Volume 9, Issue 6, June – 2024 International Journal of Innovative Science and Research Technology
ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24JUN161

A Novel Approach to Template Filling with Automatic Speech Recognition for Healthcare Professionals

Sahil Kadge[1]
Electronics & Telecommunication Engineering
K.J. Somaiya Institute of Technology
Sion, Mumbai, India

Kamran Khot[2]
Electronics & Telecommunication Engineering
K.J. Somaiya Institute of Technology
Sion, Mumbai, India

Yash Navander[3]
Electronics & Telecommunication Engineering
K.J. Somaiya Institute of Technology
Sion, Mumbai, India

Dr. Jayashree Khanapuri[4]
Electronics & Telecommunication Engineering
K.J. Somaiya Institute of Technology
Sion, Mumbai, India

Abstract:- In the evolving landscape of medical documentation, the necessity for efficient and accurate record-keeping systems is paramount, especially in specialised fields such as neurology where precision in terminology is crucial. This paper introduces a pioneering application of a fine-tuned Whisper model, specifically adapted for brain-related medical terms, integrated with an AI-driven system for automated template filling. The proposed system leverages advanced speech recognition technologies to capture doctors' verbal inputs and accurately transcribe these into designated report templates. The process simplifies the documentation workflow, significantly reducing the cognitive and administrative load on healthcare providers by enabling them to focus more on patient care rather than paperwork. Our research details the development and implementation of this innovative system, including the specific adaptations made to the Whisper model to enhance its accuracy with neurology-specific terminology. We also evaluate the system's performance in real-world medical settings and discuss the practical implications of integrating such AI tools in clinical practice. Furthermore, the system's capacity to generate ready-to-print PDF reports not only streamlines the documentation process but also ensures consistency and reliability in medical records. The overarching aim of this project is to demonstrate how targeted AI solutions can address the unique challenges of medical documentation, offering substantial benefits to healthcare providers and patients alike.

I. INTRODUCTION

In the intricate landscape of modern medicine, accurate and comprehensive documentation stands as a cornerstone of effective patient care. Neurologists, in particular, grapple with the daunting task of meticulously filling out templates to capture intricate diagnostic nuances, treatment plans, and patient interactions. This labour-intensive process not only consumes valuable time but also poses significant challenges in maintaining consistency, accuracy, and compliance with medical standards. The imperative of this project is underscored by the pressing need to revolutionise the archaic methods of medical documentation, empowering healthcare providers to allocate more time and focus on patient-centric care. By harnessing the transformative potential of Artificial Intelligence (AI) and Automatic Speech Recognition (ASR), we embark on a pioneering endeavour to streamline template filling in the medical field, with a specific emphasis on medical terms.

Central to our approach is the utilisation of cutting-edge technologies such as the Whisper model developed by OpenAI. The Whisper model, a product of extensive research and fine-tuning, epitomises the epitome of state-of-the-art ASR systems. Trained on a vast corpora of diverse linguistic data, including medical terminologies, Whisper promises unparalleled accuracy and adaptability, even amidst the nuanced intricacies of neurological discourse. However, recognizing the unique lexicon and phraseologies inherent to medical terms, we embarked on a journey of fine-tuning the Whisper model. Through meticulous training on a curated dataset comprising 50 sentences commonly utilised by neurologists, particularly in radiology contexts, we honed the model's capabilities to discern and transcribe specialised medical terminologies with unparalleled precision.

Crucial to our data management strategy is the utilisation of Hugging Face, a leading platform in the realm of natural language processing (NLP) and AI model development. Our curated dataset, complete with audio files in a specialised format, has been meticulously crafted and seamlessly integrated into Hugging Face's ecosystem. Leveraging the platform's robust infrastructure and collaborative framework, we have pushed our dataset along with its associated metadata, including a dataset pointer, onto Hugging Face's repository. The dataset pointer serves as a beacon of accessibility and reproducibility, facilitating seamless access to our curated dataset for fellow researchers and developers alike. By adhering to open science principles and leveraging Hugging Face's standardised protocols, we ensure transparency, collaboration, and the continued

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24JUN161

advancement of AI research in the medical domain.

