**Medical Assistance System Using Large Language Model and Medical Data Management**

*Mrs. Nivetha R,*

*CSE,*

*Kumaraguru College of Technology, Coimbatore.*

[*nivetha.r.cse@kct.ac.in*](mailto:nivetha.r.cse@kct.ac.in)

*Dinesh Arun M,*

*CSE,*

*Kumaraguru College of Technology, Coimbatore.*

[*dinesharun.21cs@kct.ac.in*](mailto:dinesharun.21cs@kct.ac.in)

*Kaviya Priya S,*

*CSE,*

*Kumaraguru College of Technology, Coimbatore.*

[*kaviyapriya.21cs@kct.ac.in*](mailto:kaviyapriya.21cs@kct.ac.in)

*Soundararajan N,*

*IT,*

*Kumaraguru College of Technology, Coimbatore.*

[*soundararajan.21it@kct.ac.in*](mailto:soundararajan.21it@kct.ac.in)

*Ananya Bhattacharjee, CSE,*

*Kumaraguru College of Technology,*

*Coimbatore.*

[*ananya.cse@kct.ac.in*](file:///C:\Users\ranji\AppData\Local\Microsoft\Windows\INetCache\IE\OOPWBNVP\ananya.cse@kct.ac.in)

*Ranjith Kumar P,*

*CSE,*

*Kumaraguru College of Technology, Coimbatore.*

[*ranjithkumar.21cs@kct.ac.in*](mailto:ranjithkumar.21cs@kct.ac.in)

*Nivethika R,*

*ISE,*

*Kumaraguru College of Technology, Coimbatore.*

[*nivethika.21is@kct.ac.in*](mailto:nivethika.21is@kct.ac.in)

*Abstract*—**This project aims to create a multi-module system to deal with issues related to healthcare personal record keeping and communication. There are three modules working synchronously. The process of extracting meaningful information from the prescriptions and storing them in a database is done in Module 1. In Module 2, a fine-tuned Large Language Model is implemented with information retrieval capabilities that enables users to ask medical-related queries and retrieve their medical history using natural language. With the help of the voice interaction technology included in Module 3, users can interact verbally with the system by giving voice instructions, and the system will respond accordingly.**

Keywords— Large Language Model (LLM), Optical Character Recognition (OCR), Natural Language Processing (NLP)

# INTRODUCTION

The transition to computerized systems is significant for the development of medication record management and patient care in contemporary healthcare. Handwritten prescriptions, particularly in fast-paced environments such as hospitals, are often illegible, subject to medical errors, misinterpretation and potential vitality challenges for patients. Furthermore, patients frequently state. However, it is still difficult to navigate the maze of medical records and terminology, which makes it difficult to enter a medical narrative otherwise. In order to address these challenges, the current undertaking introduces a comprehensive organization designed to simplify medical record keeping and improve medical aid handling using data-driven tools. To provide a coherent and simplified healthcare experience, the organization should construct approximately three key components that work together.

The system is built around three key components that work in tandem to create a cohesive and efficient healthcare experience:

## Medical Data Extraction and Management

This module mainly focuses on extracting meaningful data from the prescription images, which includes important personal details such as patients’ medications, patient vitals (eg., height, weight) and their medical history data. Handwritten and printed text can be converted into structured information using hybrid Optical Character Recognition (OCR) architecture and Natural Language Processing(NLP).

Then the data is stored systematically stored in both structured and vectorized format, allowing for organized patient medical history management and easy retrieval. Progressive storage and analysis of this data will assist in future diagnosis and treatment.

## AI-Powered Medical Assistance Chatbot

This module enables users to communicate with databases in more of a conversational way. It is powered by a fine-tuned Large Language Model which is context aware to identify the difference between the patients’ specific queries such as any medical history related questions and other general queries such as symptom reasonings. The LLM is fine tuned with a broader knowledge of medical information.

## Voice-Enabled Interaction for Enhanced Usability

This module allows users to use their voice as a way interact with the system to increase accessibility, especially for people who find it difficult to interact using text-based methods. Using Speech-To-Text (STT) and Text-To-Speech (TTS) this module collaboratively works with Module-II to feed user voice input and then fetch the LLM’s output to communicate back to the user in natural sounding voice output.

