Speech-to-Text Technologies in Medical Field

Dissertation Title: Speech-to-Text Technologies in Medical Analysis

Master title: MSc Data Analytics

Name: Mustafa Akgül

Year: 2025

ABSTRACT

In the medical field, speech-to-text technology has transformed the digitization of clinical processes in recent years, primarily due to the rise of electronic health records. This research investigates the challenges associated with transcribing medical terminology in healthcare, explores potential solutions, and examines how these models can be integrated with clinical decision support systems. The primary objectives of the research are to reduce transcription errors and explore their integration with real-time critical decision support systems while adhering to transparent data privacy policies and addressing ethical issues. The main findings indicate that combining Automatic Speech Recognition and Natural Language Processing models, along with the data processing techniques employed, significantly enhances medical transcription, and optimizes workflows in clinical processes. In conclusion, the research highlights that artificial intelligence models (ASR and advanced NLP) can enhance clinical processes by minimizing transcription errors.

Keywords: Automatic Speech Recognition (ASR), Natural Language Processing (NLP), Clinical Process, Speech-to-Text Technology

CONTENTS

ACKNOWLEDGEMENTS

I would like to express my deepest gratitude to the following people, without whom this dissertation would not have been possible. Their unwavering support, encouragement, and invaluable contributions have helped me complete this work with dedication and excellence. First and foremost, I am profoundly grateful to God, who has always been my source of strength and perseverance, guiding me even in the most challenging moments. I would like to give special thanks to my supervisor, Prof. Dr. Christos Lemonakis, whose insightful feedback, patience, and constant guidance have been instrumental in shaping this research. Your encouragement and willingness to provide direction whenever I needed it have been invaluable. I am deeply thankful to my parents and family, whose unconditional love and belief in me have been my strongest pillar of support. They have endured my moments of stress and self-doubt, always pushing me forward with their unwavering faith in my abilities. Finally, thank you to everyone who has supported me in ways big and small, whether through a kind word, a thoughtful suggestion, or simply by believing in me. This accomplishment is as much yours as it is mine.

With gratitude,
Mustafa Akgül

Statement of compliance with academic ethics and the avoidance of plagiarism

I honestly declare that this dissertation is entirely my own work and none of its part has been copied from printed or electronic sources, translated from foreign sources and reproduced from essays of other researchers or students. Wherever I have been based on ideas or other people texts I clearly declare it through the good use of references following academic ethics.
(In the case that is proved that part of the essay does not constitute an original work, but a copy of an already published essay or from another source, the student will be expelled permanently from the postgraduate program).

Name and Surname (Capital letters):

MUSTAFA AKGÜL

Date: 24/02/2025

DISSERTATION THESIS

1. INTRODUCTION

Artificial intelligence is at the center of today's technological developments. It leads to self-improving systems (e.g., machine learning and deep learning) with advancements in digital cloud technologies, especially over the last decade. Especially the combination of cloud technologies and artificial intelligence has been a pioneer in using health data more organized and effective. (Obermeyer & Emanuel, 2016). These developments have resulted in radical transformations across various fields, particularly in the healthcare sector. Artificial intelligence has created countless opportunities in the healthcare sector. The most important of these are clinical applications, patient monitoring, drug development, and disease diagnosis. The application of AI-based models in clinical processes has enabled doctors and healthcare professionals to make accurate and effective decisions. Recent studies show that ASR models, which convert current voice to text, reduce the error rate by 35 percent compared to traditional methods (Zhang et al., 2020). Therefore, in recent years, healthcare professionals have increasingly preferred speech recognition technology.

Aims and Objectives of the Study

The research aims to enable ASR technology used in clinical settings to recognize medical terminology and provide high-quality transcription by minimizing background noise. Thus, it identifies methods for improvement by benchmarking against word error detection techniques. Another key goal is to ensure that the ASR output is both meaningful and syntactically correct. While ASR output can classify individual words, it is crucial to maintain semantic and syntactic integrity, especially in sensitive areas where critical decisions are made. This highlights the significance and application of NLP techniques. Thus, it enhances communication among healthcare professionals by ensuring that the final transcriptions are word-based, semantically consistent, and grammatically accurate. Furthermore, the project seeks to provide insights into how critical issues such as data privacy, security, and ethics can be addressed using ASR-NLP integrated models.

- **Minimizing Error Rates**: Word error rate (WER) is one of the most critical performance measures of ASR systems and reducing this error rate is the primary goal of ASR systems.
- **Strengthening Semantic and Syntactic Analysis:** This project aims to correct linguistic differences, maintain semantic integrity, and accurately classify texts obtained from audio data by integrating advanced NLP techniques such as transformer-based models, GPT, and BERT.
- **Integration of Real-Time Applications:** This project aims to accelerate early intervention and data analysis processes by integrating ASR and NLP technologies into clinical decision support systems in emergency services and chronic disease management.
- **Ensuring Data Security and Ethical Standards:** Data anonymization techniques, storage methods and ethical standards required to protect the confidentiality of patient data will be meticulously implemented.

Research Questions

This research aims to fill the existing literature gap between case studies of ASR-NLP integration used in speech recognition systems in clinical document transcription and real-time clinical decision support development. While numerous studies have discussed the importance of AI models in healthcare, limited work has specifically investigated how modern ASR-NLP integrated architecture can work together to reduce transcription errors, improve semantic and syntactic accuracy, and facilitate rapid clinical interventions. Addressing challenges such as transcription of noisy audio data or data security is critical, especially given its importance in sensitive care settings such as emergency departments or intensive care units where decisions need to be made quickly. To address these gaps, this thesis explores the following research questions:

- *How do background noise in clinical settings and domain-specific vocabulary, such as rare diseases or complex medical terms, affect the word error rate (WER) of ASR systems? Which adaptation techniques, like audio filtering and enhancements in language modelling, are most effective at mitigating this effect?*

- *How can the semantic and syntactic accuracy of clinical data transcribed with advanced transformer-based NLP models (GPT, BERT, etc.) be improved? What is the impact of this improvement on overall documentation quality and decision-making processes in healthcare?*
- *When using ASR and NLP systems in healthcare, what data anonymisation, secure storage and ethical compliance standards should be put in place to protect the privacy of patient data? How do these approaches affect system performance and adoption?*

Methodology

The study uses a multi-stage experimental research design that includes quantitative and qualitative elements:

Research Design

A multi-stage, experimental research design was adopted to identify the challenges and opportunities presented by existing methods for automatically transcribing and analyzing medical voice recordings. This design integrated speech recognition and NLP techniques.

Data Collection

Audio data were collected in .wav format with relevant metadata such as recording time, background noise level and clinical context. The dataset contains a total of 8 and a half hours of audio recordings of patients describing their complaints, taken from patient-doctor interviews in a real hospital setting. This data contains thousands of audio utterances for common medical symptoms such as knee pain or headache.

Data Pre-processing and Implementing Techniques

Several techniques were applied during the pre-processing steps to ensure better data generalization. First, the data underwent steps such as normalization, sentence and word segmentation, data augmentation, and noise reduction. Next, the pre-processed data was transcribed using a pre-trained speech-to-text Whisper model with medical terminology.

Subsequently, semantic and syntactic techniques were applied with NLP models. Finally, the performance of the AI model was calculated using appropriate metrics.

Synopsis

The dissertation is structured as follows:

1. Introduction: This section discusses the importance of accurate clinical documentation, the challenges faced by ASR in noisy clinical environments, and the potential benefits of ASR-NLP integration. It also provides a detailed description of the study's goals and objectives, each research question, and how this project will address them.

2. Literature Review 1: This section covers the basic concepts of speech-to-text technology, the historical evolution of ASR, NLP, and recent developments in the field, including where and how speech-to-text AI models are applied in healthcare.

3. Literature Review 2: This section covers modern ASR-NLP systems, their technical details, clinical applications, limitations, and ethical and data privacy issues associated with these technologies.

4. Methodology: This section provides a detailed description of the research process, including the process of collecting and loading the dataset, pre-processing techniques such as data augmentation, filtering, normalization, AI models for ASR-NLP integration, and evaluation criteria. Additionally, case studies and techniques used in data analysis are summarized.

5. Analysis, Findings, Discussion: This section includes data analysis, visualizations, and statistical evaluations. The findings are discussed in the context of the research questions, and system performance and implementation possibilities are discussed. The discussion section also includes the current study's limitations and future recommendations.

6. Conclusion: Summarizes key insights, discusses limitations, and suggests future research directions.

This dissertation will explore the process of speech-to-text technologies and highlight the current challenges and opportunities in the medical field. By analysing the collected data, the findings will provide insights into clinical practices. The following chapters will delve into data science and artificial intelligence, focusing on identifying clinical decision support systems.

2. LITERATURE REVIEW I

The first section introduces key terms related to speech-to-text technologies in the field of health. It then explains the historical development of speech recognition technologies from the past to the present. Finally, it examines the areas in which they are widely applied in health applications.

2.1 Understanding Key Terms

The interest in understanding human sense organs, such as vision and listening through machines, which started in the 1950s, has increased even more in the process until today. Over the years, this interest has paved the way for important developments in the field of artificial intelligence and has almost approached human intelligence in areas such as speech recognition, image recognition, and language understanding. Large data sets have been created with the development of cloud systems and data transfer to digital environments. This has improved artificial intelligence models, and many models can be designed and trained. With this digitalization, increasing electronic health records have affected the field of health, and large language models have begun to be developed in the field of health.

Over the years, this interest has led to important developments in the field of artificial intelligence and accelerated the development of ASR technology. Scientists have developed ASR technology by comprehensively analyzing speech sound data, and this field has become an important focus in modern research (D. Yu and L. Deng, 2016). While initially aimed at enabling machines to understand and process human speech, rapid advances in machine learning and computational linguistics have made ASR an indispensable tool in many industries. Traditional ASR systems basically have three main components: acoustic modeling, when extracting acoustic features from speech; language modeling, when selecting words considering context; and audio pre-processing, when optimizing the data input, such as normalization, noise processing, or segmentation.

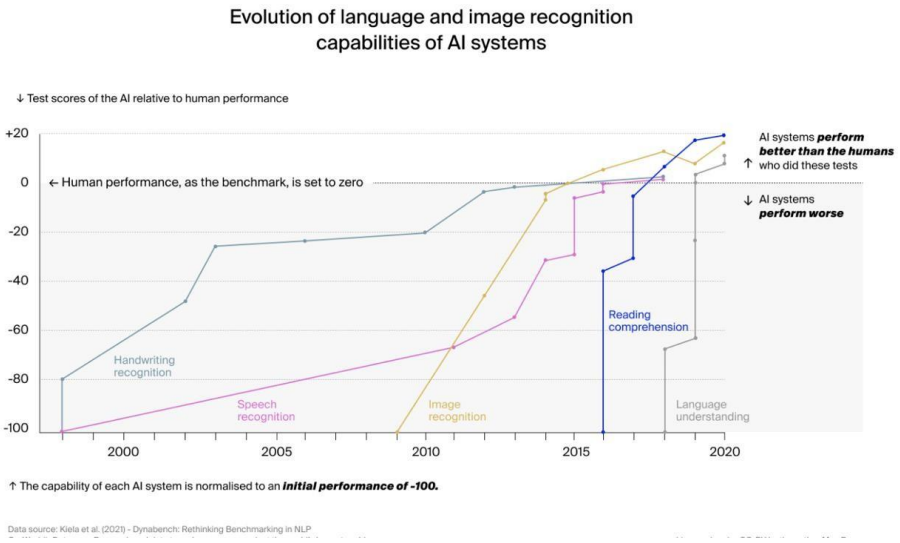

Figure 1. Evolution of language and image recognition capabilities of AI systems (2023)

One of the biggest transformations of speech-to-text technology has been the healthcare sector. With the increase in digitalisation in the healthcare sector and the widespread use of electronic health records (EHR), speech recognition systems have started to play a critical role in health informatics, electronic transcription, and automatic disease diagnosis solutions (HISA, 2014). From streamlining medical documentation to increasing efficiency in emergency departments, ASR has reshaped healthcare professionals' relationship with technology.

2.1.1. Speech-to-Text

Speech-to-text is the process of converting audio data into text transcripts. A computer program receives sound through sound wave vibrations and uses linguistic algorithms to convert the audio input into digital characters, words, and phrases. Over time, new architectures have been designed, and speech-to-text technology has made great progress. The evolution of speech-to-text technology has considerably transformed communication and accessibility. Its applications include language learning (Jeon et al., 2024) and improving user interaction in educational contexts.

2.1.2. Automatic Speech Recognition

On the other hand, ASR is the technology that converts human speech sound into readable text with artificial intelligence and machine learning algorithms. This multidisciplinary field between computer science and computational linguistics processes and transcribes the raw audio signal.

2.1.3 Natural Language Processing

NLP is a research field that investigates how computers understand human spoken language. It is an artificial intelligence technique that structures natural language text or speech in its own textual and syntactic sense.

