Journal of Multidisciplinary Healthcare Dovepress

Open Access Full Text Article ORIGINAL RESEARCH

TinyML-Based Lightweight AI Healthcare Mobile Chatbot Deployment

Anita Christaline Johnvictor[1], M Poonkodi[1], N Prem Sankar[1], Thinesh VS[2]

[1]School of Computer Science and Engineering, Vellore Institute of Technology, Chennai, Tamil Nadu, India; [2]Arista Networks Pvt Ltd, Bangalore, India

Correspondence: Anita Christaline Johnvictor, School of Computer Science and Engineering, Vellore Institute of Technology, Chennai campus, Kelambakkam-Vandalur Road, Rajan Nagar, Chennai, 600127, India, Email anitachristaline.j@vit.ac.in

Introduction: In healthcare applications, AI-driven innovations are set to revolutionise patient interactions and care, with the aim of improving patient satisfaction. Recent advancements in Artificial Intelligence have significantly affected nursing, assistive management, medical diagnoses, and other critical medical procedures.

Purpose: Many artificial intelligence (AI) solutions operate online, posing potential risks to patient data security. To address these security concerns and ensure swift operation, this study has developed a chatbot tailored for hospital environments, running on a local server, and utilising TinyML for processing patient data.

Patients and Methods: Edge computing technology enables secure on-site data processing. The implementation includes patient identification using a Histogram of Gradient (HOG)-based classification, followed by basic patient care tasks, such as temperature measurement and demographic recording.

Results: The classification accuracy of patient detection was 95.8%. An autonomous temperature-sensing unit equipped with a medical-grade infrared temperature scanner detected and recorded patient temperature. Following the temperature assessment, the tinyML-powered chatbot engaged patients in a series of questions customised by doctors to train the model for diagnostic scenarios. Patients' responses, recorded as "yes" or "no", are stored and printed in their case sheet. The accuracy of the TinyML model is 95.3% and the on-device processing time is 217 ms. The implemented TinyML model uses only 8.8Kb RAM and 50.3Kb Flash memory, with a latency of only 4 ms.

Conclusion: Each patient was assigned a unique ID, and their data were securely stored for further consultation and diagnosis via hospital management. This research demonstrates faster patient data recording and increased security compared to existing AI-based healthcare solutions, as all processes occur within the local host.

Keywords: edge computing, healthcare, HOG descriptors, assistive management, autonomous robot

Introduction

Interest in leveraging Artificial Intelligence (AI) to enhance the performance, capacity, and efficacy of health care services is increasing. Despite the enormous growth in AI-based healthcare research, only a few AI-based applications have successfully transitioned to clinical use. The main challenges impeding the adoption of AI applications under clinically validated scenarios include lack of medical records in standard formats, minimal access to curated datasets, and strict legal and ethical requirements concerning patient privacy protection. Recently, the use of chatbots has expanded beyond customer service to encompass critical life and death scenarios. Chatbots are increasingly being integrated into the health care sector, offering solutions to various health issues. In the present scenario, patients visiting hospitals are delayed in queue waiting for doctors. However, health and fitness chatbots are gaining traction in the market, allowing users to enquire about their medical concerns and receive responses. Medical chatbots have been developed to address the issue of delays in hospital queues.[1–4] Communication with a chatbot equipped with Natural Language Processing (NLP) plays a significant role in enhancing the mental and physical health of individuals owing to the instantaneous responses they receive. According to Lee and Yoon,[1] the major applications of AI in healthcare include diagnostics,

Received: 18 June 2024
Accepted: 3 October 2024
Published: 9 November 2024

nursing, and managerial assistance. In addition, AI can be used for patient engagement, disease treatment, medical error reduction, and reduced healthcare costs.

Literature Review

According to a report by the World Health Organization,[5] the major factors (around 60%) affecting the health conditions of individuals and their overall quality of life are linked to their lifestyle. These include exercise, dietary habits, sleep patterns, stress management, substance use, medication adherence, and recreational activities.[6] With the emergence of AI-driven technologies, personalised interventions and reminders tailored to individuals' daily routines can now be delivered via digital devices based on real-time vital sign data. In healthcare settings, AI-based technologies are poised to revolutionise operational processes, patient interactions, and care delivery methods, ultimately aiming to improve the overall efficiency and effectiveness of patient outcomes.

Medical Diagnosis

In medical diagnosis, AI is poised to revolutionise the diagnostic process for patients with specific illnesses. For this reason, Machine learning has been introduced as a novel technique.[7] As highlighted by Taylor,[8] "diagnostic errors make up 60% of all medical errors and are responsible for an estimated 40,000–80,000 deaths annually in US hospitals". Hence, the integration of AI-based systems across different healthcare sectors has the potential to mitigate errors resulting from human judgement.[9] The Mayo Clinic, renowned for its pioneering work in patient care, has embraced AI for cervical cancer screening. By harnessing an AI-powered solution, the clinic aims to detect precancerous changes in the cervix. This solution utilises an algorithm trained on a vast database of more than 60,000 images related to cervical cancer from the National Cancer Institute to identify early signs of precancerous conditions. Researchers have found that this algorithm achieves a significantly higher accuracy rate (91%) than that of a trained human expert (69%).[10,11] The Gachon University Gil Medical Center in South Korea evaluated the medical treatment outcomes for one year and arrived at a level of agreement between medical staff and AI-based Watson of 55.9%. However, the consensus rate dropped to 40% in patients diagnosed with stage IV stomach cancer. Furthermore, research at Konyang University Hospital in South Korea in April 2018 disclosed that the agreement rate between doctors' decisions and Watson's treatment recommendations was 48% for a cohort of 100 patients with breast cancer.[12] Manipal Hospital, a prominent Bangalore cancer care centre in India, implemented Watson for cancer treatment in 2015. Notable disparities were observed in diagnoses made by multidisciplinary medical staff and Watson's assessments for a dataset of 1000 cancer patients. This dataset contained a database of cancer cells related to breast, rectal, and colorectal cancers, and was collected by two doctors over three years. For rectal cancer cases, the agreement rate between Watson's recommendations and doctors' decisions was 85%. However, for lung cancer, the consensus rate decreased dramatically, to 17.8%. This substantial difference indicates varying levels of agreement, depending on the nature of the cancer.[13] A chatbot for radio therapy using IBM Watson[14] has been implemented with 95% accuracy.

