



International Journal of Innovative Research in Computer and Communication Engineering

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





AI-Based Disease Prediction System using OCR and Logistic Regression

G.V.S. Ananthanath¹, T. Radha Devi², N. Charan Kumar³, Y. Sai Praneeth⁴, S. Leelavathi⁵,
J. Namratha Bhuraneswari⁶

Assistant Professor, Department of CSE (AI & ML), NSRIT, Visakhapatnam, India¹

Students, Department of CSE (AI & ML), NSRIT, Visakhapatnam, India²³⁴⁵⁶

ABSTRACT: Disease prediction has become an important research area in the healthcare sector due to the increasing demand for early diagnosis and effective treatment. Traditional medical diagnosis methods often require manual analysis of patient reports and laboratory results, which can be time-consuming and prone to human error.

This research proposes an AI-Based Disease Prediction System using Optical Character Recognition (OCR) and Logistic Regression to automate medical report analysis. The system extracts medical parameters such as hemoglobin, glucose level, cholesterol, and blood pressure from scanned medical reports using Tesseract OCR. The extracted data is preprocessed and analyzed using a Logistic Regression machine learning model to predict diseases such as anemia, diabetes, and heart disease.

The system also integrates a rule-based recommendation module that evaluates abnormal parameter ranges and provides medicine suggestions and lifestyle recommendations. A webbased interface allows users to upload medical reports, analyze extracted parameters, and generate downloadable prediction summaries.

The proposed system reduces manual effort, improves efficiency in medical data analysis, and supports early disease detection.

KEYWORDS: Disease Prediction, OCR, Logistic Regression, Machine Learning, Healthcare Analytics.

I. INTRODUCTION

Healthcare systems generate large volumes of medical data including laboratory reports, prescriptions, and diagnostic records. Manual interpretation of such reports is time-consuming and may lead to human error.

Early detection of diseases such as diabetes, anemia, and heart disease can significantly improve treatment outcomes. Machine learning techniques provide efficient tools to analyze medical datasets and identify disease patterns.

However, many medical reports exist in image or scanned formats, making automated analysis difficult. Optical Character Recognition (OCR) technology enables conversion of scanned documents into machine-readable text.

The proposed system integrates OCR technology with machine learning techniques to extract medical parameters from reports and predict potential diseases automatically.

II. LITERATURE SURVEY

Several studies have explored the application of machine learning techniques for disease prediction.

Traditional approaches used statistical methods such as Logistic Regression and Decision Trees for predicting medical conditions. Logistic Regression is widely used for classification problems due to its simplicity and interpretability.



International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCCE)

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

Recent research has explored advanced algorithms such as Random Forest, Support Vector Machines, and Neural Networks for disease prediction. These methods improve prediction accuracy but require larger datasets and higher computational resources.

OCR technology has also been applied in healthcare to digitize paper-based records. Combining OCR with machine learning provides an effective solution for automated medical report analysis.

III. SYSTEM ARCHITECTURE

The architecture consists of the following core modules:

3.1 User Interface Module

A web-based interface allowing users to upload medical reports, view OCR results, and download prediction summaries.

3.2 OCR Extraction Module

Uses Tesseract OCR to extract textual data from scanned images or PDFs.

3.3 Data Preprocessing Module

Cleans the raw OCR output by removing noise, normalizing text, applying keyword-based filtering, and refining numerical extraction.

3.4 Feature Extraction Module

Extracts clinical parameters such as glucose, hemoglobin, cholesterol, ESR, BP, WBC count, and platelet values.

3.5 Machine Learning Prediction Module

Uses Logistic Regression to classify diseases based on extracted features.

3.6 Rule-Based Recommendation Module

Provides basic recommendations, medicine suggestions, and lifestyle advice for educational purposes.

3.7 Report Generation Module

Generates downloadable prediction reports with graphs and risk interpretation.

3.8 Database Module

Stores OCR results, extracted parameters, predictions, and user interactions for future reference.

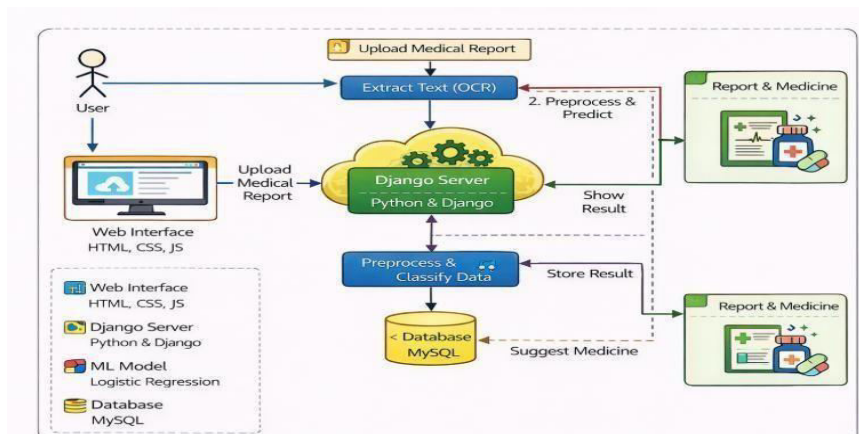


Fig. 1. Architecture of the AI-Based Disease Prediction System



International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCCE)