2.1.4 Machine Learning Algorithms

Computer programmes that aim to perform tasks such as prediction and classification for similar situations that may be encountered in the future from the patterns obtained from the data. “A computer program can be said to learn if, in a context defined by a task T and a performance measure P, it can improve its performance through experience (Mitchell, 1997)

2.1.5 Deep Learning Algorithms

Deep learning is a machine learning approach that automatically learns complex relationships in data through multilayer neural networks. ASR technologies are used to develop acoustic modelling that converts complex features from audio signals into text. (Hinton, G., Osindero, S. & Teh, Y. W., 2006).

2.1.6 Convolutional Neural Networks

It is a deep learning architecture inspired by the working principle of the visual cortex in the human brain. It is especially useful for analysing image and sound data.

2.2 Historical Development of Speech-to-Text Systems

Artificial intelligence technology has been developing since 1950 and has shown exponential development, especially in the last 40 years. This exponential development, supported by the increase in clinical information in the field of health and the storage of patient data such as EHR records in digital environments, has been beneficial for the development of some technologies. The increase in the recording of data in digital environments in the field of healthcare has increased the power of data analysis, and the field has benefited from artificial intelligence in different aspects. For example, improved AI algorithms enable real-time data analysis, improving patient outcomes and paving the way for more personalised care. As a result, healthcare professionals achieve greater efficiency in clinical decision-making (RAJ & KOS, 2023). Speech systems that understand human language have been developed to provide better and more efficient service in converting voice data into text. The historical development of speech-to-text technologies from the 1950s to the 2010s is given below.

1950s - 1960s

Since the 1950s, various institutions and organisations have developed speech-to-text technology. The Audrey (Automatic Digit Recogniser) technology, which can identify spoken digits between 0 and 9 over a telephone line, was first developed by Bell Labs in 1952. Although Audrey had a narrow vocabulary, it was almost like a prototype and was a milestone for speech recognition systems. Later, in 1962, IBM's Shoebox technology was created, and human-machine interaction was taken to the next level. This system, like Audrey, could recognize numbers between 0-9, and in addition, it could perform seven different mathematical calculations, such as addition and subtraction by voice command. It made its voice heard in academic circles and shed light on people's ideas about how to talk to the machine.

1970s

Carnegie Mellon University later developed Harpy technology, an advanced speech recognition system. The system uses a search-based method to improve the efficiency and accuracy of speech analysis (De Mori, 2019). It was able to recognize a total of 1011 words. Harpy does not use fixed transition probabilities but dynamic transition probabilities based on acoustic data. This

allows the system to reduce the error rate and more accurate recognition (Lowerre, B. T. (1976). Another speech-to-text technology developed in the 1970s was DARPA (Defence Advanced Research Projects Agency), an agency of the United States Department of Defence with the mission of developing advanced technologies. In the 70s, DARPA invested in Dynamic Time Warping (DTW) technology, which extended the vocabulary to 1011 using context-dependent phoneme modeling. DTW technology tolerates data shifts and speed differences on the time axis and, in this way, aligns and compares sound waves when people say the same word at different speeds. It has been used to align sound waves in advanced speech recognition systems such as the Harpy system. Several pilot studies were carried out in the medical field during this period. For example, transcribing doctors' commands into text or querying lab test results with voice commands (Clark, R. E., & Manion, R. L., 1974).

1980s

In the 1980s, great progress was made with HHM and HMM (Hidden Markov Model), probabilistic models that are more complex but more accurate than simple systems for analysing and processing sequential data (Davis, S., & Mermelstein, P.,1980). The Hidden Markov Model (HMM), developed by Lawrence Rabiner and colleagues, has enabled the statistical modeling of speech signals (Rabiner & Juang, 1993). The HMM approach has been widely used in speech recognition, NLP, bioinformatics and time series analysis. HHM is one of the oldest basic statistical approaches in speech technologies. Hidden Markov models (HMM) and historical Hidden Markov models (HHM) have an important place in many applications, especially in speech recognition and bioinformatics. HMMS enables modelling of sequential time data, which is essential for understanding phonetic structures in language processing (Almutiri & Nadeem, 2022). During this period, speech-to-text technology began to be used in limited but innovative ways in the medical field. Despite limited medical knowledge, doctors were taking voice notes during patient examinations and transcribing them into text. Audio conversations about specific diagnoses or treatments were recorded and then transcribed. Although it was tried to be tested in the medical field, it was not successful due to the lack of medical dictionaries and the complexity of medical terminology. Also in the 1980s, IBM's Tangora technology, named after Albert Tangora (the world's fastest typist), had a vocabulary of 20,000 words. It was revolutionary for that period with its ability to identify continuous speech. Designed to speed up voice dictation

and document preparation in the workplace, Tangora was a pioneering attempt at automatic voice recognition, enabling organisations to translate spoken language into specific data. This development has been successful in many industries due to time savings and improved productivity (Muthusamy et al., 2020). It was able to predict the correct words using probabilistic models and HMM statistical models and is considered as one of the first steps of NLP technology.

1990s

The increase in the processing power of computers in the 90s led to the development of complex location recognition algorithms. The development of sound cards made it possible to interact with voice in home and workplace environments. Universities and organisations such as Carnegie Mellon University, MIT, IBM T. J. Watson Research Centre worked on statistical Markov models, which increased the accuracy of word recognition and the accuracy of systems. Pioneers such as Kurzweil AI and Dragon Systems developed dictionaries containing disease, drug, and prescription names specific to medical terminology.

2000s

The widespread use of electronic health records has led to the generation of much data in the field of health. Statistical models and machine learning based systems have been trained with large data sets and have secured their place in the field of speech recognition. Products such as Dragon Medical have developed customised dictionaries and databases specific to areas such as cardiology, orthopaedics, and neurology. Technologies such as MedLEE (Medical Language Extraction and Encoding System) and Dragon NaturallySpeaking enabled doctors, especially in the medical field, to keep a digital record of patients instantly. This enabled speech recognition systems to recognise clinical documentation more accurately such as drug names and procedures. Towards the end of the 2000s, with the development of cloud technologies and telephones, speech recognition systems moved from desktop and corporate environments to wider use.

2010s

In the 2010s, statistical models started to be replaced by machine learning and deep learning algorithms. Training multilayer artificial neural networks on large datasets has significantly improved and accelerated speech recognition accuracy (Graves et al., 2013). Instead of a statistical approach, an end-to-end approach began to be developed. This system allowed separate components to be combined into a single deep learning model. In addition, a cloud system was developed and big data analyses were performed. In this way, medical research was accelerated and patient records were automatically entered into the system. Speech recognition technologies such as Google Voice and Apple Siri were developed. In the last 10 years, deep learning algorithms have significantly increased the accuracy of speech recognition. Specialised speech recognition models have been developed in the field of healthcare, such as Google-Cloud Speech to Text and Amazon Transcribe Medical. The vocabulary of these applications is quite high and includes specialised medical terminology such as medical terms, procedure names and drug names. Google Speech-to-Text technology has adapted medical contexts and users can add their own medical terms. Amazon Transcribe Medical app supports Domain-Specific Vocabulary and can be customised with the same feature. The accuracy rate of both is approximately 90 per cent. They use deep learning models and powerful architectures such as Transformer, RNN and CTC to increase accuracy rates. Google Speech to Text uses end-to-end deep learning models. It uses RNN to process time-dependent audio data. CTC labels time series and Transformer-based models process more complex data sets. On the other hand, Amazon Transcribe Medical technology uses advanced acoustic models for more accurate audio data processing. Domain Adaptive Models are used to better understand healthcare-specific terms. The main goal is to improve patient care processes and the efficiency of healthcare professionals.

2.3 Medical Applications of ASR

Automatic speech technology has brought important applications to the healthcare field as well as other sectors by automatically converting voice data into text. There are medical applications in many areas, such as clinical documentation and patient interaction in medical services. It has also made a significant contribution to places where critical, real-time data needs to be obtained,

such as emergency care or intensive care units, by shortening processes and reducing error rates. (Smith and Brown, 2020)

2.3.1 Clinical Documentation and Patient Record

One of the most common uses of ASR technology is to reduce the time doctors spend creating clinical notes and patient records. While traditionally manually entered data leads to errors and time loss, providing documentation with voice commands increases accuracy and speeds up clinical workflows (Chen, 2019). ASR systems allow doctors to record patient information quickly, especially in busy emergency rooms and outpatient clinics.

One of the most time-consuming tasks in healthcare is the preparation of patient appointment notes, epacrises notes, and other clinical documentation. In traditional methods, these findings are transferred manually to the hospital information system. However, this process is quite open to errors and loss of time caused by the healthcare personnel who perform the transcription. With ASR technology, the data is converted to text with the doctor's instant dictation and can be automatically added to the electronic health record (EHR) (Kumar & Singh, 2020). Especially in polyclinics and hospitals, the desired data can be found immediately thanks to pre-trained voice recognition-based applications. Huston et al In a study conducted by Houston et al. (2019), it was emphasized that ASR technology provides faster and more accurate transcription and ease of use compared to manual data entry in medical documentation.

2.3.2 Radiology and Pathology Reporting

One of the places where the most data is produced in the health sector is the radiology department. Radiologists examine and report on images such as magnetic resonance (MR), computed tomography (CT), and ultrasound hundreds of times a day. With ASR, dictated reports are instantly transferred to the digital environment, thus eliminating transcription time. In a comprehensive study conducted by Paats and his colleagues, the ASR system, which has a 5 percent error margin performance, was developed because of the analysis of the last 5 years of reports in the radiology department (2018).

2.3.3 Telemedicine and Remote Health Services

Another ASR application that has gained significant value, particularly during the Covid-19 pandemic, is remote health. This area is tremendously valuable as ASR technology facilitates interaction between doctors and patients no matter the location, allows for automatic recording of these consultations, and integrates with translation technologies to overcome language barriers. It is also extremely beneficial for individuals with hearing impairments. When considered alongside handheld medical devices and remote consultation applications, ASR technology plays a strategic role in maintaining health service continuity, especially in contexts like rural areas, disadvantaged communities, and during epidemics (Adams et al., 2019).

Summary This paper examines the Terminological development of ASR technology from the 1950s, from hidden Markov Models to the modern deep learning process. It also discusses the current problems of ASR systems (e.g., error rates, rare medical terms, and noise problems in audio data). At this point, the methods and approaches to detecting and resolving errors are analysed.

3. LITERATURE REVIEW II: ASR-NLP Integration, Challenges, and Ethics

This chapter explores the architectures of modern NLP techniques and state-of-the-art transducer-based models and the methods and techniques by which they can improve the accuracy of ASR outputs. It will also explore the integration of Named Entity Recognition (NER), semantic analysis, and medical ontologies (UMLS, SNOMED CT) with ASR systems. Finally, it will analyse international policies and procedures on data privacy and ethical issues.

3.1 NLP and Its Role in ASR Enhancement

While ASR technology enables audio signals to be converted into text, NLP technology enables ASR text to be more easily understood through a series of semantic, syntactic, and pragmatic processes (J. Lee, 2024; Jurafsky & Martin, 2020). This is especially vital for areas such as clinical settings, where small errors can cause big problems. The integration of these two technologies has been shown to be effective in analysing large data sets and achieving high accuracy rates in speaker-dependent/independent environments (S. Patel, 2022). Furthermore, NLP contributes to reasoning in medical contexts, allowing patient stories to be analysed and used predictively over time. Zhou and Hripcsak (2007) emphasize the importance of temporal reasoning in developing predictive models for patient outcomes based on past clinical data.

Named Entity Recognition (NER) is one of the most important text-processing methods in the field of NLP. The primary goal of NER is to automatically detect proper names (entities) in a text and classify them into predefined categories. In medicine, it is used to identify diseases, drug names, and other proper names in patient records. It consists of 3 steps: detection, classification, and post-processing. For example, if we take the following sentence: 'The patient responded well to treatment in the cardiology department and is scheduled for discharge from 20231205. Please arrange follow-up care with family member Emily, contact number: 045-678 34 56.' Extractable Valuable Information: Health Department: Cardiology Department, Date: 20231205, Name: Emily, Phone Number: 045-678 34. 56. To extract this information, rule or dictionary-based search systems or machine learning models such as CRF (Conditional Random Fields) and SVM

(Support Vector Machines) are applied to classify entities, or deep learning algorithms such as LSTM and BERT are used, which provide high accuracy.

