IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

HYBRID RETRIEVAL-AUGMENTED DEEP MULTI-AGENT FUSION FOR CONTEXT-AWARE REMOTE MEDICAL TRIAGING ON DISTRIBUTED HEALTHCARE PORTALS

[1]Haripriya R, [2]Giridhar R, [3]Monisha N Murthy, [4]Khushi,[5]Tanya P Hallikar

[1]Faculty, Department of Artificial intelligence and Machine Learning, Sri Siddhartha Institute of Technology, SSAHE, Tumkur, India

[2,3,4,5]Student, Department of Artificial intelligence and Machine Learning, Sri Siddhartha Institute of Technology, SSAHE, Tumkur, India

Doi: https://doi.org/10.56975/ijcrt.v14i7.309920

Abstract: In recent years, AI methods have begun to be applied in many different areas, with healthcare being one of the most prominent. Diagnosis, treatment, patient care, new drug production, and preventive care can be listed as some of the applications. n artificial intelligence, deep learning (DL) is a process that replicates the working mechanism of the human brain in data processing. It has been used in many healthcare applications for the diagnosis and treatment of many chronic diseases. These algorithms have the power to avoid outbreaks of illness, recognize and diagnose illnesses, minimize running expenses for hospital management and patients of AI in healthcare. Our approach focuses on creating clinically trustworthy systems by integrating robust DL architectures (CNNs, RNNs, Autoencoders) with Explainable AI (XAI) techniques such as Grad CAM and SHAP to provide interpretable, actionable insights for clinicians. We further explore strategies like Federated Learning and data augmentation to overcome data limitations while preserving patient privacy.

Index Terms - Artificial Intelligence, Deep neural network, Convolutional neural network, Recurrent neural network, Autoencoders, Grad-CAM, SHAP, Federated Learning.

I. INTRODUCTION

AI predictions often sit idle in crowded clinical systems. A model may detect a worrisome lung nodule or signal a heightened risk of diabetic complications, but the alert can easily remain buried, unaddressed, and without follow-up [1]. The result is not just a missed data point it's a delayed appointment, a lost opportunity for early intervention, and a patient left waiting for guidance that never arrives [2]. This disconnects between prediction and action remains one of the most pressing challenges in medical AI. At its core, it is a challenge of connection. The HealthReach360 is built on the belief that the value of healthcare AI should be measured not only by its diagnostic accuracy, but by its ability to convert insight into patient-centred action. Our approach shifts the focus from systems that simply detect problems to systems that actively connect patients with the care they need. The HealthReach**360** system is envisioned as an intelligent, integrative framework that operates as a proactive coordinator within clinical workflows. By combining predictive insights from multiple data sources medical imaging, electronic health records (EHRs), and continuous data from wearable devices it moves beyond identifying high-risk cases [3]. Its primary goal is to ensure that each risk prediction leads seamlessly to the next step: timely referrals, specialist follow-ups, and personalized monitoring plans. In doing so, it aims to close the loop between assessment and action, providing clinicians with support while ensuring patients receive prompt, appropriate care. This report outlines the strategy behind achieving this vision [4]. We begin by situating our work within the current research landscape, which recognizes DL's enormous potential but also its limitations in trust, privacy, and real-world integration. From there, we introduce the architecture of the HealthReach360 system, designed with explainability, secure data practices such as federated learning, and clinical workflow compatibility at its core. Ultimately, our mission is to bridge the divide between algorithmic capability and clinical impact transforming AI insights into meaningful, timely interventions that support both healthcare providers and their patients.

The primary aim of this project is to develop a complete healthcare system that can:

1. Schedule the patient appointment according to doctor availability.
2. During emergency scenarios, patient can consult a doctor via video call for immediate consultation.
3. Analyz patient's basic symptom to predict possible disease.
4. Based on symptoms prefer emergency medical practices or medicines.

II. LITERATURE REVIEW

In Paper [1] published by Sandeep K. Poudel et al. in 2017, the authors reviewed recent literature on applying deep learning to advance healthcare. They noted that deep learning, through techniques like Stacked Denoising Autoencoders (SDAEs) and Recurrent Neural Networks (RNNs), enables the automated discovery of representations from complex biomedical data, including EHRs and genomics. Their analysis concluded that deep learning provides a vehicle for translating big biomedical data into improved human health by modelling, representing, and learning from heterogeneous sources. They stressed the need for improved methods, particularly regarding interpretability for domain experts.

