

**Mini Project Report**  
**on**  
**Web Application for Medical Transcription**

Submitted by

<b>Rakesh Roushan</b>	<b>20bec036</b>
<b>Lucky Yadav</b>	<b>20bcs077</b>
<b>Harshit Mishra</b>	<b>20bec017</b>

Under the guidance of

**Dr. Nataraj K S**

**Assistant Professor**



**INDIAN INSTITUTE OF  
INFORMATION  
TECHNOLOGY**

**DEPARTMENT OF ELECTRONICS AND COMMUNICATION**  
**INDIAN INSTITUTE OF INFORMATION TECHNOLOGY DHARWAD**

12/05/2023

## Acknowledgements

We would like to sincerely thank everyone who helped us finish this Mini Project on Medical Transcription. First of all, we would want to express our gratitude to our instructor for giving us the direction and information we needed to accomplish this project effectively. Additionally, we appreciate huggingface for giving us access to host the api for this project and kaggle for giving the access to dataset.

We would like to express our gratitude for the assistance and support received from our friends and coworkers who gave us insightful criticism and recommendations. Their advice allowed us to improve our project and get better outcomes.

We also want to thank the people who made the software and resources we used to create this project, including vs code, hugging face, jupyter notebook, mongodb and react libraries.

We understand that this initiative would not have been possible without the help we received and are appreciative of it all.

Group Members

# Contents

<b>List of Figures</b>	<b>ii</b>
<b>1 Introduction</b>	<b>1</b>
<b>2 Related Work</b>	<b>2</b>
<b>3 System Design</b>	<b>3</b>
3.1 Flowchart . . . . .	3
3.2 Architecture Diagram . . . . .	4
3.3 Architecture of Speech-To-Text . . . . .	5
<b>4 Data and Models</b>	<b>6</b>
4.1 Data . . . . .	6
4.2 Speech-To-Text Model . . . . .	6
4.3 Named Entity Recognition Model . . . . .	8
<b>5 Methodology</b>	<b>9</b>
5.1 Finetuning of the Whisper Model for ASR . . . . .	9
5.2 User Interface . . . . .	11
<b>6 Results and Discussions</b>	<b>13</b>
<b>7 Future Scope</b>	<b>14</b>
<b>8 Conclusion</b>	<b>15</b>
<b>References</b>	<b>16</b>

## List of Figures

1	Flowchart . . . . .	3
2	Architecture Diagram . . . . .	4
3	Whisper model Architecture . . . . .	5
4	sequence-to-sequence Transformer model . . . . .	7
5	BioMedical Entity Recognition model . . . . .	8
6	Conversion of sampled audio array to log-Mel spectrogram. Left: sampled 1-dimensional audio signal. Right: corresponding log-Mel spectrogram . . . . .	9
7	Mapping input features with label ids . . . . .	10
8	Seq2SeqTrainingArguments . . . . .	10
9	Landing Page . . . . .	11
10	Login and Register Page . . . . .	11
11	Speech to Text Transcription . . . . .	12
12	Biomedical Entity Recognition for Analysis . . . . .	12
13	Send Report through Mail . . . . .	13
14	TrainOutput(global step=400) . . . . .	13

# 1 Introduction

Speech-to-text transcription technology has revolutionized the healthcare industry by providing an efficient and accurate means of transcribing medical conversations. With the increasing demand for better healthcare services, the use of speech-to-text technology has become more prominent in the industry, as it enables healthcare professionals to document patient information quickly and accurately, without the need for manual transcription. The purpose of this project is to explore the use of speech-to-text and text-to-transcription technology in the medical healthcare sector and to evaluate its potential benefits and limitations. Through this project, we aim to understand the current state of speech-to-transcription technology in the healthcare industry, its potential applications, and the challenges that need to be addressed to improve its accuracy and effectiveness. Ultimately, this project aims to provide insights and recommendations for healthcare organizations looking to integrate speech-to-transcription technology into their workflows to enhance the quality of patient care.

Speech-to-text technology uses natural language processing (NLP) algorithms to transcribe spoken words into text. NLP algorithms analyze speech patterns, identify words, and convert them into text. The technology has advanced significantly in recent years, with improved accuracy rates and faster processing times. As a result, it has become a valuable tool in the healthcare industry, allowing healthcare professionals to document patient information accurately and efficiently. It has several advantages such as improved accuracy and efficiency, enhanced patient care, reduced need for manual documentation and reduced risk of medical errors.