3.2 Specific Architectures of ASR in Healthcare

3.2.1 Transformer-Based Models

When the Transformer architecture first emerged, it introduced a framework that demonstrated advanced language processing capabilities using only attention mechanisms, encapsulated by the slogan 'Attention is all you need' (Vaswani, A., 2017). Transformers have become functional in almost every field, whether translating texts and speeches in real time or reducing the time spent by drug developers. The transformer architecture enables parallel processing using the Self-attention mechanism against the limited sequential structure of RNN and CNN structures. Self-attention is an important transformer mechanism that allows the model to focus on different parts of the input sequence while processing each element. The self-attention mechanism produces three versions of input embeddings: queries, keys, and values.

```
Queries: [Query for 'The', Query for 'cat', Query for 'sat',
          Query for 'on', Query for 'the', Query for 'mat', Query for '.']

Keys: [Key for 'The', Key for 'cat', Key for 'sat', Key for 'on',
       Key for 'the', Key for 'mat', Key for '.']

Values: [Value for 'The', Value for 'cat', Value for 'sat',
         Value for 'on', Value for 'the', Value for 'mat', Value for '.']

Random data:

Queries: [[0.23,0.4,.67,....],[0.4,0.6,.67,....],[0.2,0.2,.67,....],
          [0.5,0.3,.8,....], [0.1,0.4,.67,....], [0.2,0.4,.67,....],
          [0.7,0.4,.6,....]]

Keys: [[0.1,0.4,.5,....],[0.2,0.4,.67,....],[0.3,0.4,.67,....],
          [0.4,0.4,.67,....], [0.5,0.4,.67,....], [0.6,0.7,.8,....],
          [0.6,0.4,.8,....]]

Values: [[0.4,0.5,.67,....],[0.23,0.4,.5,....],[0.23,0.4,.8,....],
          [0.23,0.4,.45,....], [0.23,0.4,.9,....], [0.23,0.4,.6,....],
          [0.23,0.4,.10,....]]
```

Figure 2. Table of Queries, Keys and Values

Subsequently, attention scores are calculated. The dot product of a query with keys is used to calculate attention scores. These scores indicate how important or relevant each word is for the term being processed. These attention scores are then converted into probabilities by applying SoftMax. Each word in the sentence is given a value and the attention weights are summed to

one. Once this SoftMax is applied, a weighted sum of the values is determined using these scores. The resulting vector represents a context-sensitive representation of the current word that considers its relationships with other words in the sequence.

Transformer models consist of an architecture built around two components: the encoder and the decoder. The encoder processes the input data while the decoder generates the output data. The encoder's role is to analyse the input data (which can be a sentence) and transform it into a meaningful representation (hidden representation). The input words are transformed into numeric vectors in the embedding layer. This approach enables us to understand the semantic relationships between words, reduces computational costs, and minimizes acquisition efforts. The model creates a learnable embedding matrix for the words that gets updated during training. Location information is integrated through spatial coding to maintain sequential details. Spatial coding incorporates positional information into the word vectors using sine and cosine functions. When each word's position is included in the model's input, both the meaning and the position of the word are conveyed effectively by the model.

Since it is suitable for parallel processing rather than sequential learning, as seen in deep learning networks like RNNs, it integrates ranking information into the model instantly. This enables a better understanding of language by allowing the model to experience relationships between words (Hu, 2020). The computational cost is low because it is a fixed and unlearned structure. The encoder is composed of six identical layers, each containing two sub-layers. The first sub-layer is a multi-headed self-attention mechanism, and the second is a straightforward positional fully connected feed-forward network. Data passes through positional coding before entering the multi-headed self-attention mechanism, where the extent of each word's relationships with others is calculated. The multi-head self-attention mechanism is one of the most powerful features of transformers. Multiple attention heads are employed to understand word relationships within a sentence. Each head produces information outputs from different contexts, allowing the model to learn various relationships. The attention mechanism helps the model focus more on critical information in the input data, understanding which words are more important and allowing

system to assign weights in an optimized manner.

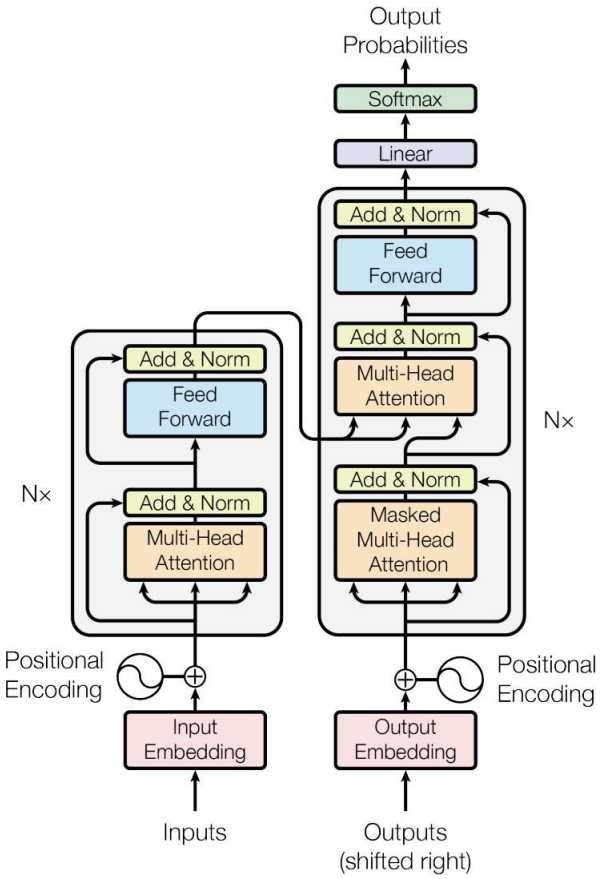

Figure 3. Self-Attention Mechanism

Self-Attention Mechanism

Self-attention mechanisms are crucial in transformer architecture because they are faster and more cost-effective than recurrent layers. Furthermore, they can connect all locations with a fixed number of operations, making them particularly beneficial for parallel information processing systems. Additionally, their ability to learn long-range dependencies is a significant advantage. A structure that can learn these dependencies by maintaining the distance between output location combinations is valuable. Self-attention mechanisms hold such importance in Transformer architecture because of their speed and cost-effectiveness compared to recurrent layers. Moreover, they can connect all locations with a fixed number of operations, positioning them as useful for parallel information processing systems. Their ability to learn long-range

dependencies is another key factor. The structure that learns long-range dependencies by maintaining the distance between output location combinations is advantageous useful.

Feed Forward Network

The Feed Forward Network (FFN) is a key component in each encoder and decoder layer. It processes the input more intricately and enhances the data processing capacity using linear transformations and activation functions. The layer consists of two fully connected layers, which means that each word of the sequence is processed independently.

$$\mathrm{FFN}(x) = \mathrm{ReLU}(xW_1 + b_1)W_2 + b_2$$

.

Figure 4. FFN Formula

- Where x: Input vector from the encoder or decoder layer
- W1 and W2 learnable weight matrices
- b1 and b2 learnable bias terms
- ReLU: Activation function (Rectified Linear Unit)
-

The FFN processes each token in the input sequence independently, complementing the self-attention mechanism that captures the relationships between tokens.

Step-by-Step Process:

1) Dense Layer

   h=xW1+b1. The size is expanded (e.g., 512 -->2048)

2) Activation Function (ReLU)

   ReLU adds nonlinearity by setting negative values to zero and leaving positive values unchanged.

3) Second Linear Layer

   In this step, the size returns to its original size (2048 → 512).

The FFN component contributes significantly to the overall performance and flexibility of the Transformer architecture. Transformers without FFN cannot learn anything. For example, FFNs facilitate parallel processing of input data, which is accelerated by these training and inference times (Yang and Su, 2022). Within the transformer architecture, the FFN and the self-attention mechanism complement each other. While the self-attention mechanism gathers information across tokens, the FFN independently enriches the representation of each token. This interaction successfully enables both contextual understanding and feature enrichment.

Embedding and SoftMax

The other two primary components of the transformer mechanism are embedding and SoftMax. Embedding allows data such as words and tokens to be represented as dense, low-dimensional vectors. A matrix with a vocabulary size is created (for example, 10,000-word variants and an embedding matrix of size 512). The vector representation of each word is obtained from the corresponding row of this matrix. The embedding learns to capture semantic relationships between words. Words with similar meanings (such as ‘cat’ and ‘dog’) are positioned close together in vector space. In the embedding process, spatial coding is applied instead of sequence information. The positional coding represents the positional information of the words in the sequence and is added to the embedding vectors. The SoftMax function transforms the score vector into a probability vector, ensuring that the sum of these probabilities equals 1. This step is known as normalization, allowing for a more meaningful analysis of the model. This enhances the significance of the model output. It is primarily used in the self-attention layer and the final layer. It determines the relationship between words and the attention weight assigned to each word. In the last layer, it is used to compute the probability of the target word. These two mechanisms work together to optimize the model’s input and output transducers.

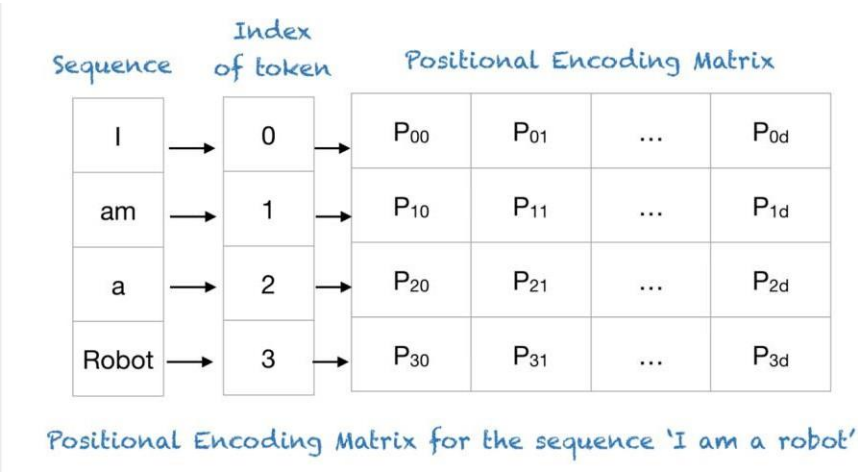

Figure 5. Positional Encoding Matrix for the Sequence 'I am a Robot'

Transformers provide effective solutions for sequential data without depending on sequential models like RNNs or Long Short-Term Memory. However, they lack positional information when processing the tokens in the input array. Positional encoding is a technique developed to address this limitation. The self-attention mechanism in transformers allows us to understand the relationships between words. Nonetheless, they cannot process any positional information during this operation. For example, the difference between the sentences "The cat chased the bird" and "The bird chased the cat" lies in the sequencing, even though the exact words are used. The goal is to make the model sensitive to sequencing. Positional encoding generates an encoding vector for each position, typically implemented using sine and cosine functions. The formula applied is as follows. It relates the position of the word directly to the sine and cosine functions so that the model can learn the ordering information.

$$P(k, 2i) = \sin\left(\frac{k}{10000^{\frac{2i}{d}}}\right)$$

$$P(k, 2i+1) = \cos\left(\frac{k}{10000^{\frac{2i}{d}}}\right)$$

Symbols	Description
k	Represents the position of an object in the input sequence. It takes a value in the range $0 \leq k < L/2$
d	Refers to the dimension of the output embedding space.
P(k,j)	A positional function used to map a position k in the input sequence to the index (k,j)(k, j)(k,j) in the positional matrix.
n	A user-defined scalar
i	Used to map column indices in the range $0 \leq i < d/2$, where a single i value maps to both sine and cosine functions.

Transformer-based models provide high performance by calculating the parallel connection between words thanks to the self-mechanism in their structure. However, they may be insufficient in studies with complicated medical terminology and rare words. In such special cases, models specially trained with medical literature, such as BioBERT or ClinicalBERT, should be used. ASR systems will perform well since such models are equipped with medical terminology. Another method is to train general transformer models on datasets with transfer learning and fine-tuning methods, which increases the model's compliance with specific terminology.