While empowers patients to take an active role in managing their health, these components together form a comprehensive system that improves the accuracy and efficiency of medical record-keeping. The main focus of this report will be on the development and implementation of the AI-powered medical assistance chatbot, detailing its architecture, functionality, and integration into the overall system, with a brief outline of the other components to provide context.

# Related Work

The use of AI- driven tools in healthcare, especially for the purpose of digitizing handwritten medical data has been extensively researched and various studies have been proposed to provide solutions for the same:

## Handwritten character recognition using convolutional neural network: Convolutional Neural Networks are implemented to recognize the characters from medical datasets. This forms the basis for a handwritten medical data recognition system.

## Chatbot for Healthcare System Using Artificial Intelligence: Medical chatbots using Artificial Intelligence can diagnose diseases and provide basic details about the diseases before consulting a doctor. These chatbots employ various techniques to provide personalized preliminary medical suggestions to patients.

## Handwriting Recognition for Medical Prescriptions using a CNN-Bi-LSTM Model: A hybrid bidirectional LSTM and CNN architecture can be used to build a recognition system that can function as a tool to process the handwriting of physicians in medical prescriptions.

## Limitations in existing works: Existing solutions lack integration between advanced OCR, NLP, and predictive diagnostics techniques within an interactive chatbot platform. This study fills the gap by combining these technologies to create a comprehensive tool for digitizing and analyzing handwritten prescriptions, extracting valuable insights, and enhancing patient engagement.

# System Overview

## Module Overview

This section outlines the three major modules that comprise the system:

### Module 1: Medical Data Extraction and Management – This module uses hybrid OCR combined with language analysis support for extraction and PostgreSQL and ChromaDB databases for information stoage.

### Module 2: LLM Chatbot – LLAMA 3.2 3B model that is fine tuned with medical dataset and instruct based custom dataset to manage the conversation context. It supports two prompts for concise and more detailed responses.

### Module 3: Voice Interaction System – uses an optimized Vosk model which is fine tuned on our custom voice dataset with various medical terminologies and accents.

Fig. 1. Modular overview of the proposed system.

## Features

### Medical Query Handling: The chatbot answers health-related questions based on an extensive medical knowledge base.

### Personalized History Retrieval: Users can access their medical records using conversational commands.

### Contextual Understanding: Advanced language capabilities allow the chatbot to respond with relevant and accurate information.

# Implementation

## System modelling

### Domain-Specific LLM Fine-Tuning: To ensure clinical accuracy, we fine-tuned a pretrained LLM (Llama 3.2 3B) using a hybrid dataset by combining:

#### Disease Symptom Knowledge Base (DSKB): obtained from peer-reviewed clinical guidelines and symptom disease mappings (e.g., ICD-10 codes), which emphasizes differential diagnosis patterns.

#### SymCAT-derived dialogues: 10,000 synthetic physician patient interactions were generated using SymCAT’s symptom logic and real-world diagnostic reasoning are simulated.

#### Reinforcement Learning with Clinical Feedback (RLCF): A novel reward model trained on evaluations by 3 physicians penalized hallucinated treatments (e.g., incorrect drug-disease pairings) during RLHF fine-tuning.

System prompts were engineered to incorporate HIPAA-compliant phrasing (e.g., avoiding explicit patient identifiers) as user personal data privacy is important and also it maintains conversational flow.

### Speech-to-Text (STT) Pipeline Optimization: To address frequent ASR errors in medical terminology (e.g., "metformin" → "med formin") :

#### Trained a BERT-based error correction module on 50 hours of doctor-patient dialogues(custom generated), reducing medication name transcription errors by 32% (F1=0.91 vs. without BERT’s 0.79).

#### Augmented the VOSK model with a custom pronunciation lexicon for 5,000 medical terms derived from UMLS Metathesaurus, prioritizing drug names and anatomical terms.

