IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Multiple Disease Prediction Using Artificial Intelligence And Machine Learning

Mrs. Manasa M, Sneha Bharti, Dhanalakshmi K, Monika M

Assistant Professor, Student, Student, Student

Department of Artificial Intelligence and Machine Learning

Vijaya Vittala Institute of Technology, Bengaluru, India

Abstract: The rapid evolution of Artificial Intelligence (AI) and Machine Learning (ML) has brought transformative advancements in healthcare. This paper presents an AI-powered web application capable of predicting multiple diseases, including Heart Disease, Parkinson's Disease, Diabetes, Jaundice, Liver Conditions, and Hepatitis. Leveraging supervised machine learning algorithms such as XGBoost and Random Forest, the system analyzes user-input medical data to generate accurate real-time predictions. The application's frontend is developed using Streamlit, providing a user-friendly interface, while the backend integrates trained models deployed using Joblib. This system is intended to support healthcare professionals in early diagnosis and decision-making, particularly in resource-constrained environments.

Index Terms

Disease Prediction, Artificial Intelligence, Machine Learning, XGBoost, Random Forest, Healthcare Informatics, Streamlit.

1. Introduction

The increasing global burden of chronic and lifestyle-related diseases has highlighted the urgent need for early diagnosis and effective health management systems. Traditional diagnostic procedures, though reliable, are often time-consuming and inaccessible in under-resourced regions. With the advent of Artificial Intelligence (AI) and Machine Learning (ML), there is a growing opportunity to automate and enhance the diagnostic process through data-driven methods.

This paper introduces a web-based multi-disease prediction system that utilizes advanced ML models to offer real-time health assessments based on user-provided medical data. Unlike many existing systems that focus on single-disease prediction, our approach consolidates the prediction of multiple diseases—Heart Disease, Diabetes, Parkinson's Disease, Jaundice, Liver Conditions, and Hepatitis—into one unified platform. The goal is to assist clinicians and users by offering quick and accurate predictions that may support early intervention and improved health outcomes.

Objective

The primary objective of this project is to design and implement an intelligent web-based system capable of predicting multiple diseases using state-of-the-art machine learning algorithms. The system aims to:

- Enable early diagnosis of critical diseases such as Heart Disease, Diabetes, Parkinson’s Disease, Jaundice, Liver Conditions, and Hepatitis.
- Provide a unified platform for predicting multiple diseases based on user-inputted health parameters.
- Utilize accurate and robust machine learning models like XGBoost and Random Forest to ensure high predictive performance.
- Deliver real-time results through an interactive and easy-to-use interface developed with Streamlit.
- Assist healthcare professionals and patients in resource-constrained settings by offering preliminary diagnostic insights.

2. Literature Survey

Numerous studies have been conducted in the field of disease prediction using Machine Learning (ML) techniques. Traditionally, most of these works have focused on single-disease prediction systems. For example, logistic regression and decision tree models have been widely used for predicting the likelihood of diabetes and heart disease based on clinical data. While these approaches have demonstrated reasonable accuracy, their scope is often limited to specific diseases and datasets.

Recent advancements have introduced more powerful ensemble models such as Random Forest and XGBoost, which offer improved accuracy and generalization across diverse health data. In one study, Random Forest was applied to predict cardiovascular disease with an accuracy exceeding 90%, indicating its suitability for medical diagnostics. Another research used XGBoost for early detection of Parkinson’s disease, showcasing high performance due to its ability to handle non-linear relationships and missing values.

However, most existing literature has failed to integrate multi-disease prediction capabilities into a single system. This gap in research and application has driven the motivation for our work, which aims to combine multiple disease prediction tasks within a unified, scalable framework. Additionally, the use of an interactive web application frontend (Streamlit) for delivering results to end-users remains underexplored in earlier works.

3. Proposed System

The proposed system is a web-based application designed to predict multiple diseases using supervised machine learning techniques. It consists of a user-friendly frontend developed with Streamlit and a backend powered by pre-trained ML models. The system architecture is modular, scalable, and capable of integrating additional diseases in the future.

3.1 Data Collection and Preprocessing

The datasets used for this system were obtained from publicly available medical repositories, such as Kaggle and UCI Machine Learning Repository. Each dataset was carefully cleaned by handling missing values, removing outliers, and normalizing numerical features to ensure uniformity. Label encoding and feature selection techniques were applied to optimize model training.

