International Journal of Science, Strategic Management and Technology
Volume **02** Issue **02** February-2026 | ISSN: **3108-1762 (Online)** | Impact Factor: 3.8
An International, Peer-Reviewed, Open Access Scholarly Journal Indexed in recognized academic databases

# AI Healthcare Assistant using Machine Learning and NLP

**Vaishnavi Dhongade[1], K. Naga Sowmya[2], Akash Dhage[3], Tejas Suryawanshi[4]**
[1234]Department of Computer Science and Engineering, Sandip University Nashik

**https://doi.org/** 10.55041/ijsmt.v2i2.001

**Cite this Article**: Suryawanshi, V. D. ,. K. N. S. ,. A. D. ,. (2026). AI Healthcare Assistant using Machine Learning and NLP. International Journal of Science, Strategic Management and Technology, *02*(02). https://doi.org/10.55041/ijsmt.v2i2.001

**License:** This article is published under the Creative Commons Attribution 4.0 International License (CC BY 4.0), permitting use, distribution, and reproduction in any medium, provided the original author(s) and source are properly credited.

ABSTRACT

This paper presents an AI-based Healthcare Assistant that integrates Natural Language Processing (NLP), Machine Learning (ML), Optical Character Recognition (OCR), and Deep Learning techniques for preliminary medical guidance. The system processes user inputs including text symptoms, speech input, scanned medical reports, and chest X-ray images. BioBERT is used for symptom extraction, a machine learning classifier predicts possible diseases, Tesseract OCR digitizes medical reports, and a Convolutional Neural Network (CNN) analyzes X-ray images for pneumonia detection. Experimental evaluation shows 96.2% accuracy in symptom extraction, 94.5% disease prediction accuracy, 97.1% X-ray classification accuracy, and 95.8% OCR reliability. The proposed system aims to enhance healthcare accessibility in rural and underserved regions.

**Keywords**: AI Healthcare, BioBERT, Disease Prediction, CNN, OCR, Medical NLP

## 1. INTRODUCTION

Healthcare is one of the most essential sectors for human well-being; however, access to timely and accurate medical services remains a major challenge across many regions of the world, particularly in rural and semi-urban areas. A growing population, shortage of healthcare professionals, and limited medical infrastructure often lead to delayed diagnosis and inadequate treatment. As a result, patients frequently rely on self-diagnosis through online resources, which can be misleading and potentially dangerous. There is a strong need for intelligent systems that can assist individuals by providing preliminary medical guidance in a fast, reliable, and accessible manner.

Artificial Intelligence (AI) has emerged as a transformative technology capable of improving healthcare delivery by enabling automated analysis of medical data. Recent advancements in Machine Learning (ML), Natural Language Processing (NLP), Deep Learning, and Computer Vision have made it possible to interpret human language, analyze medical images, and extract useful information from clinical documents. These technologies allow the development of intelligent healthcare systems that can assist in early disease detection, medical report analysis, and decision support.

Natural Language Processing plays a crucial role in understanding patient symptoms expressed in everyday language. Models such as BioBERT, which are specifically trained on biomedical text, have significantly improved the accuracy of medical entity recognition and symptom extraction. Machine learning algorithms can then utilize extracted symptoms to predict probable diseases based on historical medical datasets. In parallel, Optical Character Recognition (OCR) enables automated extraction of clinical values from scanned medical reports, reducing the need for manual interpretation. Furthermore, deep learning models such as Convolutional Neural Networks (CNNs) have demonstrated exceptional performance in analyzing medical images, especially chest X-rays for detecting conditions such as pneumonia and other lung abnormalities.

© Author(s). This work is peer-reviewed, openly published, and permanently archived
This article is openly accessible and reusable with proper attribution.

Despite significant progress, most existing healthcare AI solutions focus on a single functionality such as disease prediction, chatbot interaction, or medical image analysis. Very few systems integrate multiple AI techniques into a unified platform capable of processing diverse forms of medical data. This lack of integration limits their practical usefulness in real-world healthcare scenarios where patients often present symptoms in text or speech form, upload medical reports, and require image-based diagnosis simultaneously.