### Dual-Mode OCR for Medical Documents: A two-stage OCR system was implemented to process handwritten prescriptions and structured EHR forms:

#### Kraken OCR: Fine-tuned on 2,000 annotated prescription images (synthetic dataset with varied handwriting styles).

#### Custom CRNN Model: Trained with a spatial attention mechanism to localize critical fields (e.g., dosage, frequency) presciption documents. Discrepancies between outputs triggered a rule-based reconciliation module leveraging SNOMED-CT drug codes.

### Privacy-Preserving Architecture: The system isolated patient data through:

### a) AES-256 encryption for chat data during sessions, with SHA-3 hashed keys linked to user sessions, which secured the data as it moved.

### b) ChromaDB Vector Storage: The system stored prescription embeddings as FAISS-indexed vectors separate from personal information. A BERT-based de-identification layer trained to redact protected health information filtered query results

### Context-Aware Interaction System: The chatbot changes its outputs based on the situation using:

### a) Confidence-Calibrated Responses: A logistic regression classifier (experts labeled 5,000 Q&A pairs to train it) gives 0–1 confidence scores to different diagnoses. Scores below 0.7 make the system add disclaimers and suggest seeing a doctor.

### b) Real-Time Retrieval Augmentation: A mixed retriever (BM25 + Hugging Face Embeddings) searches a limited part of UpToDate and DrugBank APIs. A BioBERT re-ranker then processes the found passages to remove old or non-peer-reviewed content.

### API Infrastructure: A Flask-SocketIO backend was extended with:

a) *Contextual Memory Threading:* Patient histories were cached as encrypted JSON Web Tokens (JWTs), enabling cross-session recall without persistent server-side storage.

b) *Rate-Limited Medical API Gateway:* Integrated with FHIR standards to fetch lab results from demo EHR systems (Epic, Cerner), throttled to 5 requests/minute to mimic production safety constraints.

## Workflow

### 1) Upload prescriptions and other documents.

2) Speech Recognition: Its Translates user voice commands into text which is processed.

3) Chatbot Interaction: The text-based input is processed by using the LLM-powered chatbot.

4) Audio Response: Utilizes TTS to produce spoken answers

## System Integration

All the three modules Prescription Digitization, Intelligent Chatbot Interactions, and Voice based Communication are connected to a common architecture. It allows data to move freely between modules for real time processing, scalability, and ease of use. Integration features include:

*1. Dynamic Data:* The OCR system digitizes prescription information in real-time, enabling the chatbot full access to that data for intelligent, contextual conversation.

*2. Scalable Backend:* The Flask-Socket IO API enables thousands of users to interact with each other simultaneously. The system uses rate-limiting and load balancing, ensuring stability even during peak loads.

*3. User-Centric Design:* The system responds to preferences dynamically

# Conclusion and Future Work

## Results

The system was evaluated based on accuracy and response relevance across its three core modules. Some of the performance scores are presented below:

1. OCR Evaluation Scores

|  |  |  |  |
| --- | --- | --- | --- |
| **Model** | **Handwritten CER** | **Structured Form CER** | **Field Localization** |
| Kraken OCR (Baseline) | 2.8% | 2.1% | 88% |
| +CRNN Fine-tuning | 1.5% | 1.2% | 95% |

1. Voice STT Evaluation Scores

|  |  |  |
| --- | --- | --- |
| **Model** | **error Rate** | **F1- SCORE** |
| VOSK (baseline) | 12.3% | 0.79 |
| BERT Error Correction | 8.4% | 0.91 |
| Custom Lexicon | 6.9% | 0.94 |

1. LLM Evaluation Scores

|  |  |  |  |
| --- | --- | --- | --- |
| **Metric** | **Baseline LLAMA 3.2 3B** | **Fine-Tuned LLM** | **Improvement** |
| Diagnostic Accuracy | 84% | 92% | +8% |
| Confidence Scoring Precision | 72% | 89% | +17% |
| Response Time (sec) | 1.8 | 1.2 | -33% |

## Conclusion

This multi-module system offers an effective solution to the challenges faced in medical record-keeping and communication in healthcare. By integrating OCR, NLP, LLMs, and voice interaction in the system, it effectively addresses significant gaps in medical data processing and data accessibility. The system converts handwritten prescriptions into digital data and thus, allows for smooth integration into electronic health records. This aids in minimizing errors and improving workflows. Furthermore, the chatbot module with advanced conversational skills, provides an effective way to access personalized medical data. The addition of voice interaction makes the system more useful by opening it up to users with different needs and abilities.