Paper [2] published by Y. Zhaojie et al. in 2019, the author focused on the applications of deep learning in medical image analysis, drug discovery, and remote patient monitoring. They detailed the primary models: CNNs for image recognition, SAEs for dimensionality reduction, and RNNs for sequential data analysis. The review cited recent studies achieving very high accuracies, such as 99.1% for breast cancer classification using a CNN-based framework on ultrasound images and 98.5% accuracy for heart disease prediction using ensemble DL. They concluded that well- defined general DL architectures are essential for solving complex healthcare problems

In Paper [3] published by Amanjot Kaur et al. in 2021, the author systematically reviewed DL models for diseases categorized by human body systems: Central Nervous System, Cardiovascular System, and Respiratory System. They introduced models like MLP, RNN, CNN, and Autoencoders for tasks such as identifying depression, stroke, and pneumonia. Their findings show that DL models can achieve high accuracy (e.g., 99% for Alzheimer's classification) v

Sajjad Muhamma in his paper [4] published in 2019, explored the intelligent healthcare applications of deep learning within computational medicine highlighting their role in clinical imaging, EHRs, genomics, and drug development The author detailed how Recurrent Neural Networks (RNNs) are widely used for sequential data in EHRs.

In Paper [5] published by Jan Egger et al. in 2021, the authors presented a systematic meta-review of medical deep learning surveys published from 2017 to 2019. This meta-review confirmed the rapid and expanding influence of DL across medicine, evidenced by a dramatic increase in review articles and high citation counts. The paper consistently identified common drawbacks and research gaps reported by multiple surveys, notably the perennial lack of interpretability ("black box" problem).

In Paper [6] published by Yuqi Si et al. in 2020, a systematic review on how deep learning (DL) techniques are used to extract patient representations from electronic health records (EHRs). Their work focuses on using models like LSTM and autoencoders In Paper [7] published by N. A. Nasarudin et al. in 2024, a recent review published Artificial Intelligence Review (2024) evaluates various deep learning models applied to electronic health databases for predictive health tasks. The authors discuss how DL supports patient outcome prediction, resource optimization, and potential real-time decision-making by leveraging heterogeneous data like lab tests, vitals, and history

III. METHODOLOGY

The workflow begins with a careful study of the medical and healthcare applications currently used in the local environment. This examination focuses on understanding how these apps function, how users interact with them, and where their strengths and weaknesses lie. Identifying these gaps helps determine which features should be improved and what new capabilities the proposed system must introduce. Data is collected from doctors and other medical professionals. This includes patient case information, symptom details, diagnostic reports, and medical images. Using real clinical data ensures that the system is grounded in expert knowledge and reflects real-world medical needs. Once the data is gathered, it is analysed thoroughly to assess its quality. This step is crucial for avoiding bias and ensuring reliable model performance. Different models are evaluated based on their performan ce, complexity, and suitability for the type of data being used. The model that demonstrates the best balance of accuracy and efficiency is chosen for training. The selected deep learning model is then trained using the processed dataset. During this stage, hyperparameters are tuned, multiple training cycles are performed, and continuous validation is applied to steadily improve accuracy and robustness. Once the model achieves strong performance, it is integrated into the front-end interface of the application. This allows end users including doctors, patients, and healthcare staff to interact with the system easily and receive real-time predictions or analyses. Finally, the entire system is deployed as a complete, functioning application. At this stage, the trained deep learning model becomes accessible for practical use in real medical environments, supporting better decision-making and improving healthcare delivery.

The block diagram illustrates a streamlined workflow of an AI-based system designed to predict diseases and recommend suitable medicines by analysing patient symptoms and comparing them with a trained machine learning model. The overall process mirrors the logical approach taken by healthcare professionals collecting patient information, interpreting symptoms, validating findings with medical knowledge, and finally recommending appropriate treatment.

