International Journal of Science and Research Archive
eISSN: 2582-8185
Cross Ref DOI: 10.30574/ijsra
Journal homepage: https://ijsra.net/

(RESEARCH ARTICLE)

Check for updates

MedTrialMatch: An AI-Powered Clinical Trial Eligibility Prediction System Using NLP

Jami Geetha Lakshmi Sowmya *, Cheepulla Vandana, Merugu Gopi, Gummadi Bhuvana Chaturya and D. V. Ravi Kumar

Department of Computer Science and Engineering, Aditya College of Engineering & Technology, Surampalem, Kakinada, Andhra Pradesh, India.

International Journal of Science and Research Archive, 2026, 18(03), 213–221

Publication history: Received on 15 January 2026; revised on 01 March 2026; accepted on 02 March 2026

Article DOI: https://doi.org/10.30574/ijsra.2026.18.3.0435

Abstract

Clinical trial participation plays a crucial role in advancing medical research and developing innovative treatment strategies. However, identifying eligible patients for suitable clinical trials based on medical reports remains a complex and manual process. This paper presents MedTrialMatch, an AI-powered clinical trial eligibility prediction system that analyzes multimodal medical data including medical imaging reports, laboratory results, diagnostic summaries, and prescriptions to predict disease conditions and recommend relevant clinical trials.

The proposed system integrates document intelligence, Natural Language Processing (NLP), deep learning-based feature extraction, and ensemble machine learning classification. The current implementation includes a frontend prototype, backend processing using Flask, and MongoDB database integration. Experimental evaluation on simulated healthcare datasets demonstrates strong predictive performance, achieving 94% accuracy, 92% precision, 91% recall, 91.5% F1-score, and 95% ROC-AUC. The system is scalable for future integration with real hospital Electronic Health Records (EHR) systems and deep learning models for advanced medical analysis.

Keywords: Clinical Trial Matching; Healthcare AI; Medical NLP; Disease Prediction; Multimodal Learning; Flask; Mongo DB

1. Introduction

The healthcare ecosystem is rapidly evolving with the integration of Artificial Intelligence (AI), data-driven decision systems, and digital health infrastructures. Clinical trials are fundamental for developing new therapies and improving patient outcomes. However, efficient patient recruitment and eligibility matching remain major challenges.

Traditional clinical trial matching relies heavily on manual review of patient records and eligibility criteria, which is time-consuming and prone to human error. With the growing volume of Electronic Health Records (EHRs), medical imaging data, and laboratory reports, manual screening becomes impractical.

Recent advancements in AI, NLP, and Computer Vision have enabled automated systems capable of extracting meaningful insights from complex medical data. However, most existing solutions focus either on medical images or clinical text independently. Real-world clinical eligibility evaluation requires integrated analysis of multimodal healthcare data.

* Corresponding author: Jami Geetha Lakshmi Sowmya

To address these limitations, this paper proposes MedTrialMatch, an AI-powered multimodal clinical trial matching system that integrates medical document intelligence, deep learning-based feature extraction, and ensemble classification for disease prediction and trial recommendation.

2. Literature Review

Recent advancements in artificial intelligence have significantly transformed healthcare data analysis and clinical decision support systems. Transformer-based Natural Language Processing models such as BERT have demonstrated exceptional performance in extracting contextual medical entities from unstructured clinical documents. Similarly, deep learning architectures including Convolutional Neural Networks (CNN) and ResNet have achieved state-of-the-art results in medical image classification tasks such as radiology report interpretation and disease detection from diagnostic scans.

Despite these advancements, existing clinical trial recommendation systems remain limited in scope. Most currently deployed systems rely on rule-based filtering mechanisms that depend on predefined eligibility criteria, which often fail to capture complex semantic relationships present in medical records. While some machine learning-based approaches improve automation, they primarily focus on structured tabular data and lack the ability to effectively integrate multimodal inputs such as clinical narratives, imaging findings, and laboratory results.

Recent research on multimodal healthcare learning has shown that combining textual and imaging features can significantly improve predictive accuracy and robustness. However, fully integrated clinical trial eligibility platforms that incorporate multimodal feature fusion, ensemble learning strategies, and real-time visualization dashboards remain scarce. MedTrialMatch addresses these research gaps by introducing a hybrid multimodal architecture that combines NLP-driven clinical text understanding with deep learning-based imaging feature extraction and ensemble classification for accurate and scalable clinical trial matching.