3.2 Model Selection and Training

Random Forest and XGBoost were selected for their robustness and proven effectiveness in classification problems. These models were trained using the scikit-learn and XGBoost Python libraries. Each model was validated using techniques like k-fold cross-validation to ensure reliability and avoid overfitting.

3.3 System Architecture

- **Frontend:** Built using Streamlit to provide an interactive interface for entering medical parameters.
- **Backend:** Trained models are saved using Joblib and loaded during runtime to process user input.
- **Prediction Flow:** Users enter their symptoms and parameters; the model processes the input and returns the likelihood of specific diseases with confidence scores.

3.4 Evaluation Metrics

The models were evaluated using standard performance metrics:

- **Accuracy**
- **Precision**
- **Recall**
- **F1-Score**
  These metrics help assess the model's ability to correctly predict diseases while minimizing false positives and false negatives.

4. Software Requirements and Used Technologies

The implementation of the proposed system involves various tools, libraries, and frameworks that enable effective data processing, model deployment, and interface development. Below is a summary of the technologies used:

4.1 Programming Language

- Python 3.x: Used for data preprocessing, model building, training, and deployment due to its simplicity and extensive ML libraries.

4.2 Machine Learning Libraries

- Scikit-learn: For implementing standard classification models and evaluation metrics.
- XGBoost: For high-performance gradient boosting machine learning.
- Pandas & NumPy: For data manipulation and numerical operations.
- Joblib: For modelserialization and deployment.

4.3 Web Development Framework

- Streamlit: Used to build the interactive web-based frontend. It allows rapid prototyping and real-time interaction with ML models.

4.4 IDE and Tools

- Jupyter Notebook / Google Colab: For model development and testing.
- VS Code / PyCharm: For writing and debugging backend code.
- Git: For version control.

4.5 Deployment Environment

- Localhost or cloud-based hosting platforms like Heroku or Streamlit Cloud can be used to deploy the application.

5. Flow of the System

The proposed multi-disease prediction system follows a structured and logical workflow that ensures efficient data processing and accurate prediction. The system flow is described below:

Step 1: User Input

- The user accesses the web application through a browser.
- The application prompts the user to input relevant medical parameters (e.g., age, blood pressure, glucose level, tremor frequency, bilirubin levels, etc.), depending on the selected disease.

Step 2: Data Validation

- The entered data is validated in real-time to ensure completeness and correct formats (e.g., numeric values where required).

Step 3: Model Loading

- Pre-trained machine learning models for each disease are stored using Joblib.
- When the user selects a disease and submits input, the corresponding model is dynamically loaded in the backend.

Step 4: Prediction

- The model processes the input data and returns a prediction (e.g., “Disease Detected” or “No Disease Detected”) along with a probability/confidence score.

Step 5: Output Display

- The prediction result is displayed to the user on the frontend with intuitive color indicators and text (e.g., green for healthy, red for risk).
- Basic interpretation or suggestions may be provided, such as “Please consult a doctor if symptoms persist.”