Assistance Management and Nursing

In relation to nursing and assistance management, healthcare personnel frequently contend with extensive paperwork during the care journey. This tight workload has spurred hospitals to adopt electronic systems that consolidate and create digital medical records facilitated by AI-based technology. Moreover, the utilisation of online chatbots has emerged as a promising means of facilitating conversations with patients and their family members in hospital environments.[15]

AI-powered chatbots have the potential to revolutionise healthcare services in hospitals by streamlining all administrative duties and enabling medical staff or professionals to dedicate more time and attention to patient care, thereby enhancing the quality of their services. Patel and Lam[16] proposed utilising ChatGPT to create discharge summaries in hospitals, whereas Howard et al[16] underscored the capability of ChatGPT in infection consultation. AI chatbot technology could prove invaluable in assisting nurses with various cumbersome tasks, including paperwork, addressing enquiries about nurses, guiding patients to different hospital departments, and providing caregiver training.[16] This assistance could alleviate some of the burden associated with nursing staff shortages and burnout. However, caution must be exercised and robust guidelines should be proposed when integrating AI chatbots into patient healthcare systems

to safeguard patient confidentiality. These measures are essential for upholding ethical standards and preserving patient trust in the industry.[17]

After the COVID-19 pandemic, there has been growing support among researchers for a digital overhaul of mental healthcare, particularly for conditions like dementia.[18] Previous clinical investigations have revealed the effectiveness of using socially compatible assistive robots in therapy to manage behavioural and psychological symptoms of dementia (BPSD). This approach has decreased agitation, boosted social engagement, improved communication, and ultimately alleviated nurses' workload.[19] However, despite the benefits to patients, healthcare workers, and caregivers, this non-pharmacological method incurs substantial costs associated with acquiring, training, and maintaining necessary equipment.

As described by Kim,[20] AI merges nursing knowledge with information technology to manage and convey patient information. Professionals in this area focus on optimising the workflow of nurses and assuring access to necessary information to create better quality inpatient care systems. Despite its significance, nursing informatics has been overlooked in nursing education, partly due to technical limitations in IT.

One of the key features of AI chatbots is their ability to create, modify, and troubleshoot programming codes through natural language conversations. With their intuitive interfaces, even individuals with limited programming skills can overcome technical challenges and develop computer programs. This tool is beneficial for researchers in the field of nursing as it empowers them to build practical computer programs without extensive programming expertise.

Given these scenarios, the use of AI-based chatbots for assistive management, diagnostics, and therapy for various physical and mental health conditions is gaining momentum. The costs involved in acquiring and maintaining AI chatbots are equivalent to those of personal computers and the training required for patients and administrators is minimal. This presents a convenient and cost-effective solution for enhancing patients' well-being, especially in situations where behavioural and psychosocial interventions are limited or halted during epidemic-controlled measures. In view of the potential of this technology, this study has developed a voice-based AI chatbot. The proposed implementation of a voice-based assistive AI chatbot concentrates on collecting information from patients' diagnoses about their symptoms and generates a basic case sheet for the physician. This assistive chatbot would save the time and work of nurses and other frontline health care personnel, and also provide patient satisfaction. The role of Large Language Model (LLM) based AI have also been explored in the conversation of healthcare chatbots.[21]

Proposed Voice-Based AI Chatbot

This study proposes the use of TinyML-based edge computing to implement a chatbot that greets patients in a hospital environment. It has a built-in temperature sensing module. The chatbot asks a series of questions related to patient symptoms. Based on patients' answers, the chatbot records the symptom details of the probable illness and creates a basic case sheet to assist the physician. The proposed AI chatbot is planned to be built as a standalone kiosk or moving robot inside hospital premises. The chatbot (robot) was implemented using the required hardware and software to provide AI-based basic-level medical assistance. TinyML, which refers to ML models and algorithms, is tailored to achieve optimal performance on low-power, resource-limited devices, such as microcontrollers. Implementing TinyML for medical assistance entails the development of compact ML models capable of analysing the basic medical symptoms of patients. The operational sequence involved in the implementation process is illustrated in Figure 1.

Why TinyML-Based Chatbot?

TinyML belongs to the category of edge AI, also known as edge artificial intelligence, which capitalises on the benefits of edge computing, that is, performing computations locally rather than in the cloud. This approach offers several advantages:

1. Local computing ensures low latency, making it suitable for real-time applications.
2. Reduced reliance on remote communication leads to cost savings on bandwidth.
3. Local computing maintains reliability even when network connectivity is disrupted.
4. Enhanced security is achieved by minimizing data transmissions and utilizing local data storage.

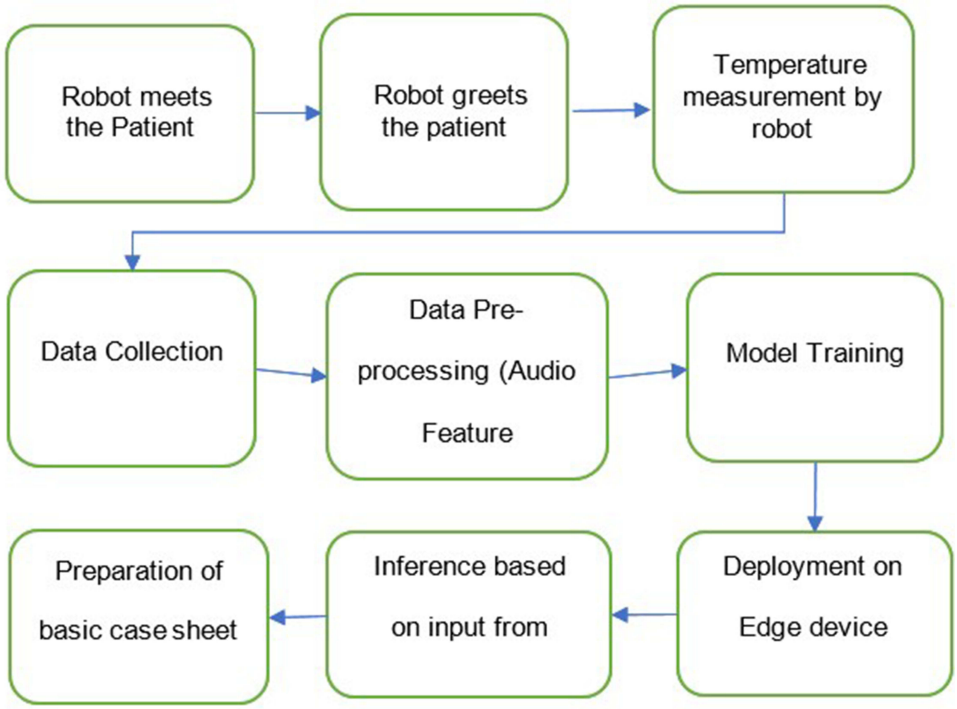

Figure 1 AI Assisted Patient Monitoring System using TinyML.