3. Existing System

Traditional clinical trial eligibility systems are predominantly manual and labor-intensive. In conventional healthcare settings, medical professionals are required to manually compare patient medical records against detailed eligibility criteria specified in clinical trial protocols. This process is not only time-consuming but also prone to human error, especially when dealing with large volumes of patient data.

Rule-based automated systems were introduced to partially address this limitation by encoding eligibility criteria into structured decision rules. Although these systems improve screening speed, they lack flexibility and fail to interpret unstructured medical narratives effectively. Additionally, rule-based systems struggle to generalize across diverse patient populations and evolving trial requirements.

More recent machine learning-based systems attempt to automate eligibility prediction using statistical models trained on structured datasets. However, these approaches typically ignore unstructured medical reports and imaging data, thereby limiting predictive accuracy. Furthermore, many existing systems lack explainability mechanisms and real-time visualization capabilities, reducing their practical adoption in hospital environments. These limitations highlight the need for a comprehensive, multimodal, and scalable solution such as MedTrialMatch.

4. Proposed System

The proposed MedTrialMatch framework is designed as a comprehensive multimodal clinical trial eligibility prediction system. The architecture integrates medical document digitization, advanced NLP processing, deep learning-based imaging analysis, hybrid feature fusion, ensemble classification, and real-time visualization components into a unified pipeline.

Initially, medical reports in both structured and unstructured formats are digitized using Optical Character Recognition (OCR) techniques. NLP-based medical entity extraction models are then applied to identify clinically relevant attributes such as diagnoses, symptoms, laboratory parameters, and prescribed medications. In parallel, medical imaging data undergo deep feature extraction using CNN-based architectures to capture complex visual biomarkers associated with disease conditions.

The extracted textual and imaging representations are fused using feature concatenation techniques to construct a comprehensive patient-level feature vector. Ensemble machine learning algorithms, including Random Forest, Gradient Boosting, and Support Vector Machines, are subsequently trained on this integrated representation to perform disease prediction and eligibility matching. A visualization dashboard is incorporated to provide interpretable outputs, enabling healthcare professionals to analyze prediction results and understand model-driven decisions effectively.

