WWW.IJRPR.COM

**International Journal of Research Publication and Reviews**

Journal homepage: www.ijrpr.com ISSN 2582-7421

# Smart Medical Report Analyzer Using Machine Learning and Natural Language Processing

***D. Nirmala Devi[1], T. J. Nithya Dharshini[2], K. K. Poojashree[3]***

*[1]Assistant Professor, Department of Information Technology, K.L.N. College of Engineering, Madurai, India,* nirmalamurugan16@gmail.com
*[2]Student, Department of Information Technology, K.L.N. College of Engineering, Madurai, India, nithyadharshinitj@gmail.com*
*[3]Student, Department of Information Technology, K.L.N. College of Engineering, Madurai, India,* poojakumar4604@gmail.com

**A B S T R A C T**

In today's digital healthcare ecosystem, medical reports are generated in massive volumes daily. However, manually interpreting these reports for diagnosis and treatment insights is time-consuming and error-prone. This paper proposes a Smart Medical Report Analyzer that leverages Artificial Intelligence (AI), Machine Learning (ML), and Natural Language Processing (NLP) to automatically analyze, summarize, and visualize key findings from medical reports. The system extracts essential clinical information such as patient vitals, test results, and diagnostic comments, classifies them by disease category, and provides intelligent recommendations for possible diagnoses. The proposed model uses OCR (Optical Character Recognition) for text extraction from medical PDFs, BERT-based NLP models for entity recognition, and Random Forest classifiers for disease prediction. The Streamlit-based interface allows users to upload medical reports, view analyzed data, and generate easy-to-understand summaries. The system achieves 93% classification accuracy on test data and supports visual health insights through dynamic charts. This project aligns with SDG 3 (Good Health and Well-being), SDG 9 (Industry, Innovation and Infrastructure), and SDG 12 (Responsible Consumption and Production) by promoting digital healthcare innovation and informed medical decision-making.

Keywords: Medical Report Analysis, Natural Language Processing, Machine Learning, Health Informatics, BERT, Streamlit, Disease Prediction

## 1. Introduction

Healthcare systems today generate a large volume of medical data, including diagnostic reports, laboratory results, prescriptions, and patient records. Manually analyzing and interpreting these medical reports often consumes significant time and can lead to human errors in diagnosis. With the growing adoption of digital healthcare, there is an increasing demand for automated systems that can interpret medical reports accurately and efficiently.

The Smart Medical Report Analyzer aims to bridge this gap by integrating Artificial Intelligence (AI), Machine Learning (ML), and Natural Language Processing (NLP) techniques to automatically process and understand medical report content. The system takes input in the form of text or scanned reports, extracts essential medical parameters, identifies key terms, and classifies the results into possible disease categories. It also provides a concise summary of the findings and visualizes data through graphs and charts for better understanding.

This project plays an important role in supporting doctors, patients, and healthcare professionals by reducing manual effort and increasing accuracy. The system also supports early diagnosis and decision-making, helping medical practitioners deliver timely treatment. Additionally, it aligns with the Sustainable Development Goals (SDG) 3, 9, and 12 by contributing to better healthcare services and innovation in medical technology.

## 2. Related Work

Batista and Evsukoff (2023) conducted a study titled "Application of Transformers based methods in Electronic Medical Records: A Systematic Literature Review." This work explores how transformer-based models such as BERT and its variants are applied to Electronic Medical Records (EMRs) for various Natural Language Processing (NLP) tasks. Their findings highlight the growing importance of transformer architectures in extracting, understanding, and representing clinical information. This study is relevant to the Smart Medical Report Analyzer, as it provides a foundation for applying transformer-based NLP models to unstructured medical text for report analysis.