Different types of chatbot were implemented using rule-based, revital-based, generative, sequence-to-sequence, and transformer-based methods. These chatbots rely on high computational requirements, whereas TinyML is lightweight and can be deployed on edge devices.

The use of TinyML for medical diagnosis offers various advantages, including enhanced healthcare accessibility in remote or underserved regions, continuous patient health monitoring, and prompt detection of medical issues. Nonetheless, it is crucial to tackle challenges such as safeguarding data privacy, ensuring model interpretability, and complying with regulations to ensure the ethical and diligent implementation of TinyML in healthcare contexts.

Implementation

The implementation includes three stages: patient face identification, temperature sensing, and chatbots. The hardware included an ARM Cortex M4 microcontroller board (Arduino Nano 33 BLE sense) with a built-in MEMS microphone and TinyML support. The majority of the questions required a yes or no response from the patient. The model was trained to recognise these keywords according to sex (ie male or female). Features such as patient name, age, and sex, which cannot be recorded using TinyML, require cloud-based deployment. These were stored as audio signals and processed to generate a printed case sheet. Training to ask various questions related to symptoms was cloud-based and deployed at the edge, and the inference part was performed in the mobile chatbot.

Patient Identification

Initially, the face of the patient was identified using an AI assistive edge robot. The robot must distinguish the human face from the other images captured by the camera. This involves a face-identification algorithm that includes data collection, training, and classification. To accomplish this task, the MobileNet V2 architecture was used. This architecture is highly efficient and can be applied to embedded devices with a limited computational capacity. The steps involved in patient identification are as follows:

Step 1: The dataset used for training was obtained from Prajna Bhandary's GitHub repository with 1376 images.

Step 2: Histogram of oriented gradients (HOG)[22] classification involves detecting HOG descriptors from a "P" positive sample.

Step 3: Sampling "N" samples that are negative from a negative training set. This set did not contain any objects required to detect and extract HOG descriptors, generally, $N \gg P$ in practical situations.

Step 4: Training a Linear Support Vector Machine (LSVM) based on the collected positive samples and negative samples.

For each image and possible scale, sliding window techniques were applied, and HOG descriptors were computed and applied to the classifier at each window. If a given window is incorrectly classified, the false-positive patch and its classification probability are recorded using the hard-negative mining approach. During this hard-negative mining, if a few false-positive samples are found with a good classification probability, these false-positive samples are used to re-train the classifier. After training, testing was performed using the sliding window approach, where the HOG features were extracted and presented to the classifier. If the classification probability is high, the bounding box of the window is recorded. If overlapping bounding boxes exist, nonmaximum suppression is used to remove them. Thus, a custom and unique object identification classifier was trained and built to identify the patients.

Temperature Measurement

An autonomous temperature-sensing unit was integrated into the robot to obtain the temperature readings of the patients. The unit uses an infrared contactless temperature sensor so that there is no physical contact between the robot and the patient. The temperature-sensing unit is in sleep mode to conserve energy. When a patient stands in front of the robot and is detected, temperature sensing is possible. The steps involved in the temperature measurement are as follows:

Step1: When a human (patient) face is detected, Raspberry Pi serially communicates with the Arduino uno through the Firmata protocol.

Step 2: When data are received serially through the Firmata protocol, the Arduino Uno makes a pin high for 1s, and then decreases.

Step 3: Arduino nano BLE 33 is connected to the Arduino uno. Arduino nano BLE 33 contains a temperature-sensing code, and if there is any change in the Arduino uno-pin, it activates the temperature-sensing unit.

Step 4: Once triggered, the code provides a 10-second time slot for the patient to place his/her forehead before the sensor.

Step 5: After 10s, the sensor records the patient's temperature.

Step 6: Once the temperature is obtained, the Arduino nano BLE 33 sends the data to the second Raspberry Pi 2 via serial communication.

Step 7: Temperature data are converted into JSON format and sent to a Google cloud spreadsheet.

The code used is the Adafruit mlx90614 library. The library provides two sets of readings obtained from the sensor: the ambient temperature and the object temperature. Only the object temperature reading is used because it takes the temperature readings from an object (ie the patient). The sensor measuring range is approximately 3–5 cm and has an accuracy of ±0.5 °C. The measured temperature reading is sent to the second Raspberry Pi 2 through serial communication using the Arduino Nano BLE 33. This triggers the chatbot.

Chatbot Implementation

After recording the patient's temperature, the chatbot begins its sequence of operations. The steps involved in the implementation of the chatbot are as follows.

Step 1: Speech Dataset

Speech commands v0.01 are the datasets used in this study.[23] The dataset is obtained from Google and contains a list of common words. Yes, no, male, female, and background noises were used to train the TinyML model. The background noise dataset is included such that the model predicts it as noise when other words or noises other than the four keywords are heard via the microphone. Each dataset lasts for 1s. waw file. The datasets were randomly split into an 80:20 ratio for the training and testing of the model. The following are the number of datasets (ie files) for each keyword used to train the TinyML model: "Yes" – 2377 files, "No" – 2375, "Male" – 4095, "Female" – 3976, and "Background noises" – 1897.

The sampling rate of the audio files is 44.1 kHz, but the development board only supports a maximum sampling rate of 16 kHz. This does not pose any problem because the edge impulse studio downsamples audio files for us to the appropriate 16 kHz.