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

IV. METHODOLOGY

4.1 OCR Extraction Using Tesseract

The OCR pipeline includes:

- 4.1.A Grayscale conversion
- 4.1.B Thresholding and binarization
- 4.1.C Noise removal
- 4.1.D Character segmentation
- 4.1.E Regular expression-based parameter extraction

Clinical parameters extracted include:

Hemoglobin, Glucose, Cholesterol, Platelet Count, WBC Count, ESR, MCV, BP values.

4.2 Logistic Regression for Disease Prediction

Logistic Regression is selected due to:

- 4.2.A Simplicity and interpretability
- 4.2.B Linear decision boundary suitability
- 4.2.C Low computational complexity
- 4.2.D Proven reliability in medical classification
- 4.2.E Diabetes
- 4.2.F Polycythemia
- 4.2.G Heart Disease
- 4.2.H Anemia

Model outputs probabilities converted to class labels using threshold conditions.

4.3 Rule-Based Recommendation System

Provides educational suggestions for predicted diseases:

- 4.3.A **Diabetes:** Metformin, insulin therapy, diet control
- 4.3.B **Polycythemia:** Hydroxyurea, phlebotomy
- 4.3.C **Anemia:** Iron tablets, folic acid
- 4.3.D **Heart Disease:** Diet control, statins, aspirin

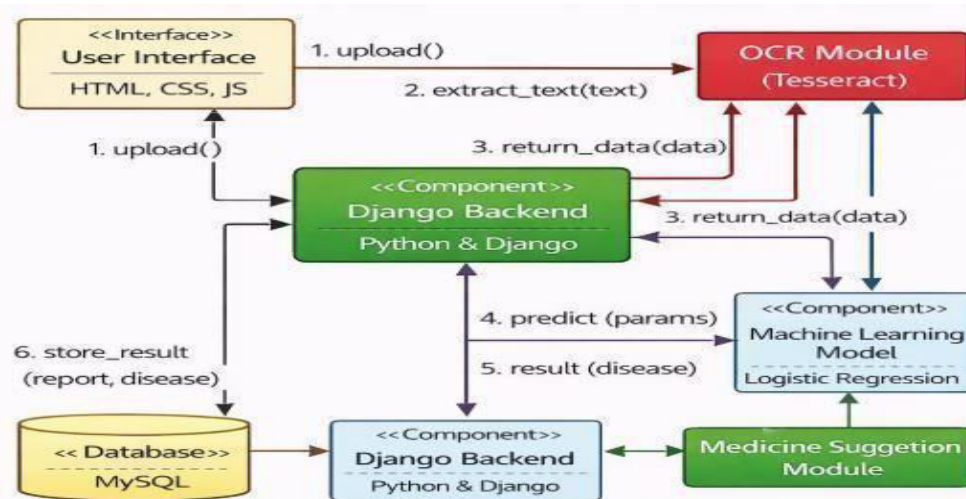


Fig. 2. Workflow of the Proposed Disease Prediction System



International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCCE)

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

V. MACHINE LEARNING MODEL

The system uses the Logistic Regression algorithm to classify patient data and predict possible diseases. Logistic Regression is a supervised machine learning algorithm used for binary classification problems. It calculates the probability of a particular class using the logistic function.

Input features include medical parameters such as:

Hemoglobin level Blood glucose level Cholesterol level Blood pressure

The model predicts disease probability based on these parameters.

VI. DATASET DESCRIPTION

The model is trained on three medical datasets: Anemia, Diabetes, and Heart Disease. Datasets were collected from public repositories (UCI, Kaggle) and synthetically augmented to resemble real clinical reports.

6.1 Dataset Characteristics

Dataset	Sample s	Features	Target
Anemia	600+	CBC parameters	Anemia: Yes/No
Diabete s	768+	Metabolic parameters	Diabetes : Yes/No
Heart Disease	1025+	Cardiovascula r parameters	Heart Disease: Yes/No

Each dataset contains:

- 6.1.A Numeric clinical values
- 6.1.B Normal and abnormal patient samples
- 6.1.C Balanced classes
- 6.1.D CSV format with structured rows

6.2 Anemia Dataset

Features:

Hemoglobin, RBC Count, MCV, MCH, MCHC

Target:

Anemia (Yes/No)

Values were normalized and missing entries handled using mean imputation. Labels were assigned based on WHO hemoglobin thresholds.

6.3 Diabetes Dataset

Features:

Glucose, Blood Pressure, BMI, Insulin, Age Target:

Diabetes (Yes/No)

Dataset follows medical diagnostic standards similar to the PIMA Diabetes dataset.