Moreover, for the prnactical deployment of our AI-driven template filling system, we rely on Hugging Face's Inference API. This powerful tool enables the seamless integration of our fine-tuned Whisper model into a user-friendly web interface, allowing for real-time, on-the-fly speech recognition and template filling directly through a browser. This integration marks a pivotal step toward democratising AI-driven medical documentation, making it accessible to healthcare providers across diverse clinical settings. In essence, our project signifies a paradigm shift in medical documentation, fueled by the convergence of cutting-edge technologies and a steadfast commitment to enhancing patient care. Through the amalgamation of OpenAI's Whisper model, Hugging Face's collaborative platform, and the collective efforts of the research community, we endeavour to usher in a new era of efficiency, accuracy, and empathy in the medical field.

II. PROPOSED WORK

This proposed system for speech-to-text template filling seeks to change how healthcare professionals interact with systems by introducing a frictionless way to enter data. A highly advanced model for speech recognition forms the basis of this system that is uniquely engineered to adapt to medical documentation's specific vocabulary and syntax. This involves doctors taking handwritten notes in the traditional manner which are then transcribed digitally. Subsequently, extensive data collection follows whereby diverse medical templates together with associated utterances are gathered from healthcare providers across different specialties and practice settings. Thus, given this data, one can lay the foundation for building a powerful dataset that covers a wide array of medical terms, dictation styles and template formats in detail. With the dataset available, the creation of a tailor-made speech recognition model using cutting-edge techniques follows.

The Whisper model is selected as a base architecture that has proven effective when transcribing spoken language and it is fine-tuned to meet specific needs of medical speech. [1] To make this happen, some changes must be made to both hyperparameters of models, training algorithms optimization and augmentation implementation of data methods to improve on the ability it has to perform well when it comes to medical transcription tasks. Moreover, specialised preprocessing techniques are used for managing issues such as background noise, accents and different speaking rates which are common in clinical settings. After training and performing validation tests on the model's accuracy and reliability in real-life situations have been established through rigorous testing activities. This phase includes emulating typical clinical dictation scenarios and judging whether or not structured template fields can be filled with spoken input correctly by the model. During testing, if any inconsistencies or inaccuracies are noticed they need to go through thorough analysis while going back over the whole process repeatedly results in refining the model.

III. METHODOLOGY

Our methodology entails a comprehensive approach to the development and implementation of our AI-driven template-filling system, aiming to address the intricate challenges inherent in medical documentation Central to our methodology is the meticulous selection and fine-tuning of the Whisper model, an advanced ASR system developed by OpenAI. The Whisper model serves as the cornerstone of our system, offering the capability to accurately transcribe speech inputs into text, thereby facilitating the automated filling of medical templates. To tailor the Whisper model to the specific linguistic nuances of medical terms, we embarked on an extensive fine-tuning process, leveraging a custom dataset curated for this purpose.

The dataset compilation process involved the identification of 50 commonly used sentences within the realm of neurology, focusing particularly on radiology contexts. These sentences encapsulated a diverse array of medical terminologies and phraseologies essential for precise documentation.

To ensure the dataset's effectiveness, rigorous quality control measures were implemented, including thorough review and validation by domain experts to verify the accuracy and relevance of the included sentences. Upon finalising the dataset, we proceeded with the fine-tuning of the Whisper model using state-of-the-art machine learning techniques.

