Journal of Industrial Information Technology and Application *ISSN(Online): 2586-0852*
Vol. No. ()

Medichat - A Medical Chatbot with Multilingual Support

Sang Suh[1,*)] and Rama Krishna Kamma[2)]
[1,2)] Computer Science Department
Texas A&M University-Commerce, Texas, U.S.A.

Abstract. In the rapidly evolving realm of AI-driven medical consultations, linguistic inclusivity is paramount. The exchanges between patients and doctors obtained from a popular internet resource for medical consultations. In AI-driven medical consultations, the lack of multilingual support in online platforms for patient-doctor interactions poses a significant challenge to achieving linguistic inclusivity. This project introduces Medichat, a revolutionary platform designed to enhance the accessibility, comprehensibility, and reliability of medical advice. By fine-tuning language models, integrating multilingual support, and implementing a self-directed information retrieval system, Medichat empowers users with a universal and secure tool for healthcare guidance. The model's responsiveness is significantly improved through real-time data retrieval from trusted sources, while privacy concerns are addressed with anonymized patient- doctor interactions. Moreover, the project achieves multilingual accessibility by integrating a high-precision translation model and successfully leverages the strengths of various AI components. Medichat represents a significant leap forward in medical AI, offering a comprehensive and inclusive solution for modern healthcare challenges.

Keywords; Chatbot; Artificial Intelligence; Medical.

1. Introduction

A medical chatbot represents a technological marvel, a software creation endowed with artificial intelligence [1] [2] and the prowess of natural language processing [4]. Its mission? To act as a virtual medical guide, extending a helping hand to users navigating the intricate complex of health-related queries. These digital companions excel in deciphering symptoms, suggesting potential maladies, offering insights into treatment options and lifestyle adjustments, and even give pearls of wisdom regarding overall well-

* Corresponding author: Sang.Suh@tamuc.edu
Received: Feb. 09, 2024; Accepted: May. 15, 2024; Published: Jun. 30, 2024

being. They're the genie in the healthcare lamp, capable of aiding users in locating nearby healthcare havens or securing coveted appointments [16]. Operating seamlessly across various digital domains like messaging apps, social media realms, and web-based platforms, medical chatbots possess an inherent user friendliness and interactivity [6], that closely mirrors human conversation. They form part of a broader trend aimed at elevating the realm of healthcare services and ameliorating the patient's journey.

As illuminated by Peter Luba [23], these ingenious bots undertake a multifaceted mission encompassing appointment scheduling, healthcare facility reconnaissance, and even encouragement to participate in clinical trials [5]. Increasingly, medical chatbots are taking on the roles of initial symptom scrutineers and expert triage consultants [6]. They engage patients in dialogue, probing for relevant symptom related details and offering preliminary insights into potential diagnoses. The next logical step? Guiding the user towards arranging appointments with healthcare professionals, ensuring that the journey from ailment to remedy is a seamless one. A shining example of this digital healthcare frontier is the chatbot Ada Health [24], which deftly guides users towards a diagnosis based on their symptoms and prescribes the next steps to take. But the prowess of medical chatbots extends beyond these initial interactions. They are quick learners, absorbing user data through the magic of machine learning algorithms, allowing them to tailor responses to individual preferences [8]. By harnessing the incredible capabilities of artificial intelligence and natural language processing, these chatbots constantly refine their ability to comprehend and generate human-like language, ultimately delivering more effective responses to user queries.

In the realm of chatbot implementation, several pivotal phases of Language Processing (NLP) [11] [13] reign supreme. Tokenization takes the lead, converting a stream of characters into a harmonious symphony of tokens. These tokens could be identifiers, words, numerals, or punctuation marks, each getting the lowercase treatment. Then comes the art of stemming, where word variations are harmonized into simpler, root words, guided by the wisdom of language dictionaries. Finally, the tokens find their place through sorting, meticulously arranged based on shared properties or sorted according to a specific criterion. The medical field, characterized by its intricate terminologies, necessitates clear and universally accessible communication. While there have been significant strides in AI-driven medical consultations, a substantial segment of the global population remains potentially underserved due to language barriers. This project seeks to bridge this gap by incorporating multi-lingual support, ensuring that accurate medical advice is available to all, irrespective of their linguistic background.