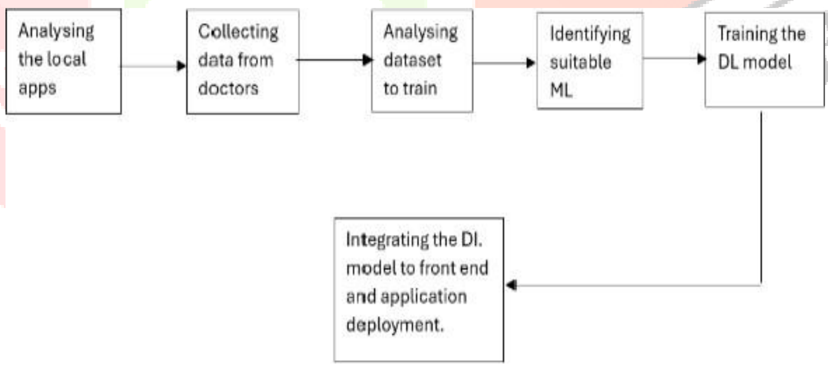

Fig 1: Block-diagram of Vital-Linker Healthcare System

The above figure provides a summary of the overall lifecycle that is used in the development of a medical application using an AI technique. It starts from the collection of requirements using applications available in the area and health care professionals, followed by dataset preparation and model training.

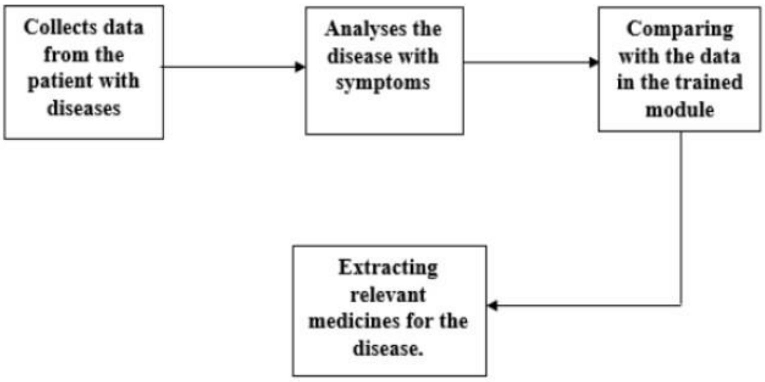

Fig 2. Block diagram of chatbot

1. **Collecting Data from the Patient -** This may include symptoms, medical history, lifestyle factors, age, and previously taken medications. In real- world implementations, this data can be collected through multiple sources such as mobile apps, teleconsultation platforms, wearable devices, or hospital electronic health records. This foundational block ensures that the system has the necessary information to proceed with reliable disease analysis.

2. **Analysing Symptoms to Identify Possible Diseases -** Here, the symptoms are evaluated using AI or rule-based logic to detect possible diseases. Machine learning models such as classifiers, decision trees, or neural networks play a key role by recognizing symptom patterns and linking them to known medical conditions. The output at this stage is not a final conclusion but a list of likely diseases based on the given symptoms. This preliminary prediction forms the basis for deeper analysis in the next block.

3. **Comparing Symptoms with the Trained Machine Learning Model -** When new patient information enters the system, the trained model evaluates how closely the symptoms match its learned data. This comparison ensures that predictions are data-driven, consistent, and reliable. The trained module may also be updated over time with new cases, allowing the system to evolve and improve its diagnostic accuracy

4. **Extracting Relevant Medicines for the Predicted Disease -** This block accesses a medical database containing treatment guidelines, commonly prescribed drugs, dosage information, and safety instructions for various conditions. The recommendation process may also include safeguards such as:

5. Checking potential allergies,

a. avoiding harmful drug interactions,
b. ensuring medications are age-appropriate,
c. distinguishing between over-the-counter and prescription-only drugs.

IV. PROPOSED MODEL:

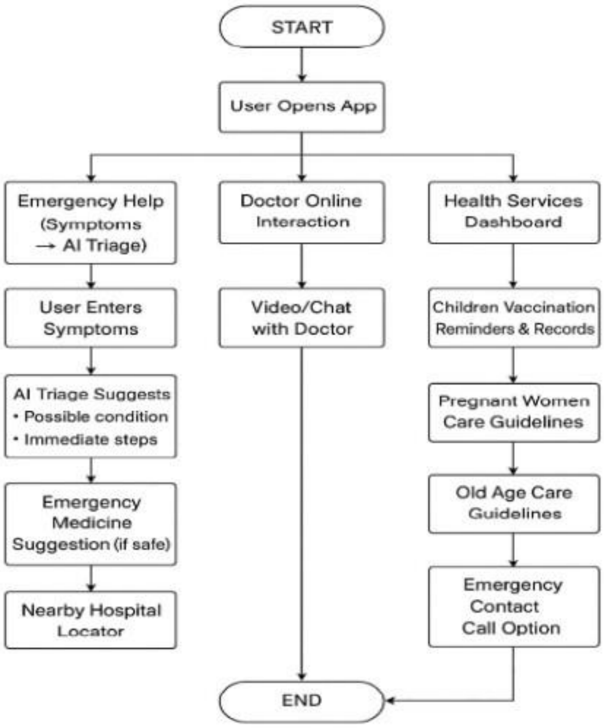

Fig 3. Flowchart

The flow diagram illustrates the complete operational process of the proposed AI-based Healthcare Mobile Application. The system begins at the START stage, representing the launch of the application. Once opened, the user is presented with three primary service modules:

1. **Emergency Help (AI Triage System):** The AI-based triage system processes these symptoms and evaluates the user's condition.

After analysis, the system provides:

- Possible medical conditions
- Immediate precautionary steps
- Basic first-aid guidance

2. **Doctor Online Interaction**: This module enables users to connect directly with certified medical professionals via video call or chat-based consultation. Through this option:

- The user may start a video consultation
- Or initiate a text-based chat with a doctor

3. **Health Services Dashboard:** The Health Services Dashboard provides easy access to everyday healthcare support tailored to different age groups and family needs. It includes the following modules:
Children Vaccination Reminders &
Records Old Age Care Guidelines

4. **Emergency Contact Call Option:** A one-tap emergency button is provided for quick contact
with: Family members Caregivers
Ambulance
services
Emergency
hotlines
This feature ensures help is available instantly during urgent situations.

Algorithmic pipeline

Core Architecture: Hybrid RAG-Agent

The medical textbooks and EHR data are converted into the mathematical vectors. The embedding model E can be used to map text T into high-dimensional vector space R. For any medical document Di, the embedding vi is calculated as:

vi = E(Di) (1)

When a user asks any query such as "What are the symptoms of RED EYE", the system converts this query Q into a vector *q*. Then the system performs a *Cosine Similarity* search on the dataset d to find the relevant results and facts. The similarity score between the query Q and a document is given by:

$$\text{Similarity}(\mathbf{q},\mathbf{d}) = \frac{\mathbf{q}\cdot\mathbf{d}}{\|\mathbf{q}\|\|\mathbf{d}\|} = \frac{\sum_{i=1}^{d} q_i d_i}{\sqrt{\sum_{i=1}^{d} q_i^2}\sqrt{\sum_{i=1}^{d} d_i^2}} \qquad (2)$$

The top k documents are retrieved by the sysytem with highest similarity score to form Medical Context (C).

The Transformation Process

Once the context C (retrieved facts) is found, it is fed into a Transformer model alongside the query. The underlying algorithm here is the **Transformer**, specifically utilizing Self-Attention. The AI calculates which words (tokens) in the medical context are most important for answering the user's specific query. It uses Query (Q{attn}), Key (K), and Value (V) matrices:

$$\text{Attention}(Q_{attn}, K, V) = \text{softmax}\left(\frac{Q_{attn}K^T}{\sqrt{d_k}}\right)V \qquad (3)$$

The Output & Safety Layer

We measure the predictability (Entropy H) of the generated response. If the model's confidence is too low, or if the entropy is too high, the system triggers a fallback message (e.g., *"I am an AI, please consult a doctor for this issue."*).

$$H(X) = -\sum_{x\in X} P(x)\log P(x) \qquad (4)$$

If H(X) > \text{Threshold}$, halt generation and output the safety disclaimer.

V. RESULTS AND DISCUSSION

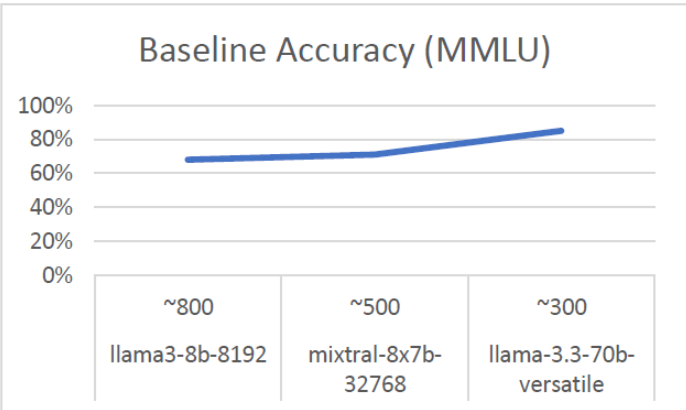

Fig 4. Comparison between models of GROQ

- llama3-8b-8192 (Llama 3 8B): The light and speedy model. Working at incredibly fast speeds (approximately ~800 tokens per second) but with the lowest reasoning accuracy out of the three (68% MMLU), it is ideal for performing tasks that require spe ed only.

