A Project Report

on

From Findings to Final Notes: Hybrid AI for Clinical Report Summarization

Submitted in partial fulfillment for the requirements of BE(CSE) VII Semester

Project Part-1

BACHELOR OF ENGINEERING

in

COMPUTER SCIENCE AND ENGINEERING

by

KOTAGIRIWAR SRIYA (160122733012)

M. S. L. AASHRITHA (160122733014)

Under the Supervision of

Smt. CH. MADHAVI SUDHA, Asst. Prof., CSE Dept, CBIT

Department of Computer Science and Engineering,

Chaitanya Bharathi Institute of Technology (Autonomous),

(Affiliated to Osmania University, Hyderabad)

Hyderabad, TELANGANA (INDIA) –500 075

[2025-2026]

DECLARATION

We hereby declare that the report entitled “From Findings to Final Notes: Hybrid AI for Clinical Report Summarization” submitted for the requirements of B.E (CSE) VII Semester Project Part-1 is our original work and it has not formed the basis for the award of any other degree, diploma, fellowship or any other similar titles.

Kotagiriwar Sriya (160122733012)

M. S. L. Aashritha (160122733014)

Place: CBIT, Hyderabad

Date:

ACKNOWLEDGEMENT

We would like to take this opportunity to express our sincere gratitude to our **Supervisor, Smt. Ch. Madhavi Sudha,** for her constant support, invaluable guidance, and expert insights throughout this project. Her encouragement, constructive feedback, and patience have been instrumental in shaping our work and helping us complete it successfully.

We extend our heartfelt thanks to our **Project Coordinators-Dr. M. Swamy Das, Dr. G. Vanitha,** and **Dr. K. Spandana,** for their valuable guidance, timely suggestions, and continuous encouragement during the course of this project.

We also wish to express our deep appreciation to **Prof. C.V. Narasimhulu**, Principal of our institute, for his inspiring leadership and for nurturing a culture of academic excellence among students. Our sincere thanks to **Dr. S. China Ramu**, Head of the Department of Computer Science and Engineering, for his encouragement and support throughout this dissertation process.

We are also grateful to all the **faculty members and staff of the Department of Computer Science and Engineering, CBIT**, for their cooperation, assistance, and motivation during this project.

Lastly, we express our heartfelt gratitude to our **parents**, whose unconditional love, emotional strength, and constant support both moral and financial have been the driving force behind our success. Their faith in us has been our greatest motivation throughout this journey.

ABSTRACT

Healthcare professionals spend considerable time on manual documentation, which reduces direct patient interaction and contributes to burnout. Existing systems such as Electronic Health Records (EHRs) and basic speech-to-text tools often fail to generate structured, context-aware clinical notes and lack the ability to integrate multimodal inputs like audio consultations and laboratory data. The purpose of this study is to design and develop an AI-powered Clinical Assistant that automates clinical documentation, enhances data accessibility, and reduces the administrative burden on healthcare providers.

The proposed system leverages Generative AI (GenAI) and Large Language Models (LLMs) to transcribe doctor–patient conversations, process laboratory reports, and synthesize structured SOAP (Subjective, Objective, Assessment, Plan) notes. Transformer-based architectures such as BERT and BART are employed for summarization, while Named Entity Recognition (NER) ensures precise extraction of medical entities. A context-aware interface further enables real-time querying and retrieval of patient information. By integrating conversational and diagnostic data, the proposed Clinical Assistant enhances documentation accuracy, improves workflow efficiency, and supports better clinical decision-making, ultimately contributing to improved healthcare outcomes.