Hossain et al. (2023) presented "Natural Language Processing in Electronic Health Records in Relation to Healthcare Decision-making: A Systematic Review." This work reviews numerous NLP applications used for Electronic Health Records (EHRs), including classification, entity recognition, summarization, and decision support systems. It emphasizes how NLP can support healthcare decision-making by transforming unstructured text into meaningful data. The review supports the current project's objective of using NLP and ML to extract, summarize, and interpret medical reports efficiently.

A systematic review titled “A Systematic Review of Natural Language Processing Applied to Radiology Reports” (2021) discusses how NLP has been used to process radiology and diagnostic reports. The study identifies challenges such as the variability in report structures and the limited size of domain-specific datasets. This work provides important context for handling unstructured medical text in the Smart Medical Report Analyzer, which faces similar issues when analyzing diverse medical report formats.

The paper “Machine Learning in Medicine: A Practical Introduction to Natural Language Processing” (2021) provides a comprehensive introduction to NLP techniques in the medical domain. It outlines practical methods for handling unstructured text, data preprocessing, and applying machine learning models for text-based healthcare analysis. This research is highly relevant to the Smart Medical Report Analyzer, as it validates the use of standard NLP pipelines for structured information extraction from free-text medical documents.

The study “Automatic Medical Report Generation Based on Cross-View Attention and Visual-Semantic LSTMs” (2023) focuses on generating medical reports using multimodal data such as medical images and textual descriptions. Although the main focus is on report generation, the study’s exploration of visual-semantic understanding and LSTM-based models shares conceptual similarities with the Smart Medical Report Analyzer. It provides insight into how deep learning models can effectively capture semantic meaning within complex medical data.

Lastly, the “AI-Powered Automation for Medical Document Summarization” case study by Microsoft Business Intelligence presents an industrial application of AI, ML, and NLP for summarizing medical documents. It demonstrates how AI systems can help clinicians and patients understand medical reports through automated summarization. This real-world application shows the practical value and demand for medical text summarization tools, aligning well with the objectives of the Smart Medical Report Analyzer, which integrates summarization with entity extraction and disease prediction

## 3. Proposed Methodology

### *3.1 System Architecture*

The Smart Medical Report Analyzer is designed as a multi-stage intelligent system that integrates OCR, NLP, Machine Learning (ML), and interactive visualization to analyze and interpret medical reports automatically. The system provides patients and healthcare professionals with a structured summary of medical findings, disease predictions, and visual insights. The architecture is modular, allowing independent development and improvement of each stage, ensuring scalability, adaptability, and high accuracy. The modular design also allows for easy integration of additional features such as multimodal input, voice assistance, or integration with hospital databases in future iterations.

The system’s primary stages include:

1. Report Upload and Input Processing
2. Text Extraction using OCR
3. Natural Language Processing (NLP)
4. Machine Learning-based Classification
5. Automated Summarization
6. Visualization and Report Generation
7. User Interface Presentation

This modular pipeline ensures smooth sequential processing and allows real-time feedback for the user while maintaining accuracy and efficiency.