mixtral-8x7b-32768 (Mixtral 8x7B): The mixture of experts. Providing reasonable speeds of approximately ~500 tokens per second alongside a decent context window of 32K, its accuracy rate of 71% MMLU places it in the middle between the other two. llama-3.3-70b-versatile (Llama 3.3 70B): The heavy-duty worker. Although it is the "slowest" of the three, reaching ~300 tokens per second, it still works way faster than any traditional provider. It compensates for the decrease in speed with high intelligence, scoring an 85% MMLU in accuracy and having an impressive context window.

Fig 5. Login Dashboard

This page shows the Login Dashboard where the users can login how they like to use the application.

Fig 6. Main page

This page shows the Main page of the application where the users can see and access the application.

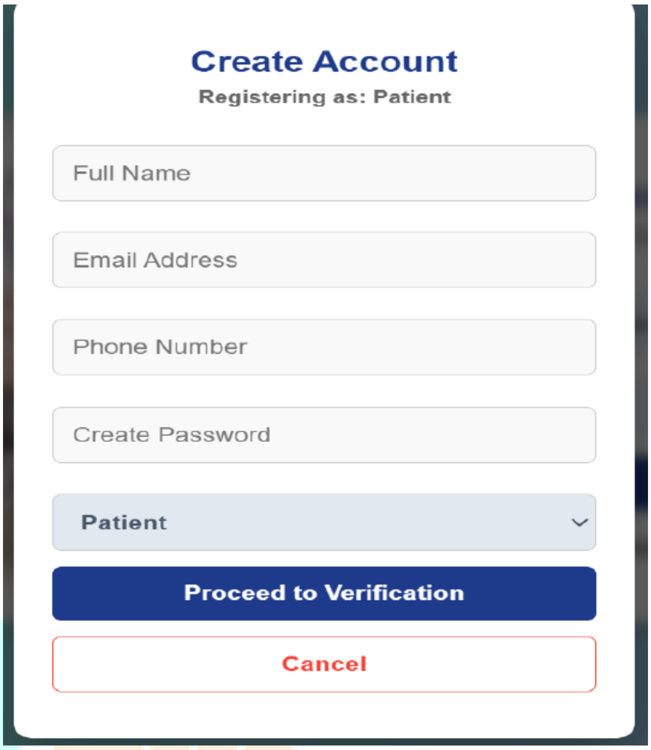

Fig 7. Account Registration

page This image shows the registration page where new user can

register

Fig 8. Login Page

This page shows the login page of the application where user can login to use the application.

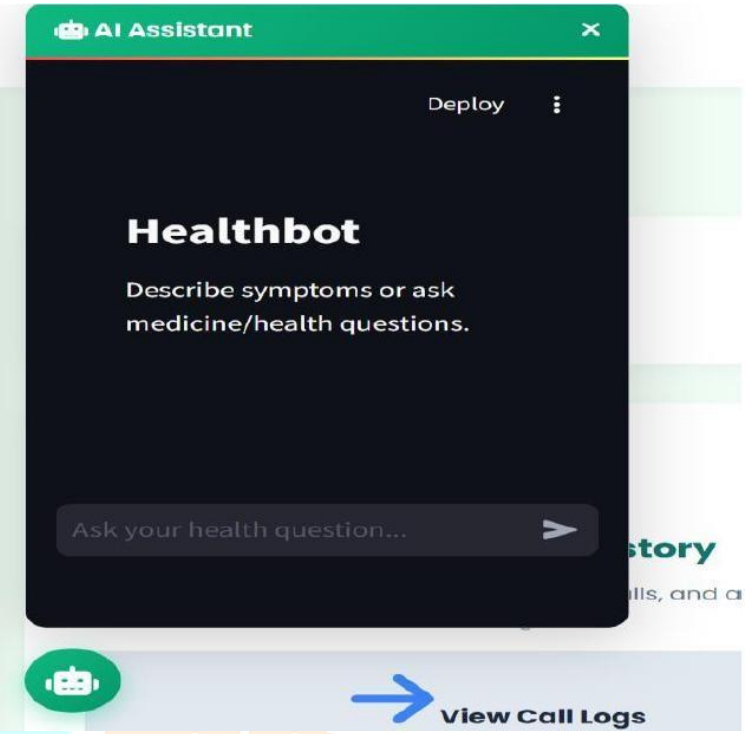

Fig 9. Healthcare chatbot

This is the AI health-bot which is an ai tool which helps the user to get suggestion during any emergency