2. Related Work

A reasonably AI chatbot that can respond to frequently requested medical inquiries concerning cardiac problems is developed in [22]. Many query patterns and the associated results are stored in a JSON file to form a dataset. The dataset is then preprocessed using NLP techniques to provide responses in natural language that are human-like. On the cleaned-up data, an artificial neural network model is trained. The ANN achieves a 99% accuracy rating with an epoch of 1000. An AI chatbot that can respond to frequently requested medical queries concerning cardiac problems has been developed [22]. Many query patterns and the associated results are stored in a JSON file to form a dataset. The dataset is then preprocessed using NLP techniques to provide responses in natural language that are human-like. On the cleaned-up data, an artificial neural network model is trained. The ANN achieves a 99% accuracy rating with an epoch of 1000.

In a study by P. Srivastava and N. Singh [14], they build a diagnosis bot that converses with patients about their medical problems and asking them questions [15] in order to provide a customized diagnosis based on their profile and identified symptom. With a typical precision of 65%, the diagnosis chatbot can recognize symptoms from user inputs. Correct symptoms were recognized with a recall of 65% and a precision of 71% using these extracted diagnosed symptoms. The chatbot processes a user's message and extracts pertinent patterns that can be used to diagnose probable ailments using AIML (Artificial Intelligence Mark-up Language), an XML-based language (Extensible Mark-up Language), in order to achieve these results.

In a work by R. B. Mathew et al. [9], uses Python's Natural Language Toolkit (NLTK) to build a chatbot system for disease prediction by doing natural language processing (NLP) on messages. A chatbot's ability to comprehend human speech and respond appropriately is made possible by natural language processing (NLP). In a textbox, a user enters their symptom. After processing the content, the chatbot displays a list of symptoms that closely match the user's input. The user verifies which symptom from the list of symptoms the chatbot has returned appropriately corresponds to their symptom. The chatbot outputs the diagnosis in a textbox page that looks similar after anticipating the illness. In a study by P. I. Prayitno [20], a health chatbot employing NLP for disease prediction is also developed. The user's input is contrasted with the symptoms that are present in a database using cosine, though. An ID3 decision tree method is then utilized to determine the user's ailment using the symptom that is the most comparable. The dataset used in this study was generated from the official Alodoketer website and provides extensive information about the disease, including its symptoms and therapies. With an accuracy of 87.5%, the chatbot can diagnose a user's illness based on their

symptoms. P. I. Prayitno wants to improve the model's precision and incorporate more creatures, like the ability to purchase medications through a chatbot, in future development.

The use of machine learning algorithms for disease prediction [9] and categorization has been extensively researched utilizing a variety of datasets and methodologies. The Pima Indians Diabetes dataset, the UCI Heart illness dataset, and the UCI Thyroid Disease dataset are among the datasets that may be found in many illness prediction studies. We will group the categorization methods applied to each dataset in this section. A common machine-learning technique called Random Forest is used to identify different diseases. The Random Forest method is used to predict breast cancer early in a work by K. Mridha [18] using the Wisconsin Breast Cancer dataset. According to K. Mridha, the Random Forest model has the highest accuracy (98.83%), while the K-nearest Neighbors model has the lowest accuracy (91.22%). In other illness datasets, the Random Forest method also produces good accuracy ratings.