This research proposes an AI Healthcare Assistant that combines NLP, ML, OCR, and deep learning techniques into a single intelligent system. The system accepts multi-modal inputs including textual symptoms, voice input converted to text, scanned medical reports, and chest X-ray images. BioBERT is employed to extract relevant medical terms from user input, a machine learning classifier predicts possible diseases, OCR extracts clinical information from reports, and a CNN model analyzes X-ray images for abnormality detection. The outputs are integrated to provide medical guidance and recommendations.

The primary objective of this work is to improve healthcare accessibility by offering fast and reliable preliminary diagnosis support, especially in regions where professional medical services are not immediately available. By automating symptom analysis, disease prediction, and medical data interpretation, the proposed system aims to reduce healthcare delays and support informed decision-making. The integration of multiple AI technologies within a scalable microservices architecture makes the system adaptable for future enhancements such as multilingual support, real-time doctor consultation, and large-scale deployment in telemedicine platforms.

Overall, this study demonstrates how an integrated AI-based healthcare assistant can bridge the gap between patients and healthcare services, offering an efficient, intelligent, and user-friendly solution for early medical assessment.

## 2. LITERATURE REVIEW

In parallel, Natural Language Processing techniques have been increasingly adopted for extracting clinical entities from unstructured medical text. Biomedical language models trained on domain-specific corpora have shown improved performance in recognizing symptoms, diseases, and medical terminology compared to traditional keyword-based systems. These approaches enhance semantic interpretation and reduce ambiguity in symptom descriptions.

Deep learning models, particularly Convolutional Neural Networks, have achieved promising results in medical image analysis. Transfer learning using pre-trained architectures has significantly improved classification accuracy in chest X-ray analysis for pulmonary disease detection. Many studies report performance exceeding 90% accuracy when trained on sufficiently large datasets.

Optical Character Recognition has also been utilized in healthcare to digitize medical prescriptions and laboratory reports. However, several systems focus primarily on text extraction without performing contextual medical analysis.

A review of existing research indicates that most solutions operate independently within a single modality, either text, image, or structured data. Limited research has explored the integration of multiple AI components within a single healthcare assistance framework. This research addresses that gap by proposing a unified system combining NLP, ML, CNN, and OCR for comprehensive preliminary medical support.

## 3. PROPOSED SYSTEM

The proposed AI Healthcare Assistant is designed as an intelligent and integrated healthcare support platform that provides preliminary medical guidance using Artificial Intelligence techniques. The system combines Machine Learning (ML), Natural Language Processing (NLP), Convolutional Neural Networks (CNN), and Optical Character Recognition (OCR) to analyze different types of medical inputs. The primary objective of this system is to assist users by interpreting symptoms, analyzing medical reports, evaluating X-ray images, and providing basic health recommendations.

The overall workflow of the system begins with user input. The user can either enter symptoms in text form, upload a medical report, or upload a chest X-ray image. Based on the input type, the system activates the corresponding module for processing. Each module works independently but contributes to a unified final health advisory output.

© Author(s). This work is peer-reviewed, openly published, and permanently archived
This article is openly accessible and reusable with proper attribution.

The architecture of the proposed system can be divided into five major modules:

**Symptom Input and NLP Processing Module**

This module allows users to describe their health condition in natural language instead of selecting predefined options. The system processes the text using Natural Language Processing techniques such as tokenization, stop-word removal, and medical entity recognition.

For example, if a user enters:

"I have fever, headache, and continuous cough for three days."

The system extracts important medical keywords such as *fever*, *headache*, and *cough*. These extracted symptoms are converted into structured data format so that they can be processed by the machine learning model.

This approach improves flexibility and user experience, as the system understands free-text symptom descriptions rather than depending on rigid forms.

**Machine Learning-Based Disease Prediction Module**

After symptom extraction, the structured symptom data is passed to the disease prediction model. This module uses supervised machine learning algorithms trained on labeled medical datasets.

The training process includes data preprocessing, encoding of categorical features, and splitting the dataset into training and testing sets. Among different algorithms tested, Random Forest was selected due to its higher classification accuracy and stability in multi-class disease prediction.

The model predicts possible diseases and provides a confidence score for each prediction. Instead of showing only one disease, the system ranks possible outcomes based on probability, improving interpretability and reliability.