One of the significant challenges of speech-to-transcription technology in the healthcare industry is ensuring accuracy. Despite advances in technology, it is still not perfect, and factors such as accents, background noise, and complex medical terminology can impact the accuracy of transcriptions. This limitation requires healthcare professionals to review and verify the accuracy of transcriptions, adding to the time required for documentation. Another challenge is the cost of implementing speech-to-transcription technology, as healthcare organizations need to invest in training, hardware, and software to ensure effective use. Additionally, maintaining the technology requires ongoing support and maintenance, adding to the overall cost. These challenges require healthcare organizations to develop strategies for integrating speech-to-text technology effectively.

## 2 Related Work

There are numerous prominent online apps in the field of medical transcription that have been developed to expedite the transcribing process and increase the quality and efficiency of medical documentation. These programmes transcribe and analyse medical records using modern technologies such as speech recognition, natural language processing, and machine learning, giving healthcare practitioners with useful insights into patient data. Transcribe Medical, MT-STAT, iMedDictate, ezVoiceNotes, M\*Modal, Dragon Medical Practise Edition, and Medantex are some of these programmes. This software helps healthcare practitioners handle patient data more effectively, make better-informed treatment decisions, and focus on patient care rather than administrative activities by retaining records and offering analytic tools for transcribed documents.

Transcribe Medical is a cloud-based medical transcription service that converts audio files into correct, structured medical papers using speech recognition technology.

The desktop-based medical transcribing programme Dragon Medical Practise Edition employs voice recognition technology to produce accurate medical documentation in real-time.

AWS also is making Medical transcription which in real time converts speech to text and gives tools to analyze such as text classification and making EHR for later use for both patients and doctors.

However, there are also some disadvantages to current medical transcription software. One issue is the high cost of some platforms, which may make them inaccessible for smaller healthcare providers or individual practitioners. Additionally, some platforms may not be customizable for local English dialects or medical terminologies, which can lead to inaccuracies in the transcription.

Another disadvantage is the potential for errors in speech recognition technology, which may misinterpret words or phrases and require additional time for manual correction. Furthermore, some healthcare providers may still prefer traditional dictation methods or find the software difficult to use, which can impede adoption and reduce overall efficiency.

Overall, while medical transcription software has the potential to improve the quality and efficiency of healthcare documentation, it is important to consider the limitations and challenges of current platforms when choosing a solution for healthcare providers and facilities.