Another machine-learning [21] technique that can be used to identify various diseases is support vector machines. With the UCI Thyroid Disease dataset, it has been demonstrated to produce highly accurate rates. The SVM model produced an accuracy score of 99.63% in A. Tyagi's paper [17]. Using the PIMA Indians Diabetes dataset, D. S. Sisodia and R. Agrawal suggested a method in paper [10] that consists of four basic steps: handling missing values, instance reduction by k-means clustering, dimension reduction, and classification as well as the Gradient Boost method [3]. The model's accuracy score was 97.10%, which was higher than that of earlier methods like Naive Bayes, Logistic Regression, and SVM.

3. Methodology

A. LLaMA Model

Large language model meta-AI (LLaMA) has been built by Meta as a foundational large language model to aid in AI research. LLaMA is being made available in various sizes, ranging from 7B to 65B parameters. It's designed to be versatile and can be fine-tuned for various tasks. The model is based on text data from 20 languages with a focus on those with Latin and Cyrillic alphabets. Large language models like LLaMA have shown remarkable capabilities in text generation, mathematical problem-solving, and more. However, full research access has been limited due to resource requirements. Access to LLaMA is provided under a noncommercial research license, and access will be granted on a case-by- case basis to academic researchers, government and civil society organizations, and industry research laboratories. Meta emphasizes the need for

the AI community to collaborate on responsible AI guidelines [19], particularly for large language models, and looks forward to the contributions and applications that the community can develop using LLaMA. Meta acknowledges that there is ongoing research needed to address issues like bias, toxicity, and hallucinations in large language models. By releasing LLaMA, the company aims to encourage researchers to find ways to mitigate these problems.

The Fine-tuning the LLaMA Model contains the three phases Data Augmentation, Transfer Learning and Multi-Lingual Generation Layer. During Data Augmentation Phase (From Fig 1) used to artificially increase the size of the dataset by involving the following operations like synonym replacement, random word insertion, word rearrangement, and paraphrasing. These techniques are used to generate additional text samples for tasks like text classification and language modeling. By doing this, the goal is to make sure that the dataset contains a wide variety of linguistic examples and representations, even if the original data is limited. In other words, data augmentation is a way to create more diverse and comprehensive training data for a given task or model.

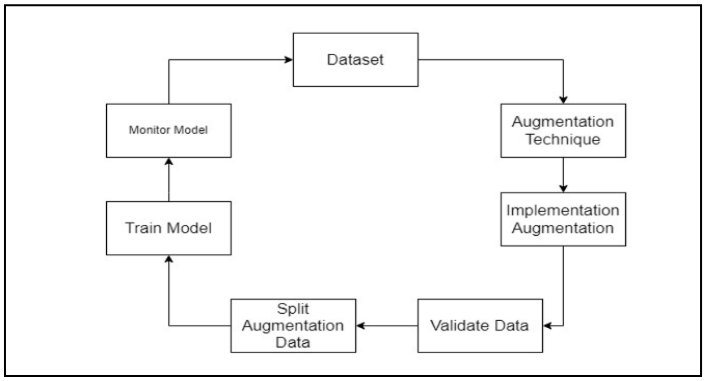

Figure 1. Data Flow of Augmentation

In the Transfer Learning Phase is already knowledgeable about medical language, will be further trained using additional data that has been artificially created. This additional training aims to make the model better at understanding and using medical terminology in various languages, including the subtle details and differences. It's a way to enhance the model's performance in multilingual medical contexts.

During the final phase i.e., Multi-Lingual Generation Layer an integrated module will first identify the language of the incoming query, ensuring the response is generated in the same language and generate responses based on the detected language, ensuring linguistic accuracy and relevance.

a) Input Question (Symptoms): From Fig 2, in this input a user provides a description of their symptoms or other health related worries.

b) Language Detection: The system determines the language a user has supplied while describing their symptoms in this stage. This is essential to guaranteeing that the user's question is correctly understood.

c) Machine Translation: The system may translate the input into a common language for processing once it has determined the user's language. This stage guarantees that the system can comprehend and react to questions from people speaking different languages.

d) Tokenization: It is the process of breaking up the translated text into smaller chunks, usually words or phrases. Tokenization is necessary in order to divide the text into digestible parts for examination.