Step 2: Audio Feature Extraction

The discrete Fourier transform (DFT) and Fast Fourier transform (FFT) are employed for analysing audio data; it seems to be good for non-vocal noises such as breaking glass and knocking. Because this research concentrates on acquiring human voices, the Mel frequency cepstral coefficients (MFCC) that convert raw audio into numbers are employed here. The energy of each filter output is obtained by the summation of the area under the curve after multiplying the filter output by the FFT, which is stored in an array. MFCCs are Discrete Cosine Transform (DCT) of these overlapping energy levels. DCT also allows for compression of the information and decorrelation of the energy values. With regard to voice recognition, DCT and the overall spectrogram shape are helpful.

Step 3: Training the Model

To classify spectrogram images, a convolutional neural network (CNN) was used to identify patterns in the input data. The filters in the CNN layers are updated according to the image features that are MFCCs. The size of each filtered image was reduced by max-pooling, which is crucial for memory efficiency and processor usage. After training the model, the Edge Impulse Studio was used to analyse on-device performance based on the processing power of the cortex M4 chip.

Step 4: Deployment of the Chatbot

Deployment includes conversion into a C++ library or binary firmware after various optimisations. This eases integration with the Raspberry Pi because the Raspberry Pi allows serial communication through its USB. The Arduino C++ library is preferred because the code generated by Edge Impulse Studio provides data such as the loss, latency, and predicted values of each keyword. As these data are not necessary for the Raspberry Pi, the library has been edited to provide only the keyword that the user said if the model is predicting above a threshold of 0.8. This ensures the communication of the keyword to the hardware instead of all words. A few more optimisations of the model and the NN classifier were performed with the enable edge optimised neural (EON) compiler, which can run with 25–55% less RAM, with up to 35% less flash, despite retaining the same accuracy as that of TensorFlow Lite for Microcontrollers. The selection of the quantised (int8) optimisation further reduces RAM and Flash memory usage.

Integration

Integration of these three integral sections is easy because the temperature-sensing unit is already using the serial communication method to send the temperature data through the USB as in Figure 2. The same procedure can be followed to send keywords from the chatbot to Raspberry Pi via serial communication. The data were temporarily stored in Raspberry Pi. After the screening of a patient is completed, the data are sent to a Google cloud spreadsheet, where the answers to the chatbot questions and the temperature readings of the patients are stored in table format. Subsequently, the data were erased in Raspberry Pi.

Results

As discussed in the previous section, there are three operational stages in the proposed AI assistive chatbot with TinyML. They include patient identification, temperature sensing, and chatbots asking a series of questions of the patients and recording their symptoms.

Results of Face Identification

The face identification classifier was trained using TensorFlow and Histogram of Oriented Gradients (HOG) was used to extract the region of interest (ROI), and the classifier was applied only to the ROI. The classification accuracy varies from 90.4% to 95.8%. HOG identification is shown in Figure 3; it shows the suppression of overlapping bounding boxes into a suppressed version (single bounding box).

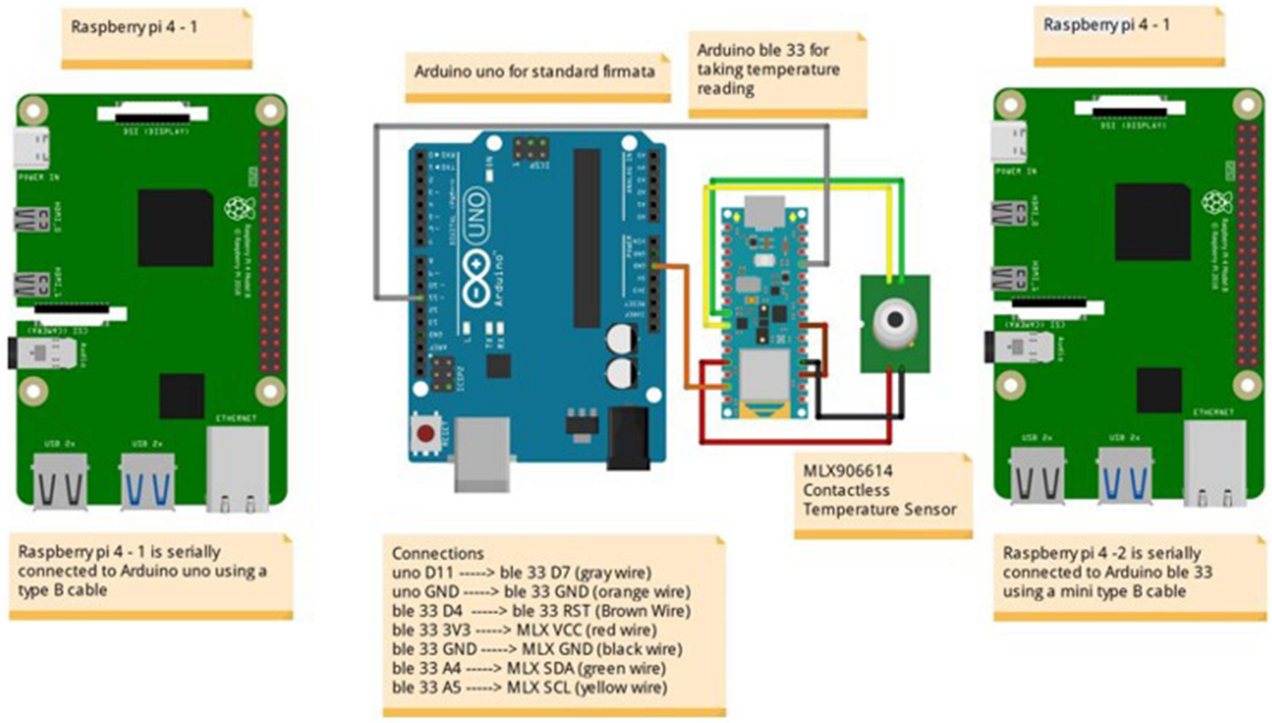

Figure 2 Components setup for AI healthcare chatbot with Tiny ML.