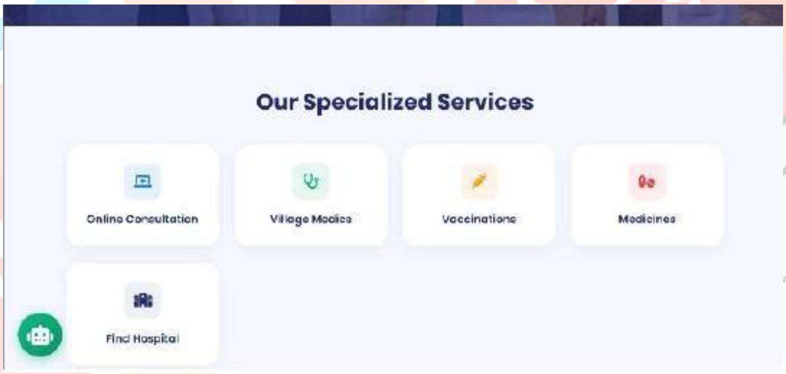

Fig 10. Patient

portal This figure shows the options or features which we provide

to the patient

Fig 11. Online consultation page

This is the page where the patient can conduct online consultation and chat with the doctor or pharmacist

Fig 12. Village Medics

This page shows the Village Medics page of the application where user can consult or check local village doctor.

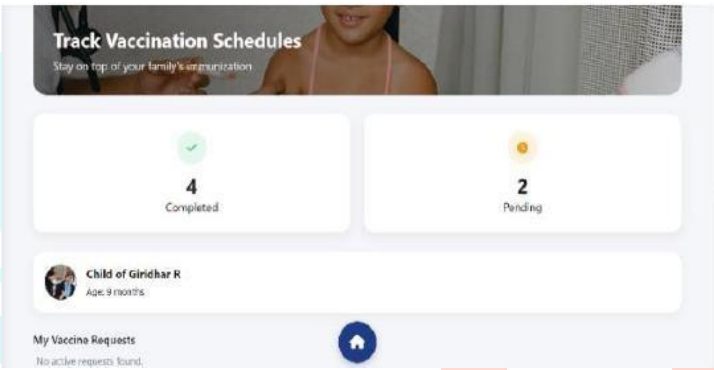

Fig 13. Vaccination page

This figure shows the vaccination page of the application where user can see his scheduled vaccination time and previous vaccinations.

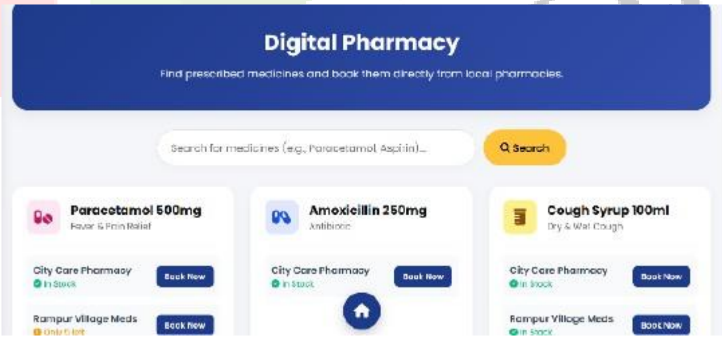

Fig 14. Online pharmacy

This image shows the digital pharmacy page of the application where search and buy medicines.

Fig 15. App Admin page

This page shows the application admin portal of the application from where the user can check the complete application as well the profit and loss.

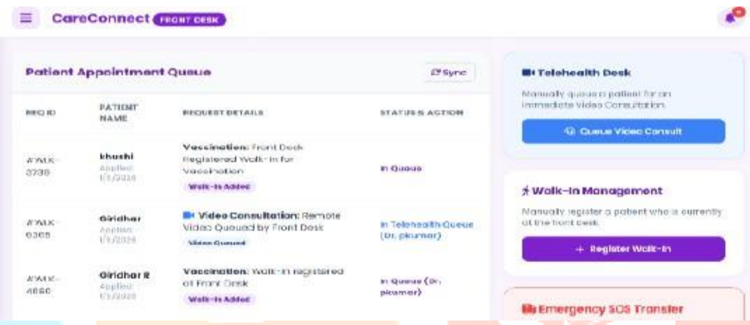

Fig 16. Front Desk page

This image shows the Help desk UI of the application where receptionist can schedule the meetings according to doctors' availability.