LIST OF FIGURES

LIST OF ABBREVIATIONS

Abbreviation	Full Form
ASR	Automatic Speech Recognition
BART	Bidirectional and Auto-Regressive Transformers
BERT	Bidirectional Encoder Representations from Transformers
EHR	Electronic Health Record
EMR	Electronic Medical Record
HIPAA	Health Insurance Portability and Accountability Act
NER	Named Entity Recognition
QLoRA	Quantized Low-Rank Adaptation
RAG	Retrieval-Augmented Generation
ROUGE	Recall-Oriented Understudy for Gisting Evaluation
RRS	Radiology Report Summarization
SOAP	Subjective, Objective, Assessment, Plan
UMLS	Unified Medical Language System
ViT	Vision Transformer
WER	Word Error Rate

Table of Contents

1. INTRODUCTION

1.1 Background and Motivation

Healthcare keeps changing fast with all this digital stuff. It is meant to make services better and run things more smoothly. Even so, doctors and nurses still spend way too much time on paperwork. That cuts into time they could use talking directly with patients. It also wears them out over time. Electronic Health Record systems help store data easier now. They do not really get what the info means on their own though. They cannot sort it out or put it in context automatically. So providers end up typing everything in by hand. They have to check it all too. That takes forever and repeats a lot.

Doing docs by hand wastes precious minutes. It leads to mistakes too. Things get left out or do not match up. This happens a lot in busy spots like emergency rooms or ICUs. Those slip-ups mean patient files are not full. Decisions get held up because of it. Safety for patients can take a hit as well. On top of that, all the mental effort for these tasks burns people out. Burnout is becoming a big issue in healthcare today.

AI and Natural Language Processing have come a long way lately. They could automate a bunch of this. Still, old-school tools like simple voice-to-text do not cut it. They miss the medical side of things. They cannot pull out important details from notes. Nor can they format everything neatly for standards. That leaves a real hole. We need tools that grasp context well. They should be accurate and easy to get patient info from.

This work comes from wanting to fix clinical routines. It aims to lighten the admin load. Patient care should get better overall. Tackling the drag of long documentation helps. So does sorting out scattered data. Then staff can focus more on patients face to face. Outcomes improve that way. People feel happier too. Plus, sharper records mean teams talk better. Decisions come from solid info. In the end, the whole system gets tougher and works smoother.

1.2 Problem Statement

Clinical documentation plays a big role in healthcare. But it eats up a lot of time. That cuts into the moments doctors get with their patients. Electronic Health Records and simple transcription options are out there. Still, conversations between doctors and patients hardly

ever turn into organized notes that mean something in a clinical way. This leads to records that feel scattered and not quite whole. On top of that, things like lab reports, diagnostic tests, and other patient details do not blend well with what comes from those talks. It all leaves patient files less thorough than they should be.

Systems for handling documentation right now fall short on smart features. They do not pick up on context when you search for info. Healthcare workers find it tough to pull up the right details fast and spot on from different sources. All these gaps pile on extra admin work. They add mental pressure for clinicians too. In the end, that threatens how safe patients stay. It affects the quality of choices made in care.

We need a real fix for these issues. One that handles patient interactions with notes that are spot on, well structured, and cover everything. It should give easy access to patient data all tied together with context. Tackling this matters a ton. It boosts how efficiently clinics run. It helps ease burnout for those in healthcare. And it makes sure care stays informed, comes when needed, and keeps things safe.

1.3 Objectives

This study focuses on building a smart Clinical Assistant. The goal is to improve clinical documentation in several ways. It makes records more accurate and complete. It also makes them easier to access. At the same time, it lightens the administrative load for healthcare workers. The system helps both doctors and patients. It does this through context-aware searches for information. It protects data privacy with strong access controls.

The key objectives cover a few main areas. One is converting conversations between doctors and patients into structured SOAP notes. This uses speech-to-text tools along with natural language processing techniques. The process ensures high accuracy. It cuts down on manual work. It also supports quick updates to patient records.

Another objective involves blending conversational data with lab reports and diagnostic results. This creates a fuller picture of each patient. In turn, it aids better decision-making. It raises the overall quality of care provided.

The plan includes creating a querying system that understands context. This allows for fast retrieval of relevant medical information. Such a feature boosts usability. It streamlines efficiency in clinical workflows.

Finally, the system prioritizes secure and privacy-conscious documentation. It relies on Role-Based Access Control, or RBAC. This setup permits access to sensitive patient details only for authorized users.