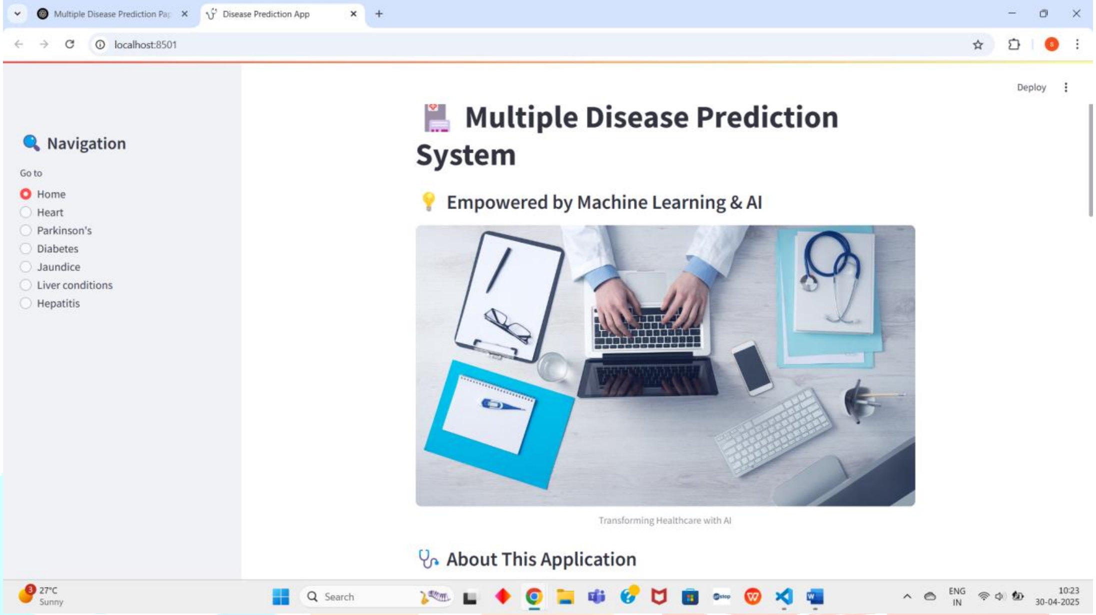
Step 6: Restart/Exit

- The user can either test another disease or exit the application.

6. Results

1) Main Dashboard

The main dashboard of the “Multiple Disease Prediction using Artificial Intelligence and Machine Learning”. At the top left side is a Navigation button where you can select a particular disease to predict it.



Navigation

Go to

● Home
○ Heart
○ Parkinson's
○ Diabetes
○ Jaundice
○ Liver conditions
○ Hepatitis

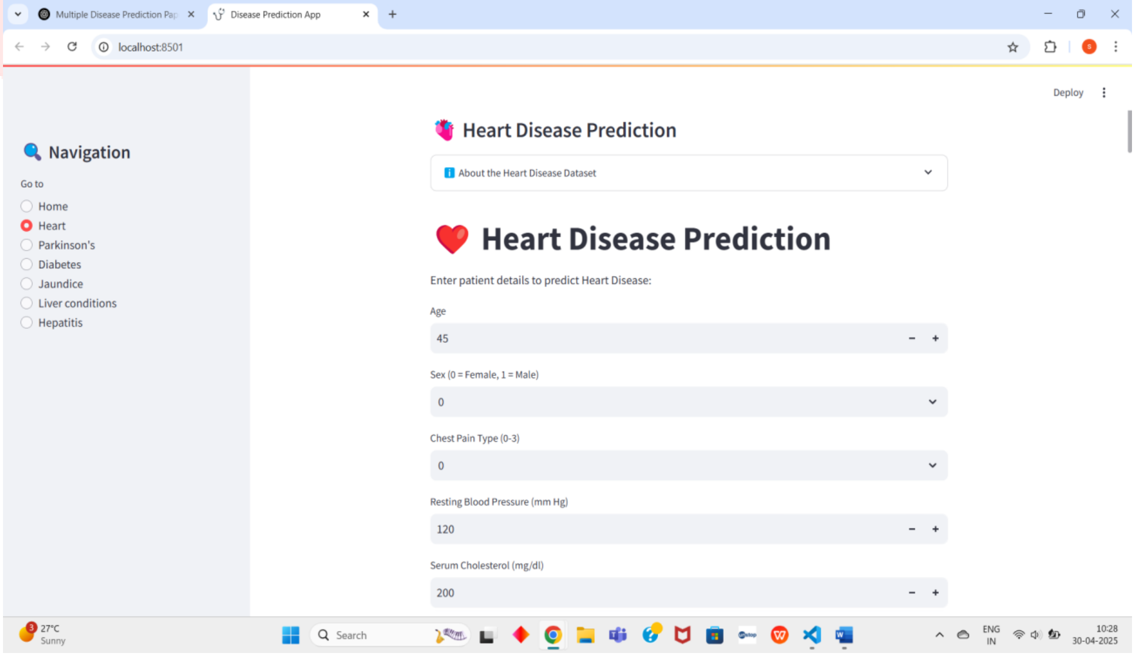
Multiple Disease Prediction System

Empowered by Machine Learning & AI

Transforming Healthcare with AI

About This Application

2. Disease Prediction



Navigation

Go to

○ Home
● Heart
○ Parkinson's
○ Diabetes
○ Jaundice
○ Liver conditions
○ Hepatitis