3.2.2 Recent Advancements of Speech-to-Text

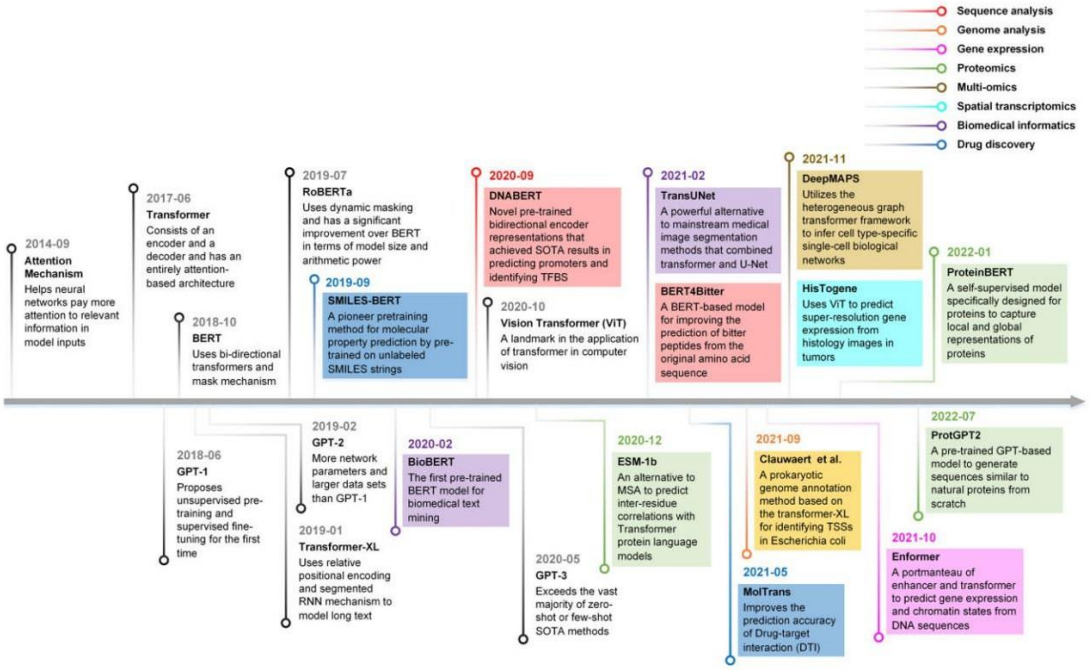

Figure 6. Timeline of Transformer and AI Advancements in Biomedical Research (2014–2023)

The introduction of BERT (Bidirectional Encoder Representations from Transformers) has led to significant advancements in the field of NLP. This model has transformed development through the Transformer architecture, which differs from traditional methods (Min et al., 2023). BERT's applications encompass sentiment analysis, text classification, and language translation, providing a comprehensive analysis. BERT models emphasize the relationship of a word with other related words in a sentence rather than interpreting the word in isolation. Notable improvements in NER performance have been achieved, particularly with deep learning and transducer-based models. The latest NER systems include BERT, optimized BERT, RoBERTa, SpaCy, and Flair. Gozalo-Brituela and Gonzalez-Carvajal (2023) conducted an extensive comparison of BERT with traditional text classification approaches. They found that BERT outperforms traditional models in precision and efficiency. BERT, a bidirectional architecture developed by Google, has two main attributes. The first is Contextual Understanding. For instance, while older models only recognize the word ‘apple’ as a fruit name, BERT can comprehend it in various contexts; it can also identify it as the name of a technology company

depending on its meaning within a sentence. The second attribute is Bidirectional Analysis, allowing text to be analysed from right to left and left to right

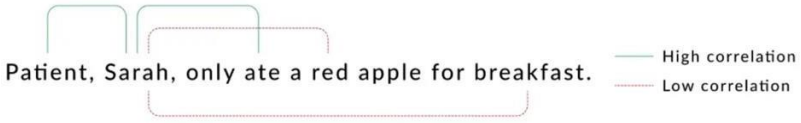

Figure 7. Visualization of Word Correlations in the Sentence

Pre-trained BERT models train to understand medical terminology. It can be trained on medical journals, patient records, or other health-related texts. BioBERT is fine-tuned to biomedical literature, and ClinicalBERT is designed to interpret clinical notes.

Several ontology-based NLP applications have been developed to provide consistent and familiar medical documentation. Structured biomedical ontologies such as UMLS (Unified Medical Language System) and SNOMED CT (Systematised Nomenclature of Medicine-Clinical Terms) improve knowledge extraction and semantic understanding from clinical data, especially in the health and biomedical fields. These ontological structures organize concepts such as diseases, drugs, treatment methods, and the relationships between them. UMLS is a system developed by the American National Library. It ensures the correct use of medical terms in the analysis of clinical texts in NLP applications. The components of UMLS are the Meta thesaurus, Semantic Network, and SPECIALIST Lexicon. Meta thesaurus is a comprehensive database that combines many word and code systems. The Semantic Network categorizes medical words and helps to discover semantic links between these categories. SPECIALIST Lexicon is a comprehensive syntactic dictionary for biomedical and general English. SNOMED CT is the most widely used clinical terminology worldwide. Its structure consists of concepts, definitions, and relationships. It makes Electronic Health Records (EHR) more understandable and improves treatment processes in clinical decision support systems.

3.2.3 ASR Error Detection and Correction Strategies

While ASR technology transforms audio signals into text, NLP technology enhances the ASR text's readability through various semantic, syntactic, and pragmatic processes (J. Lee, 2024; Jurafsky & Martin, 2020). This enhancement is especially crucial in fields like clinical settings, where minor errors can lead to significant issues. The integration of these two technologies has proven effective in analyzing large data sets and achieving high accuracy rates in both speaker-dependent and speaker-independent environments (S. Patel, 2022).

ASR technology undertakes the task of converting audio data into text by analysing different stages such as pre-processing, feature extraction, acoustic and language modelling. Traditional ASR systems have the following components: pre-processing, feature extraction, acoustic model, language model, and dictionary. McCowen et al. defined the ideal performance measures of ASR as direct, objective, interpretable and modular (McCowan, 2024). The Word Error Rate (WER) metric measures the model's performance. It measures the percentage of incorrect words relative to the total number of words and uses the following formula.

$$WER = \frac{S + D + I}{N1} = \frac{S + D + I}{H + S + D}$$

Figure 8. Formula of Word Error Rate

I: Total insertions
D: Total deletions
S: Total substitutions
H: Correctly detected words
N1: Total number of input words

However, it should be noted that the WER formula does not indicate the accuracy of the system's performance but the extent to which it is better than the other system being compared. ASR error detection aims to increase accuracy, improve system performance, and strengthen language and acoustic models. It is a necessary and critical step in increasing the reliability and usability of the system. Default words are divided into two groups: correct words and errors. There are two approaches to ASR error detection:

1) Approaches that use features (e.g., grammar) generated from the ASR Decoder

2) Approaches using additional features (e.g., n-grams, syntactic and semantic). These two approaches use different methodologies and techniques to detect and correct errors.

In the following, both different approaches are analysed in detail.

1.a) Decoder-Based Fault Detection

Decoder-based error detection systems correct errors by analysing the output of ASR systems. These methods include language models, acoustic models, and output probabilities. Decoder-based techniques provide access to the system's internal structure and outputs. Markov Chains and DSR (Dynamic Speaker Adaptation) are essential among these techniques.

1.b) Markov Chains and Decoder Based Error Detection

Hidden Markov Models are stochastic models that assume that an event depends only on the preceding event, and these Markov chains are represented by states and transition probabilities**.** If the difference between the expected transition probabilities and the actual values is too large, the system suspects that there is an erroneous word.

1.c) DSR (Dynamic Speaker Adaptation) and Decoder Based Error Detection

DSR (Dynamic Speaker Adaptation) recognizes speaker-specific voice features by considering factors such as accent, pitch, speech rate, and ambient noise and is effective in improving the accuracy of ASR systems.

The second technique used for ASR error detection is the non-decoding approach. These methods usually use grammar rules, statistical models, or machine learning techniques. They are more general-purpose and do not access the system's internal structure. Non-decoding methods help to correct sentences that contain incorrect or missing grammar rules. They are particularly effective with structured texts. They can analyse the output of any ASR system. However, they produce less accurate results since they cannot access the system's internal structure. Below are three examples of non-decoder-assisted methods.

2.a) Confidence Scoring:

ASR systems generate a confidence score for each word and fact. With threshold-based filtering, scores below a certain threshold can be marked as errors. This score distribution is statistically observable in the error estimation.

2.b) Acoustic Only Approaches:

Words or phenomena with low scores by the acoustic model are marked as having a high probability of error. In this method, only acoustic scores are used for error detection.

2.c) Embedding Models without Language Models:

The audio signal is transformed into vector space with pre-trained audio embedding models such as wav2vec2. The embeddings in the ASR output are compared with the pre-computed embeddings and if any vector is far from a certain threshold value, it is treated as an error. This error detection method is effective in detecting word and phenomenon substitution errors, noise-induced distortions, and errors in texts with missing or redundant words.

Decoder-based methods provide more accurate and integrated solutions by accessing the system's internal structure. However, they tend to be complex and expensive. Conversely, non-decoder-based methods are general-purpose and less expensive, but their accuracy is limited. Academic studies indicate that combining these two approaches can yield the best outcomes. While decoder-assisted methods can identify errors early on, non-decoder-assisted methods

facilitate the correction of those identified errors. This collaboration can enhance the performance of ASR systems.

The incorrect recognition of rare medical terms by ASR systems leads to errors. In such instances, NLP steps in to conduct advanced semantic and syntactic analyses. During the initial stage, Named Entity Recognition techniques are employed to identify and classify the terms in the texts. The effectiveness of the correction mechanisms largely relies on the quality and breadth of the datasets used to train the model. When trained on extensive datasets, the model can analyse both common and rare terms with high accuracy. At this stage, the dataset must be continually updated and supported by feedback from field experts. In addition to error detection, the noisy incoming audio data reduces the performance of ASR systems. Especially in hospital environments, the presence of numerous people, device sounds, and other noises negatively affects the audio data. At the same time, different accents, speech rates, and voice tones cause deviations from the model's training data, making it difficult to accurately recognize rare terms. To overcome these challenges, Wiener filtering and deep learning noise suppression methods have been applied. Research by Wang and Chen has shown that Wiener filtering and Mel spectrogram analysis provide significant improvements in high-noise audio recordings (Wang & Chen, 2018), and in recent years, deep learning approaches have reduced error rates by processing complex acoustic data more efficiently (Hinton G 2012).

In the pre-processing process, background noise can be minimized, and the speaker's voice can be clearly heard. To overcome speaker differences, a dynamic speaker adaptation technique can be applied. The system can be trained according to the characteristics of different speakers, such as accent, speed, stress, and tone. Although background noise or speaker differences make rare medical transcription difficult, these problems can be significantly overcome with these techniques. Continuous testing and improvement with real-world data are critical for the system to cope with these difficulties.

3.3 Ethics and Data Privacy

3.3.1 Ethical Challenges

Although artificial intelligence technologies in the field of medicine facilitate patient care and increase workflow efficiency, the use of these technologies also raises serious questions about data privacy and ethics. Since patient-doctor conversations and clinical reports used by technologies contain sensitive information about patients, they carry the risk of privacy breach and data misuse. As Siti Hawa emphasized in her 2024 study, patient information used in the field of medicine should be protected with sensitive encryption methods and differentiated privacy techniques. (Awa S. 2024) This health data stored in the cloud environments can become the target of cyber attackers. Therefore, it is necessary to implement role-based access controls (RBAC) that only authorised persons can access.

3.3.2 Informed Consent and Transparency

The ethical dimension of using ASR technologies in medical applications is important not only in data privacy. In addition, it is necessary to focus on issues such as patient rights, informed consent, and algorithmic justice and to develop specific procedures and protocols. It is essential to inform patients about how this technology will use patient data in the future and to obtain their consent. The principle of transparency should be adopted in collecting and analysing patients' data (Shoghli, A, 2024).

3.3.3 Data Minimization and Fitness for Purpose

Data minization principle stipulates that only the data required for a specific purpose should be collected (European Parliament (2016). Technologies are better trained with large amounts of data and accelerate the innovation process. However, data taken without a consent process may increase the risk of misuse (Lee et al., 2020)

3.3.4 Legal Regulations

In the United States, HIPAA (Health Insurance Portability and Accountability Act) sets standards for protecting sensitive patient health information. This regulation mandates measures

such as confidentiality and accessibility of electronically protected health data (U.S. Department of Health & Human Services, 1996). The same missions are the General Data Protection Regulation (GDPR) in the European Union. Voice data used in the field of health is considered as personal data and is very strict on issues such as data processing and storage.

3.3.5 Overcoming Challenges

Since health data has a valuable cost in the black market, it is targeted by cyber criminals, even though legal regulations are being made (Ferry, Q. et al. (2019). Therefore, the design of speech-to-text technologies should be integrated with privacy and ethical principles from the very beginning. Privacy-oriented design approach prevents data from going to cloud systems and data is not transferred to third parties. An ethical approach, on the other hand, prioritises fairness, transparency, and accountability over compliance (Floridi & Taddeo, 2016). In this regard, technology developers should transparently explain how the patient's data is processed.

4. METHODOLOGY

This section explains the methodology used to collect primary data. The study aims to explore patients' diseases by transcribing their voice recordings. It adopted a multi-stage and experimental research design.

4.1 Research Design

This study aims to determine the challenges and opportunities offered by existing methods in the automatic transcription and analyzing medical voice recordings by integrating speech recognition and NLP techniques in the medical field. In this way, it investigates how to improve the efficiency of clinical decision support systems by using artificial intelligence models is investigated. Considering the study's exploratory nature and experimental approach, a multi-stage and experimental research design was adopted.

4.2 Research Method

This research addresses the process of analysing texts obtained from medical audio data using NLP techniques. The research followed an experimental methodology, and the characteristics of the audio recordings (e.g., recording duration, speaker characteristics, ambient noise) were brought into compliance with certain criteria. At this stage, techniques such as cleaning, normalizing, and noise reduction of the raw audio data were applied. In addition, data augmentation techniques (e.g., synonym replacement) were used to ensure that the model could perform better, allowing it to generalize better.