1.4 Methodology Overview

The system they propose uses a mix of AI approaches to handle clinical documentation automatically and make it better. It pulls in both talk-based and diagnostic details to create organized patient files that include plenty of context. Interactions between doctors and patients get recorded through high-tech speech-to-text tools. Those tools turn spoken words into reliable written forms. Then the written info goes through natural language processing methods and models based on transformers. This helps sum up the discussions, spot important clinical details, and put together structured SOAP notes. They also apply named entity recognition to pull out and sort medical items like symptoms, diagnoses, medications, and treatment plans. That way the notes stay relevant to clinical needs and stay accurate.

Beyond just the conversation parts, the system looks at lab reports and other diagnostic results too. It blends those with the summarized talk data to build full patient records. Bringing together this kind of varied data gives a complete picture of patient health. It helps with better decisions and cuts down on chances of spotty or missing records. They include a querying setup that pays attention to context. This lets doctors and patients find medical info easily and fast. The system figures out what queries mean by looking at the surrounding data details. Users can then pull up patient files, lab outcomes, or summary notes from consultations without much hassle or error.

Putting all these AI methods together means the system produces documentation that is spot-on, well-organized, and mindful of privacy. It lightens the load of paperwork for healthcare workers quite a bit. Workflow gets smoother overall. Clinical choices become

more timely and based on solid info. In the end, this mixed AI approach offers a way to scale up that works reliably. It boosts both the quality and speed of clinical documentation in different healthcare places.

1.5 Scope and Limitations

The system mainly handles automating structured SOAP notes pulled straight from conversations between doctors and patients. It also weaves in laboratory reports to build out full patient records that cover everything needed. This setup lets both doctors and patients run context-aware queries pretty easily. They pull up relevant medical details without much hassle that way. Privacy stays solid along with security thanks to Role-Based Access Control, or RBAC, keeping things locked down properly. The whole thing cuts back on administrative tasks and helps avoid mistakes in documentation. That pushes clinical workflows to run more smoothly in the end. You can adjust it for hospitals, clinics, or telemedicine setups. Those spots depend on solid, organized records to keep operations going right.

Still, the system runs into some limits along the way. How well it performs ties right into the audio quality coming in. Bad sound or noise in the background can mess with transcription accuracy and the notes it spits out. It does integrate lab reports, sure. But it sticks to standard digital files mostly. Unstructured stuff or handwritten ones might not process well at all. The context-aware queries only dig into data already inside the system. They do not reach out to outside medical records or other databases. The NLP models and summarization tools train on particular medical data sets. That can cause slip-ups now and then with rare terms or local dialects. Real-time speed might slow down too because of computing power limits. This happens especially with long talks or big batches of lab info.

1.6 Organization of the report

This report is organized into five chapters that collectively describe the development and evaluation of **“From Findings to Final Notes: Hybrid AI for Clinical Report**

Summarization." Each chapter presents a distinct phase of the system's design and implementation in a logical sequence.

The first part introduces the research background, objectives, and significance of the proposed system. It establishes the motivation for automating clinical documentation through hybrid AI techniques that combine speech-to-text and natural language processing.

The following section reviews related research and existing solutions in clinical transcription and summarization. It discusses key challenges in current approaches, evaluates various models, and identifies the research gaps that led to the development of the proposed system.

Subsequently, the design phase outlines the overall system architecture and the major functional modules. It explains the workflows and algorithms involved, including preprocessing, transcript cleaning, token-based chunking, Whisper-based transcription, and transformer-based summarization.

The implementation part focuses on the practical realization of the system. It covers the hardware and software needs right from the start. Then it goes into the dataset that was involved. It breaks down every step of the implementation too. All of this gets backed up with things like Word Error Rate, which is WER, along with some graphical breakdowns.

The report wraps up in the end with a quick summary of the results and main points. It points out ways the system boosts accuracy in transcription. It also makes sure documentation stays organized. Plus it helps with efficiency in clinical work. That last part talks about what could come next. Things like putting it into real-time use. Or tweaking it more for certain medical areas.

