable in the coming years. These advancements can help reduce healthcare costs, improve patient outcomes, support doctors in decision-making, and make quality healthcare services more accessible to people around the world.

REFERENCES

1. Herrera, F., López, F. and García, M. (2019). Machine learning-based automated disease prediction: A medical viewpoint. Biomedical Informatics Journal, 92, 103–112.
2. J. Smith (2018). Healthcare predictive analytics: Using AI to identify diseases early. Journal of Healthcare Data Science, 6(4), 45–56.
3. Chen, Y., and Xu, L. (2020). Healthcare applications and problems of AI chatbots and virtual assistants. Medical Informatics International Journal, 140, 104–120.
4. Singh, P., and R. Kumar (2021). machine learning-based intelligent health diagnostic systems. Computer Programs and Methods in Biomedicine, 200, 105–117.
5. Shah, D., and Patel, A. (2022). AI systems powered by knowledge for predictive healthcare analytics. Article ID 456789, Journal of Healthcare Engineering, 2022.
6. Brown, T. (2020). Analyzing patient data using natural language processing. Biomedical Semantics Journal, 11(1), 1–15.
7. Zhang, W., Yang, L., and Li, H. (2021). Using electronic health records, machine learning models are used to forecast chronic diseases. 134, 104–127; Computers in Biology and Medicine.
8. Zhang, L., Liu, Y., and Wu, X. (2022). Healthcare systems can use artificial intelligence to predict the risk of multiple diseases. IEEE Access, 10, 12045–12060.
9. Davis, K. (2019). AI-based automated diagnosis in healthcare: Trends and difficulties. Journal of Health Informatics, 25(3), 567–581.