Heart Disease Prediction

About the Heart Disease Dataset

Heart Disease Prediction

Enter patient details to predict Heart Disease:

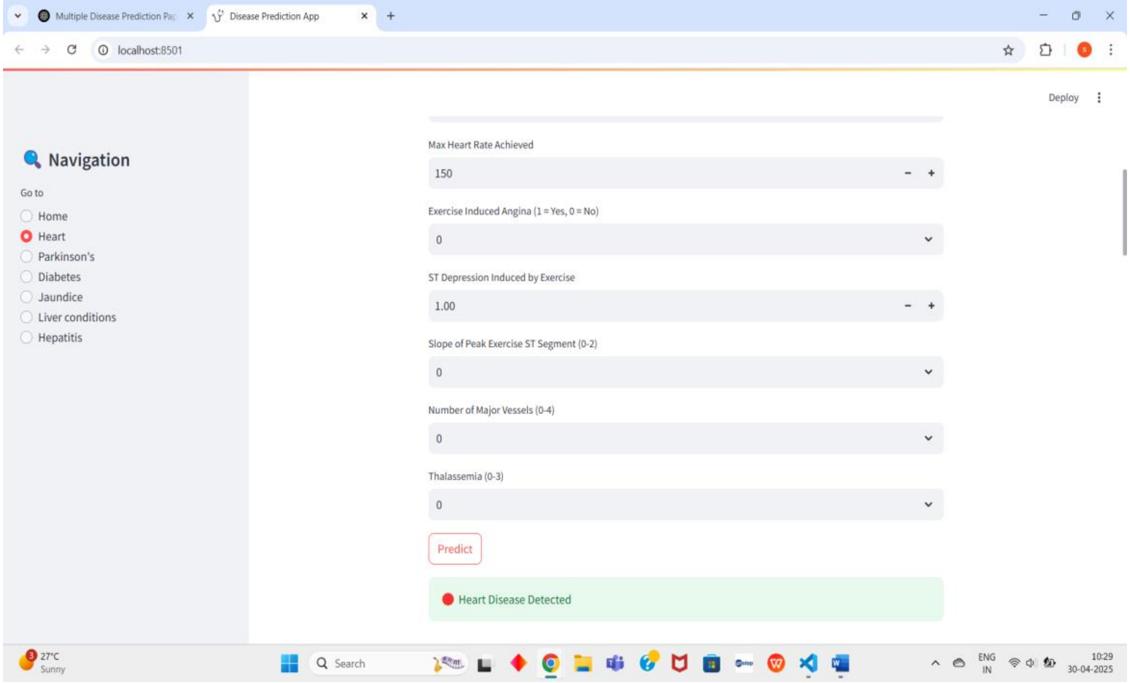
Age: 45

Sex (0 = Female, 1 = Male): 0

Chest Pain Type (0-3): 0

Resting Blood Pressure (mm Hg): 120

Serum Cholesterol (mg/dl): 200



Max Heart Rate Achieved: 150

Exercise Induced Angina (1 = Yes, 0 = No): 0

ST Depression Induced by Exercise: 1.00

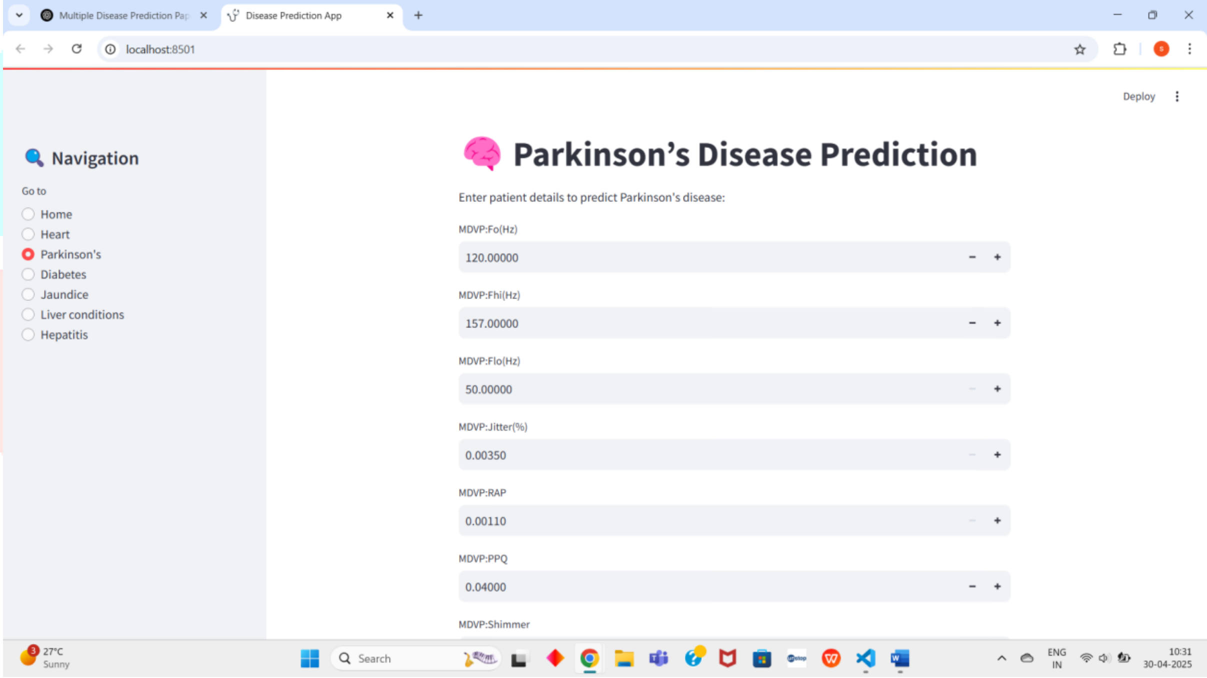
Slope of Peak Exercise ST Segment (0-2): 0

Number of Major Vessels (0-4): 0

Thalassemia (0-3): 0

Predict

● Heart Disease Detected



Parkinson's Disease Prediction

Enter patient details to predict Parkinson's disease:

MDVP:Fo(Hz): 120.00000

MDVP:Fhi(Hz): 157.00000

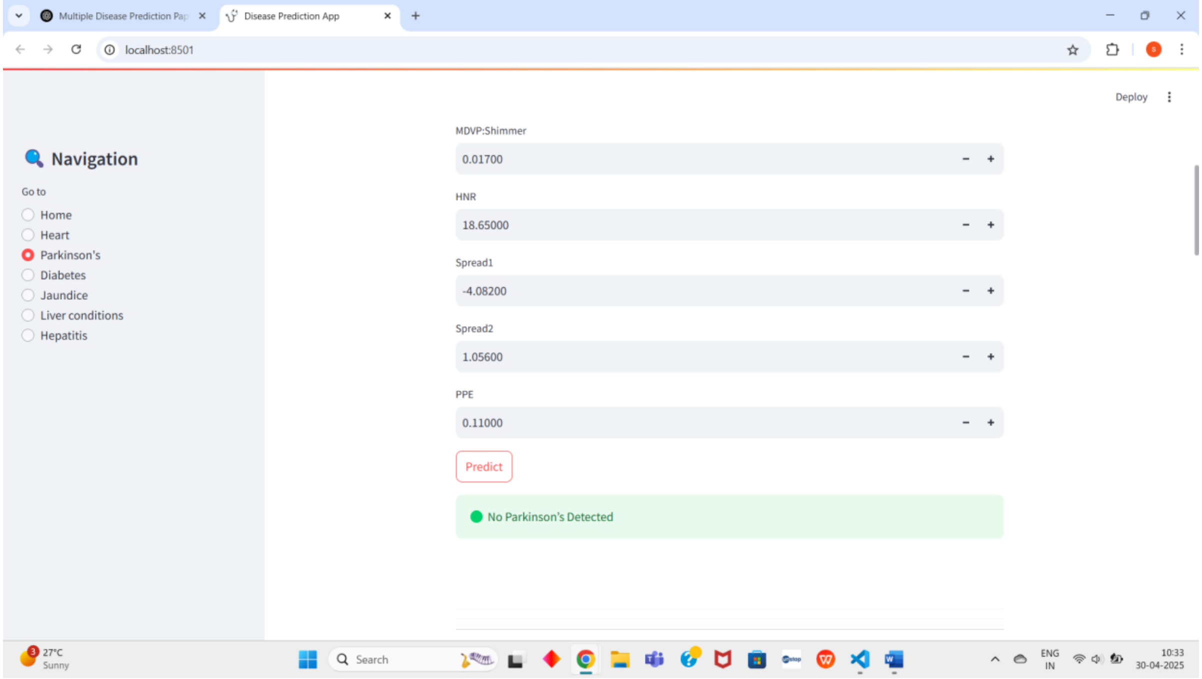
MDVP:Flo(Hz): 50.00000

MDVP:Jitter(%): 0.00350

MDVP:RAP: 0.00110

MDVP:PPQ: 0.04000

MDVP:Shimmer



Navigation

Go to

○ Home
○ Heart
● Parkinson's
○ Diabetes
○ Jaundice
○ Liver conditions
○ Hepatitis

MDVP:Shimmer
0.01700

HNR
18.65000

Spread1
-4.08200

Spread2
1.05600

PPE
0.11000

Predict

No Parkinson's Detected

8. Conclusion and Future Scope:

In this study, we presented a machine learning-based approach for predicting multiple diseases using various classification algorithms such as Random Forest, XGBoost, and Logistic Regression. By utilizing a comprehensive dataset, we demonstrated that machine learning techniques can significantly enhance the accuracy and reliability of disease prediction systems. The integration of these models not only provides early detection of diseases but also aids in personalized treatment plans for patients.

Our results indicate that the models achieved satisfactory performance in terms of accuracy, precision, and recall, highlighting their potential in the healthcare domain. The predictive models were able to learn patterns from historical medical data, enabling effective decision-making. Moreover, the versatility of machine learning allows the system to be adapted to various healthcare settings, making it a valuable tool for doctors, healthcare providers, and researchers.

In conclusion, the study confirms that machine learning techniques can be a powerful tool in the healthcare industry, supporting diagnostic processes and improving overall patient care.

The current study focuses on a limited dataset and uses a few machine learning models. However, there is significant scope for expanding and improving the system in future work. Some potential areas for future development include:

1. **Integration with Real-time Data**: Future work could involve integrating real-timemedical data from wearable devices and sensors, allowing the prediction system to update dynamically and provide more accurate predictions.
2. **Expanding the Dataset**: Incorporating a larger, more diverse dataset with additional health conditions and variables could improve the model's generalization ability and accuracy across different patient demographics.