Using the pre-trained Whisper model, the ASR model was fine-tuned to understand clinical terminology. The Whisper architecture is particularly notable for being robust to a wide variety of language and accent patterns (Radford et al., 2022). The performance of the ASR model was measured with metrics such as word error rate and sentence error rate during the transcription process.

Transformer-based models trained with medical terminology, such as BioBERT and ClinicalBERT, were employed during the integration of the ASR model with NLP. Utilizing pre-

trained models like BioBERT and ClinicalBERT enhances the model's understanding of medical terminology, which is crucial for grasping complex medical concepts (Devlin et al., 2019; Lee et al., 2020). Therefore, Named Entity Recognition, relationship extraction, and text classification processes were performed.

The research methodology, along with its experimental design, has improved the efficiency of processing large data sets with parallel and batch processing techniques. The experimental setup was carried out using GPU-supported high-performance computing units and multi-core CPUs, thus enabling real-time ASR and NLP operations.

This research has experimentally demonstrated the automatic transcription and interpretation of patients' voice recordings through the integration of ASR and NLP. During this process, the system's performance and applicability to real cases were considered. Finally, the research provides a solid foundation for future studies on model development, error analysis, and data privacy and ethics in clinical settings.

4.3 Sampling Design

The sampling design in this study is based on a large and diverse data set created with voice recordings of real patients. Using a random sampling method, 6,500 voice recordings were selected, and stratification techniques were applied. The data were collected from the hospital environment, but each audio recording has different background noise levels based on the varying devices used and ambient sounds. In the dataset, the patients' IDs were used instead of their names to protect data confidentiality. Additionally, the CSV file of the voice recordings contains two columns labeled 'phrase' and 'prompt.' These columns indicate what the patient said and their specific complaint.

The general purpose of the sample design is to process the ASR-NLP integration on a large data set and ensure the generalizability of the results obtained. The aim of the sample design is to ensure that the voice data are both high-quality and noisy and to measure the system's performance under different accents, speech rates, and background noise conditions. Each voice recording was processed with a standard 30-second time limit, thus normalizing all data within a certain period.

As a result, the sample design increases the representativeness of the data obtained from the large voice recording data set, while ensuring the integration of real-world scenarios encountered in

clinical settings into the system. This approach increases the reliability of the results obtained in the clinical medical field and makes the model's performance reliable. Thus, the research findings are a reference in both academic and applied fields.

4.4 Data Collection

In this study, voice recording data collected from different clinics and totaling over 8 hours were used. These files, in which each patient describes their own disease, are in .wav format and each file is limited to 30 seconds. This time limitation will ensure that the data is analysed in a certain standard and works in harmony with the algorithms to be applied later. During the data collection, metadata information such as recording time and background noise level was also obtained in addition to clinical information about each patient. With this metadata, elements such as segmentation and quality control of the data are useful in terms of analysing the performance of the system. In addition, during the data analysis, it was aimed for the system to cope with accent analysis and voice quality variations due to the diversity of environmental conditions within the dataset and factors such as different recording devices. As a result, the data collection phase is one of the cornerstones of the research. The richness and diversity of the dataset provide a basic basis for the training of ASR systems and fine-tuning of NLP models. The collected data reflects real-world scenarios.

4.5 Execution of the Research

The research was conducted according to the established methodological frameworks and carried out in multiple stages. In the first stage, the collected information was organized and categorized into different subfolders labeled train, test, and validate. During this phase, the verification of metadata and the storage of each file in the standard .wav format were ensured. The data were transferred to a secure repository along with the information gathered during the collection process, and anonymization was performed. In the second stage, pre-processing was carried out on the raw audio data. Audio recordings from various sources were converted to a single format with a sampling rate of 16 kHz. This step ensured consistency across the information obtained from different devices. Each audio recording was then limited to 30 seconds while maintaining time harmony throughout the processing. In this stage, the Wiener filtering method was applied to reduce noise in the audio recordings. This method clarifies the system by removing unwanted

noise, thereby increasing the transcription accuracy. Following that, the audio signal was normalized to ensure all recordings had similar dynamics. Mel Spectrograms were extracted to enhance the model's understanding of the audio data's acoustic properties, specifically aiming for better comprehension of rare medical terms. The long audio recordings were segmented into sentence and word-based units during the segmentation process, allowing the model to analyse the audio more effectively. Another technique employed in data pre-processing was data augmentation. Methods such as pitch shifting (both raising and lowering the tone) were applied to the audio data, enhancing the model's ability to handle different tones and accents. Consequently, the aim in the data pre-processing phase was to create cleaner, more stable raw data suitable for model training. This section directly impacts the performance of the ASR-NLP integrated model, laying the foundation for the data to be utilized in the subsequent research steps and maximizing the overall accuracy of the project.

4.6 Model Adaptation and Training

The model adaptation process applies various techniques to make the model work better adapted to medical terminology. Firstly, data augmentation techniques were applied, and the model was provided to generalize between different variations and expressions. Thus, it was provided to give better performance against new and unexpected data. Synonym Replacement techniques were applied for text data augmentation, and Shift Pitching techniques were applied for voice data augmentation against various accents and intonations. Then, the Weiner filtering method was used to prevent background noise. This filtering method helps to clean the voice signal and provides more accurate transcription. In addition, normalization and feature extraction were performed with the Librosa technique, and the model was provided to learn more effectively. Considering the size of the dataset, it is important to reduce the processing time. Therefore, Parallel Processing techniques (ProcessPoolExecutor and Multiprocessing) were applied. Transfer learning and fine-tuning techniques are intensively applied in this phase. In this project phase, hyperparameter settings such as learning rate or chunk size were optimized, and low rates were targeted in the word error rate and sentence error rate metrics. In this way, the ASR model aims to reduce the error rate of the transcript output from Whisper. In addition, pre-trained models in the field of medicine named BioBERT and ClinicalBERT were used in the integration phase with NLP. Since these models are trained by medical terminology, drug names, and other

clinical terms, they increase the semantic and syntactic accuracy of the texts obtained after transcription. The Whisper model is a modern language model capable of analysing a wide variety of languages and acoustic features well. Its good performance, even in noisy clinical environments, is the main reason for its use in this project.

On the other hand, weighted cross-entry and entropy-loss functions were used to overcome possible imbalance problems in the dataset. Afterward, validations were measured for continuous improvement, and the results were optimized by making hyperparameter adjustments. As a result, the model adaptation and training phase is a critical step that ensures the project's success and allows ASR and NLP systems to demonstrate high performance on medical data.

4.7 Evaluation Metrics

This section of the research includes metrics that measure the performance of the ASR-NLP system. These metrics provide insights into the system's accuracy, error rate, and application in clinical settings. The Word Error Rate (WER) and Sentence Error Rate (SER) metrics were employed to assess ASR system performance, while precision, recall, and F1 scores were utilized to evaluate NLP system performance. ASR performance metrics indicate the percentage of correctly transcribed data from the audio recordings, both on a sentence and word basis. NLP performance metrics assess the sensitivity and comprehensiveness of the model, demonstrating the percentage of entities the model correctly predicted as well as those it missed or misclassified. This allows for a detailed evaluation of the NLP process's performance on clinical notes. Additionally, a clinical relevance metric was used to gauge the system's utility in clinical applications. Consequently, these metrics enabled a comprehensive assessment of the system's performance and potential for usage.

4.8 Ethical Issues and Data Privacy

In this project, all processes were designed and implemented in accordance with international data protection standards such as the General Data Protection Regulation (GDPR) of the European Union and the Health Insurance Portability and Accountability Act (HIPAA) of the United States. In accordance with the data minimization principle of the GDPR (Article 5(1)(c)), characteristics of patients such as name, age, and gender were not used in the analysis phase. IDs were assigned to distinguish patients from each other. This practice is also compatible with

HIPAA's de-identification standards (45 CFR §164.514). The data retention policy was defined at the beginning of the project in accordance with the GDPR guidelines (Article 5(1)(e)) and the privacy requirements of HIPAA. The data were used only in the research and were subsequently destroyed according to data destruction protocols. As a result, the research adopted an approach that complies with sensitive patient confidentiality and ethical protocols.

Summary

As a result, the methodology used in the research conducts a thorough examination of the challenges and opportunities within the field of speech-to-text technology. It illustrates the processes involved in analysing voice recording data in a manner that meets specific standards. Additionally, it highlights that the voice signal can be understood more clearly thanks to normalization, segmentation, and noise cancellation techniques applied during the data pre-processing phase. This section of the research also addresses practical strategies to overcome challenges such as the complexities of rare medical terms and background noise. In doing so, it underscores how artificial intelligence models interact with real-world scenarios, the methods they analyze, and their appropriateness for clinical decision support systems. The integration of artificial intelligence and clinical decision support systems is becoming increasingly crucial in the healthcare ecosystem (Jiang et al., 2017). Furthermore, it illuminates the importance of collecting real data that not only assesses current system performance but also identifies areas for future improvements. This comprehensive approach offers a solid foundation for understanding how advanced ASR and NLP technologies can be further refined to enhance clinical decision support systems and ultimately improve patient care.

CHAPTER FIVE – FINDINGS / ANALYSIS / DISCUSSION

This section aims to present the dataset formed by the collected audio recordings through visualizations. The discussion section will provide a critical analysis and combine the results with the literature review. It will discuss to what extent the ASR-NLP system works correctly and whether it can be used in a clinical case study.

5.1 Findings

This study's sample consisted of voice recordings of 6661 patients over a total of 6 hours. Audio clipping, background noise and their confidence levels, overall quality of audio, quiet speaker confidence level, speaker ID, file path, file name, phrase, prompt, writer ID, and 13 features were analysed.

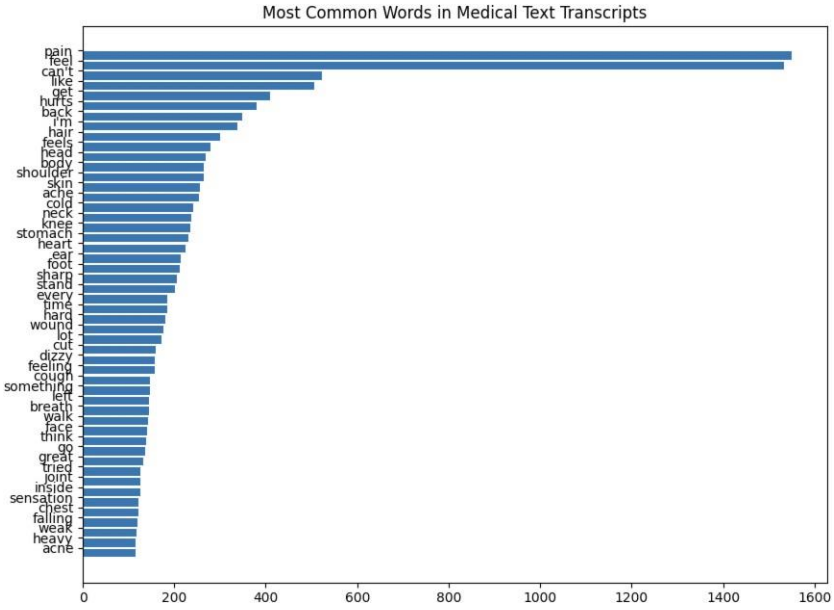

Figure 9. Most Common Words in Medical Transcripts

In the dataset of 6661 patients, the visualisation of the most frequently found words with a bar graph is shown. The most common words are ‘pain’ and ‘feel’, each of which is used approximately 1600 times.

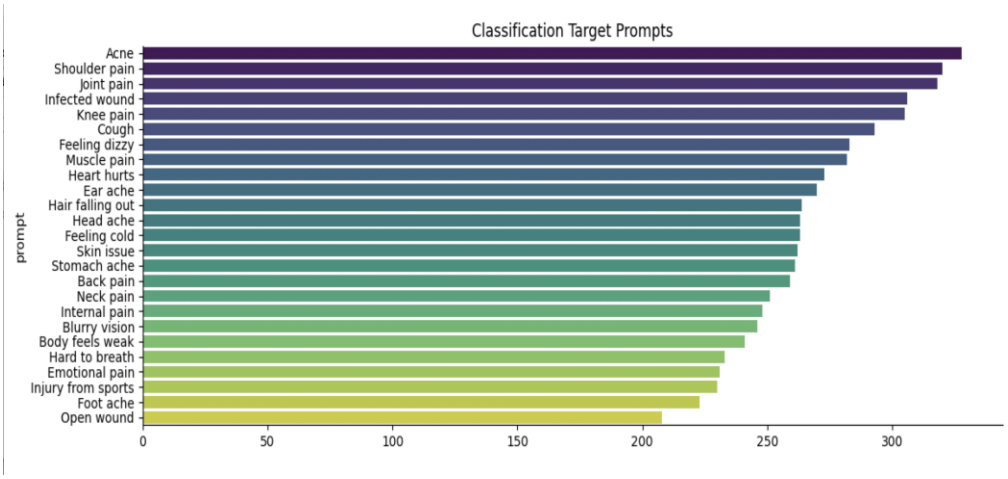

Figure 10. Distribution of Classification Target Expressions

This bar chart shows the frequency of classification target expressions in the dataset. The top ranking is ‘acne’ followed by dermatological and orthopedic problems such as ‘shoulder pain’ and ‘joint pain.’ At the same time, the graph is useful for reading the medical fields that the project is focused on. It is a guide to which complaints are most common when developing a classification model. This distribution shows that some types of complaints are more frequent and important. For example, the fact that word groups such as ‘acne,’ ‘shoulder pain,’ ‘joint pain,’ ‘feeling dizzy,’ and ‘muscle pain’ are at the upper levels of dermatological, orthopedic, and musculoskeletal levels gives findings on where the research can be focused.