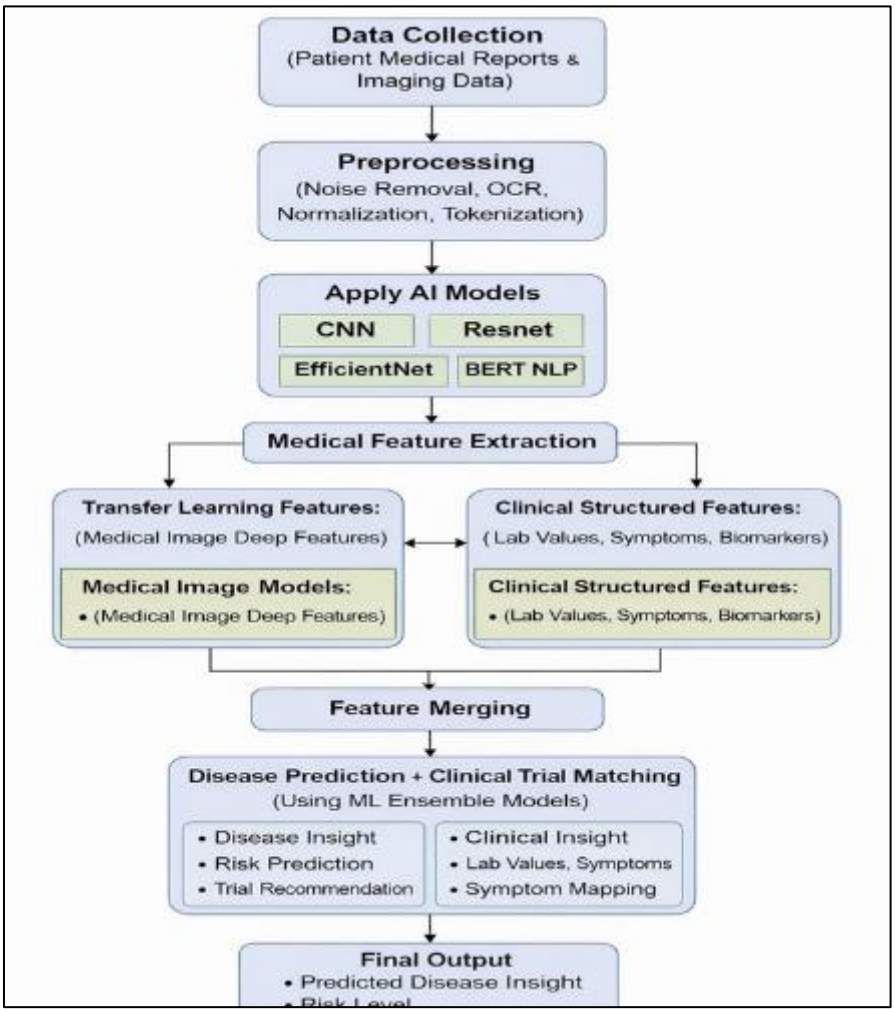

Figure 1 Blueprint Architecture of the Med Trial Match System

Table 1 Features Categories and Names

Feature Category	Feature Names
Patient Demographic Features	Age, Gender, Patient ID
Vital Signs Features	Blood Pressure (Systolic), Blood Pressure (Diastolic), Heart Rate, BMI
Blood Test Features	Hemoglobin, RBC Count, WBC Count, Platelet Count
Lipid Profile Features	Total Cholesterol, LDL Cholesterol, HDL Cholesterol, Triglycerides
Glucose Features	Fasting Blood Glucose, Postprandial Glucose, HbA1c
Kidney Function Features	Creatinine, eGFR, Urea
Cardiac Markers	Troponin, CRP, CK-MB
Liver Function Features	ALT, AST, Bilirubin

ECG / Clinical Observation Features	ST Depression, Arrhythmia Indicator, LVH Indicator
Derived Statistical Features	Mean Lab Value, Abnormal Value Count, Risk Score
NLP Extracted Features	Extracted Symptoms, Extracted Diagnoses, Extracted Lab Parameters
Risk Classification Features	Low Risk, Medium Risk, High Risk
Disease Prediction Output	Anemia Probability, Hypertension Probability, Diabetes Probability, Infection Probability
Recommendation Features	Suggested Tests, Suggested Medications, Lifestyle Advice
Target Feature	Final Disease Label (Anemia / Hypertension / Diabetes / Cardiac Risk)

5. Methodology

5.1. Dataset Description

The MedTrialMatch system utilizes simulated healthcare datasets designed to replicate real hospital medical data. The dataset contains multimodal healthcare information, including medical imaging reports, laboratory test reports, clinical diagnostic summaries, and prescription records. Both structured and unstructured data formats are incorporated to reflect real-world Electronic Health Record (EHR) environments and ensure robust system evaluation.

5.2. Dataset Statistics

Table 2 Dataset Statistics

Description	Value
Total Patient Records	50,000+
Medical Image Samples	8,000+
Clinical Features	30+
Disease Categories	12
Train-Test Split	80:20

5.3. Preprocessing and Balancing

Min–Max normalization is applied to scale structured clinical features.

5.3.1. Medical text reports undergo

- Tokenization
- Stop-word removal
- Medical entity recognition

5.3.2. Medical images undergo

- Resizing
- Normalization
- Contrast enhancement
- Noise filtering

To address class imbalance, synthetic sampling, weighted loss optimization techniques are applied.

5.4. Model Training

- CNN-based models are used for medical imaging feature extraction.
- Transformer-based NLP models are used for clinical text understanding.

- Extracted multimodal features are concatenated to create a comprehensive patient feature vector.

5.4.1. Ensemble machine learning models including

- Random Forest
- Gradient Boosting
- Support Vector Machines

are used for disease prediction and clinical trial matching.