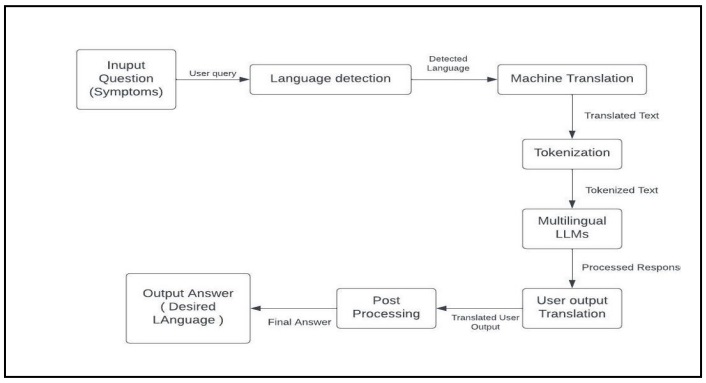

Figure 2. Data Flow of LLaMA Model

e) Multilingual Language Models: In this instance, the machine learning model to comprehend the symptoms of the user and responding properly is made feasible by multilingual language models, which can interpret and produce text in various languages that is similar to that of a person.

f) User Output Translation: The system may translate a response that was generated in a common language back into the language that was identified by the user. This stage guarantees that the user can comprehend the response from the system.

g) Postprocessing: To make the system's answer more logical and legible, postprocessing entails organizing and cleaning it up. It could entail editing the text's grammar, eliminating superfluous details, or rewording it.

h) Output Answer: Lastly, the user gets the system's response in the language of their choice in the output answer. In response, you should offer details, guidance, or suggestions on their symptoms or medical issues.

B. Proposed Method - Approaches for Multi-Lingual Implementation:

The Multilingual implementation contains the following approaches i.e. Multilingual BERT (mBERT), Cross-lingual Language Model (XLM), Translation-Based Approach with Transformer Model, Zero-shot Learning LASER (Language- Agnostic Sentence Representation), Cross-Lingual Transfer Learning with NLLB-200. The purpose of medical consultations where accuracy and reliability are paramount, the Cross- Lingual Transfer Learning with NLLB-200 is recommended as the best approach. NLLB-200 is trained on a diverse set of languages, making it robust for various linguistic scenarios [6]. Its capacity to be optimized for high-resource languages and then implemented for low-resource languages guarantees that appropriate medical consultations can be provided even for languages that are spoken less often. Furthermore, its architecture, which is an extension of the powerful model, ensures state-of-the-art performance across multiple NLP tasks [4].

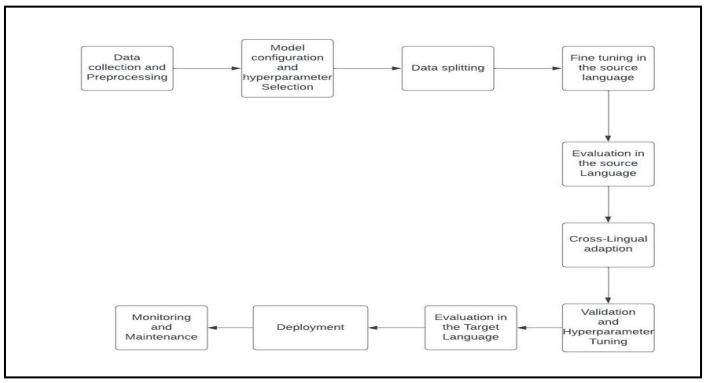

Figure 3. Data Flow of LLaMA Model

By integrating multi-lingual support through the fine-tuned LLaMA model, the dedicated generation layer, and the application of the NLLB-200 model, the system is set to become a universally accessible tool for medical consultations [22], emphasizing the importance of linguistic inclusivity in AI- driven healthcare.