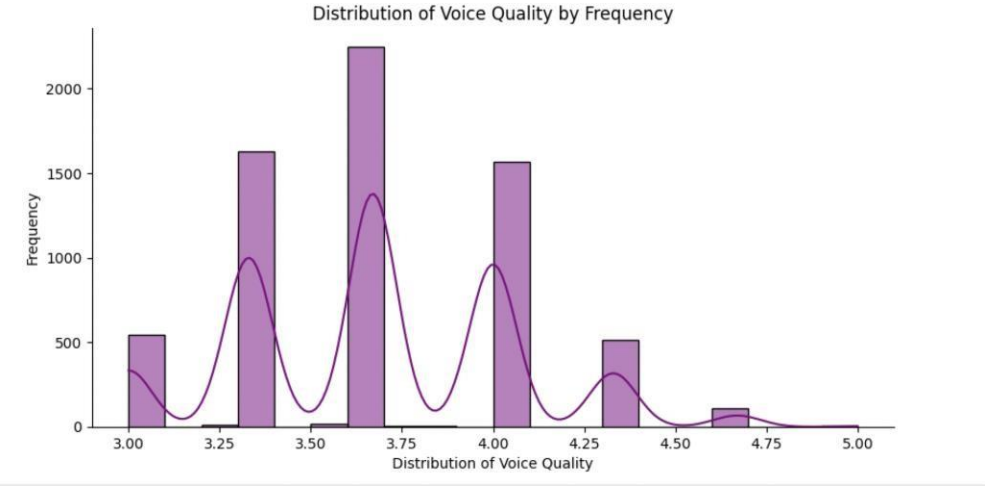

Figure 11. Frequency Graph of Sound Quality Distribution

This multi-peaked distribution indicates that the voice recordings in the dataset feature various device types and speech styles. The kernel density curve overlaying this histogram illustrates how the sound quality of the audio recordings is distributed between 3.0 and 4.75. It is noted that there are multiple modes, particularly at 3.25, 3.50, 3.75, and 4.0, visible in the plot. The peaks of low-quality sound recordings are concentrated within the 3.32-3.50 range, while recordings made with better equipment or in quieter environments are found in the 3.75-4 range. Additionally, there is a variety of sounds.

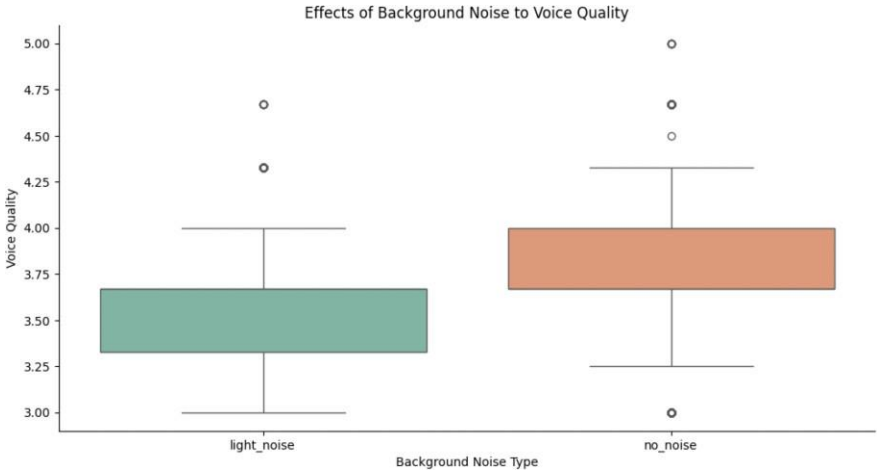

Figure 12. Effect of Background Noise on Sound Quality

This box plot shows the distribution of two different categories of background noise (‘light noise’ and ‘no noise’) on the sound quality. Comparing the medians of the two categories, it can be observed that a higher number of audio recordings are ‘no noise’. The outlier distribution of the two categories is also shown. The graph shows that background noise, even if slight, has a negative effect on sound quality. The ‘no noise’ category has a higher median and more outliers, indicating that sound recordings from quiet environments are of better quality.

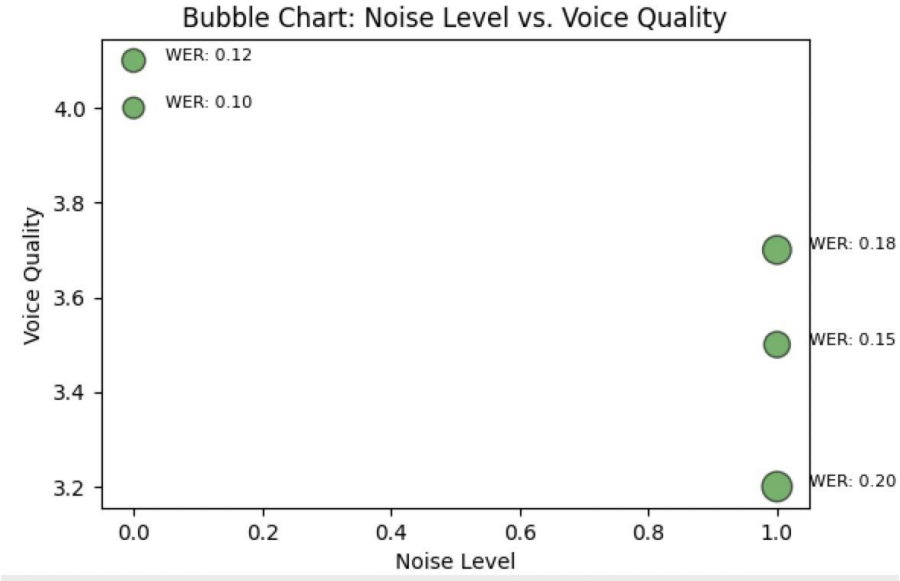

Figure 13. The Relationship Between Noise Level and Sound Quality

This bubble chart displays Noise *Level* on the x-axis and *Voice Quality* on the y-axis. The size of the bubbles represents the Word Error Rate. After evaluating the performance of the ASP-NLP system, this graph indicates that the voice quality of recordings with a zero-noise level is 4 or higher. At these levels, it is observed that the Word Error Rate is relatively low, ranging from 0.1 to 0.12. Conversely, the Word Error Rate for audio recordings with very high noise levels was measured between 0.18 and 0.2. Furthermore, as the noise level increases, the average sound quality decreases to a range of 3.2 to 3.8. In short, the graph shows that as the noise level increases, the sound quality decreases and the word error rate increases in parallel.

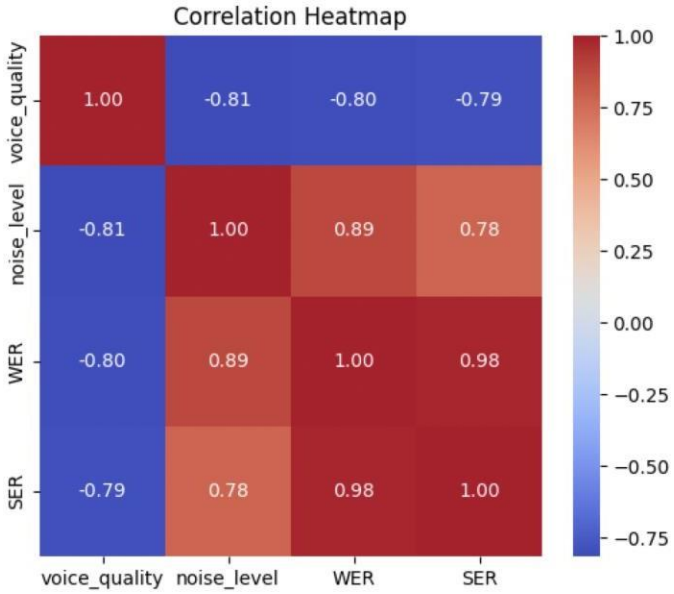

Figure 14. Correlation Heat Map Between Variables

This heatmap shows the correlations between *voice quality* (voice quality), *noise level* (noise level), *Word Error Rate* (WER), and *Sentence Error Rate* (SER) variables. There is a high negative correlation of -0.81 between voice quality and noise level. Similarly, there is a significant negative correlation between voice quality and *WER* (-0.80) and *SER* (-0.79). On the other hand, a positive and strong correlation of 0.89 is observed between noise level and WER. Finally, a high positive correlation of 0.98 is observed between WER and Sentence Error Rate.

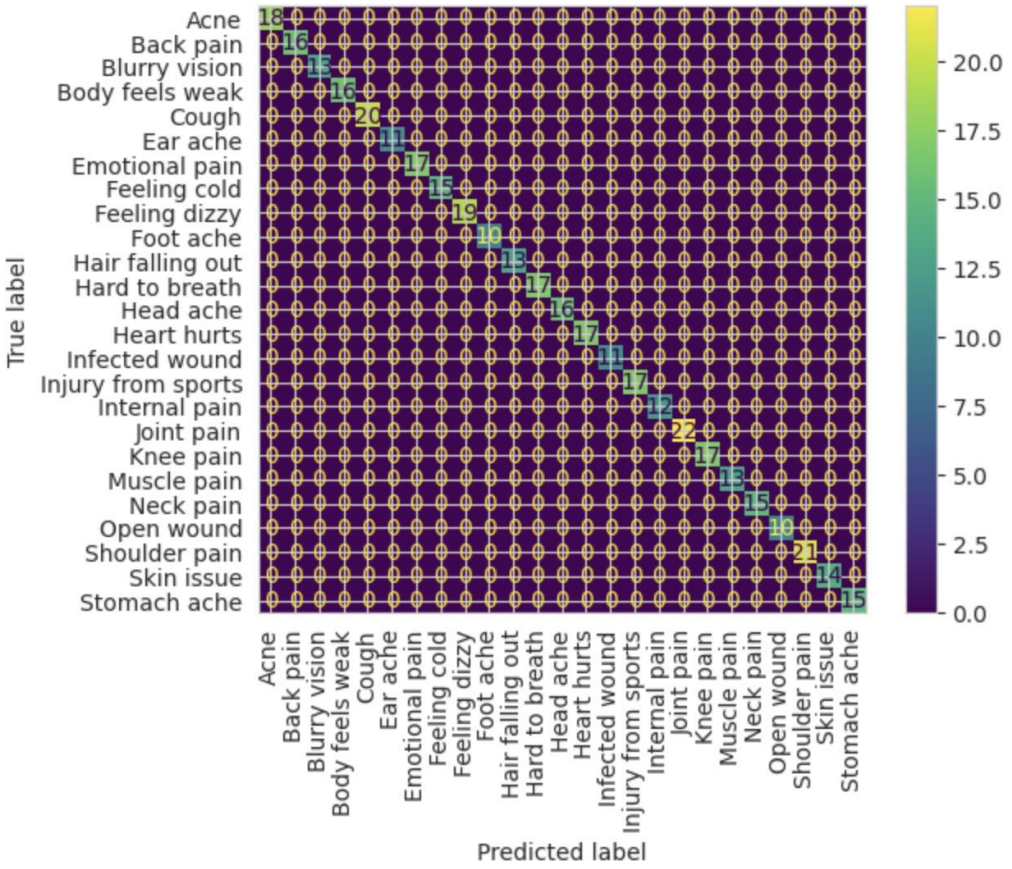

Figure 15. Confusion Matrix for Medical Complaint Classes

The matrix table is the confusion matrix showing the model's performance across different classes of medical complaints. The Y-axis represents the actual labels, and the X-axis represents the labels predicted by the model. The bar chart of the colors of the intersection of the predicted labels with the true values is on the right-hand side. These colors indicate how accurately the system knows the actual labels. There are significantly more brightly colored values along the diagonal, indicating that the model recognizes each disease class to a large extent. For example, for classes such as ‘Acne,’ ‘Back pain,’ ‘Earache,’ ‘Shoulder pain,’ ‘Knee pain,’ ‘Open wound,’ and ‘Muscle pain,’ the predictions mostly matched their class. This shows that the model is successful in classification.