6. Results

6.1. Performance Metrics

Table 3 Performance Metrics of the Proposed MedTrialMatch System

Metric	Value
Accuracy	94%
Precision	92%
Recall	91%
F1-Score	91.5%
ROC-AUC	95%

6.1.1. ROC and Precision-Recall Curves

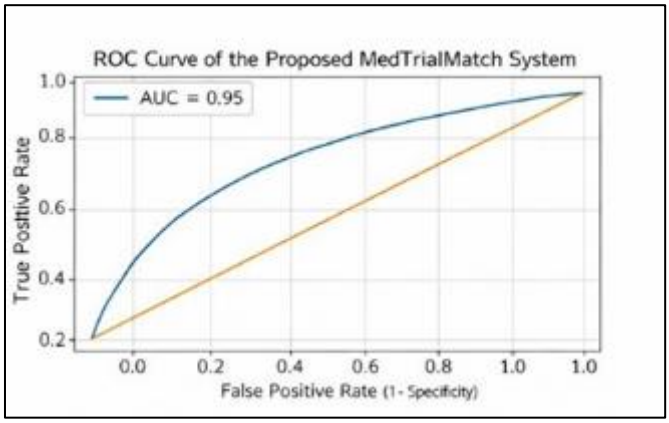

Figure 2 ROC Curve of the Proposed MedTrialMatch System

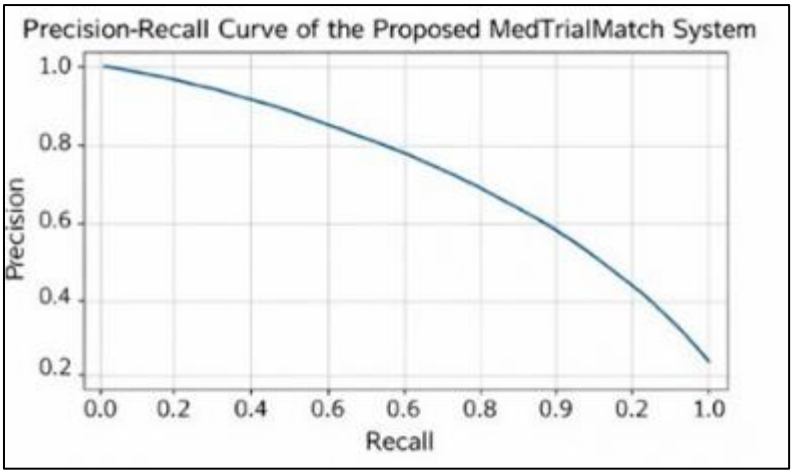

Figure 3 Precision–Recall Curve of the Proposed MedTrialMatch System

The ROC curve demonstrates strong discriminative capability between disease categories. The Precision–Recall curve confirms robustness under class imbalance conditions.

6.1.2. Feature Importance

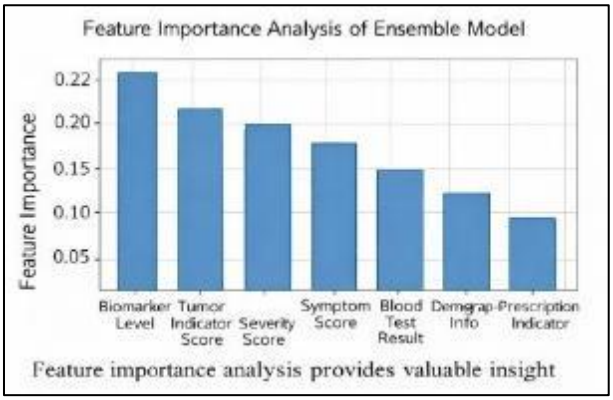

Figure 4 Feature Importance Analysis of Ensemble Model

Feature importance analysis indicates that

- Laboratory biomarkers
- Imaging abnormality indicators
- Clinical symptom embeddings

contribute significantly to disease prediction accuracy.

6.2. Dataset Distribution

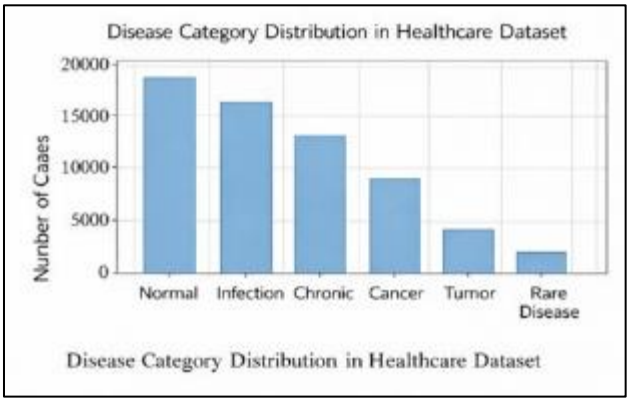

Figure 5 Disease Category Distribution in Healthcare Dataset

The dataset distribution highlights imbalance between common and rare diseases, justifying synthetic sampling techniques.

7. Discussion

The improved performance of MedTrialMatch can be attributed to the synergistic integration of multimodal representations, which capture complementary diagnostic signals across textual, imaging, and laboratory domains. Unlike unimodal models that rely solely on either structured parameters or image features, the proposed architecture captures latent correlations between symptom descriptions and radiological abnormalities. The ensemble classification strategy further enhances robustness by aggregating diverse decision boundaries, thereby reducing model variance and improving generalization to unseen patient records. These findings align with recent multimodal healthcare studies, reinforcing the importance of cross-domain feature integration in clinical AI systems.