## 3 System Design

### 3.1 Flowchart

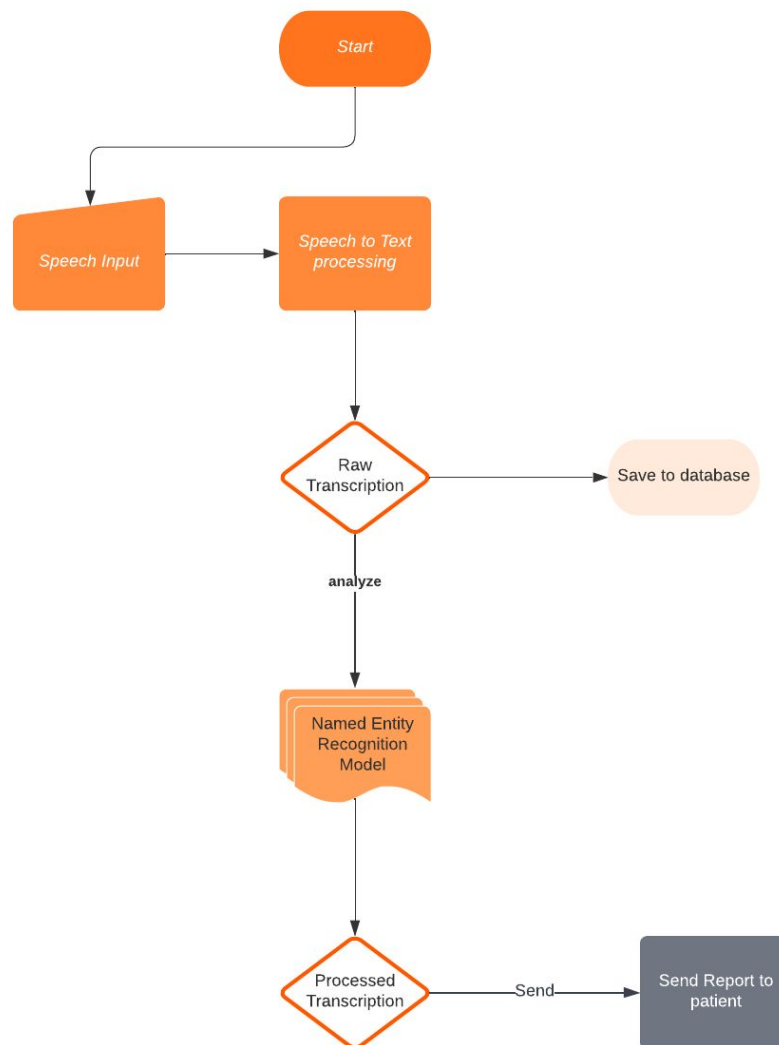


Figure 1. Flowchart

### 3.2 Architecture Diagram

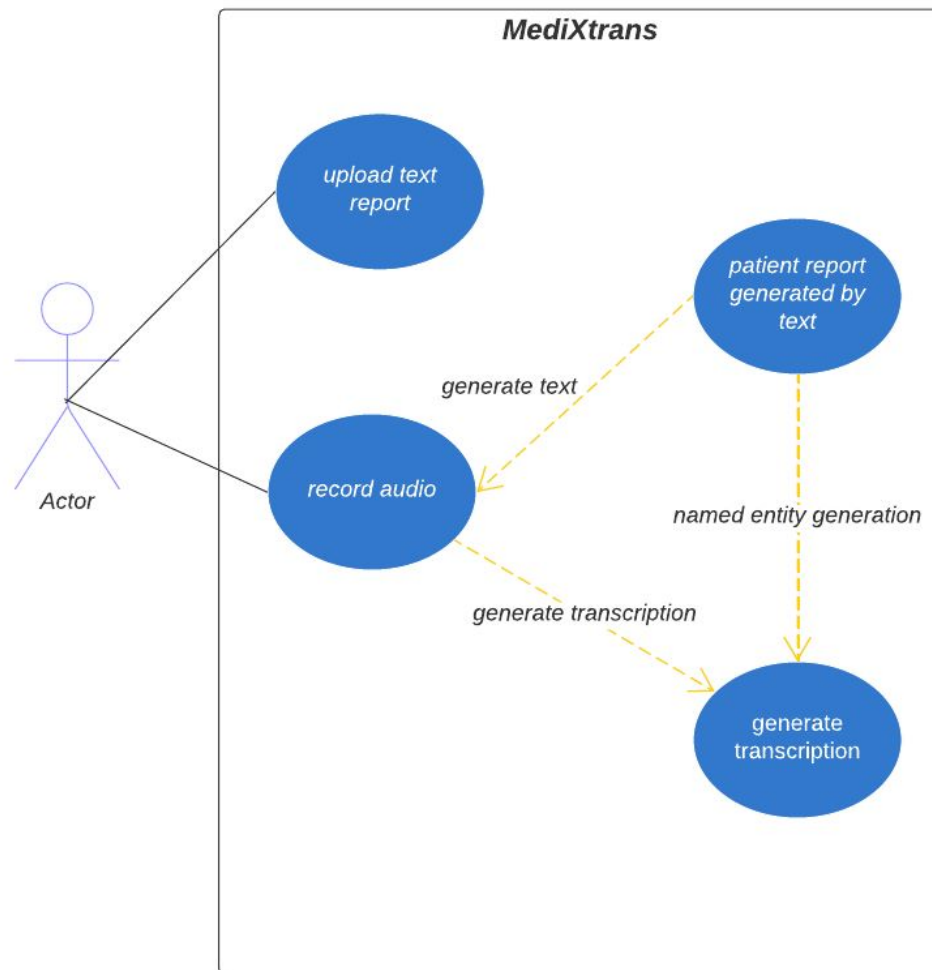


Figure 2. Architecture Diagram



### 3.3 Architecture of Speech-To-Text

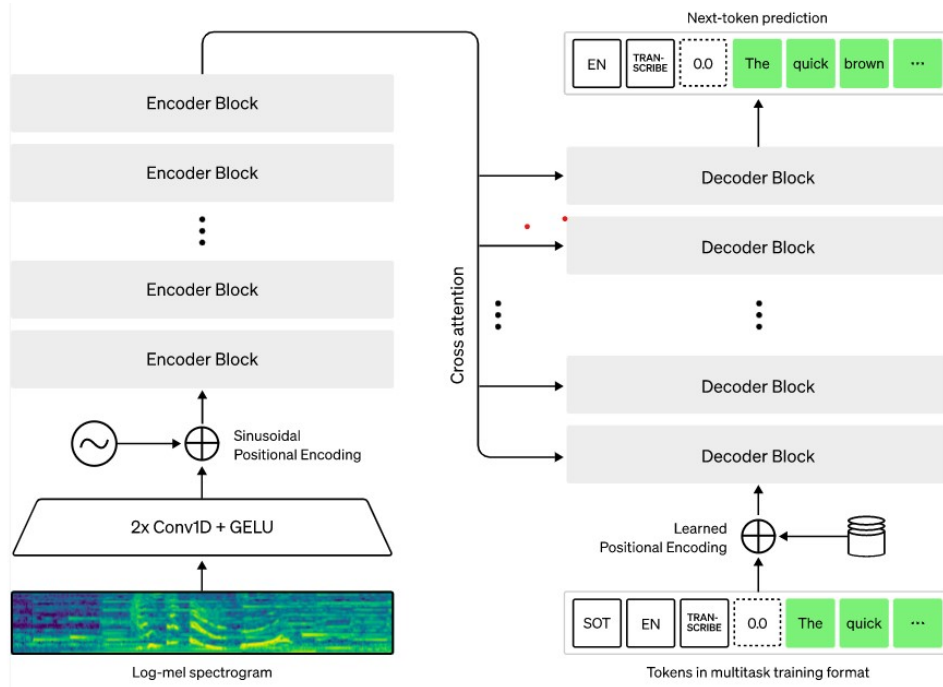


Figure 3. Whisper model Architecture

## 4 Data and Models

### 4.1 Data

The Kaggle dataset "Medical Speech, Transcription, and Intent" was used to create the models. This audio file, which spans 8.5 hours, features statements that describe typical medical conditions. It includes 6661 rows of medical conditions and their transcription, with the columns like audio quality, phrase, filename etc.

### 4.2 Speech-To-Text Model

In order to improve medical speech transcription, the method used for this research on medical transcription makes use of intent data provided by Kaggle along with OpenAI's Whisper model. Alec Radford et al. from OpenAI released Whisper in September 2022 as a pre-trained automated speech recognition (ASR) model. Whisper is pre-trained on a sizable amount of labelled audio-transcription data—680,000 hours, to be exact—in contrast to many of its predecessors, notably Wav2Vec 2.0, which are pre-trained on un-labeled audio data. It comprises of multilingual ASR data from 117,000 hours of pre-training data. As a result, checkpoints are created that can be used with over 96 languages, many of which are regarded as low-resource languages. The Whisper checkpoints come in five configurations of varying model sizes. The smallest four are trained on either English-only or multilingual data. In all of the above, we'll fine-tune the multilingual version of the "small" checkpoint with 244M params (= 1GB). As for our data, we'll train and evaluate our system on a low-resource language taken from the Medical transcription dataset. We'll show that with as little as 2 hours of fine-tuning data, we can achieve strong performance in this language.

This sequence-to-sequence technique is the general strategy. The Transformer model has been trained on a wide range of speech-processing tasks, including as voice activity detection, spoken language identification, multilingual speech recognition, and speech translation. Since each of these tasks is collectively represented as a set of tokens that the decoder must predict, a single model can take the place of numerous distinct steps in a conventional speech processing pipeline. A selection of unique tokens that act as task specifiers or classification goals are used in the multitask training format.