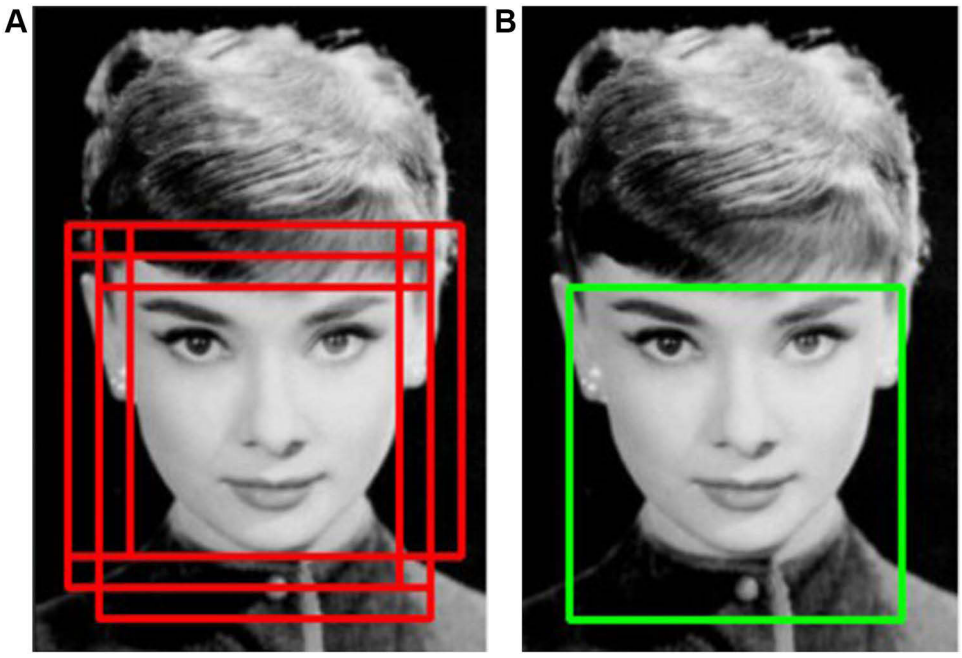

Figure 3 (**A**) The overlapping bounding boxes. (**B**) The NMR suppressed version with single bounding box.

The six overlapping bounding boxes of the same face in Figure 3 were subjected to max pooling to confine them to the largest one while suppressing the five smallest bounding boxes. The training accuracies for the 30 trials are listed in Table 1.

Table 1 The Training Accuracies

Trained Dataset	Accuracy
200	70.2%
600	82.7%
1000	90.4%
1300	95.8%

The parameters were tested and trained with 1300 datasets and a minimum of 10 trials each. For prolonged video streams, the accuracy of the models was approximately 90.2% for the histogram of gradients combined with the region-based convolutional neural network. The classification accuracies are presented in Table 2.

After the face has been detected, the temperature sensor and chatbot are triggered, and the patient has to deliver the appropriate information (answers) to the robot.

Results of Temperature Measurement

The following outputs and results were obtained at room temperature. Table 3 shows the temperature readings recorded for the 15 test subjects to assess their accuracy.

The sensor was calibrated using a medical-grade infrared temperature scanner. After calibration, it was found that the temperature-sensing unit of the robot had an error of approximately 3.8 °F. The 3.8 °F has been added to the temperature sensing algorithm to rectify the error. The outputs were recorded after the error rectification. The time taken by the robot's temperature-sensing unit to record the temperature of the patient is presented in Table 4.

Results of Chatbot

As discussed in the previous section, the processing of speech signals yielded 49 sets of 13 MFCC values. These provide an image spectrogram, as shown in Figure 4. Neural networks were also used for image classification. All calculations were performed automatically using an edge impulse studio. Before generating the features, some FFT parameters may be changed if required or maintained as a default. Because it is nearly impossible to visualise all 637 dimensions, edge impulses tend to combine the MFCCs of each sample and create a 3D plot, as shown in Figure 5.

After training the model, the Edge Impulse Studio was used to analyse on-device performance based on the processing power of the cortex M4 chip. The implemented tinyML model used 17kb of peak RAM and a processing time of approximately 217 ms, as shown in Figure 6, which is very fast.

After testing the model with the test data, the TinyML model yielded an accuracy of 95.3% and loss of 0.20. This accuracy is much better than that of the conventional method implemented using the cloud speech recogniser API tool, which is based on NLP. However, on-device accuracy may vary by 1–2% depending on the board's capability. The training performance of the last training dataset is shown in Figure 7.

Table 2 The Classification Accuracies

Tested Dataset	Accuracy
10	95.8%
20	90%
40	87.6%
50	70%

Table 3 Temperature Reading for 15 Test Subjects

S.No	Medical Grade Infrared Temperature Scanner (in Fahrenheit)	Robot's Temperature Sensing Unit (in Fahrenheit)
1	97.4	97.34
2	97.4	97.49
3	97.4	97.31
4	97.3	97.46
5	97.4	97.35
6	97.4	97.31
7	97.5	97.41
8	97.4	97.33
9	97.1	96.97
10	97.2	97.13
11	97.2	97.16
12	97.2	97.11
13	97.1	96.94
14	97.2	97.26
15	97.2	97.22

Table 4 Average Time Taken by Temperature Sensing Unit

Medical Grade Infrared Temperature Scanner (in Seconds)	Robot's Temperature Sensing Unit (in Seconds)
3.8	0.6

The implemented TinyML model uses only 8.8Kb RAM, 50.3Kb Flash memory with a latency of only 4 ms, as shown in Figure 8. The board has kb RAM of 256Kb and Mb flash memory of 1Mb. The model does not utilise 10% of the board's capabilities. This ensures that the model operates smoothly onboard.

The chatbot was designed to record answers to questions (about symptoms) asked of the patient. If the patient answers yes, that symptom is stored and passed on to the case sheet. If the answer is "no", then that symptom is not recorded. This enables the creation of a basic case sheet for patients. The times taken to record individual answers from patients are shown in Table 5.

Case Sheet Preparation

Finally, both codes (interactive chatbot and Google database sheet update) were merged, and a new table was generated with a unique patient ID. This ID can be used in hospital management to further guide patients for consultation and diagnosis. Access to stored data by hospital management is provided by software developed using the pyqt5 cross-platform to access patient data when patient ID is provided. The code will have a JSON file to connect to the Google cloud spreadsheet. Data were obtained from the cloud. It automatically creates a new word file according to the form provided by hospital management and updates that file with the data. The word file can then be converted into a pdf and printed.