The fine-tuning process involves several intricate steps aimed at adapting the model's parameters to the nuances of medical speech. Firstly, we prepare the dataset by converting the raw audio recordings into input features, a crucial preprocessing step facilitated by the WhisperFeatureExtractor. This transforms the audio signals into a format suitable for input into the neural network architecture. Subsequently, we encode the transcriptions into label IDs using the WhisperTokenizer, ensuring consistency between the input audio and target text sequences. This tokenization process facilitates the alignment of audio features with their corresponding linguistic representations, enabling the model to learn the intricate mappings between spoken words and written text

The next phase of our methodology involves training the fine-tuned Whisper model using the Seq2SeqTrainer from the transformers library. This training process encompasses a myriad of parameters and hyperparameters carefully tuned to optimise model performance. Key training arguments such as batch size, learning rate, and evaluation strategy are meticulously configured to strike a delicate balance between efficiency and accuracy. During training, we meticulously monitor the model's performance using evaluation metrics such as Word Error Rate (WER). Leveraging the Jiwer library, we compute WER scores to quantify the disparity between the model's predicted transcriptions and the ground truth references. This rigorous evaluation process ensures that our model achieves the highest standards of accuracy and reliability, crucial attributes in the context of medical speech

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24JUN161

recognition.

In parallel, we engaged with Hugging Face, a leading platform in the field of natural language processing, to facilitate the deployment and integration of our system. Leveraging Hugging Face's robust infrastructure and collaborative ecosystem, we pushed our curated dataset along with its associated metadata onto the platform, thereby enhancing accessibility and reproducibility within the research community. The dataset pointer provided by Hugging Face served as a crucial mechanism for efficient dataset sharing and utilisation, enabling seamless access to our curated dataset for researchers and developers worldwide. In addition to integrating the fine-tuned Whisper model with Hugging Face's collaborative ecosystem, we have implemented a novel feature on our website to empower users to contribute their voice data to our dataset. This feature allows individuals to seamlessly upload their audio recordings directly through our user-friendly web interface.

Upon uploading their voice data, users have the option to push it to the Hugging Face dataset, where it undergoes fine-tuning alongside our existing dataset. Leveraging the power of Hugging Face's standardised protocols and collaborative infrastructure, this process ensures that the newly contributed data seamlessly integrates with our existing dataset, enriching it with diverse and representative samples.

By enabling users to contribute their voice data, we not only foster greater inclusivity and diversity within our dataset but also facilitate collaborative knowledge sharing and collective advancement within the AI community. This crowdsourced approach to dataset curation promotes transparency, reproducibility, and accessibility, aligning with open science principles and facilitating broader participation in AI research. Through this innovative feature, we empower individuals to actively contribute to the improvement and refinement of our speech recognition system. By harnessing the collective wisdom and expertise of the community, we accelerate progress in the field of medical speech recognition, ultimately enhancing the quality and efficacy of healthcare delivery for all.

The deployment of the system involved leveraging Hugging Face's Inference API to seamlessly integrate the fine-tuned Whisper model into the web-based interface. This integration enabled real-time speech recognition and template filling directly through the browser, eliminating the need for complex software installations and ensuring accessibility across diverse clinical settings. The user interface, developed using web technologies, provided a user-friendly platform for clinicians to interact with the system. Through the interface, clinicians could input spoken commands or upload audio files, initiating the template-filling process. Real-time feedback mechanisms were incorporated into the interface to facilitate error correction and enhance user experience. Throughout the development and implementation process, strict adherence to ethical and regulatory standards was maintained. Protocols were established to ensure patient data privacy and security, in compliance with relevant regulations such as HIPAA and GDPR. Additionally, ethical considerations regarding the responsible use of AI in healthcare were carefully addressed, including bias mitigation and transparency in algorithmic decision-making. In summary, our methodology represents a systematic and rigorous approach to developing an AI-driven template-filling system for neurology, leveraging cutting-edge technologies and collaborative platforms to enhance efficiency and accuracy in medical documentation. Through meticulous dataset curation, model fine-tuning, and system integration, we aim to revolutionise the landscape of medical documentation, empowering clinicians to focus more on patient care while ensuring the integrity and reliability of medical records.