C. Cross-Lingual Transfer Learning with NLLB-200:

a) Data Preparation and Collection: Assemble data in both the source and target languages (From Fig 3), whether labeled or unlabeled. Process the data, ensuring it adheres to the tokenization scheme of NLLB-200 [11] through cleaning, formatting, and tokenization.

b) Preliminary Model Setup and Parameter Tuning: Commence with the versatile NLLB-200 model, which comes pretrained on multilingual data. Establish essential hyperparameters, including learning rate, batch size, and training epochs, tailoring them to the specific task.

c) Data Segregation: Partition your labeled data into training, validation, and test sets, a fundamental step for training and evaluating the model.

d) Source Language Performance Evaluation: Appraise the model's performance on the source language test set [13]. Employ task-specific metrics, such as accuracy, F1 score, or others, to gauge performance.

e) Cross-Lingual Adaptation: Fine-tune the model using data from target languages while retaining the same task- specific configuration [7].

f) Evaluation in the Target Language: Gauge the model's performance on test sets in the target languages to assess its adaptability [14].

g) Model Integration and Deployment: Integrate the fine-tuned NLLB-200 model into your application or system for real-world data inference.

h) Continuous Monitoring and Maintenance: Maintain a vigilant watch on the model's performance in a production environment [12]. Be prepared to retrain or fine-tune the model as new data surfaces or if performance starts to deteriorate.

4. Implementation

This paper describes the procedures and results of our most recent machine learning studies, with an emphasis on language models and their uses in translation and healthcare. The experiments included investigating zero-shot learning strategies, tackling computational resource constraints, and fine-tuning already-existing models. In addition, we have developed APIs for project implementation and integrating a multi-language translation approach.

In our endeavor to enhance the performance of the HealthCareMagic-100k GPT-2 model, we embarked on the journey of fine-tuning. However, despite rigorous efforts, the finalized scores for ROUGE-1, ROUGE-2, and ROUGE-L of 0.90 fell short of our desired specifications, signifying the need for further refinement and alternative approaches. Meanwhile, fine-tuning the LLAMA-2 model posed challenges, primarily due to memory constraints on the platforms we attempted. Initially, on a 32GB M1 chip CPU, we faced a memory outage issue, which prompted us to explore Google Collab, equipped with ample resources. Unfortunately, even with 50GB RAM and a 15GB shared GPU in V100, memory outage issues persisted, ultimately preventing the fine-tuning of the LLAMA-2 model. As a result, we pivoted to a zero-shot learning approach, successfully writing the necessary code and generating outputs. The subsequent step involves evaluating this approach to gauge its performance and accuracy.

Additionally, inspired by Facebook's paper," No Language Left Behind: Scaling Human Centered Machine Translation" [25] on scaling human centered machine

translation, we embarked on the implementation of a multilingual translation model. The objective was to create a model capable of accurately translating across 200 languages, a feat that has shown promise with good BLEU scores, indicating its effectiveness in facilitating multi-language communication (As Observe in Table I). To lay the groundwork for the project's integration, foundational Skeleton APIs were developed. These APIs serve as a critical building block for further development and seamless integration with the existing system.

Table I. QUANTITATIVE COMPARISON WITH BERT SCORE BETWEEN MEDICHAT CHATDOCTOR AND CHATGPT

Models	Precision	Recall	F1 Score
Medichat	0.845±0.018	0.850±0.016	0.847±0.015
ChatDoctor	0.837±0.0188	0.8445±0.0164	0.8406±0.0143
ChatGPT	0.8444±0.018	0.8451±0.0157	0.8446±0.0138

5. Results

Fig 4 displays the outputs generated by ChatGPT and ChatDoctor in response to a given input. Notably, ChatGPT provided a comprehensive but broad response regarding otitis, while ChatDoctor offered a more targeted answer, emphasizing the application of external knowledge from the brain to treat otitis. However, from Fig 5 and Fig 6 MediChat surpassed both ChatGPT and ChatDoctor in terms of precision, delivering a step-by-step and lucid explanation for the queried topic. This underscores the superior performance of MediChat, showcasing its ability to provide a more detailed and accurate response compared to its counterparts.