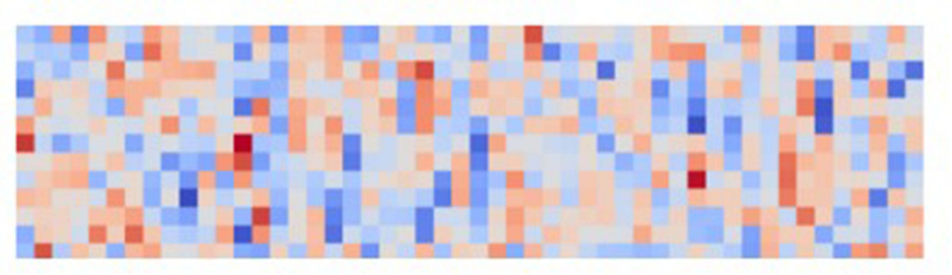

Figure 4 Spectrogram of the processed features.

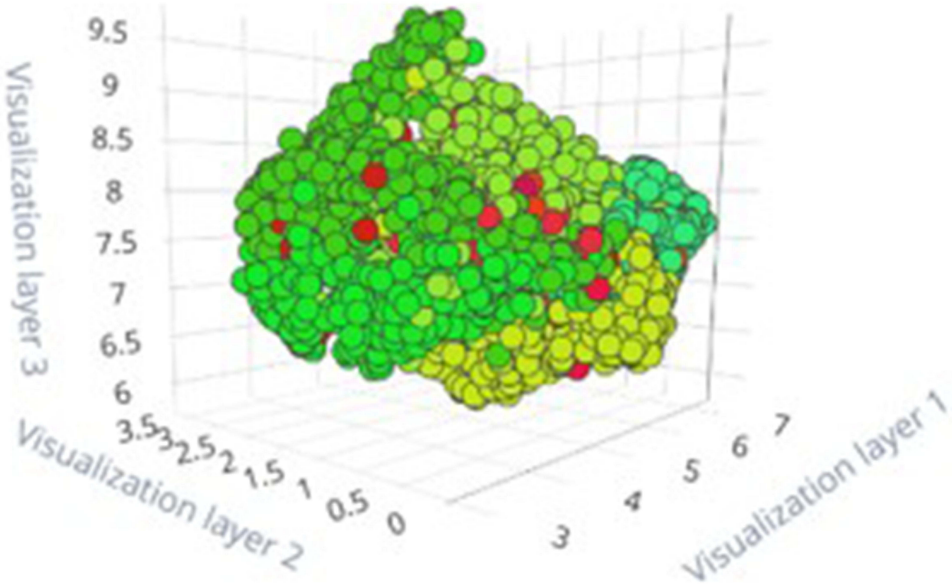

Figure 5 3D plot of features.

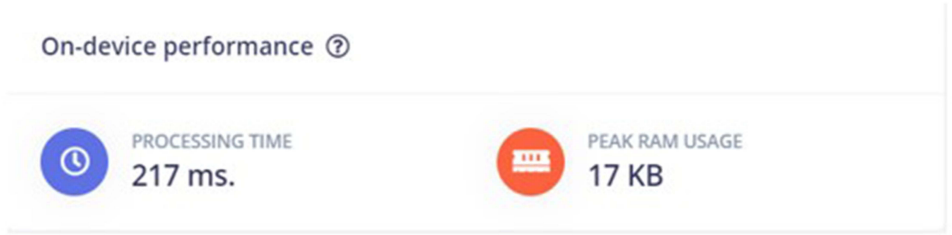

Figure 6 On Device Performance in Tiny ML.

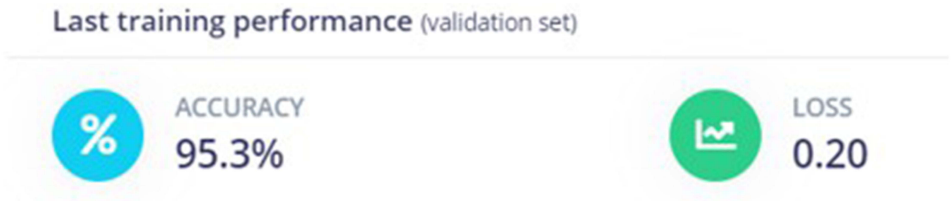

Figure 7 Training performance for the last run data.

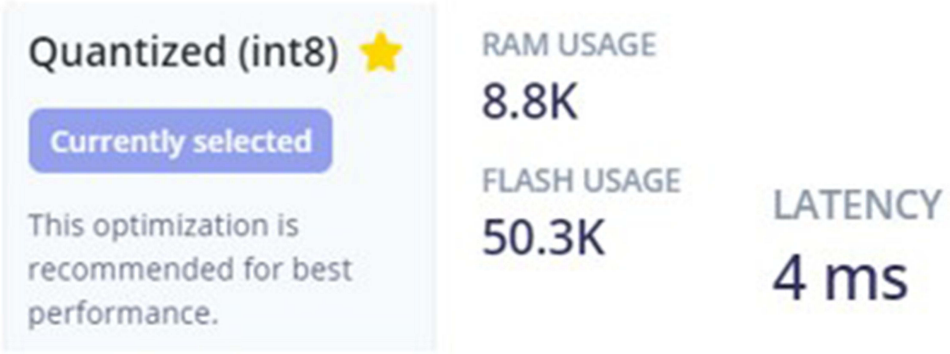

Figure 8 Performance in terms of Memory usage and Latency.

Discussion

The voice-based AI chatbot implemented in this study is on an edge device and hence provides the utmost security with respect to patient information. Although there are many online health assistance chatbots, they lack secrecy of patient data. This demands the implementation of cryptographic algorithms along with algorithms that process patient data,

Table 5 Time Taken to Record the Answers from the Patients

Questions	Time Taken
Name	9 seconds
Age	9 seconds
Gender	7 seconds
First yes or no question	7 seconds
Second yes or no question	7 seconds
Third yes or no question	7 seconds
Fourth yes or no question	7 seconds
Fifth yes or no question	7 seconds
Sixth yes or no question	7 seconds
Seventh yes or no question	7 seconds

Table 6 Comparison of This Research Work and Related Existing Research Work

Author Reference	Privacy Technique Used	Implementation	Performance
Kumar et al[25]	Secured Multiparty Computation	Communication between the patient and hospital.	Sluggish
Sarosh et al[26]	Rivest Cipher 6 (RC6) encryption algorithm	For secured storage of patient images. The key is shared according to Perfect Secret Sharing.	The pixel change rate is 99.55%
Paul et al[27]	Collective learning protocol for classified time series data exchange	Used for in-hospital mortality prediction in terms of precision, recall.	Good score of precision and recall.