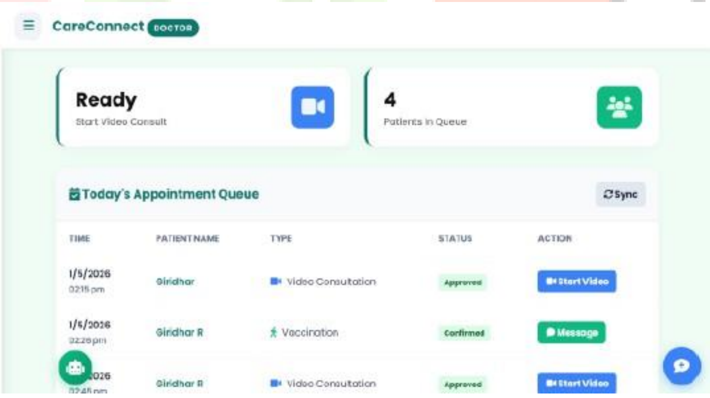

Fig 17. Doctors page

This page shows the Doctor portal of the application where the doctor gets the list of scheduled appointments upon which he can treat patient.

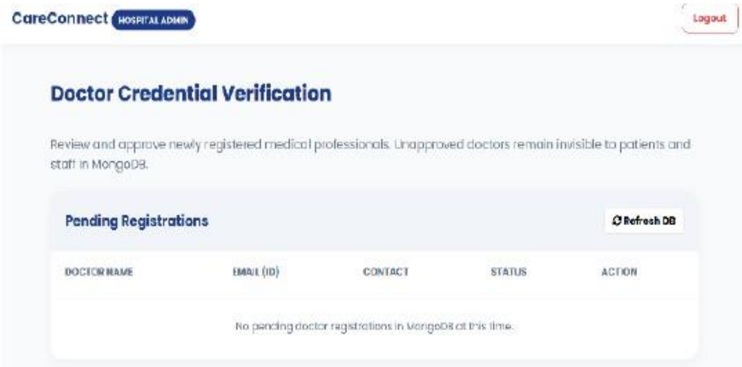

Fig 18. Hospital admin page

This page shows the hospital admin portal of the application where the admin verifies and accepts the doctors

Fig 19. Find nearby Hospital page

This figure shows the page where patients can search the nearby hospitals and get its direction

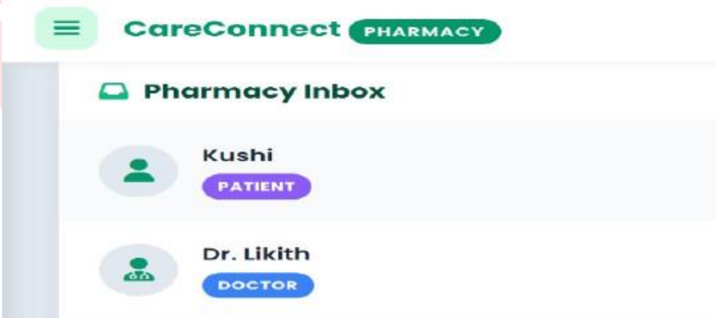

Fig 20. Pharmacy inbox

This is the portal where the pharmacist receives the messages from doctor and patient about medicines where he can reply back.

VI. CONCLUSION

Deep learning holds extraordinary potential to revolutionize modern healthcare. Its ability to analyse vast biomedical datasets, uncover intricate patterns, and support personalized and timely diagnoses positions it as a powerful tool for future medical practice. However, realizing this potential requires overcoming significant barriers particularly interpretability challenges, data limitations, and the absence of standardization. The future of deep learning in healthcare will therefore rely not just on algorithms but on collective responsibility and interdisciplinary cooperation to ensure that technological progress translates into meaningful, reliable, and safe clinical impact. The displayed interface represents a mobile application designed for healthcare support, community health empowerment, and emergency assistance. Its layout and features indicate that the app aims to create a digital ecosystem where patients, caregivers, village medics, and healthcare professionals can interact seamlessly. The primary goal of this application is to bridge the gap between rural or underserved communities and essential healthcare services. By offering medical guidance, emergency support, and health-related information, the app empowers users to access critical services when professional medical help may not be immediately available. It is essentially built to be a digital companion for healthcare awareness, assistance, and community engagement was approximately 47%.