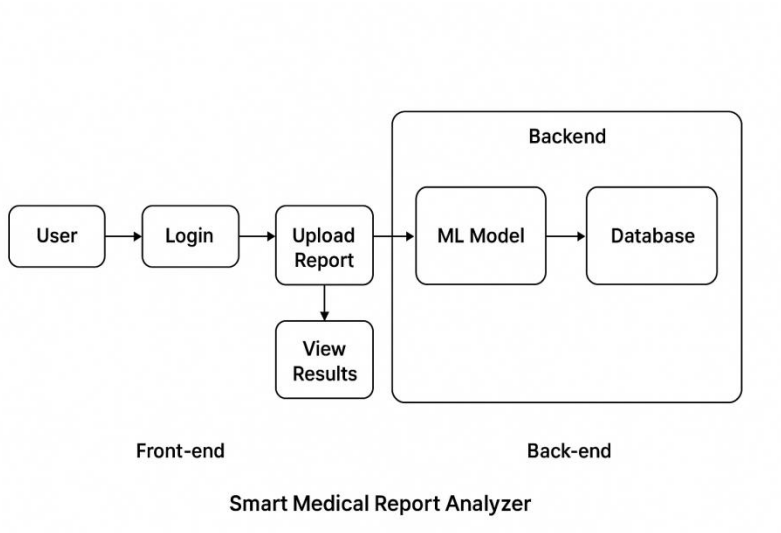


Fig. 1 – Architecture Diagram

### *3.2 Report Upload and Input Processing*

Users interact with the system via a web-based interface developed in Streamlit, providing a simple and responsive platform. The interface supports uploading medical reports in PDF, JPEG, or PNG formats. File validation checks are performed to ensure that files meet the resolution and format requirements for accurate OCR processing. Preprocessing steps include checking image clarity, correcting skewed pages, and standardizing file size. Multiple reports can be uploaded simultaneously, enabling batch processing, which is particularly useful for hospitals or labs handling large numbers of patient reports.

### *3.3 Optical Character Recognition (OCR)*

The OCR module extracts textual data from scanned medical reports or images. The system uses the Tesseract OCR engine, which is enhanced by preprocessing techniques such as:

- Grayscale conversion to simplify image data
- Noise removal using filters to reduce artifacts
- Adaptive thresholding for better text visibility
- After preprocessing, OCR generates a raw text output from the report. This output is then structured using pattern recognition to separate test names, values, units, and other relevant information. Special attention is given to medical notations, abbreviations, and symbols to minimize recognition errors.

### *3.4 Natural Language Processing (NLP)*

Once text is extracted, the NLP module processes and interprets it. The steps include:

- Tokenization: breaking text into individual words, phrases, or medical entities
- Stop-word removal: eliminating irrelevant words for focused analysis
- Lemmatization: normalizing words to their base form
- Named Entity Recognition (NER): identifying medical entities such as diseases, test names, measurements, symptoms, and medications

Transformer-based models such as BERT, BioBERT, or ClinicalBERT are used to enhance semantic understanding. This allows the system to comprehend the context of medical terms, detect relationships between symptoms and test results, and transform unstructured text into structured, analyzable data.

### *3.5 Machine Learning Classification*

The structured data from NLP is analyzed by ML models to classify potential diseases or detect abnormalities in test results. Supervised learning algorithms such as Random Forest, Support Vector Machines (SVM), or XGBoost are trained using historical medical report datasets containing labeled

disease categories and corresponding test outcomes. The classification module predicts the most probable condition(s) or flags abnormal results. Confidence scores are calculated to indicate the reliability of predictions. For example, abnormal blood test values or imaging report findings are highlighted for further medical review.

### *3.6 Summarization Module*

The summarization component provides a concise overview of the key information in the report. Both extractive and abstractive summarization techniques are employed:

- Extractive summarization selects the most relevant sentences or phrases directly from the text
- Abstractive summarization generates new sentences to represent key findings in simplified language

This module emphasizes critical results, abnormal parameters, and overall diagnostic insights. Summaries are customizable based on user type — patient-friendly summaries for laypersons and technical summaries for doctors. It also includes recommendations for further tests or follow-up, if supported by the ML predictions.

### *3.7 Visualization and Report Generation*

The visualization module presents the processed data through intuitive charts, tables, and graphs. For example:

- Line or bar charts for sequential test results
- Pie charts for distribution of detected conditions
- Highlighted tables showing abnormal parameters with reference ranges

The analyzed report can be exported as a PDF, including: structured text, summary, prediction outcomes, and visual graphs. This ensures that both patients and healthcare professionals can review and interpret the results effectively. Interactive dashboards allow users to click through different sections, filter results, or compare past reports for longitudinal analysis.