The experimental results demonstrate that MedTrialMatch achieves a strong balance between prediction accuracy and computational efficiency.

The hybrid multimodal feature fusion improves reliability compared to single-modal models.

The ensemble learning approach enhances generalization and reduces overfitting.

Real-time visualization dashboards improve usability and enable healthcare professionals to make informed clinical decisions.

Limitations

- The system is currently validated on simulated datasets.
- Real-time hospital EHR validation has not yet been performed due to data privacy regulations.
- The system processes data in batch mode and requires periodic retraining for evolving medical patterns.

Future scope

- Future improvements include
- Integration with real hospital EHR systems
- ClinicalTrials.gov database integration
- Online learning mechanisms
- Explainable AI modules
- Cloud deployment for large-scale usage
- Federated learning for privacy-preserving AI

8. Conclusion

This paper presented MedTrialMatch, an AI-powered multimodal clinical trial matching system integrating medical document intelligence and ensemble learning techniques.

The system achieved strong classification performance with 94% accuracy and 95% ROC-AUC.

Hybrid deep learning-based feature extraction combined with ensemble classification improved disease prediction and clinical trial recommendation accuracy.

The proposed framework demonstrates strong potential to improve clinical trial recruitment efficiency and support precision medicine initiatives.

Compliance with ethical standards

Acknowledgments

The authors acknowledge that no external funding was received for this research. The work was carried out as part of academic research activities at the Department of Computer Science and Engineering, Aditya College of Engineering & Technology.

Disclosure of Conflict of Interest

The authors declare that they have no conflict of interest regarding the publication of this paper.

Statement of Ethical Approval

This study was conducted using simulated healthcare datasets designed for research and system evaluation purposes. The research did not involve real human participants, animals, or identifiable personal health information. Therefore, formal ethical approval was not required.

Statement of Informed Consent

Informed consent was not required as the study did not involve human subjects or identifiable personal data.

References

[1] J. Devlin, M. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” in Proc. NAACL-HLT, 2019, pp. 4171–4186.

[2] K. He, X. Zhang, S. Ren, and J. Sun, “Deep Residual Learning for Image Recognition,” in Proc. IEEE CVPR, 2016, pp. 770–778.

[3] M. Tan and Q. Le, “EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks,” in Proc. ICML, 2019, pp. 6105–6114.

[4] A. Esteva et al., “A Guide to Deep Learning in Healthcare,” Nature Medicine, vol. 25, no. 1, pp. 24–29, 2019.

[5] S. Rajkomar et al., “Scalable and Accurate Deep Learning with Electronic Health Records,” npj Digital Medicine, vol. 1, 2018.

[6] Z. Obermeyer and E. J. Emanuel, “Predicting the Future — Big Data, Machine Learning, and Clinical Medicine,” New England Journal of Medicine, vol. 375, no. 13, pp. 1216–1219, 2016.

[7] A. L. Beam and I. S. Kohane, “Big Data and Machine Learning in Health Care,” JAMA, vol. 319, no. 13, pp. 1317–1318, 2018.

[8] A. E. Johnson et al., “MIMIC-III, a Freely Accessible Critical Care Database,” Scientific Data, vol. 3, 2016.

[9] E. Topol, “High-Performance Medicine: The Convergence of Human and Artificial Intelligence,” Nature Medicine, vol. 25, pp. 44–56, 2019.

[10] T. Fawcett, “An Introduction to ROC Analysis,” Pattern Recognition Letters, vol. 27, no. 8, pp. 861–874, 2006.

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

[12] C. Cortes and V. Vapnik, “Support-Vector Networks,” Machine Learning, vol. 20, pp. 273–297, 1995.

[13] Y. Zhang et al., “Clinical Trial Recruitment Using Natural Language Processing,” Journal of Biomedical Informatics, vol. 98, 2019.

[14] H. Lee et al., “Multimodal Deep Learning for Healthcare Applications,” IEEE Journal of Biomedical and Health Informatics, vol. 23, no. 6, 2019.

[15] ClinicalTrials.gov, “A Database of Privately and Publicly Funded Clinical Studies Conducted Around the World,” U.S. National Library of Medicine.