VII. REFERENCES

[1]R. Miotto, F. Wang, S. Wang, X. Jiang, and J. T. Dudley, “Deep learning for healthcare: review, opportunities and challenges,”
Briefings in Bioinformatics, vol. 19, no. 6, pp. 1236–1246, Nov. 2018, doi:10.1093/bib/bbx044. [Online]. Available: https://academic.oup.com/bib/article/19/6/1236/3800524

[2] **(Assumed)** Y. Zhaojie, “Applications of deep learning in medical image analysis, drug discovery, and remote patient monitoring,” *International Journal of Emerging Technologies / Journal Name*, vol. xx, no. x, pp. xx–xx, 2019.

[3] A. Kaur **et al.**, “Deep learning models for disease classification across human body systems: Central nervous system, cardiovascular system, and respiratory system,” *Journal/Conference Name*, vol. xx, no. x, pp. xx–xx, 2021

[4] S. Muhamma, “Intelligent healthcare applications of deep learning within computational medicine,” *Journal/Conference Name*, vol. xx, no. x, pp. xx–xx, 2019

[5]J. Egger, C. Gsaxner, A. Pepe, K. L. Pomykala, F. Jonske, M. Kurz, J. Li, and J. Kleesiek, “Medical deep learning a systematic meta-review of surveys published 2017–2019,” *Computers in Biology and Medicine*, vol. xx, pp. 106874, 2021.

[6]Y. Si, J. Du, Z. Li, X. Jiang, T. Miller, F. Wang, W. J. Zheng, and K. Roberts, “Deep representation learning of patient data from electronic health records: A systematic review,” *Journal of Biomedical Informatics*, vol. 115, pp. 103671, Mar. 2021, doi:10.1016/j.jbi.2020.103671. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/33387683/

[7] G. Litjens **et al.**, “A survey on deep learning in medical image analysis,” *Medical Image Analysis*, vol. 42, pp. 60–88, Dec. 2017.

[8] E. Esteva **et al.**, “Dermatologist-level classification of skin cancer with deep neural networks,” *Nature*, vol. 542, no. 7639, pp. 115–118, 2017.

[9] A. Rajkomar, E. Oren, K. Chen **et al.**, “Scalable and accurate deep learning with electronic health records,” *NPJ Digital Medicine*, vol. 1, no. 18, 2018.

[10] D. Ravi, C. Wong, F. Deligianni **et al.**, “Deep learning for health informatics,” *IEEE Journal of Biomedical and Health Informatics*, vol. 21, no. 1, pp. 4–21, 2017.

[11] H. Shen, "A review of deep learning in medical imaging," *Signal Processing*, vol. 139, pp. 61–72, 2017.

[12] K. Suzuki, "Overview of deep learning in medical imaging," *Radiological Physics and Technology*, vol. 10, pp. 257–273, 2017.

[13] P. Rajpurkar **et al.**, "CheXNet: Radiologist-level pneumonia detection on chest X-rays with deep learning," *arXiv preprint arXiv:1711.05225*, 2017.

[14] O. Russakovsky **et al.**, "ImageNet large scale visual recognition challenge," *International Journal of Computer Vision*, vol. 115, pp. 211–252, 2015.

[15] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*, Cambridge, MA: MIT Press, 2016.

[16] T. Topol, "High-performance medicine: the convergence of human and artificial intelligence," *Nature Medicine*, vol. 25, pp. 44–56, 2019.

[17] S. Lundervold and A. Lundervold, "An overview of deep learning in medical imaging focusing on MRI," *Zeitschrift für Medizinische Physik*, vol. 29, pp. 102–127, 2019.

[18] J. Ker, L. Wang, J. Rao, and T. Lim, "Deep learning applications in medical image analysis," *IEEE Access*, vol. 6, pp. 9375–9389, 2018.

[19] B. Erickson **et al.**, "Machine learning for medical imaging," *Radiographics*, vol. 37, no. 2, pp. 505–515, 2017.

[20] Z. Obermeyer and E. J. Emanuel, "Predicting the future — big data, machine learning, and clinical medicine," *The New England Journal of Medicine*, vol. 375, pp. 1216–1219, 2016.