MediChat exhibits a remarkable multilingual capability by seamlessly processing input from 200 different languages and delivering output in the same language. This versatility allows MediChat to provide responses in various linguistic contexts, ensuring effective communication across a wide spectrum of language preferences. As demonstrated by a sample output in French (From Fig 7), Spanish (Fig 8) and Portuguese (Fig 9) MediChat's proficiency extends beyond a singular language, emphasizing its ability to cater to diverse linguistic needs with ease and accuracy.

Figure 4. Comparison of ChatGPT vs Ours (Medichat)

Figure 5. Sample output of Medichat in English

Figure 6. Sample output of Medichat in French

Figure 7. Sample output of Medichat in Spanish

Figure 8. Sample output of Medichat in Portuguese

6. Conclusion and Future Works

In parallel, worked diligently on the front-end development, with a keen focus on optimizing the user experience. Ensuring that the interface is intuitive and accessible is our top priority, as we aim to provide users with seamless interaction with the models and their outputs. However, some issues have arisen during this process. Notably, the front-end interface has been deemed less user- friendly, impacting the overall user experience. In response, we have initiated improvements in the front-end design to enhance usability, striving to create an intuitive interface that facilitates easy interaction is shown in Fig 6.

Furthermore, the time taken to generate text on CPU has proven to be excessively long, approximately 30 minutes. To address this issue, we are actively exploring strategies to decrease text generation time. This includes considerations such as optimizing algorithms, parallelizing processes, and potentially acquiring a GPU server for faster processing. Our goal is to identify and implement the most efficient and cost-

effective solution to significantly improve overall performance, ensuring that users receive timely and reliable responses from the system.

References

[1] D. Shnavi, et al., ”A Self-Diagnosis Medical Chatbot Using Artificial Intelligence”, Journal of Web Development and Web Designing, vol.3, pp. 1, MAT Journals, 2018.

[2] S. Ghare, et al., ”Self-Diagnosis Medical Chat-Bot Using Artificial Intelligence”, pp. 1, February 2020.

[3] K. L. Kumar and B. E. Reddy, ”Heart Disease Detection System Using Gradient Boosting Technique,” in 2021 International Conference on Computing Sciences (ICCS), Phagwara, 2021, n. pag.

[4] P. I. Prayitno et al., ”Health Chatbot Using Natural Language Processing for Disease Prediction and Treatment,” in 2021 1st International Conference on Computer Science and Artificial Intelligence (ICCSAI).

[5] D. Madhu, et al., “A novel approach for medical assistance using trained chatbot,” in International Conference on Inventive Communication and Computational Technologies, pp. 1, March 2017.

[6] N. Haristiani, “Artificial Intelligence (AI) Chatbot as Language Learning Medium: An inquiry,” in International Conference on Education, Science and Technology, pp. 1-5, March 2019.

[7] T. Nadarzynski, et al., “Acceptability of artificial intelligence (AI)-led chatbot services in healthcare: A mixed-methods study,” Digital Health, vol. 5, pp. 1, January 2019.

[8] M. Adam, M. Wessel, and A. Benlian, “AI-based chatbots in customer service and their effects on user compliance,” in The International Journal on Networked Business, pp. 1, February 2020.

[9] R. B. Mathew, et al., “Chatbot for Disease Prediction and Treatment Recommendation using Machine Learning,” in International Conference on Trends in Electronics and Informatics, pp. 853, October 2019.

[10] D. S. Sisodia and R. Agrawal, ”Data Imputation-Based Learning Models for Prediction of Diabetes,” in 2020 International Conference on Deci- sion Aid Sciences and Application (DASA), Sakheer, Bahrain, 2020.