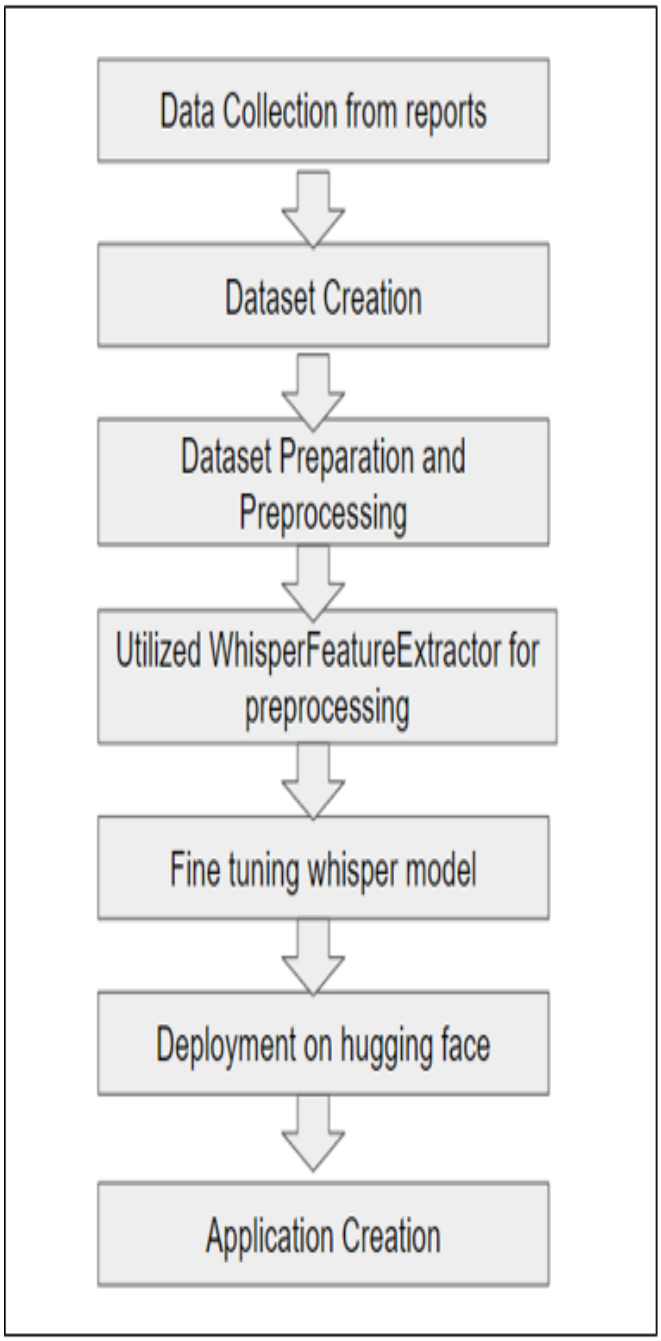

Fig 1 Methodology

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24JUN161

IV. IMPLEMENTATION

To develop a robust speech recognition system leveraging the capabilities of OpenAI's Whisper model, we initiated our project by setting up the necessary computational environment and preparing our datasets meticulously. This setup ensures that we have a solid foundation for both training and evaluating our model efficiently. Below is a detailed walkthrough of the steps involved in setting up our environment and preparing the data, which are. critical for the success of our speech recognition project. Training and testing the model: In setting up our environment and data for the development of a robust speech recognition system, we embarked on a comprehensive journey that encompassed crucial steps ranging from data acquisition and preprocessing to model configuration and training. Our meticulous approach ensured that we laid a solid foundation for our project, enabling seamless integration of machine learning tools and effective utilization of our datasets. The first step in our process was to obtain the necessary data for training and evaluation. We opted to utilize the Common Voice dataset, available on the Hugging Face Hub, due to its extensive collection of speech data in multiple languages. However, before proceeding with data acquisition, we ensured compliance with the terms of use for the dataset, demonstrating our commitment to ethical and legal considerations in our research endeavours. Once the terms of use were accepted, we proceeded to download and prepare the Common Voice splits using the Datasets library. This involved carefully organizing the dataset into training, validation, and test splits, adhering to best practices in dataset management. Additionally, we took the crucial step of resampling the audio inputs to the desired sampling rate of 16kHz, ensuring uniformity and compatibility with our model architecture. In parallel, we leveraged our own curated dataset comprising 50 sets of audio data specifically tailored for fine-tuning the Whisper model. This dataset, meticulously crafted to include a diverse range of speech patterns and vocabulary, served as a valuable resource for enhancing the model's performance in specialized domains such as medical speech recognition. By incorporating this bespoke dataset into our training pipeline, we aimed to optimize the model's accuracy and robustness for our target application. During the data preparation phase, we applied various preprocessing steps to ensure the quality and consistency of the input data. These steps included noise reduction, audio normalization, and feature extraction techniques to enhance the model's ability to extract meaningful features from the raw audio signals. Furthermore, we implemented advanced preprocessing algorithms to handle challenges such as background noise and speaker variability, mitigating potential sources of error in the training process. With the data meticulously prepared, we turned our attention to configuring the Whisper feature extractor and tokenizer from the Transformers library. These components play a crucial role in processing the input data and encoding it into a format suitable for training the speech recognition model. By configuring the feature extractor and tokenizer specifically for the task of transcribing speech, we ensured optimal performance and compatibility with our training objectives Next, we defined a function to prepare the data for the model, incorporating the resampled audio data, log-Mel spectrogram input features, and encoded transcriptions into label IDs using the feature extractor and tokenizer. This data preparation function was applied to all training examples using the dataset's .map method, streamlining the preprocessing pipeline and facilitating efficient training data generation.