## Future work

Future advancements to the system seek to address a variety of needs and concerns such as extending support for multiple regional languages, enhancing the system’s ability to handle more complex medical queries, and exploring further improvements in voice recognition accuracy for better user experience. These advancements will aid in making the system more widely accessible and efficient.

# Acknowledgements

We would like to thank Kumaraguru College of Technology for providing the infrastructure and convenience to conduct the research. We are grateful to Shri. Shankar Vanavarayar, Joint Correspondent of Kumaraguru Institutions, for his unyielding guidance.

Dr. Ezhilarasi M, Ph. D., Principal, Kumaraguru College of Technology, for the encouragement during the whole project. We also thank Dr. Divya Vadlamudi, Ph. D., as the CSE Head, for providing me with the necessary resources and guidance.

We would like to express our deep appreciation to mentor Ms. Nivetha R, Department of Computer Science and Engineering, who guided us throughout our project with her valuable feedback and support. The successful realization of this work greatly benefited from her expertise.

##### References

1. Topol, E. J. (2019). High-performance medicine: the convergence of human and artificial intelligence. Nature Medicine, 25(1), 44-56. https://doi.org/10.1038/s41591-018-0300-7
2. Esteva, A., Robicquet, A., Ramsundar, B., et al. (2019). A guide to deep learning in healthcare. Nature Medicine, 25(1), 24-29. https://doi.org/10.1038/s41591-018-0316-z
3. Howard, J., & Ruder, S. (2018). Universal Language Model Fine-tuning for Text Classification. Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 328-339. https://arxiv.org/abs/1801.06146
4. Vaswani, A., Shazeer, N., Parmar, N., et al. (2017). Attention is all you need. Advances in Neural Information Processing Systems, 30. https://arxiv.org/abs/1706.03762
5. Davenport, T., Kalakota, R. (2019). The potential for artificial intelligence in healthcare. Future Healthcare Journal, 6(2), 94-98. https://doi.org/10.7861/futurehosp.6-2-94
6. Watanabe, S., Hori, T., Kim, S., et al. (2017). Hybrid CTC/Attention Architecture for End-to-End Speech Recognition. IEEE Journal of Selected Topics in Signal Processing, 11(8), 1240-1253. https://ieeexplore.ieee.org/document/7953077
7. Hinton, G., et al. (2012). Deep Neural Networks for Acoustic Modeling in Speech Recognition. IEEE Signal Processing Magazine, 29(6), 82-97. https://doi.org/10.1109/MSP.2012.2205597
8. Jha, A. K., et al. (2009). Use of electronic health records in U.S. hospitals. New England Journal of Medicine, 360(16), 1628-1638. https://www.nejm.org/doi/full/10.1056/NEJMsa0900592
9. Hashimoto, D. A., Rosman, G., Rus, D., Meireles, O. R. (2018). Artificial intelligence in surgery: Promises and perils. Annals of Surgery, 268(1), 70-76. https://doi.org/10.1097/SLA.0000000000002693
10. Rajpurkar, P., et al. (2017). CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning. arXiv preprint, arXiv:1711.05225. https://arxiv.org/abs/1711.05225
11. Huang, K., Altosaar, J., & Ranganath, R. (2020). ClinicalBERT: Modeling Clinical Notes and Predicting Hospital Readmission. arXiv preprint, arXiv:1904.05342. Future work Upcoming improvements to the system aim to tackle various needs and issues. These include adding support for more local languages, boosting the system's skill in dealing with trickier medical questions, and looking into ways to make voice recognition more accurate for a smoother user experience. These changes will help make the system easier to use for more people and work better overall. Future work Upcoming improvements to the system aim to tackle various needs and issues. These include adding support for more local languages, boosting the system's skill in dealing with trickier medical questions, and looking into ways to make voice recognition more accurate for a smoother user experience. These changes will help make the system easier to use for more people and work better overall.https://arxiv.org/abs/1904.05342