[11] Li, Y., Li, Z., Zhang, K., Dan, R., Jiang, S., & Zhang, Y. ”ChatDoctor: A Medical Chat Model Fine-Tuned on a Large Language Model Meta-AI (LLaMA) Using Medical Domain Knowledge,” Cureus, vol. 15, no. 6, p. e40895, Jun. 24, 2023.

[12] B. Su, et al., “Health Care Spoken Dialogue System for Diagnostic Reasoning and Medical Product Recommendation,” in International Conference on Orange Technologies, pp. 4, October 2018.

[13] M. Virkar, V. Honmane, and S. U. Rao, “Humanizing the Chatbot with Semantics based Natural Language Generation,” in International Conference on Intelligent Computing and Control Systems, pp. 893, May 2019.

[14] P. Srivastava and N. Singh, ”Automatized Medical Chatbot (Medibot),” in 2020 International Conference on Power Electronics & IoT Applications in Renewable Energy and its Control (PARC), Mathura, India, 2020.

[15] S. Mujeeb, M. H. Javed, and T. Arshad, “Aquabot: A Diagnostic Chatbot for Achluophobia and Autism,” in International Journal of Advanced Computer Science and Applications, pp. 209-216, January 2017.

[16] H. Gertz, C. C. Pollack, M. D. Schultheiss and J. S. Brownstein, ”Delayed medical care and underlying health in the United States during the COVID-19 pandemic: A cross-sectional study,” Preventive medicine reports, vol. 28, 2022.

[17] Tyagi, R. Mehra and A. Saxena, ”Interactive Thyroid Disease Prediction System Using Machine Learning Technique,” in 2018 Fifth International Conference on Parallel, Distributed and Grid Computing (PDGC), Solan, India, 2018.

[18] K. Mridha, ”Early Prediction of Breast Cancer by using Artificial Neural Network and Machine Learning Techniques,” 2021 10th IEEE International Conference on Communication Systems and Network Technologies (CSNT), Bhopal, India, 2021, pp. 582-587.

[19] Aggarwal, C. C. Tam, D. Wu, . X. Li and S. Qiao, ”Artificial Intelligence-Based Chatbots for Promoting Health Behavioral Changes: Systematic Review,” Journal of Medical Internet Research, vol. 25, p. e40789, 2023.

[20] P. I. Prayitno, R. P. Pujo Leksono, F. Chai, R. Aldy and W. Budiharto, ”Health Chatbot Using Natural Language Processing for Disease Predic- tion and Treatment,” in 2021 1st International Conference on Computer Science and Artificial Intelligence (ICCSAI), Jakarta, Indonesia, 2021.

[21] Gupta and M. K. Gupta, ”Prediction of Diseases Using Different Machine Learning Approaches,” in 2022 3rd International Conference on Intelligent Engineering and Management (ICIEM), London, United Kingdom, 2022.

[22] J. N. Jothi, S. Poongodi, V. Chinnammal, L. Kannagi, M. Panneerselvam and R. T. Prabu, ”AI Based Humanoid Chatbot for Medical Application,” in 2022 3rd International Conference on Smart Electronics and Communication (ICOSEC), Trichy, India, 2022.

[23] P. Luba,” Healthcare Chatbots Can Help With the Pandemic,” Towards Data Science, May 14, 2020. [Online]. Available: https://towardsdatascience.com/healthcare-chatbots-can-help-with-the-pandemic-bcc07fc606c9. [Accessed: Oct. 10, 2023].

[24] “About Ada: Personal Health Companion,” [Online]. Available: https://ada.com/about/. [Accessed: Oct. 10, 2023].

[25] NLLB Team, M. R. Costa-jussa, J. Cross, O. Celebi, M. El- bayad, K. Heafield, ... J. Wang, ”No Language Left Behind: Scal- ing Human-Centered Machine Translation,” 2022. [Online]. Available: https://arxiv.org/abs/2207.04672.