3. **Model Enhancement**: More advanced techniques such as deep learning and neural networks could be explored to improve prediction accuracy, especially for complex and rare diseases.
4. **Multi-modal Data**: Incorporating multi-modal data, including medical imaging (such as X-rays, MRIs) and patient history, could provide a more comprehensive approach to disease prediction.
5. **Clinical Validation**: For real-world application, it is essential to validate the predictions through clinical trials and ensure the system's reliability and effectiveness in actual healthcare settings.

6. **Personalized Healthcare**: Future research could focus on personalized disease prediction models that take into account an individual's genetics, lifestyle, and other unique factors, enabling more accurate predictions tailored to each patient.

By pursuing these avenues, the potential for machine learning in healthcare applications can be fully realized, leading to more effective, efficient, and accessible healthcare solutions.

9. References

1. L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001. DOI: 10.1023/A:1010933404324
2. 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*, pp. 785–794, 2016.
   DOI: 10.1145/2939672.2939785
3. D. Dua and C. Graff, "UCI Machine Learning Repository," https://archive.ics.uci.edu/ml, University of California, Irvine, School of Information and Computer Sciences, 2019.
4. J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*, 3rd ed., Elsevier, 2011.
5. S. Hochreiter and J. Schmidhuber, “Long short-term memory,” *Neural Computation*, vol.9,no. 8, pp. 1735–1780, 1997.
   DOI: 10.1162/neco.1997.9.8.1735
6. M. J. Kowsari et al., "HDLTex: Hierarchical Deep Learning for Text Classification," in *2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pp. 364–371. DOI: 10.1109/ICMLA.2017.0-128
7. Rajkomar, A., et al. “Scalable and accurate deep learning with electronic health records,” *npj Digital Medicine*, vol. 1, no. 1, 2018.
   DOI: 10.1038/s41746-018-0029-1
8. S. Ravi and D. Wong, “AI in Healthcare: Past, Present and Future,” *Journal of Healthcare Informatics Research*, vol. 5, pp. 1–25, 2021.
   DOI: 10.1007/s41666-020-00080-4