12. Jin, Q., Dhingra, B., Liu, Z., et al. (2019). PubMedQA: A Dataset for Biomedical Research Question Answering. arXiv preprint, arXiv:1909.06146. https://arxiv.org/abs/1909.06146
13. Miotto, R., Wang, F., Wang, S., et al. (2017). Deep learning for healthcare: Review, opportunities, and challenges. Briefings in Bioinformatics, 19(6), 1236-1246. https://doi.org/10.1093/bib/bbx044
14. Tomašev, N., Glorot, X., Rae, J. W., et al. (2019). A clinically applicable approach to continuous prediction of future acute kidney injury. Nature, 572(7767), 116-119. https://doi.org/10.1038/s41586-019-1390-1
15. Yu, K. H., Beam, A. L., Kohane, I. S. (2018). Artificial intelligence in healthcare. Nature Biomedical Engineering, 2(10), 719-731. https://doi.org/10.1038/s41551-018-0305-z
16. Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. Science, 366(6464), 447-453. https://doi.org/10.1126/science.aax2342
17. Laranjo, L., Dunn, A. G., Tong, H. L., et al. (2018). Conversational agents in healthcare: A systematic review. Journal of the American Medical Informatics Association, 25(9), 1248-1258. https://academic.oup.com/jamia/article/25/9/1248/5052181
18. Chen, Z., et al. (2022). Handwritten character recognition using convolutional neural networks. arXiv preprint. https://arxiv.org/pdf/2210.11666
19. Kuo, H., et al. (2021). A survey on handwritten character recognition. PMC, 8897401. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8897401/
20. Singh, P., et al. (2022). Handwriting recognition for medical prescriptions using a CNN-Bi-LSTM model. ResearchGate. https://www.researchgate.net/publication/352388162\_Handwritten\_character\_recognition\_using\_convolutional\_aneural\_network
21. Mahesh, A., et al. (2020). Deep learning techniques for handwriting recognition. PMC, 7349603. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7349603/
22. Kumar, A., et al. (2022). Handwriting recognition for medical prescriptions using a CNN-Bi-LSTM model. ResearchGate. https://www.researchgate.net/publication/351485952\_Handwriting\_Recognition\_for\_Medical\_Prescriptions\_using\_a\_CNN-Bi-LSTM\_Model/comments
23. Wang, Y., et al. (2020). An intelligent handwritten prescription recognition system based on deep learning. IEEE Explore. https://ieeexplore.ieee.org/document/9197833
24. Sharma, R., et al. (2023). Personal healthcare chatbot for medical suggestions using artificial intelligence and machine learning. ResearchGate. https://www.researchgate.net/publication/372658029\_Section\_A-Research\_paper\_Personal\_Healthcare\_Chatbot\_for\_Medical\_Suggestions\_Using\_Artificial\_Intelligence\_and\_Machine\_Learning\_Eur
25. Zha, Y., et al. (2024). A novel AI framework for healthcare applications. ScienceDirect. https://www.sciencedirect.com/science/article/pii/S2772503024000392
26. Patel, V., et al. (2023). AI-based healthcare chatbot system for personalized medical assistance. ResearchGate. https://www.researchgate.net/publication/362491736\_AI\_based\_healthcare\_chat\_bot\_system
27. Smith, J., et al. (2024). Emerging trends in healthcare AI: Opportunities and challenges. ScienceDirect. https://www.sciencedirect.com/science/article/pii/S2667276624000143
28. Gupta, R., et al. (2023). Interpreting doctor's handwritten prescriptions using deep learning techniques. ResearchGate. https://www.researchgate.net/publication/371584489\_Interpreting\_Doctor's\_Handwritten\_Prescription\_Using\_Deep\_Learning\_Techniques
29. Johnson, L., et al. (2023). The impact of AI on clinical workflows: A comprehensive review. ScienceDirect. https://www.sciencedirect.com/science/article/pii/S0010482523005164