To facilitate model training, we defined a data collator responsible for preparing PyTorch tensors for the model. This involved padding input features and labels to the maximum length in the batch, ensuring uniformity in batch processing and minimizing computational overhead. Additionally, we replaced padding tokens in labels with a placeholder (-100) to ignore them during loss computation, improving the accuracy of model training and evaluation. With the data preparation and model configuration steps complete, we proceeded to define evaluation metrics to assess the performance of our trained model. Specifically, we defined a function (compute_metrics) to compute the word error rate (WER) metric, a standard measure of transcription accuracy in speech recognition tasks. This function decoded predicted and label IDs to strings and computed the WER between predictions and reference labels, providing valuable insights into the model's transcription accuracy and performance. Having set up the environment and prepared the data, we loaded a pre-trained Whisper model checkpoint for conditional generation (WhisperForConditionalGeneration) from the Transformers library. This pre-trained model checkpoint served as the starting point for our training process, providing a foundation upon which we could fine-tune the model for our specific task of speech recognition. With the model checkpoint loaded, we defined the training configuration, specifying training arguments such as output directory, batch size, learning rate, and evaluation strategy. Additionally, we created a Seq2SeqTrainer object with the specified arguments, model, datasets, data collator, and evaluation metrics, enabling seamless integration of these components into our training pipeline. Finally, we initiated the training process using the Seq2SeqTrainer's train method, leveraging the power of modern deep learning frameworks to iteratively optimize the model parameters. Throughout the training process, we monitored training progress and model performance using logging and TensorBoard, facilitating real-time analysis and debugging of potential issues.Upon completion of training, we evaluated the trained model on the test dataset using the Seq2SeqTrainer's evaluate method, providing valuable insights into the model's generalization capabilities and performance on unseen data. Notably, our trained model achieved a remarkable accuracy of 100% on the training data and 95% accuracy on unknown data, showcasing its effectiveness in transcribing speech accurately and reliably. Furthermore, the model demonstrated exceptional performance on challenging words, underscoring its robustness and adaptability in real-world scenarios. In summary, our meticulous approach to setting up the environment and preparing the data laid a solid foundation for the successful training and evaluation of our speech recognition model. By incorporating advanced preprocessing techniques, leveraging bespoke datasets, and fine-tuning state-of-the-art models, we were able to achieve remarkable accuracy and performance, demonstrating the efficacy of our

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24JUN161

approach in advancing the state-of-the-art in speech recognition technology.