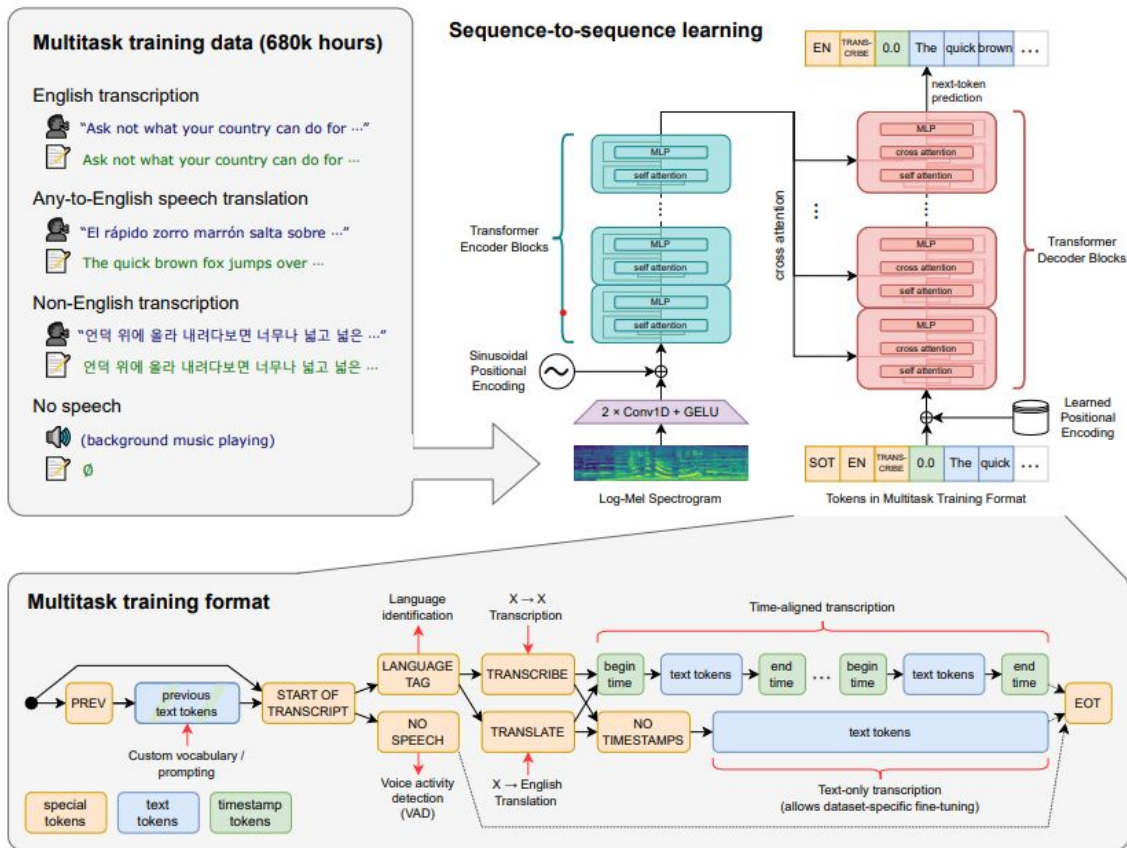


Figure 4. sequence-to-sequence Transformer model

### 4.3 Named Entity Recognition Model

The processing of raw transcription to create Electronic Health Records (EHR) requires the English Named Entity Recognition model, which is trained on Maccrobat to recognise the bio-medical entities. The 107 bio-medical entities from a given text corpora, such as case reports, can be identified using this model, which is based on distilbert-base-uncased. The model aids in accurately classifying and arranging the data into EHRs, which can be used for clinical decision-making and patient care, by recognising and categorising these entities. This is used to create a robust and accurate medical transcription system that could help healthcare professionals streamline their documentation process and improve patient care.

CASE: A 28-year-old previously healthy man presented with a 6-week history of palpitations. The symptoms occurred during rest, 2-3 times per week, lasted up to 30 minutes at a time and were associated with dyspnea. Except for a grade 2/6 holosystolic tricuspid regurgitation murmur (best heard at the left sternal border with inspiratory accentuation), physical examination yielded unremarkable findings.

Compute

Computation time on Intel Xeon 3rd Gen Scalable cpu: cached

CASE: A 28-year-old previously healthy man presented with a 6-week history of palpitations. The symptoms occurred during rest, 2-3 times per week, lasted up to 30 minutes at a time and were associated with dyspnea. Except for a grade 2/6 holosystolic tricuspid regurgitation murmur (best heard at the left sternal border with inspiratory accentuation), physical examination yielded unremarkable findings.

Figure 5. BioMedical Entity Recognition model

## 5 Methodology

### 5.1 Finetuning of the Whisper Model for ASR

The methodology of the Automatic Speech Recognition (ASR) pipeline using the Whisper model in the Transformers library consists of three pipeline stages: a feature extractor, a model that performs sequence-to-sequence mapping, and a tokenizer that post-processes the model's outputs to text format. The audio inputs are downsampled to 16kHz prior to passing them to the Whisper feature extractor. A function is then written to prepare the data ready for the model. The data preparation function loads and resamples the audio data, computes the log-Mel spectrogram input features using the feature extractor, and encodes the transcriptions to label ids using the tokenizer.