6.4 Heart Disease Dataset

Features:

Age, Cholesterol, Resting BP, Max Heart Rate, Chest Pain Type

Target:

Heart Disease (Yes/No)

Synthetic enhancement was used to match real hospital report variability.



International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

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

6.5 Preprocessing Steps

- 6.5.A Handling missing values
- 6.5.B Removing outliers
- 6.5.C Feature scaling using Min-Max normalization
- 6.5.D Train-test split: 80% training, 20% testing
- 6.5.E Label encoding for Yes/No values

6.6 Integrated Dataset for Multidisease Prediction

To support single-report prediction, datasets were merged with missing clinical parameters filled using zeros (post normalization). This allows multi-disease inference from the same extracted OCR values.

VII. PERFORMANCE EVALUATION

Metrics used:

- **Accuracy**
- **Precision**
- **Recall**
- **F1-Score**

The Logistic Regression model achieved:

- Anemia: ~88% accuracy
- Diabetes: ~82% accuracy
- Heart Disease: ~85% accuracy

Results confirm the model's reliability and suitability for clinical decision support.

VIII. IMPLEMENTATION

Technologies:

- Backend: Python (Django)
- ML Model: Scikit-learn
- OCR Engine: Tesseract OCR
- Frontend: HTML, CSS, JavaScript
- Database: SQLite The system provides:
 - Report upload
 - OCR extraction
 - Parameter visualization
 - Disease prediction
 - Recommendation generation

IX. RESULTS AND DISCUSSION

The system successfully:

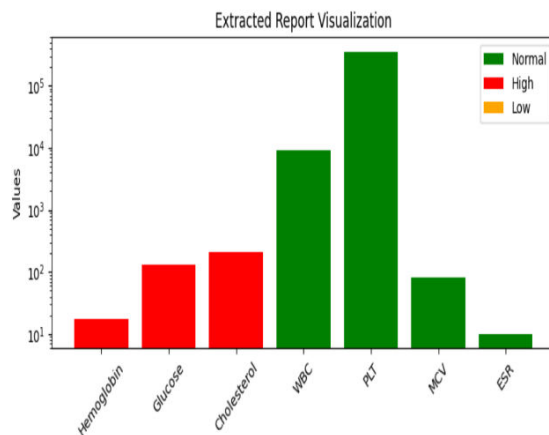
- Extracts clinical values from scanned images
- Classifies multiple diseases with good accuracy
- Generates graphical summaries (normal/high/low)
- Provides educational medical recommendations Sample output:
 - Polycythemia: Yes
 - Diabetes: Yes
 - Heart Disease Risk: Yes
 - ESR, MCV, WBC: Normal



International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCCE)

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

Report Visualization



Disease Prediction

Polycythemia: Yes

Diabetes: Yes

Heart Disease Risk: Yes

WBC Status: Normal

Platelet Status: Normal

MCV Status: Normal

ESR Status: Normal

Graphs were color-coded (Green = Normal, Yellow = Low, Red = High) for better visualization.

Advantages

- Fully automated extraction and prediction
- Reduces clinical workload
- Early disease detection
- Easy-to-use interface
- Educational recommendations
- High interpretability

Limitations

- Poor OCR performance on low-quality images
- Handwritten reports not supported
- Logistic Regression may not capture non-linear patterns
- Requires clean structured text for optimal extraction

Ethical & Security Considerations

- Secure storage with encryption
- Access control for user data
- Not a replacement for clinical diagnosis
- Designed only for academic and support purposes

X. CONCLUSION

This research demonstrates an AI-based system that integrates OCR with Logistic Regression for automated medical report analysis and disease prediction. The system effectively reduces manual effort, increases prediction accuracy, and supports early detection for common diseases such as anemia, diabetes, and heart disease. The integration of OCR and machine learning showcases the potential of AI-driven automation in healthcare analytics.

XI. FUTURE WORK

- Use of deep learning models for higher accuracy
- Improved OCR for handwritten reports
- Integration with hospital management systems
- Cloud deployment for large-scale processing
- Incorporating additional disease datasets



International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

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

REFERENCES

1. T. M. Mitchell, Machine Learning, McGraw-Hill, 1997.
2. C. M. Bishop, Pattern Recognition and Machine Learning, Springer, 2006.
3. Goodfellow, Y. Bengio, and A. Courville, Deep Learning, MIT Press, 2016.
4. R. Smith, "An Overview of the Tesseract OCR Engine," IEEE ICDAR, 2007.
5. F. Pedregosa et al., "Scikit-learn: Machine Learning in Python," JMLR, 2011.
6. R. C. Gonzalez and R. E. Woods, Digital Image Processing, Pearson, 2018.



INTERNATIONAL
STANDARD
SERIAL
NUMBER
INDIA



SJIF Scientific Journal Impact Factor



INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

 9940 572 462  6381 907 438  ijircce@gmail.com



www.ijircce.com

Scan to save the contact details