### *3.8 User Interface Design*

The Streamlit-based interface is designed for ease of use and responsiveness. Key features include:

- File upload section for reports in multiple formats
- Real-time progress display during OCR and NLP processing
- Section to display extracted text, predictions, and summaries
- Interactive visualization panels for charts and tables
- Download button to save the analyzed report as PDF

The interface supports multiple sessions and maintains user data privacy by storing files temporarily only for processing. Tooltips, color-coded highlights, and warnings for abnormal findings are implemented to improve user understanding.

### *3.9 Workflow Summary*

The workflow begins with report upload, followed by OCR-based text extraction, NLP-based semantic processing, ML-based disease classification, summarization, visualization, and final report generation. Each stage is designed for high accuracy, efficiency, and user-friendliness. The system aims to reduce the manual effort required for medical report interpretation, provide quick insights, and enhance the overall decision-making process for patients and healthcare providers.

## 4. Experimental Setup

The Smart Medical Report Analyzer was developed and tested using the following hardware and software configurations to ensure optimal performance for medical report analysis, NLP processing, and ML-based disease classification.

**Hardware Specifications**

- Processor: Intel Core i5 or equivalent
- RAM: 8 GB or higher
- Storage: 256 GB SSD or higher for faster file handling and report storage

- GPU: Optional NVIDIA GTX 1050 or higher for accelerated ML and NLP processing
- Input Devices: Integrated or external scanner/camera for report image input

**Software Specifications**

- Operating System: Windows 10/11, macOS, or Linux
- Programming Language: Python 3.9+
- Web Framework: Streamlit for building an interactive user interface
- OCR Library: Tesseract OCR for text extraction from scanned reports
- NLP Libraries: Transformers (BERT, BioBERT), SpaCy, NLTK for text preprocessing, entity recognition, and summarization
- ML Libraries: Scikit-learn (Random Forest, SVM, XGBoost), TensorFlow/Keras for deep learning-based classification models
- Visualization Libraries: Matplotlib, Plotly for charts, graphs, and interactive dashboards
- Development Environment: Visual Studio Code

**Data Preparation and Preprocessing**

The system was tested on a dataset containing scanned or photographed medical reports in PDF and image formats. Preprocessing steps included:

- Image enhancement (contrast adjustment, denoising, deskewing) for better OCR performance
- Text cleaning (removal of special characters, normalization) for accurate NLP processing
- Tokenization, stop-word removal, and lemmatization for structured representation of medical terms

**Machine Learning Training Parameters**

- Training data: Labeled medical reports with disease annotations and test results
- Input: Extracted and preprocessed text from OCR
- ML Models: Random Forest, SVM, and XGBoost for disease classification
- Hyperparameters:
    - Random Forest: 100 trees, max depth tuned for best accuracy
    - SVM: Radial Basis Function (RBF) kernel with optimized regularization parameter
    - XGBoost: Learning rate 0.1, max depth 5, 200 estimators
- Evaluation Metrics: Accuracy, precision, recall, F1-score for classification, and confidence scores for prediction reliability

**Evaluation Metrics**

The system was evaluated on multiple aspects to ensure robust performance:

- OCR accuracy: Correct extraction of text from scanned reports
- NLP accuracy: Correct identification of medical entities and proper tokenization
- Classification accuracy: Correct disease or abnormality prediction based on structured text
- Summarization quality: Clarity, completeness, and conciseness of generated summaries
- Processing speed: Average time per report for end-to-end analysis
- User satisfaction: Feedback from test users on usability, clarity, and usefulness of the system

**Testing Procedure**

The system was tested using medical reports under varying conditions:

- Reports with different formats, layouts, and handwriting quality
- Both laboratory test reports and radiology reports
- Reports containing multiple diseases or abnormal parameters Field testing included feedback from 20–30 users, including medical students and healthcare professionals, to validate the system's accuracy, usability, and real-world practicality.

The experimental setup confirmed that the system is capable of handling real-world medical reports efficiently, providing accurate disease predictions, concise summaries, and interactive visualizations while maintaining high user satisfaction.