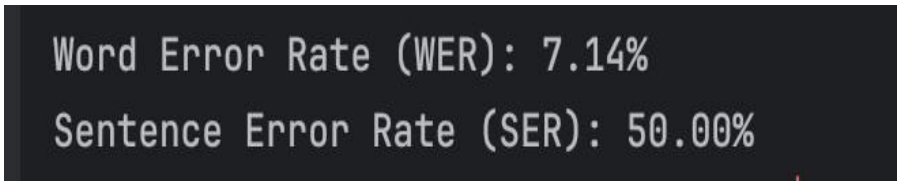

Figure 16. Comparison of Word and Sentence Error Rates

This figure shows the word error rate and sentence rate measured by the ASR system on the dataset. The WER rate of 7.14 percent indicates that the model is mostly correct word-by-word. On the other hand, the Sentence Error Rate of 50 percent indicates that the model misrecognises one out of every two sentences. On a word basis, the model works well. However, even 1 or more errors in short sentences cause the sentence to be incorrectly evaluated. This difference emphasizes the importance of corrective methods or additional language model improvements at the sentence level.

Figure 17. NLP Performance Metrics: Precision, Recall and F1-Score

Evaluation metrics evaluate the success of the model from various perspectives. The precision rate (77.08%) shows the percentage of truly positive examples that the model labels as positive; the Recall rate (77.31%) indicates the percentage of positive examples that the model correctly detects among the total positive examples. The F1 Score (77.18%) is a metric that summarizes the balance between Precision and Recall. An F1 value of around 77% indicates that the model can identify the relevant classes with reasonable accuracy.

5.2 Analysis

This section analyses the results of the analysis of approximately 8 hours of records of 6661 patients and the findings obtained from the visualizations. The aim of the analysis section is to obtain clues about the usability of the findings obtained from ASR and NLP models in clinical settings.

In Figure 9, the most common words in the transcripts obtained from 6661 patients were ‘pain’ and ‘feel’. Both words were used 1600 times each, showing that patients predominantly described pain and feeling. In addition, the fact that words such as ‘shoulder,’ ‘back’, and ‘arm’ ranked high indicates that most of the patients had problems in the musculoskeletal system. The NLP model can explain this situation. In Figure 9, the most common words in the transcripts obtained from 6661 patients were ‘pain’ and ‘feel’. Both words were used 1600 times each, showing that the patients predominantly described pain and feeling. In addition, the fact that words such as ‘shoulder,’ ‘back,’ and ‘arm’ are ranked high shows that most patients have problems in the musculoskeletal system. This means that the NLP model can give a high success rate when trained on pain and muscular systems. It is also helpful for decision support systems focusing on pain management and symptom analysis. The bar graph in Figure 10 shows the patients' complaints and frequency, which is ideal for categorizing the target diseases. Looking at the most frequently transcribed medical terms in the dataset, it can be observed that dermatology and orthopedics problems are frequent. It is also helpful for decision support systems focusing on pain management and symptom analysis. The bar graph in Figure 10 shows the complaints and frequency of the patients, which is ideal for identifying the target diseases. Looking at the most frequently transcribed medical terms in the dataset, it can be observed that dermatology and orthopaedics problems are frequent. In Figure 11, the presence of many peaks between 3.0 and 4.75 indicates that the audio recordings are from multiple speakers and different devices. Values between 3.25 and 3.50 indicate low-quality audio recordings, while peaks between 3.75 and 4 indicate higher-quality recordings. The diversity shown by this visualization emphasizes the importance of pre-processing operations such as normalization and noise reduction. Furthermore, the low number of values of 4.5 and above indicates the limited data obtained from very high-quality environments. This figure may reflect a realistic clinical scenario. The bubble graph in Figure 12 highlights the importance of noise-cancellation techniques in audio recordings. It

clearly shows that as the noise level (x-axis) increases, the sound quality (y-axis) decreases, and the word error rate (bubble size) increases. The heat graph in Figure 14 shows the effect of noise and sound quality in audio recordings on word and sentence error rates. A strong negative correlation exists between sound quality and noise, indicating that sound quality decreases when noise level increases. There is also a strong negative correlation between sound quality and WER and SER. Sound quality affects these metrics by 80 percent. Similarly, noise level and these metrics have a similar negative correlation. This correlation analysis shows how important the efforts to reduce noise and improve sound quality in the project are in improving the model's overall performance. When the ASR metrics WER and SER results are examined, the WER result is a very good score of 7.14 percent. Of course, this rate can be reduced even further with improvements. However, the high SER rate indicates that a more controlled approach should be taken to areas such as language model integration, grammar control, or upper-lower case. Post-processing stages are needed for the reflection of word-based success on sentence integrity. Finally, the model found the true positive values among the positively predicted examples to be 77 percent accurate, which is acceptable. It was also observed that the model correctly found positive examples at 77 percent. It is observed that the model has similar performance in both metrics and shows an average accuracy.

5.2.1 Critical Analysis

The expression "light noise" may include categories of different types of noise (e.g. human speech, humming sound, keyboard clicks). Therefore, the sub-divisions are also examined to show which type of noise makes the system more difficult to use. The multi-peak distribution in Figure 11 shows that the voice recordings were recorded with different devices. Additional sources related to the activity of recording devices can be evaluated by comparing them with the acoustics of the model. The most frequently used dermatological and orthopedic complaints in the model dataset have been evaluated partially well. However, is the same performance achieved in rare transmissions or complaints with fewer records? For this, the control condition of specialist physicians should be specified. In order to take into account data confidentiality, the age, gender, and additional disease history of the patients are not given in the dataset. It should be discussed how the model will be integrated with these additional sources and what they can be. The WER rate shown in Figure 16 is successful, but the SER rate is unsuccessful. This

situation shows that the sentence-based language model has an incomplete structure. Therefore, in subsequent research, language models specially prepared for the medical field should be used.

5.3 Discussion

A total of 8 hours of data from 6,661 patients were explained and analysed using various visualizations. It was observed that the analysed dataset primarily related to dermatological and musculoskeletal disorders (Figures 9 and 10). This aligns with previous literature emphasizing that pain is one of the most frequently reported symptoms in clinical settings (Chen, 2019). However, the frequent appearance of medical terms such as 'acne,' 'shoulder pain,' and 'joint pain' in other graphs suggests that it may not fully represent other important areas such as cardiology, neurology, and psychiatry. Therefore, future studies could benefit from a broader dataset scope to encompass patient complaints.

The analysis of audio quality and background noise (Figures 11, 12, and 13) indicates that noise levels and audio quality significantly impact approximately 80 percent of the Word and Sentence Error rates. The Word Error Rate stands at a promising 7 percent, which is adequate for real-world case studies. However, the 50 percent Sentence Error Rate suggests that some post-processing is needed for the model. Integrating advanced language modelling or post-processing techniques (e.g., grammar checking and contextual error correction) would be beneficial in lowering the sentence error rate. The study shows that the calculated recall, precision, and score of 77 percent exhibit a reasonable level of accuracy. This indicates that the model meets the basic objectives but considering that false positive and false negative predictions can lead to critical results in the healthcare field, it is necessary to increase these rates even further. In future studies, combining the outputs of the model with additional NLP-based processing steps (e.g. medical term definition, grammar correction) will further increase the NLP metrics. Although the obtained performance metrics are promising in a challenging field such as the healthcare field, additional improvements are required.

In summary, the current ASR-NLP system effectively identifies diseases and accurately recognizes words in various noise conditions. It has also provided insights into how noise and sound quality affect speech-to-text technology. However, generalizing its application in real

clinical practice presents challenges. Balancing data diversity, enhancing noise reduction techniques, and advancing NLP integration further are crucial for developing a system that can reliably contribute to clinical decision-support systems.

Chapter 6 - Conclusion

Over the last 50 years, significant advancements in speech-to-text technologies have opened numerous development opportunities in medicine. This thesis explores how speech transcription systems are integrated into crucial areas of the medical sector, such as clinical documentation, patient care, and disease diagnosis. Eight hours of voice recording data from 6,661 patients were analysed using ASR-NLP integrated models and assessed based on system performance, error rates, data quality, noise, data privacy, and ethics. The conclusion section summarizes the findings from the analysed data comprehensively and references literature that investigates the impact of artificial intelligence models in healthcare research.

Overview and Significance of the Research

Clinical documentation increases the workload for healthcare professionals and results in time loss. Traditional methods are also susceptible to errors and inefficiency. Therefore, the need for advanced data collection and analysis tools in healthcare is growing daily. However, speech-to-text technology should be viewed not only as a development but also as an operational and ethical necessity. This research focuses on three main interrelated issues: the performance and challenges of current ASR systems in comprehending medical terminology, addressing obstacles such as background noise and sound quality, and their application in decision support systems, data privacy, security, and ethics. To tackle these challenges, analyse the data using experimental methodologies, and explore its applicability in health technologies

Contributions to Clinical Decision Support

The thesis examines how and to what extent ASR systems can be applied to decision support mechanisms. It emphasizes that accurate and rapid documentation in today’s healthcare can enhance the patient care process, reduce erroneous findings, and assist doctors in analysing challenging issues. One of the project's primary objectives is to develop applicable systems for clinical support, including highly accurate transcripts and alert systems that highlight abnormal results. Advanced NLP techniques can also automatically extract rare medical terms, drug

names, and procedure codes. The findings reveal the significant impact that clinical decision support systems can have on clinical workflows, doctor-patient interactions, disease diagnosis, and treatment timelines by utilizing more structured datasets.

Ethical Considerations and Data Privacy

Although the development of speech recognition technologies adds significant value to healthcare services, data privacy and ethical issues remain sensitive. With regulations such as HIPAA and GDPR, countries have established laws prioritizing patient privacy. In this thesis, the data were anonymized with the participants' consent and analysed solely for this project. A consistent level of transparency was maintained at every stage of model design.

Limitations

In this study, the feasibility of the ASR-NLP integration system was demonstrated through audio recordings, but the methodology has some limitations and potential sources of bias. Identifying and investigating these limitations sheds light on future research. Firstly, although the dataset is a large dataset of 8 and a half hours, the lack of patient data from different geographies, socioeconomic levels and ethnic origins from similar hospitals may limit the applicability of the model in real world scenarios. In addition, the fact that all of the patients speak English may limit the use of the model in other languages. Another point is that due to data confidentiality, specific information such as age, name or gender of the patients is not included in the dataset, which may limit the generalisability of the research when additional clinical information is needed at later stages. Since the diseases encountered in the model analysis are generally common diseases, the model may not analyse well when encountering rare diseases. Finally, the model should be constantly updated with up-to-date medical libraries. This is a process that requires additional resources and time.

CONCLUDING REMARKS

Research shows that combining ASR and NLP technologies can significantly enhance the accuracy and efficiency of clinical documentation within the healthcare sector. The main findings indicate that advanced ASR techniques (like Whisper) and NLP models (such as BioBERT and ClinicalBERT) notably reduce the Word Error Rate, even amidst background noise. This is especially critical when processing sensitive patient data, where anonymization is essential in compliance with regulations. Additionally, these technologies accelerate translation processes and lessen workloads in real-time clinical settings. The study's dataset primarily focuses on medical terminology in fields like dermatology and orthopaedics. Future research directions could involve applying this ASR-NLP system in various specialties, including neurology and psychiatry, thereby yielding new insights. Another potential area of study is to explore transcription in diverse environments by strategically investigating alternative noise cancellation methods in key clinical settings. Furthermore, embedding real-time ASR-NLP outputs into electronic health record systems offers a promising avenue for enabling automated alerts, instant error correction, and advanced analytics, which empower physicians to make evidence-based decisions quickly.

This study addresses the applicability of speech recognition techniques in the healthcare sector and highlights how these technologies can be utilized more effectively and efficiently. It anticipates becoming a critical and indispensable component of future healthcare systems, where data is managed and evaluated with transparency and an ethical perspective.

BIBLIOGRAPHY

1. Adams, T.T., Morgan, J.R. and Kim, E., 2019. Implementation strategies for speech recognition in clinical settings. *Medicine*, 98(31), p.e16612.
2. Al Mansour, A.G.M., Alshomrani, F., Alfahaid, A. and Almutairi, A.T.M., 2025. MammoViT: A Custom Vision Transformer Architecture for Accurate BIRADS Classification in Mammogram Analysis. *Diagnostics*, 15(285).
3. Almutiri, T. and Nadeem, F., 2022. Markov models applications in natural language processing: A survey. *International Journal of Information Technology and Computer Science*, 2, pp.1–16.
4. Ardila, D., Kiraly, A.P., Bharadwaj, S., Choi, B., Reicher, J.J., Peng, L., Tse, D., Etemadi, M., Ye, W., Corrado, G. and Naidich, D., 2019. End-to-end lung cancer screening with three-dimensional deep learning in low-dose chest computed tomography. *Nature Medicine*, 25(6), pp.954–961.
5. Awa, S., 2024. Privacy Risks and Mitigation Strategies in AI-Driven Healthcare Systems: Ensuring Confidentiality in Sensitive Data. *Computer Science Department, University Malaya, Malaysia*.
6. Beauchamp, T.L. and Childress, J.F., 2012. *Principles of Biomedical Ethics*. 7th ed. Oxford: Oxford University Press.
7. Ben Khalfallah, H., Jelassi, M., Demongeot, J. and Bellamine Ben Saoud, N., 2023. Decision support systems in healthcare: Systematic review, meta-analysis and forecasting with the example of COVID-19. *AIMS Bioengineering*, 10(1), pp.27–52. doi:10.3934/bioeng.2023004.
8. Buchanan, W.J., Pandeeswari, S. and Padmavathi, S., 2021. QR-based paperless outpatient health and consultation records sharing system. In: *Proceedings of the International Conference on Machine Learning, Big Data, Cloud and Data Science (ICMLBDA)*. Springer.