(*Continued*)

Table 6 (Continued).

Author Reference	Privacy Technique Used	Implementation	Performance
Li et al[28]	Paillier encryption for data in server.	Encryption of data in hospital server.	Can withstand variety of threats towards data in server. It is a Low-cost implementation.
Dey et al[29]	Session key based on Perceptron and intermediate keys based on logistic mapping.	Lossless secret information sharing in Electronic Health Records (EHR).	Significant computation time for the cryptographic operations involved, leading to delay in health record processing.
Brisimi et al[30]	Primal-Dual splitting method for large scale SVM problems.	Decentralized cluster mechanism for healthcare record access.	Increased communication cost.
Rouhani et al[31]	GC protocol with CNN	With 58 folds throughput, health record learning process is secured.	Increased run time.
Choudhury et al[32]	Differential privacy-enabled federated learning.	Two layers of privacy protection for protecting raw data.	Considerable loss in performance in health care record keeping
Lee et al[33]	KNN with 3-fold validation.	Secured patient similarity learning	Accuracy around 80% and 90% for balanced and unbalanced data respectively.

leading to additional overheads and increased time for providing healthcare assistance. Khalid et al[24] surveyed and elaborated on privacy protection techniques such as federated learning and hybrid techniques in various AI healthcare applications. They discussed cryptographic and noncryptographic security techniques related to AI. Table 6 shows a comparison of this research work and existing research work related to electronic medical health care.

Based on the above comparison, it can be noted that, for electronic healthcare, many encryption techniques have been used for various scenarios. All of them have inherent disadvantages in terms of performance in record-keeping and recording processing time. Our research avoids all of these hassles in terms of local or edge computing with a Tiny ML. The total time to generate the case sheet of patients was found to be 3.8 sec for temperature measurement, 18s for recording the patient's personal details, such as name and sex, and 7s for each question relating to symptoms.

Thus, it can be validated that this research work is time efficient, is processing patient data, and is highly secure, as all the processing is performed on a local server without the use of cloud-based services. In future, the chatbot may be implemented in wearable healthcare devices.

Also recent technologies like Retrieval Augmented Generation AI methods may be used for more accuracy.

Conclusion

AI-powered innovations in healthcare settings focus on how patients interact and how care is delivered with the ultimate aim of enhancing the overall efficiency and effectiveness of patient outcomes. AI has recently played a major role in nursing, assistive management, medical diagnosis, and other critical medical procedures. Most of these AI implementations are online, and hence run the risk of patient data security. To overcome the security issues related to patient data and to achieve a high speed of operation, this research work has implemented a chatbot in a hospital scenario that runs on a local server with tinyML-based processing of patient data. Tiny ML is an edge computing technology that processes patient data on a local server, ensuring the high security of patient data. The implementation includes patient identification based on Histogram Of Gradient (HOG)-based classification. This was followed by basic patient care, such as temperature measurement and recording of name and gender. The autonomous temperature-sensing unit, integrated into the robot, is triggered when a patient is detected and uses a medical-grade infrared temperature scanner. After temperature measurement, the chatbot implemented with Tiny ML is activated and the patients are asked a series of

questions. The questions may be prescribed or defined by the doctor and used to train the Tiny model according to the diagnostic scenario. The answers given by the patients in "Yes" or "No" format to each question have been recorded, stored and printed out in the case sheet. Each patient is provided with a unique ID and the information is stored which can be used by hospital management to further guide the patient for consultation and diagnosis. A comparison with existing research in the field of AI-based healthcare shows that this research implementation recodes patient data very swiftly and is more secure because all patient data is stored and processed by the local host. Further enhancements in this research work may be in terms of including this chatbot in wearable devices.

Ethics and Consent Statements

This research work involved only a chatbot without any invasive procedure and hence was exempted from approval from Institutional Review Board. Written consent was obtained from the participants to participate in the chat at the health centre of the institution.

Funding

The authors report that this research was financially supported by the Vellore Institute of Technology, Chennai campus, Tamil Nadu, India.

Disclosure

The authors report no conflicts of interest in this work.