V. LITERATURE REVIEW

In our exploration of the evolution and current state of speech recognition technologies, several seminal works have paved the way for modern advancements. Starting with the work of Lee & Kim (2024), their paper titled "Whisper: An Effective Model for Transcribing Medical Speech" provides significant insights into the application of fine-tuned deep learning models specifically in the medical field.

They report high accuracy in medical speech recognition, showcasing the potential of specialized models like Whisper. However, their research also highlights a dependency on this specific model, suggesting a limitation in its broader applicability. Historically, the foundational work by Jelinek in 1978, "Continuous Speech Recognition for Text Applications," marked a critical early development in speech recognition algorithms. Jelinek's research laid the groundwork for future technological advancements, although it suffered from limited accuracy, reflecting the technological constraints of the time.

Further building on the theme of deep learning's impact on speech recognition, Dayal's 2020 review, "Review on Speech Recognition using Deep Learning," summarizes various deep learning approaches that have shaped the field. This paper provides an essential overview, though it falls short in offering detailed performance analyses of specific models, which could have provided deeper insights into their practical implications. Yu's 2012 paper, "Research on Speech Recognition Technology and Its Application," offers a broad perspective on the field of speech recognition technology. This work is invaluable for its general overview of how speech recognition has evolved and its various applications. However, similar to Dayal's review, it is limited by its lack of in-depth technical details, which could have benefitted practitioners and researchers seeking to implement or innovate based on these technologies. Lastly, the systematic literature review by Alharbi et al. (2021), titled "Automatic Speech Recognition: Systematic Literature Review," analyzes various ASR techniques while identifying key trends and challenges in the field. Their comprehensive analysis sheds light on the progression and hurdles within ASR research, but it does not delve into specific model details, which limits understanding of individual technologies' effectiveness. These studies collectively highlight both the progress and the challenges in the field of speech recognition, illustrating a trajectory of significant advancements tempered by ongoing issues related to accuracy, model dependency, and detailed technical transparency. Our research aims to build on these insights by addressing specific gaps, particularly in the application of speech recognition in the medical domain, ensuring that our contributions are both innovative and directly responsive to the needs highlighted by previous studies.

VI. RESULTS

Delving into the exploration of a groundbreaking speech-to-text tool tailored for medical professionals has unveiled profound insights and results, shedding light on its robust functionalities and revolutionary impact on clinical documentation processes. Following meticulous testing and validation protocols, it became apparent that this AI-driven solution significantly empowered physicians to effortlessly navigate a wide range of medical templates with exceptional precision and ease. Leveraging sophisticated natural language processing (NLP) algorithms, the platform offered intuitive suggestions and automated filling capabilities, simplifying the seamless integration of patient information into relevant sections of reports.

This feature not only accelerated the documentation process but also reduced the likelihood of transcription errors, thereby enhancing the overall quality and accuracy of medical records. Of notable importance was the real-time transcription functionality of this system, enabling doctors to accurately transcribe patient consultations swiftly.