2. LITERATURE SURVEY

2.1 Introduction to the Problem and basic Terminology

The healthcare industry generates vast amounts of unstructured textual data through doctor-patient conversations, clinical notes, radiology reports, and electronic medical records (EMRs). Clinical documentation, while essential for patient care, diagnosis, and treatment planning, imposes a significant administrative burden on healthcare professionals. The process of manually creating comprehensive clinical notes and summaries is time-consuming, labor-intensive, and prone to inconsistencies, contributing substantially to physician burnout [1].

Clinical Note Generation refers to the automated process of converting doctor-patient conversations or medical data into structured, concise clinical documentation. The most common format is the SOAP (Subjective, Objective, Assessment, and Plan) note, which organizes patient information into distinct sections for better comprehension and continuity of care [2-3].

Text Summarization in the medical domain involves condensing lengthy medical documents into shorter versions while preserving essential information. This can be achieved through two primary approaches: extractive summarization, which selects important sentences from the source text, and abstractive summarization, which generates new sentences to capture the essence of the original content [2][4].

Radiology Report Summarization (RRS) specifically focuses on generating concise "Impression" sections from detailed "Findings" sections in radiological reports. This task is particularly critical as radiologists often need to quickly communicate key diagnostic information to referring physicians . Natural Language Processing (NLP) and Large Language Models (LLMs) form the technological foundation for automated clinical documentation. Recent advancements in transformer-based architectures, such as BERT, GPT, and T5, have significantly improved the quality and accuracy of medical text generation [5-8].

Multimodal Learning represents an emerging paradigm that integrates multiple data types - such as text, images, and structured data - to enhance the comprehensiveness and accuracy of clinical summaries [8][9].

2.2 Issues and Challenges

2.2.1 Data Quality and Structure

Clinical conversations and medical records often contain unstructured, inconsistent data with specialized medical terminology, abbreviations, and jargon [10]. Spoken language in doctor-patient dialogues includes disfluencies, incomplete sentences, and informal expressions that require normalization before processing [2]. Additionally, medical records may contain handwritten notes that are difficult to digitize accurately using OCR technology [1].

2.2.2 Domain Specificity

Medical language differs significantly from general language, requiring domain-specific knowledge and understanding of clinical context [11]. General-purpose language models often struggle with medical terminology, disease classifications, and the relationships between symptoms, diagnoses, and treatments [7]. The risk of generating factually incorrect information (hallucination) is particularly problematic in healthcare, where inaccuracies can have serious consequences [1][12].

2.2.3 Semantic Understanding and Context

Capturing the complete clinical context from conversations is challenging, particularly when relevant information is scattered throughout lengthy dialogues [2][3]. Understanding semantic relationships between different sections of clinical notes (e.g., linking medications to specific problems) requires sophisticated reasoning capabilities [13]. The integration of temporal information, such as changes in patient condition over time or comparison with previous medical examinations, adds another layer of complexity [5][6].

2.2.4 Computational Resources

Training and deploying large language models for medical applications require substantial computational resources and infrastructure [7][12]. Fine-tuning pre-trained models on domain-specific datasets demands significant time and specialized hardware, making it challenging for smaller healthcare institutions to adopt these technologies [1][8].

2.2.5 Data Privacy and Security

Patient data privacy is a paramount concern in healthcare applications [7][12]. The use of cloud-based LLM services raises questions about data security, especially when handling sensitive medical information. Ensuring compliance with regulations such as HIPAA while leveraging powerful AI models presents significant challenges.

2.2.6 Evaluation and Validation

Evaluating the quality of generated clinical summaries is complex, as traditional metrics like ROUGE scores may not fully capture clinical accuracy, completeness, or usability [3][6]. Human evaluation by medical professionals is often necessary but time-consuming and subjective. Ensuring that generated summaries align with clinical standards and are actionable for healthcare practitioners requires rigorous validation and verification for better performance [9].

2.2.7 Integration with Existing Systems

Integrating AI-powered summarization tools with existing healthcare IT infrastructure, such as Electronic Health Record (EHR) systems, Picture Archiving and Communication Systems (PACS), and Radiology Information Systems (RIS), poses technical and organizational challenges [8][12].