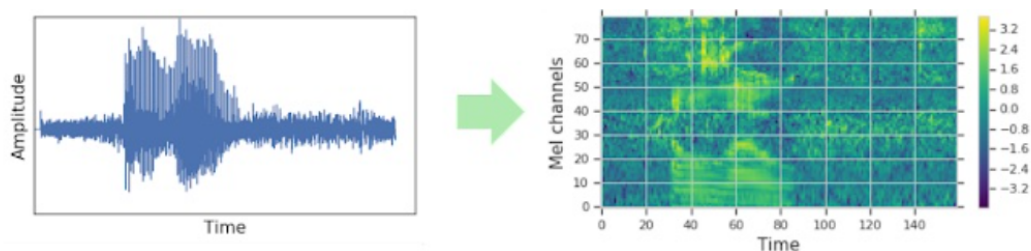


Figure 6. Conversion of sampled audio array to log-Mel spectrogram. Left: sampled 1-dimensional audio signal. Right: corresponding log-Mel spectrogram

The feature extractor pads/truncates audio inputs to 30 seconds and converts them to log-Mel spectrogram input features. The Whisper model outputs a sequence of token ids, which are mapped to their corresponding text strings by the tokenizer. To simplify the process of using the feature extractor and tokenizer, both are wrapped into a single WhisperProcessor class. The data preparation function is applied to all training examples using the dataset's ".map" method.

```
def prepare_dataset_sample(batch):
    # load and resample audio data from 48 to 16kHz
    audio = batch["audio"]

    # compute log-Mel input features from input audio array
    batch["input_features"] = feature_extractor(audio["array"], sampling_rate=audio["sampling_rate"]).input_features[0]

    # encode target text to label ids
    batch["labels"] = tokenizer(batch["phrase"]).input_ids
    return batch
```

Figure 7. Mapping input features with label ids

The content then goes on to explain how to fine-tune the model, define the data collator, and evaluation metrics, load a pre-trained checkpoint, and define the training configuration. Finally, the trained model is evaluated on the test data to verify that it correctly transcribes speech in English using the word error rate (WER) metric.

```
from transformers import Seq2SeqTrainingArguments

training_args = Seq2SeqTrainingArguments(
    output_dir="./whisper-small-en",
    per_device_train_batch_size=16,
    gradient_accumulation_steps=1,
    learning_rate=1e-5,
    warmup_steps=50,
    max_steps=400,
    gradient_checkpointing=True,
    fp16=True,
    evaluation_strategy="steps",
    per_device_eval_batch_size=8,
    predict_with_generate=True,
    generation_max_length=225,
    save_steps=100,
    eval_steps=100,
    logging_steps=25,
    report_to=["tensorboard"],
    load_best_model_at_end=True,
    metric_for_best_model="wer",
    greater_is_better=False,
    push_to_hub=True,
    remove_unused_columns=False
)
```