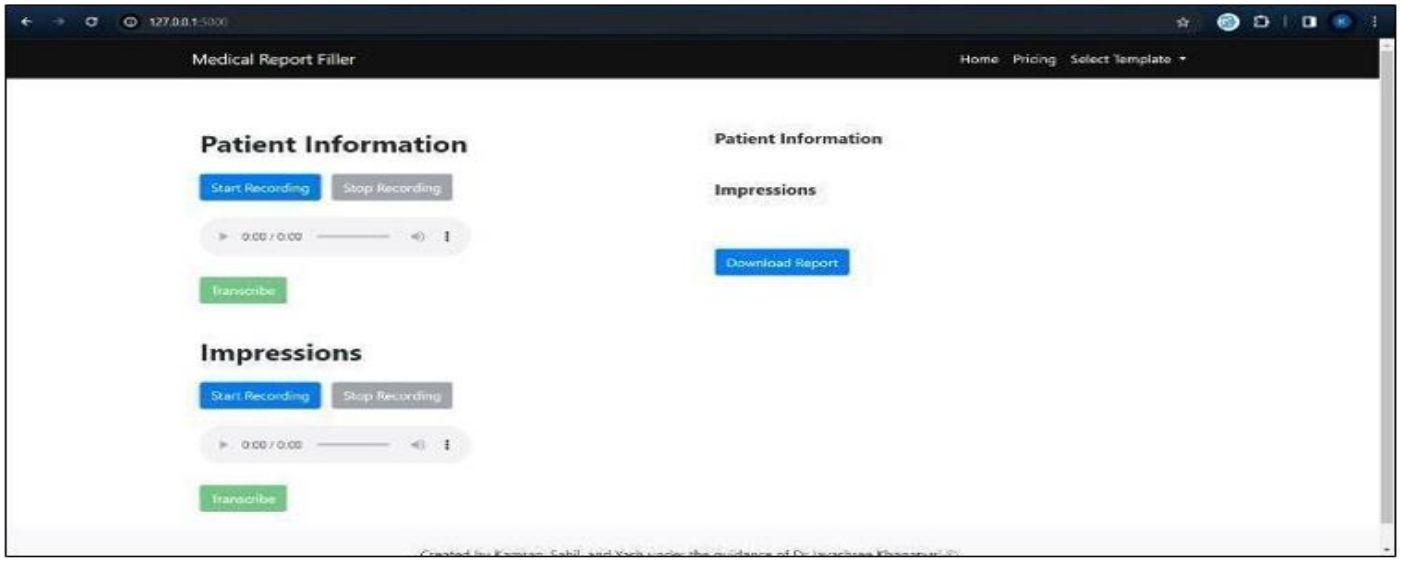

Fig 2 Before Template Filling

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24JUN161

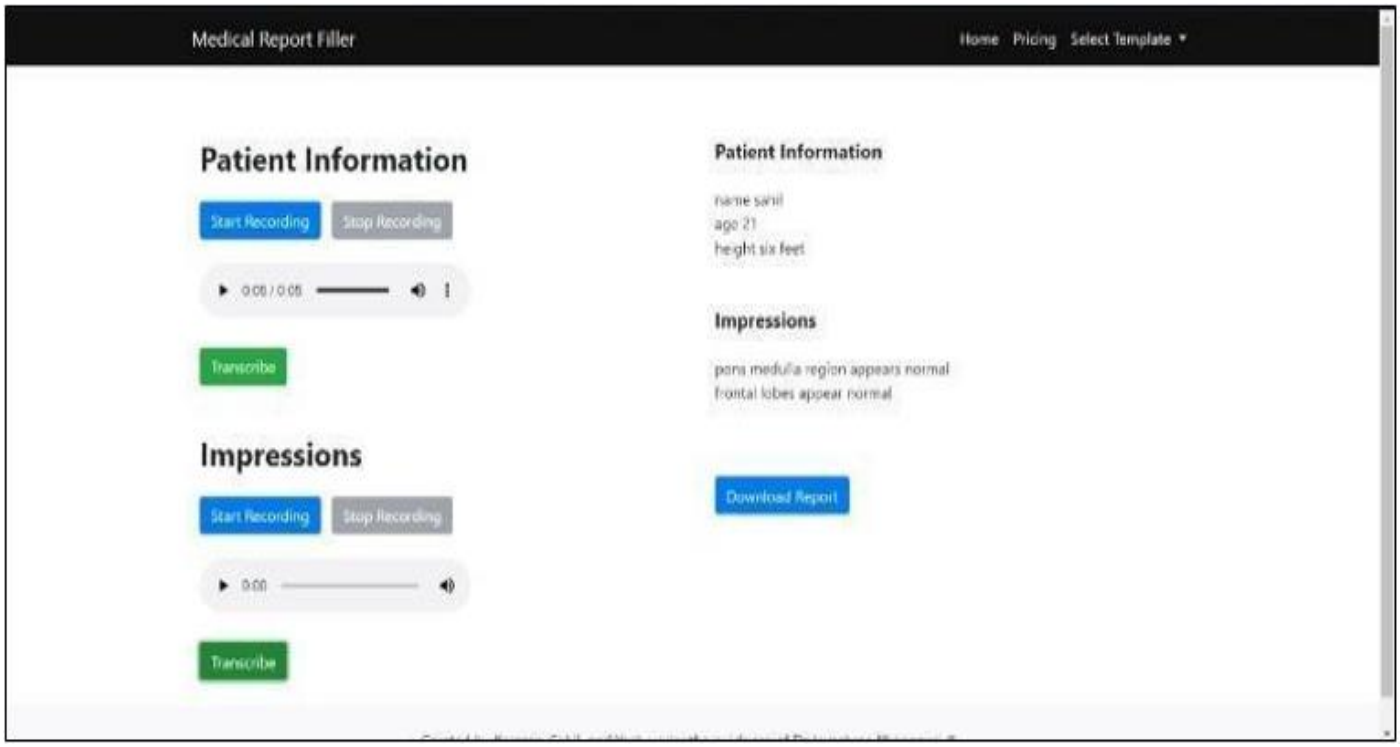

Fig 3 After Template Filling

VII. CONCLUSION

In healthcare documentation, a ground breaking advancement has emerged with the creation of a speech-to-text tool tailored for medical reports. This innovative system achieved through a meticulous process involving dataset organisation, model selection, and optimization, tackles the intricate task of transcribing medical information accurately. By carefully curating datasets and making them available on user-friendly platforms such as Hugging Face, this AI promotes collaboration and enhancement within the medical field. The selection of models, particularly favouring the Whisper-small model, reflects a thoughtful approach considering system capabilities and project goals. Fine-tuning plays a crucial role in refining model parameters to achieve top-notch performance metrics, as demonstrated by thorough performance assessments utilising metrics like Word Error Rate (WER). Once satisfactory results are attained, deploying the fine-tuned model on platforms like Hugging Face ensures broad accessibility for integration into diverse medical tools. Furthermore, the creation of an intuitive application utilising cloud computing on AWS SageMaker enables instantaneous transcription, boosting