References

1. Lee D, Yoon SN. Application of artificial intelligence-based technologies in the healthcare industry: opportunities and challenges. *Int J Environ Res Public Health*. 2021;18(1):271. doi:10.3390/ijerph18010271
2. Lee DH. Strategies for technology-driven service encounters for patient experience satisfaction in hospitals. *Technol Forecast Soc Change*. 2018;137:118–127. doi:10.1016/j.techfore.2018.06.050
3. Yoon SN, Lee D. Artificial intelligence and robots in healthcare: what are the success factors for technology-based service encounters? *Int J Healthc Manage*. 2019;12(3):218–225. doi:10.1080/20479700.2018.1498220
4. Lee S, Lim S. *Living Innovation: From Value Creation to the Greater Good*. Bingley, UK: Emerald Publishing Limited; 2018.
5. World Health Organization. The WHO cross-national study of health behaviour in school aged children from 35 countries: findings from 2001–2002. *J Sch Health*. 2009;74:204–206. doi:10.1111/j.1746-1561.2004.tb07933.x
6. Abe M, Abe H. Lifestyle medicine—an evidence based approach to nutrition, sleep, physical activity, and stress management on health and chronic illness. *Pers Med Universe*. 2019;8:3–9. doi:10.1016/j.pmu.2019.05.002
7. Siddique S, Chow JC. Machine learning in healthcare communication. *Encyclopedia*. 2021;1:220–239. doi:10.3390/encyclopedia1010021
8. Taylor N. Duke report identifies barriers to adoption of AI healthcare systems. MedTech Dive. 2019. Available from: https://www.medtechdive.com/news/duke-report-identifies-barriers-to-adoption-of-ai-healthcare-systems/546739/. Accessed November 1, 2024.
9. Uzialko A. Artificial Intelligence will change healthcare as we know it. Business News Daily. 2019. Available from: https://www.businessnewsdaily.com/15096-artificial-intelligence-in-healthcare.html. Accessed November 1, 2024.
10. Arsene C. Artificial Intelligence in healthcare: the future is amazing. Healthcare Weekly. 2019. Available from: https://healthcareweekly.com/artificial-intelligence-in-healthcare/. Accessed November 1, 2024.
11. MDDI Staff. Can AI really be a game changer in cervical cancer screenings? Medical Device and Diagnostic Industry (MDDI). 2019. Available from: https://www.mddionline.com/can-ai-really-be-game-changer-cervical-cancer-screenings. Accessed November 1, 2024.
12. ChosunBiz. The world's first medical AI Watson 'Chanbap'. Both South Korea and the U.S. "lack of training". 2018. Available from: http://biz.chosun.com/site/data/html_dir/2018/11/23/2018112302467.html. Accessed November 1, 2024.
13. Somashekhar S, Kumar R, Kumar A, Patil P, Rauthan A. Validation study to assess performance of IBM cognitive computing system Watson for oncology with Manipal multidisciplinary tumour board for 1000 consecutive cases: an Indian experience. *Ann Oncol*. 2016;27:1–2. doi:10.1093/annonc/mdw601.002
14. Chow JCL, Wong V, Sanders L, Li K. Developing an AI-assisted educational chatbot for radiotherapy using the IBM Watson assistant platform. *Healthcare*. 2023;11:2417. doi:10.3390/healthcare11172417
15. Palanica A, Flaschner P, Thommandram A, Li M, Fossat Y. Physicians' perceptions of chatbots in healthcare: cross-sectional web-based survey. *J Med Internet Res*. 2019;21:e12887. doi:10.2196/12887
16. Patel SB, Lam K. ChatGPT: the future of discharge summaries? *Lancet Digit Health*. 2023;5(3):e107–e108. doi:10.1016/S2589-7500(23)00021-3
17. Chow JCL, Sanders L, Li K. Impact of ChatGPT on medical chatbots as a disruptive technology. *Front Artif Intell*. 2023;6:1166014. PMID: 37091303; PMCID: PMC10113434. doi:10.3389/frai.2023.1166014
18. Cuffaro L, Di Lorenzo F, Bonavita S, Tedeschi G, Leocani L, Lavorgna L. Dementia care and COVID-19 pandemic: a necessary digital revolution. *Neurol Sci*. 2020;41(8):1977–1979. doi:10.1007/s10072-020-04512-4
19. Ong Y, Tang A, Tam W. Effectiveness of robot therapy in the management of behavioural and psychological symptoms for individuals with dementia: a systematic review and meta-analysis. *J Psychiatr Res*. 2021;140:381–394. doi:10.1016/j.jpsychires.2021.05.077

20. Kim HN. A conceptual framework for interdisciplinary education in engineering and nursing health informatics. *Nurse Educ Today*. 2019;74:91–93. doi:10.1016/j.nedt.2018.12.010
21. Chow JCL, Wong V, Li K. Generative pre-trained transformer-empowered healthcare conversations: current trends, challenges, and future directions in large language model-enabled medical chatbots. *BioMedInformatics*. 2024;4:837–852. doi:10.3390/biomedinformatics4010047
22. Dalal N, Triggs B. Histograms of oriented gradients for human detection. In: 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05). San Diego, CA, USA; 2005:886–893. doi:10.1109/CVPR.2005.177
23. speech_commands • Datasets at Hugging Face. Available from: https://huggingface.co/datasets/google/speech_commands. Accessed 08 November 2024.
24. Khalid N, Qayyum A, Bilal M, Al-Fuqaha A, Qadir J. Privacy-preserving artificial intelligence in healthcare: techniques and applications. *Comput Biol Med*. 2023;158:106848. doi:10.1016/j.compbiomed.2023.106848
25. Kumar AV, Sujith MS, Sai KT, Rajesh G, Yashwanth DJS. Secure multiparty computation enabled E-healthcare system with homomorphic encryption. *IOP Conf Ser*. 2020;981(2):022079. doi:10.1088/1757-899X/981/2/022079
26. Sarosh P, Parah SA, Bhat GM, Heidari AA, Muhammad K. Secret sharing based personal health records management for the internet of health things. *Sustainable Cities Soc*. 2021;74:103129. doi:10.1016/j.scs.2021.103129
27. Paul J, Annamalai MSMS, Ming W, Al Badawi A, Veeravalli B, Aung KMM. Privacy-preserving collective learning with homomorphic encryption. *IEEE Access*. 2021;9:132084–132096. doi:10.1109/ACCESS.2021.3114581
28. Li D, Liao X, Xiang T, Wu J, Le J. Privacy-preserving self-serviced medical diagnosis scheme based on secure multi-party computation. *Comput Secur*. 2020;90:101701. doi:10.1016/j.cose.2019.101701
29. Dey J, Bhowmik A, Karforma S. Neural perceptron & strict lossless secret sharing oriented cryptographic science: fostering patients' security in the "new normal" COVID-19 E-health. *Multimedia Tools Appl*. 2022;1–32. doi:10.1007/s11042-022-12440-y
30. Brisimi TS, Chen R, Mela T, Olshevsky A, Paschalidis IC, Shi W. Federated learning of predictive models from federated electronic health records. *Int J Med Inf*. 2018;112:59–67. doi:10.1016/j.ijmedinf.2018.01.007
31. Rouhani BD, Riazi MS, Koushanfar F. Deepsecure: scalable provably secure deep learning. In: *Proceedings of the 55th Annual Design Automation Conference*. 2018:1–6. doi:10.1109/DAC.2018.8465894.
32. Choudhury O, Gkoulalas-Divanis A, Salonidis T, et al. Differential privacy-enabled federated learning for sensitive health data. *arXiv preprint arXiv:1910 02578*. 2019. doi:10.48550/arXiv.1910.02578
33. Lee J, Sun J, Wang F, Wang S, Jun C-H, Jiang X. Privacy-preserving patient similarity learning in a federated environment: development and analysis. *JMIR Med Inform*. 2018;6(2):e7744. doi:10.2196/medinform.7744