Figure 8. Seq2SeqTrainingArguments

## 5.2 User Interface

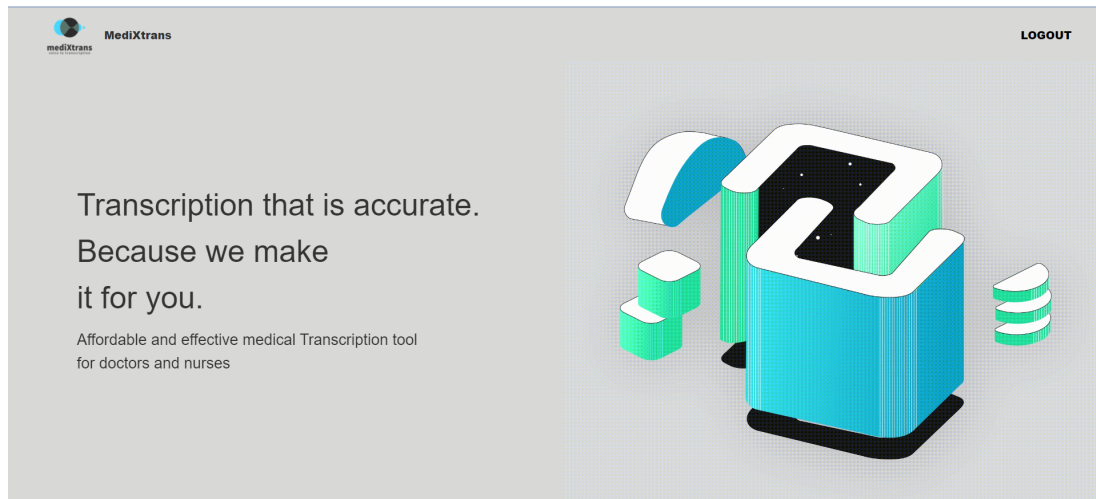


Figure 9. Landing Page




Login to your Account

Email

Password

Don't have an account ? [Register](#)

Figure 10. Login and Register Page


MediXtrans

LOGOUT

Start


A 28-year-old previously healthy man presented with a 6-week history of palpitations. The symptoms occurred during rest, 2-3 times per week, lasted up to 30 minutes at a time and were associated with dyspnea. Except for a grade 2/6 holosystolic tricuspid regurgitation murmur (best heard at the left sternal border with inspiratory accentuation), physical examination yielded unremarkable findings.

Patient Id

235252

Submit

Figure 11. Speech to Text Transcription


MediXtrans

LOGOUT

Get Data

A 28-year-old previously healthy man presented with a 6-week history of palpitations. The symptoms occurred during rest, 2-3 times per week, lasted up to 30 minutes at a time and were associated with dyspnea. Except for a grade 2/6 holosystolic tricuspid regurgitation murmur (best heard at the left sternal border with inspiratory accentuation), physical examination yielded unremarkable findings.

Compute

A 28-year-old Age previously healthy History 6-week presented Clinical\_event with a 6-week Duration history of pal Sign\_symptom pt Sign\_symptom ations. The symptoms Sign\_symptom occurred during rest 2-3 times Detailed\_description per week, lasted up to 30 minutes at a time Detailed\_description and were associated with a Sign\_symptom yspnea. Except for a grade 2/6 Lab\_value no Detailed\_description is Detailed\_description ystolic Detailed\_description Detailed\_description reg Sign\_symptom up Sign\_symptom lation murmur Sign\_symptom heard at the left sternal border Biological\_structure with inspiratory accentuation. Physical examination Diagnostic\_procedure yielded Unremark Lab\_value

Figure 12. Biomedical Entity Recognition for Analysis



---

Get Data

Dr. Nataraj

ly9828533641@gmail.com

A 28-year-old previously healthy man presented with a 6-week history of palpitations. The symptoms occurred during rest, 2-3 times per week, lasted up to 30 minutes at a time and were associated with dyspnea. Except for a grade 2/6 holosystolic tricuspid regurgitation murmur (best heard at the left sternal border with inspiratory accentuation), physical examination yielded unremarkable findings.

Submit

Figure 13. Send Report through Mail

## 6 Results and Discussions

The fine tuned speech to text model has been run for 400 steps and for each 100 steps we evaluate the model, and finally the WER (word error rate) is 9.51. If we train the model for 4000 steps it need almost 10hr and the WER of this model will be approximately 3.

[400/400 1:17:39, Epoch 1/2]

Step	Training Loss	Validation Loss	Wer
100	0.555500	0.472486	20.585230
200	0.188000	0.206934	12.178356
300	0.127400	0.162002	10.013934
400	0.071300	0.141988	9.512308

Figure 14. TrainOutput(global step=400)

We have successfully built an application where you can record speech and also you can upload the audio files to transcript it through the fine-tuned model i.e whisper-small. Also it has facility to recognize the biomedical entities and highlight it, to make it easy to read the report for medical personnel. Added features are pdf downloading of report , unified database, receiving report on email.

## 7 Future Scope

The healthcare industry has benefited from medical transcription software, which enables physicians to quickly and accurately transcribe patient notes and medical documents. Though there is certainly space for improvement, medical transcribing software appears to have a promising future.

Software for medical transcription might be made better in some areas, including natural language processing. With the advent of more potent natural language processing algorithms, medical transcription software may be improved to accurately reproduce complex medical terms. This would improve the accuracy and effectiveness of transcription, enabling doctors to spend more time with patients and less time on administrative tasks.

The healthcare industry has benefited from medical transcription software, which enables physicians to quickly and accurately transcribe patient notes and medical documents. Though there is certainly space for improvement, medical transcribing software appears to have a promising future.