2.3 Related Research Works/Studies

2.3.1 Clinical Note Generation from Conversations

Nguyen et al.(2023) proposed a semantic partition-oriented summarization approach for generating clinical notes from doctor-patient conversations. Their method employs an

extractive module using Sentence-BERT embeddings and semantic-based partitioning to cluster sentences relevant to specific SOAP sections, followed by an abstractive module using BART and DistilBERT. This approach achieved a ROUGE-1 score of 0.512 on the MEDIQA-Sum 2023 dataset, demonstrating improved performance over baseline models[2].

Krishna et al.(2021) introduced CLUSTER2SENT, an algorithm that extracts important utterances from doctor-patient transcripts, clusters related utterances, and generates one summary sentence per cluster. The study emphasized the importance of pre-training (particularly with T5) and the benefits of structuring summaries into sections with supporting evidence annotations[3].

Kumar et al.(2023), developed an automated discharge summary generation system using Quantized Low-Rank Adaptation (QLoRA) and LLMs. Their system integrates image-to-text (using PyTesseract OCR and t5-base), speech-to-text (using AssemblyAI API), and radiology report summarization modules. By fine-tuning the t5-base model with domain-specific datasets like MeQSum, they achieved improved performance in generating standardized summaries with sections such as "Patient History," "Diagnosis," and "Medications."[1].

2.3.2 Medical Domain Knowledge Integration

The MedicalSum model introduced by Nuance researchers incorporated medical domain knowledge from the Unified Medical Language System (UMLS) through three mechanisms: guidance signals with medical words, semantic type embeddings, and weighted loss functions prioritizing medical term prediction. This approach achieved state-of-the-art ROUGE score improvements (0.8-2.1 points) with a 6.2% ROUGE-1 error reduction in the Physical Examination section when evaluated on Family Medicine conversations and radiology reports[11].

2.3.3 Radiology Report Summarization

Mei et al.(2024)developed RadChat, which reframes radiology report summarization as a conversational question-answering task by creating "temporal radiological conversations"

where findings are questions and impressions are answers. By fine-tuning the Vicuna chat model and incorporating patient history, RadChat outperformed state-of-the-art baselines on the MIMIC-CXR dataset and surpassed human references in diversity according to human evaluation[5] .

Ma et al.(2024) introduced ImpressionGPT, an iterative optimization framework for radiology report summarization using ChatGPT with dynamic prompts. The framework uses similarity search to identify relevant reports from a corpus, constructs dynamic prompts with similar examples, and employs an iterative optimization algorithm using ROUGE-1 scores to refine outputs. ImpressionGPT achieved superior performance on MIMIC-CXR and OpenI datasets without additional training or fine-tuning[6] .

Sultan et al.(2024) proposed SumGPT, a multimodal framework combining T5 with Vision Transformer (ViT) for radiology report summarization. This approach integrates textual and visual information through cross-modal fusion, achieving ROUGE-1, ROUGE-2, and ROUGE-L scores around 0.85 and a BLEU score of 0.8470 on a dataset of radiology images and reports. Ablation studies confirmed the critical role of the cross-attention mechanism and multimodal fusion layer[8] .

2.3.4 Clinical Text Extraction and Segmentation

Zelina et al., in 2025, developed an unsupervised method for extraction, labeling, and clustering of textual segments from Czech clinical notes. Their approach involved record splitting using hard-coded algorithms, label extraction with regular expressions, segment classification using RobeCzech (fine-tuned RoBERTa), title clustering with Doc2Vec, and semantic mapping to predefined ontologies. The method demonstrated practical relevance through improved patient similarity analysis and faster information retrieval [10].

Sabariram et al.(2024) introduced an end-to-end pipeline integrating segmentation, anomaly detection, and report generation for medical imaging. The system uses a modified Visual Large Language Model (VLM) with medical-specific adapters, a ResNet50-based segmentation model, and a fine-tuned Gemma 2B LLM for report generation. The pipeline achieved over 90% AUC for anomaly detection and segmentation in few-shot learning settings across Brain MRI, Chest X-ray, and Retina scans[9] .