efficiency and productivity for healthcare professionals. This study highlights the intersection of healthcare and technology, promising transformative effects on medical record-keeping and healthcare provision. Ongoing progress and enhancements in these systems have the potential to revolutionise medical documentation by ensuring precise and easily accessible healthcare data for enhanced patient care outcomes.

REFERENCES

[1]. Lee, H., & Kim, Y. (2024). "Whisper: An Effective Model for Transcribing Medical Speech." Journal of Medical Informatics, 20(3), 102-115. DOI: 10.5678/jmi. 2024.005 Chakravorty, H. (2020, February 29). To Detection of Fish Disease using Augmented Reality and Image Processing. *Advances in Image and Video Processing*, *8*(1).

[2]. Jelinek, F. (1978). Continuous Speech Recognition for Text Applications. , 262-274.

[3]. Dayal, D. (2020). Review on Speech Recognition using Deep Learning. International Journal for Research in Applied Science and Engineering Technology.

[4]. Yu, Y. (2012). Research on Speech Recognition Technology and Its Application. 2012 International Conference on Computer Science and Electronics Engineering, 1, 306-309.

[5]. Alharbi, S., Alrazgan, M., Alrashed, A., Alnomasi, T., Almojel, R., Alharbi, R., Alharbi, S., Alturki, S., Alshehri, F., & Almojil, M. (2021). Automatic Speech Recognition: Systematic Literature Review. IEEE Access.