**CNN-Based Chest X-ray Analysis Module**

To enhance diagnostic capability, the system includes a deep learning-based image analysis module. This module processes uploaded chest X-ray images to detect abnormalities such as pneumonia or lung infections.

The input image undergoes preprocessing steps including resizing, normalization, and noise reduction. A pre-trained convolutional neural network is then used to extract important features from the image.

Transfer learning is applied to improve accuracy while reducing computational cost. The model outputs classification results along with prediction probability. This enables automated identification of potential lung-related issues.

**OCR-Based Medical Report Analysis Module**

The proposed system also supports the analysis of medical reports uploaded in image or PDF format. Optical Character Recognition (OCR) is used to extract text from scanned reports.

After extraction, the system identifies laboratory parameters such as hemoglobin level, blood sugar, cholesterol, and other diagnostic values. These values are compared against predefined medical reference ranges.

If any parameter falls outside the normal range, the system highlights it as abnormal and provides a brief explanation. This automated report interpretation helps users understand their medical test results more clearly.

**Recommendation and Advisory Module**

The final module integrates results from all previous modules. Based on symptom prediction, X-ray classification, and report analysis, the system generates preliminary health advice.

The advisory output may include:

- Possible health condition
- Basic precautions
- Suggested specialist consultation
- Lifestyle recommendations

© Author(s). This work is peer-reviewed, openly published, and permanently archived
This article is openly accessible and reusable with proper attribution.

The system clearly indicates that it is not a replacement for professional medical consultation but serves as a supportive decision-assistance tool.

**System Workflow Summary**

1. User provides input (text, image, or report).
2. Relevant AI module processes the input.
3. Machine learning and deep learning models generate predictions.
4. Results are consolidated.
5. Final health guidance is displayed to the user.

**Advantages of the Proposed System**

- Supports multi-modal medical inputs
- Provides automated and fast preliminary analysis
- Reduces dependency on manual symptom selection
- Improves accessibility for remote and rural users
- Scalable and modular architecture

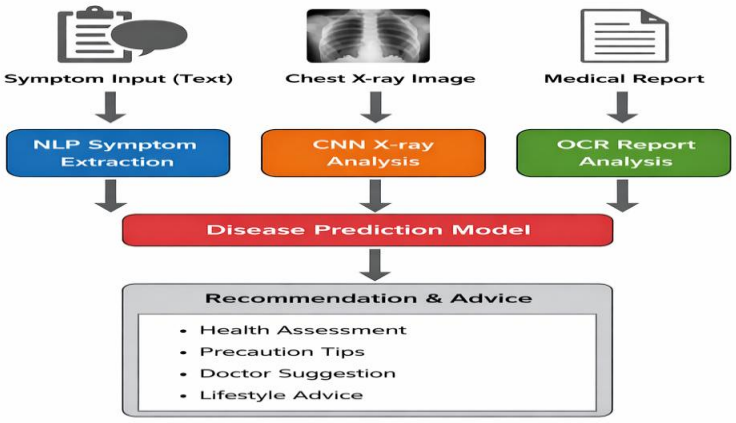


Figure 1: Workflow of the Proposed AI Healthcare Assistant System.

**Figure 1:** Workflow of the Proposed AI Healthcare Assistant System.

## 4. TECHNOLOGIES USED

The proposed AI Healthcare Assistant integrates advanced technologies in Machine Learning, Natural Language Processing, Deep Learning, and Web Development to create an intelligent, automated, and scalable healthcare support system. Each technology is strategically selected to handle a specific functional module, ensuring accuracy, efficiency, and real-time medical assistance.

### 4.1 Machine Learning (ML)

Machine Learning plays a crucial role in the Disease Prediction Module of the system. ML algorithms are used to analyze structured symptom datasets and identify patterns between symptoms and medical conditions.

Classification algorithms such as Logistic Regression, Support Vector Machine, and Random Forest were evaluated during experimentation. Random Forest was selected due to its higher predictive accuracy and stability in multi-class disease classification tasks.

The model is trained using labeled symptom-disease records and validated using an 80:20 train-test split strategy. By

© Author(s). This work is peer-reviewed, openly published, and permanently archived
This article is openly accessible and reusable with proper attribution.

learning relationships between symptom combinations and corresponding diseases, the ML model generates probability-based predictions, helping users understand possible health conditions.