Software for medical transcription might be made better in some areas, including natural language processing. With the advent of more potent natural language processing algorithms, medical transcription software may be improved to accurately reproduce complex medical terms. This would improve the accuracy and effectiveness of transcription, enabling doctors to spend more time with patients and less time on administrative tasks.

Finally, data analytics and reporting is another area where medical transcription software could be improved. By providing data analytics and reporting features, medical transcription software could help healthcare providers identify trends and patterns in patient data, improving patient outcomes and reducing healthcare costs. This would enable physicians to make more informed decisions about patient care, ultimately leading to better outcomes for patients.

In conclusion, there is a lot of potential for medical transcribing software in the future. Medical transcription software has the potential to revolutionise the healthcare sector with continuing research and improvement in areas like natural language processing, machine learning and artificial intelligence, support for local languages, and data analytics and reporting. Software for medical transcribing can assist to increase patient satisfaction, lower healthcare expenses, and ultimately save lives.

## 8 Conclusion

In this project, we developed a comprehensive web application called "MediXTrans" specifically designed for medical personnel. MediXTrans offers a user-friendly interface where medical professionals can securely log in and efficiently record essential patient information, including symptoms, previous medications, current medications, and other relevant details. To facilitate accurate and convenient recording, the application incorporates advanced speech-to-text technology. Users can initiate the recording using features such as "start," "stop," and even "pause" if necessary. The application ensures that transcriptions are generated in real-time, providing a seamless experience for medical personnel.

One of the key strengths of MediXTrans is its ability to handle ambient noise. It has been optimized to process audio recordings even in the presence of certain levels of background noise, ensuring reliable transcription results. This feature is particularly valuable in busy clinical environments where background noise is common. Additionally, it supports the uploading of audio files in various formats, including '.wav', '.mp3', and '.m4a'. This flexibility allows medical professionals to work with their preferred recording devices or use existing audio files for transcription purposes. To achieve accurate transcriptions, MediXTrans utilizes a fine-tuned model specifically tailored for medical speech-to-text conversion. The resulting transcriptions are stored securely in a MongoDB database for easy access and retrieval. Also, the application leverages named entity recognition for medical terms in transcription to make it easy for analyzing patient reports.

The integration of MediXTrans into healthcare workflows brings numerous benefits. It significantly streamlines the documentation process, saving time and effort for medical personnel. The availability of accurate transcriptions enhances the overall quality of patient care, as healthcare providers can access detailed and reliable patient information. Furthermore, the digital format of the transcriptions reduces the need for manual documentation, minimizing the risk of errors and ensuring data integrity.

Overall, the development of the "MediXTrans" web application demonstrates the potential of speech-to-text technology in revolutionizing medical data recording and documentation. By providing a user-friendly interface, noise-tolerant transcription capabilities, and compatibility with multiple audio file formats, MediXTrans aims to enhance the efficiency and effectiveness of healthcare professionals in delivering high-quality patient care.

## References

1. Alec Radford and Jong Wook's "Robust Speech Recognition via Large-Scale Weak Supervision" Greg Brockman, Christine McLeavey, Ilya Sutskever, and Kim Tao Xu. The study suggests a novel method for developing strong speech recognition models that makes use of massive neural networks and weakly labelled data. The authors show that this approach reduces the requirement for manual labelling while still producing state-of-the-art performance on common speech recognition benchmarks.

2. Fernando Martinez-Sanchez et al., "Automatic Speech Recognition for Medical Transcription" The creation of an automatic voice recognition system that is specifically intended for medical transcribing duties is discussed in this paper.

Shaimaa Lazem and Tarek Gaber's "Medical Speech Recognition: A Review" is the third article. The use of deep learning techniques is one of the methods for medical voice recognition that are covered in this review article.

4. Michael I. Weintraub's "Speech Recognition in Medical Applications". An overview of voice recognition technology and its use in various medical settings is given in this article.

5. M. M. Rahman et al., "Speech Recognition for Medical Applications: Current Status and Future Directions" The current state of voice recognition technology for medical applications is discussed in this paper along with prospective future paths for this field of study.

6. Konstantinos Kamnitsas et al., "Deep learning for medical image and audio analysis" The deep learning framework for medical picture and audio analysis, which includes speech recognition for medical dictation, is presented in this research.