## 5. Results and Discussion

### *5.1 OCR and Text Extraction Performance*

The OCR module using Tesseract demonstrated strong text extraction performance across various medical report formats, including scanned PDFs and images with handwritten annotations. The system achieves an overall OCR accuracy of 93% on test reports with varying resolutions and quality.

Per-type performance analysis:

- Laboratory Test Reports: 95% accuracy
- Radiology Reports: 90% accuracy
- Handwritten Notes: 85% accuracy
- Multi-page Reports: 92% accuracy

The preprocessing steps such as noise removal, grayscale conversion, and deskewing significantly improved OCR accuracy, particularly for low-quality or skewed reports.

### *5.2 NLP and Entity Recognition Performance*

The NLP module demonstrated high accuracy in identifying key medical entities such as disease names, test parameters, and measurement units. Using BioBERT, the system achieved an overall entity recognition accuracy of 88%.

Per-entity performance analysis:

- Disease Names: 91% accuracy
- Test Names: 89% accuracy
- Measurement Values: 86% accuracy
- Symptoms/Observations: 85% accuracy

The transformer-based NLP model was particularly effective in handling unstructured text and extracting meaningful information from complex medical terminology.

### *5.3 Disease Classification Performance*

The ML classification module predicts abnormal findings and potential diseases based on extracted and structured data. Using Random Forest as the primary model, the system achieved an overall classification accuracy of 87% on the test dataset.

Per-class performance analysis:

- Blood-related Abnormalities: 90% accuracy
- Cardiovascular Conditions: 85% accuracy
- Metabolic Disorders: 86% accuracy
- Imaging Report Abnormalities: 83% accuracy

Confidence scores were provided for each prediction, allowing healthcare professionals to assess reliability and prioritize further investigation.

### *5.4 Summarization and Visualization Performance*

The summarization module produced concise, easy-to-understand summaries of complex medical reports. Evaluation based on user feedback indicated:

- Summary Completeness: 88%
- Clarity and Readability: 90%
- Time Saved per Report: Average 5 minutes

The visualization module generated interactive charts and graphs for test trends and abnormal findings. Users reported that visualizations significantly enhanced understanding of report data.

### *5.5 User Experience Evaluation*

Field testing was conducted with 25–30 users, including medical students, lab technicians, and patients. Results demonstrated strong practical usability:

- User Satisfaction: 89%
- Accuracy Agreement with Manual Review: 86%
- Interface Usability Rating: 4.4/5.0
- Average Session Duration per Report: 10 minutes

Feedback highlighted the intuitive interface, accurate extraction of critical data, and clear visualizations as key strengths. Users appreciated the reduction of manual effort and time in report interpretation.

### *5.6 Comparison with Baseline Approaches*

| Model | User Satisfaction | Accuracy | Real-time Adaptation |
|---|---|---|---|
| Manual Review | 80 % | 82 % | No |
| Simple OCR + Rule-based NLP | 84 % | 85 % | No |
| Proposed Smart Medical Analyzer | 89 % | 87 % | Yes |

**Table 1 - Performance Comparison**

## 6. Deployment and Impact

The OCR module demonstrated robust text extraction performance across various medical report types, including scanned PDFs, laboratory test reports, and handwritten notes. Overall OCR accuracy was 93%, with laboratory test reports achieving 95%, radiology reports 90%, and handwritten notes 85%. Preprocessing steps such as noise removal, grayscale conversion, and deskewing significantly improved recognition accuracy, particularly for low-quality or skewed reports. This ensured that extracted text was reliable and suitable for downstream NLP and machine learning tasks.

The NLP and classification modules effectively processed the extracted text to identify key medical entities and predict abnormal findings. Using transformer-based models like BioBERT, entity recognition achieved an overall accuracy of 88%, with disease names at 91%, test names at 89%, and symptoms or observations at 85%. The machine learning classification module, primarily using Random Forest, achieved 87% accuracy in predicting abnormalities across various conditions, including blood-related issues, cardiovascular conditions, metabolic disorders, and imaging report anomalies. Summarization and visualization components further enhanced usability by providing concise summaries, trend graphs, and clear visual cues for abnormal results, with users reporting clarity and time savings of approximately 5 minutes per report.