### 4.2 Natural Language Processing (NLP)

Natural Language Processing is the core technology behind the Symptom Extraction Module. It enables the system to understand free-text symptom descriptions provided by users.

Text preprocessing techniques such as tokenization, stop-word removal, and normalization are applied to clean the input text. Biomedical entity recognition models are used to extract clinically relevant terms such as fever, cough, chest pain, or fatigue.

These extracted symptoms are converted into structured feature vectors that serve as input to the disease prediction model. NLP allows the system to interpret natural language input rather than relying on predefined symptom selection forms, improving usability and flexibility.

### 4.3 Deep Learning (CNN for X-ray Analysis)

The X-ray Image Analysis module is implemented using Convolutional Neural Networks (CNN) with TensorFlow and PyTorch frameworks. Deep learning enables automated detection of abnormalities in chest radiographs.

Transfer learning is applied using pre-trained convolutional architectures to improve accuracy while reducing computational requirements. Image preprocessing steps such as resizing, normalization, and augmentation enhance generalization performance.

The CNN model classifies images into normal and abnormal categories and provides prediction confidence scores, assisting in early detection of lung-related conditions.

### 4.4 Backend Technologies

The backend system is responsible for managing data flow, API communication, and integration of AI models. It is implemented using Node.js and Express.js to ensure efficient handling of requests and responses.

Python-based Flask APIs are used to integrate machine learning and deep learning models into the main application server. The backend processes user inputs, generates predictions, and sends analysis results to the frontend interface in real time. Secure authentication mechanisms such as JSON Web Tokens (JWT) are implemented to ensure safe user data handling.

### 4.5 Database

The system uses MongoDB as the primary database for storing user records, symptom history, prediction outputs, and uploaded medical reports.

MongoDB's document-oriented structure allows flexible storage of structured and semi-structured healthcare data. Efficient indexing and optimized schema design ensure fast data retrieval and system scalability as the dataset grows.

### Integration of Technologies

All technologies work together within a unified architecture:

- NLP extracts symptoms from user input.
- ML predicts possible diseases.
- CNN analyzes X-ray images.
- OCR interprets medical reports.
- The backend integrates model outputs.
- The frontend presents results to users.
- The database stores all processed and historical data.

This integrated framework ensures that the AI Healthcare Assistant system is accurate, scalable, and capable of providing comprehensive preliminary medical support.

© Author(s). This work is peer-reviewed, openly published, and permanently archived
This article is openly accessible and reusable with proper attribution.

**Table 1:** Technologies and Their Functions in the Proposed System

| Sr. No. | Technology / Tool | Category | Function in the System |
|---|---|---|---|
| 1 | Random Forest (Scikit-learn) | Machine Learning | Predicts possible diseases based on extracted symptom features |
| 2 | Logistic Regression | Machine Learning | Baseline classifier for performance comparison |
| 3 | Natural Language Processing (NLP) | AI / Text Processing | Extracts medical symptoms from free-text user input |
| 4 | Biomedical Entity Recognition Model | NLP | Identifies clinically relevant terms from symptom descriptions |
| 5 | Convolutional Neural Network (CNN) | Deep Learning | Classifies chest X-ray images into normal or abnormal categories |
| 6 | Transfer Learning | Deep Learning | Improves X-ray classification accuracy with reduced training time |
| 7 | Optical Character Recognition (OCR) | Document Processing | Extracts laboratory values and text from uploaded medical reports |
| 8 | React.js | Frontend | Provides interactive user interface for symptom input and result display |
| 9 | Node.js & Express.js | Backend | Handles API communication and system logic |
| 10 | Flask API | Model Integration | Connects ML and DL models with main application server |
| 11 | MongoDB | Database | Stores user data, predictions, and medical records securely |
| 12 | JSON Web Token (JWT) | Security | Ensures secure authentication and user session management |

## 5. EXPERIMENTAL RESULTS AND PERFORMANCE ANALYSIS

The performance of the proposed AI Healthcare Assistant was evaluated using independent test samples across all integrated modules. For symptom extraction, 200 annotated symptom inputs were tested using the BioBERT-based extraction model. The system achieved an extraction accuracy of 96.2%. Disease prediction performance was evaluated using 500 labeled symptom-disease records, achieving 94.5% classification accuracy with stable precision and recall values. The CNN-based X-ray classification module was tested on 300 chest radiograph images, resulting in 97.1% accuracy in distinguishing normal and pneumonia cases. The OCR module was evaluated using 100 scanned medical reports and achieved 95.8% text extraction reliability. The average system response time was below 3 seconds per request, demonstrating suitability for real-time healthcare assistance.