9. Chandramouli, N.A., Natarajan, S., Alharbi, A.H., Kannan, S., Khafaga, D.S., Raju, S.K., Eid, M.M. and El-Kenawy, E.-S.M., 2024. [Title not provided]. [Unpublished/Incomplete reference].
10. Chen, X., 2019. Improving clinical documentation with voice recognition technology. *Journal of Healthcare Informatics*, 12(3), pp.45–55.
11. Cheng, P., Gilchrist, A., Robinson, K.M. et al., 2009. The risk and consequences of clinical miscoding due to inadequate medical documentation: a case study of the impact on health services funding. *HIM Journal*, 38(1), pp.35–46. doi:10.1177/183335830903800105.
12. Clark, R.E. and Manion, R.L., 1974. Speech recognition in a biomedical application. *Computers and Biomedical Research*, 7(2), pp.171–181.
13. Dasgupta, S., Brodsky, J. and Widdershoven, J., 2025. Editorial: Paediatric vestibular disorders – a focussed diagnostic approach for best management outcomes, 2024. [No detailed publication info].
14. Davis, S. and Mermelstein, P., 1980. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 28(4), pp.357–366.
15. De Mori, R., 2019. Automatic speech recognition. In: *Applications of Pattern Recognition*. Boca Raton, FL: CRC Press, pp.237–265.
16. Devlin, J., Chang, M.-W., Lee, K. and Toutanova, K., 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of NAACL-HLT*, pp.4171–4186.
17. Elhadad, A., Hamad, S., Elfiky, N., Alanazi, F., Taloba, A.I. and El-Aziz, R.M.A., 2024. Advancing Healthcare: Intelligent Speech Technology for Transcription, Disease Diagnosis, and Interactive Control of Medical Equipment in Smart Hospitals. *AI*, 5(4), pp.2497–2517.
18. European Parliament, 2016. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 (General Data Protection Regulation). *Official Journal of the European Union*.
19. Ferry, Q. et al., 2019. Assessing the Cybersecurity Threat in Healthcare. *Health Informatics Journal*, 25(4), pp.1656–1667.

20. Florida, L., Taddeo, M. and Turilli, M., 2016. What is data ethics? *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 374(2083), pp.1–5.
21. Friedman, C.P. et al., 1994. MedLEE: A practical natural language system for clinical information extraction. *Proceedings of the AMIA Annual Fall Symposium*.
22. Gallardo, L.F., Möller, S. and Beerends, J., 2017. Predicting automatic speech recognition performance over communication channels from instrumental speech quality and intelligibility scores. *INTERSPEECH*, pp.2939–2943.
23. Garrido-Merchan, E.C., Gozalo-Brizuela, R. and Gonzalez-Carvajal, S., 2023. Comparing BERT against traditional machine learning models in text classification. *Journal of Computational and Cognitive Engineering*, 2(4), pp.352–356.
24. Goodfellow, I., Bengio, Y. and Courville, A., 2016. *Deep Learning*. Cambridge, MA: MIT Press.
25. Graves, A., Mohamed, A.-r. and Hinton, G., 2013. Speech recognition with deep recurrent neural networks. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp.6645–6649.
26. Heymans, W., 2022. *Automatic speech recognition of poor-quality audio using generative adversarial networks*. Doctoral dissertation, North-West University (South Africa).
27. Hinton, G., Deng, L., Yu, D., Dahl, G.E., Mohamed, A.R., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T.N. and Kingsbury, B., 2012. Deep neural networks for acoustic modelling in speech recognition: The shared views of four research groups. *IEEE Signal Processing Magazine*, 29(6), pp.82–97.
28. Hinton, G., Osindero, S. and Teh, Y.W., 2006. A Fast Learning Algorithm for Deep Belief Nets. *Neural Computation*, 18(7), pp.1527–1554.
29. HISA, 2013. Health Informatics Society of Australia. Available at: www.hisa.org.au [Accessed 14 January 2014].
30. Hu, D., 2020. An introductory survey on attention mechanisms in NLP problems. In: *Proceedings of the 2019 Intelligent Systems Conference (IntelliSys)*. Cham: Springer, pp.432–448.

31. Huang, X., Baker, J. and Reddy, R., 2014. A historical perspective of speech recognition. *Communications of the ACM*, 57(1), pp.94–103.
32. Huston, J. L., Lee, N. H., & Otto, J. L. (2019). Evaluating the impact of voice recognition in a clinical environment. Health Informatics Journal, 25(3), 703-710.
33. Jiang, F., Jiang, Y., Zhi, H., Dong, Y., Li, H., Ma, S., Wang, Y., et al., 2017. Artificial intelligence in healthcare: past, present and future. *Stroke and Vascular Neurology*, 2(4), e000101.
34. Khan, W., Daud, A., Khan, K., Muhammad, S. and Haq, R., 2023. Exploring the frontiers of deep learning and natural language processing: A comprehensive overview of key challenges and emerging trends. *Natural Language Processing Journal*, p.100026.
35. Kumar, A., & Singh, A. (2020). Integration of automatic speech recognition with EHR systems. Journal of Biomedical Informatics, 108, 103484.
36. Latif, S., Qadir, J., Qayyum, A., Usama, M. and Younis, S., 2020. Speech technology for healthcare: Opportunities, challenges, and state of the art. *IEEE Reviews in Biomedical Engineering*, 14, pp.342–356.
37. Lee, J., Wu, Y. and Yang, S., 2020. Ethical Considerations of AI in Healthcare: Challenges and Solutions. *Health Policy and Technology*, 9(4), pp.396–405.
38. Lee, J. et al., 2024. Streamlining IP Protection with AI-Integrated Management Systems. *IEEE Transactions on Engineering Management*, 71(5), pp.2109–2122. doi:10.1109/TEM.2024.3290123.
39. Liew, C.-H., Ong, S.-Q. and Ng, D.C.-E., 2025. Leveraging machine learning to predict hospitalizations of pediatric COVID-19 patients (PrepCOVID-Machine). *Scientific Reports*, 15, p.3131.
40. Lowerre, B.T., 1976. *The Harpy Speech Recognition System*. Doctoral dissertation, Carnegie-Mellon University.
41. McCowan, I.A., Moore, D., Dines, J., Gatica-Perez, D., Flynn, M., Wellner, P. and Bourlard, H., 2004. On the use of information retrieval metrics for speech recognition evaluation. *Technical Report*, IDIAP.

42. McKinney, S.M., Sieniek, M., Godbole, V., Godwin, J., Antropova, N., Ashrafian, H., Back, T. et al., 2020. International evaluation of an artificial intelligence system for breast cancer screening. *Nature*, 577(7788), pp.89–94.
43. Min, B., Ross, H., Sulem, E., Veyseh, A.P.B., Nguyen, T.H., Sainz, O., Agirre, E., Heintz, I. and Roth, D., 2023. Recent advances in natural language processing via large pre-trained language models: A survey. *ACM Computing Surveys*, 56(2), pp.1–40.
44. Mittelstadt, B. and Floridi, L., 2016. The ethics of big data: Current and foreseeable issues in biomedical contexts. In: *The Ethics of Biomedical Big Data*. Cham: Springer, pp.445–480.
45. Mourid, M.R., Irfan, H. and Oduoye, M.O., 2025. Artificial Intelligence in Paediatric Epilepsy Detection: Balancing Effectiveness with Ethical Considerations for Wellbeing. *Health Sciences Reports*, 8, e70372.
46. Muthusamy, V. et al., 2020. Evaluation of Dragon NaturallySpeaking for medical transcription. *Journal of Medical Systems*, 44(4), p.68.
47. Nautsch, A., Saeidi, R., Rathgeb, C. and Busch, C., 2016. Robustness of quality-based score calibration of speaker recognition systems concerning low-SNR and short-duration conditions. *Odyssey*, pp.358–365.
48. NIHMS, 2019. Privacy and data sharing in biomedical research: Ethical and technical challenges. *National Institutes of Health Article Submission System*.
49. Obermeyer, Z. and Emanuel, E., 2016. Predicting the Future—Big Data, Machine Learning, and Clinical Medicine. *New England Journal of Medicine*, 375(13), pp.1216–1219.
50. Paats, A., Alumäe, T., Meister, E. and Fridolin, I., 2018. Retrospective Analysis of Clinical Performance of an Estonian Speech Recognition System for Radiology: Effects of Different Acoustic and Language Models. *Journal of Digital Imaging*, 31, pp.615–621. doi:10.1007/s10278-018-0085-8.
51. Patel, S. et al., n.d. Integrating AI into IP Management Workflows: Benefits and Best Practices. *Journal of Intellectual Property Management* [Forthcoming].
52. Rabiner, L. and Juang, B.H., 1993. *Fundamentals of Speech Recognition*. Englewood Cliffs, NJ: Prentice Hall.

53. Radford, A. et al., 2022. Robust Speech Recognition via Large-Scale Weak Supervision. *arXiv preprint* arXiv:2212.04356.
54. Rule, A., Florig, S., Bedrick, S., Mohan, V., Gold, J. and Hribar, M., 2022. Comparing Scribed and Non-scribed Outpatient Progress Notes. *AMIA Annual Symposium Proceedings 2021*, pp.1059–1068.
55. Shoghli, A., Darvish, M. and Sadeghian, Y., 2024. Balancing Innovation and Privacy: Ethical Challenges in AI-Driven Healthcare. *Journal of Reviews in Medical Sciences*, 4(1), e31.
56. Smith, L. and Brown, A., 2020. Automatic Speech Recognition in Healthcare: Applications and Challenges. *Journal of Medical Informatics*.
57. SSRN, 2021. Legal and ethical perspectives on artificial intelligence in healthcare. *Social Science Research Network*.
58. Tiralongo, F. and Dragičević, B., 2024. Editorial: Biology and ecology of Mediterranean coastal fishes. *Frontiers in Marine Science*, 11, p.1512836.
59. U.S. Department of Health & Human Services, 1996. Health Insurance Portability and Accountability Act (HIPAA).
60. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L. and Polosukhin, I., 2017. Attention is all you need. *Advances in Neural Information Processing Systems (NeurIPS)*, pp.5998–6008.
61. Wang, D. and Chen, J., 2018. Supervised Speech Separation Based on Deep Learning: An Overview. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 26(10), pp.1702–1726.
62. Wright, A., Sittig, D.F., Ash, J.S., Bates, D.W., Feblowitz, J., Fraser, G., Maviglia, S.M., McMullen, C., Nichol, W.P., Pang, J.E., Starmer, J. and Middleton, B., 2011. Governance for clinical decision support: Case studies and recommended practices from leading institutions. *Journal of the American Medical Informatics Association*, 18(2), pp.187–194. doi:10.1136/jamia.2009.002030.
63. Xu, J., Tan, X., Ren, Y., Qin, T., Li, J., Zhao, S. and Liu, T.Y., 2020. Lrspeech: Extremely low-resource speech synthesis and recognition. *Proceedings of the 26th ACM SIGKDD International Conference*, pp.2802–2812.

64. Xu, Y., Du, J., Dai, L. and Lee, C.-H.H., 2015. An experimental study on speech enhancement based on deep neural networks. *IEEE Signal Processing Letters*, 21(1), pp.65–68.
65. Yang, B., Wang, L., Wong, D.F., Shi, S. and Tu, Z., 2021. Context-aware self-attention networks for natural language processing. *Neurocomputing*, 458, pp.157–169.
66. Yang, X. and Su, T., 2022. Efa-trans: An efficient and flexible acceleration architecture for transformers. *Electronics*, 11(21), p.3550.
67. Yu, D. and Deng, L., 2016. *Automatic Speech Recognition*. London: Springer.
68. Zhang, X., Wu, J. and Chen, H., 2020. Deep Learning Approaches for Enhanced Clinical Documentation. *IEEE Transactions on Neural Networks and Learning Systems*, 31(8), pp.3060–3072.
69. Zhang, Y., Wu, J., Qiu, Y., Song, A., Li, W., Li, X. and Liu, Y., 2023. Intelligent speech technologies for transcription, disease diagnosis, and medical equipment interactive control in smart hospitals: A review. *Computers in Biology and Medicine*, 153, p.106517.
70. Zhou, L. and Hripcsak, G., 2007. Temporal reasoning with medical data—a review emphasizing medical natural language processing. *Journal of Biomedical Informatics*, 40(2), pp.183–202.

TABLE OF FIGURES