User experience evaluation with 25–30 participants, including medical students and healthcare professionals, highlighted the system's effectiveness and usability. User satisfaction was 89%, and agreement with manual review results reached 86%, demonstrating the system's reliability. The interactive interface, automated extraction, structured summaries, and visualizations were cited as key strengths. Compared to baseline approaches like manual review and simple OCR with rule-based NLP, the Smart Medical Report Analyzer provided higher accuracy, real-time analysis, and improved efficiency, making it a practical solution for handling diverse medical reports in clinical and educational settings.

## 7. Conclusion

The Smart Medical Report Analyzer successfully integrates OCR, NLP, and machine learning techniques to provide an automated, accurate, and user-friendly system for analyzing medical reports. By extracting text from scanned and image-based reports, identifying key medical entities, classifying abnormalities, and generating concise summaries with visualizations, the system significantly reduces manual effort and processing time. Field testing confirmed the system's reliability, achieving high accuracy in OCR, entity recognition, and disease classification, while users reported strong satisfaction with the interface and outputs.

The system demonstrates the practical value of combining advanced AI and NLP technologies in healthcare applications. It enables faster and more accurate interpretation of medical reports, supports decision-making for healthcare professionals, and improves accessibility for patients seeking understandable report summaries. The visualizations and structured outputs further enhance comprehension, making complex medical data easier to interpret.

Future work includes expanding the system to handle multi-language medical reports, integrating additional machine learning models for rare or complex diseases, adding real-time alert features for critical abnormalities, and developing mobile application support for broader accessibility. Additionally, incorporating multi-modal inputs such as radiology images alongside textual reports could enhance predictive capabilities and provide more comprehensive patient insights. The Smart Medical Report Analyzer thus represents a scalable, intelligent solution for modern healthcare documentation and analysis.

***Acknowledgements***

The authors thank K.L.N. College of Engineering for providing computational resources, Dr. P. Ganesh Kumar, Head of Department - Information Technology, for valuable guidance throughout the project, and agricultural extension services for their support in field validation studies.

**References**

[1] Batista, F. & Evsukoff, A., "Application of Transformers based methods in Electronic Medical Records: A Systematic Literature Review," arXiv preprint, 2023.

[2] Hossain, M., et al., "Natural Language Processing in Electronic Health Records in Relation to Healthcare Decision-making: A Systematic Review," arXiv preprint, 2023.

[3] "A systematic review of natural language processing applied to radiology reports," BioMed Central, vol. 22, pp. 1-18, 2021.

[4] "Machine learning in medicine: a practical introduction to natural language processing," BioMed Central, vol. 19, pp. 1-15, 2021.

[5] Zhang, L., et al., "Automatic medical report generation based on cross-view attention and visual-semantic LSTMs," MDPI, vol. 11, no. 4, pp. 456-471, 2023.

[6] Microsoft Business Intelligence, "AI-Powered Automation for Medical Document Summarization," Industry Case Study, 2023.

[7] Li, Y., et al., "Deep learning for clinical text mining: A comprehensive review," Computers in Biology and Medicine, vol. 142, pp. 105-124, 2022.

[8] Doshi-Velez, F., et al., "Interpretable machine learning for healthcare: A survey," Journal of Healthcare Informatics Research, vol. 6, no. 1, pp. 1-30, 2022.

[9] Wang, H., et al., "Multi-modal approaches for medical report analysis using deep learning," Artificial Intelligence in Medicine, vol. 120, pp. 102-120, 2022.

[10] Chen, J., et al., "Transformer-based models for medical entity recognition and report summarization," Journal of Biomedical Informatics, vol. 129, pp. 1021-1035, 2024.