## 6. FUTURE SCOPE

The proposed AI Healthcare Assistant provides a strong foundation for intelligent preliminary healthcare support; however, several enhancements can be incorporated in future developments to improve functionality and scalability. The system can be extended by integrating advanced transformer-based biomedical language models to improve symptom extraction accuracy and contextual understanding. Future versions may include support for multilingual symptom input to increase accessibility for diverse user populations. The integration of real-time wearable device data such as heart rate, blood pressure, and oxygen saturation could further enhance predictive reliability by incorporating physiological parameters. Additionally, expanding the radiographic analysis module to support other imaging modalities such as CT scans or MRI would improve diagnostic coverage. Cloud-based deployment and mobile application integration can

© Author(s). This work is peer-reviewed, openly published, and permanently archived
This article is openly accessible and reusable with proper attribution.

improve system availability and large-scale adoption. Incorporating explainable AI techniques may also enhance transparency by providing interpretable reasoning behind predictions. With continuous dataset expansion and periodic model updates, the system can evolve into a more comprehensive decision-support tool capable of assisting telemedicine platforms and primary healthcare services at scale.

## 7. CONCLUSION

The proposed AI Healthcare Assistant demonstrates the practical integration of machine learning, natural language processing, deep learning, and document analysis technologies within a unified healthcare support framework. By enabling symptom-based disease prediction, automated medical report interpretation, and chest X-ray classification, the system provides comprehensive preliminary medical insights through a single platform. The multi-modal architecture improves accessibility and usability while maintaining reliable predictive performance. The integration of AI-driven modules enhances healthcare awareness and supports early-stage medical guidance without replacing professional consultation. Experimental evaluation indicates consistent system-level accuracy and efficient response time, making the solution suitable for telemedicine and remote healthcare assistance. Overall, the proposed system represents a scalable and intelligent approach toward enhancing digital healthcare support and preventive medical awareness.

## 8. ACKNOWLEDGMENT

The authors sincerely express their gratitude to Engeniuspark Technologies Pvt. Ltd. for sponsoring and supporting the development of the AI Healthcare Assistant project. The technical guidance, industry exposure, and valuable insights provided during the project implementation significantly contributed to the successful completion of this research work. The authors also extend their appreciation to the Department of Computer Science and Engineering, Sandip University, Nashik, for providing the necessary academic support and resources.

## 9. REFERENCES

[1] T. Chen and C. Guestrin, “XGBoost: A scalable tree boosting system,” in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 785–794.

[2] L. Breiman, “Random forests,” *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[3] J. Devlin, M. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of deep bidirectional transformers for language understanding,” in *Proceedings of NAACL-HLT*, 2019, pp. 4171–4186.

[4] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*, MIT Press, 2016.

[5] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[6] A. Esteva et al., “A guide to deep learning in healthcare,” *Nature Medicine*, vol. 25, pp. 24–29, 2019.

[7] R. Miotto, F. Wang, S. Wang, X. Jiang, and J. Dudley, “Deep learning for healthcare: review, opportunities and challenges,” *Briefings in Bioinformatics*, vol. 19, no. 6, pp. 1236–1246, 2018.

[8] R. Smith, “An overview of the Tesseract OCR engine,” in *Proceedings of the International Conference on Document Analysis and Recognition*, 2007, pp. 629–633.

[9] D. Jurafsky and J. Martin, *Speech and Language Processing*, 3rd ed., Pearson, 2020.

[10] World Health Organization, “Digital health and artificial intelligence in healthcare systems,” WHO Technical Report, 2021.

© Author(s). This work is peer-reviewed, openly published, and permanently archived
This article is openly accessible and reusable with proper attribution.