2.3.5 Problem-Oriented and Structured Summarization

Devarakonda et al. (2014)developed a problem-oriented patient record summary system using Watson analytics. The system generates problem lists from clinical notes, organizes clinical data into meaningful aggregates (medications, lab tests, procedures), identifies semantic relationships between problems and treatments, and presents information in a dashboard-style visualization. The approach uses UMLS concepts, Latent Semantic Analysis, and Distributional Semantics for relationship scoring[13].

2.3.6 Speech Summarization Techniques

Liu et al. (2011) proposed a supervised framework for keyword extraction from meeting transcripts with application to medical conversations. They introduced novel features including term specificity, decision-making sentence features, prosodic prominence scores, and summary-derived features. A feedback loop mechanism leveraging the relationship between keywords and summary sentences demonstrated superior performance over unsupervised TF-IDF and traditional keyphrase extraction systems[14].

Chen and Lin (2012) developed a risk-aware modeling framework for extractive speech summarization, formulating sentence selection as a Bayes risk minimization problem. Their approach incorporates list-wise selection strategies, multiple loss functions (VSM, KL-divergence, MMR), and combines generative and direct modeling paradigms. The method achieved substantial improvements over traditional approaches on broadcast news corpora, with performance gains of 4-5% over baseline supervised methods[4].

2.3.7 Survey and Comprehensive Studies

Palanisamy et al.(2025) conducted a comprehensive survey on leveraging GPTs for medical and biomedical document summarization. The survey explored applications including information extraction, automated summarization, sentiment analysis, and clinical decision support. It identified critical challenges such as ensuring faithfulness and accuracy, interpretability, data privacy, and computational resource requirements, while suggesting future directions including domain-specific refinement and multimodal functionalities[7].

Sai et al. (2024) provided a comprehensive study of generative AI in healthcare, covering applications in medical imaging, drug discovery, personalized treatment, clinical trial optimization, and text generation. The paper discussed healthcare-customized LLMs such as Med-PaLM, BioGPT, and BioBERT, while addressing significant limitations including attribution problems, data quality and bias, patient data privacy, false information generation, and integration challenges with existing healthcare technologies[12].

2.4 Comparative Analysis of Existing Solutions

2.4.1 Summarization Approaches

Extractive vs. Abstractive Methods: Extractive approaches [2][3][4] select relevant sentences from source documents, offering higher factual accuracy and reduced hallucination risk. However, they may produce less coherent summaries with redundant information. Abstractive methods [1][6][8][11] generate new text, producing more natural and concise summaries but with higher risk of factual errors. Hybrid approaches combining both paradigms [2][3] have shown superior performance by leveraging the strengths of each method.

Modular vs. End-to-End Architectures: Modular systems [1][2][3] separate extraction, clustering, and generation stages, offering interpretability and easier debugging. End-to-end models [5][6][8] learn the entire summarization pipeline jointly, potentially capturing more complex relationships but with reduced transparency. Recent studies [3] demonstrate that modular approaches with pre-trained language models (especially T5) achieve competitive or superior performance compared to end-to-end methods.

2.4.2 Knowledge Integration Strategies

Domain-Specific Pre-training: Models fine-tuned on medical corpora [1][10][11] show improved understanding of clinical terminology and context compared to general-purpose models. The integration of UMLS knowledge [11][13] through semantic type embeddings and weighted loss functions provides additional performance gains, particularly for recognizing and generating medical concepts.

Prompt Engineering vs. Fine-tuning: ImpressionGPT [6] demonstrates that carefully designed dynamic prompts with iterative optimization can achieve state-of-the-art results without model fine-tuning, offering a cost-effective alternative to traditional training approaches. However, fine-tuned models [1][5][8] generally provide more consistent performance across diverse inputs and better integration of domain-specific knowledge.

2.4.3 Multimodal Integration

Text-Only vs. Multimodal Approaches: Text-only models [2][3][6] are computationally efficient and work well for conversation-based summarization. Multimodal approaches [8][9] incorporating visual information demonstrate superior performance for tasks like radiology report generation, where image content provides critical diagnostic information. SumGPT [8] achieved approximately 85% ROUGE scores by effectively fusing T5 text embeddings with ViT image features through cross-attention mechanisms.

Feature Fusion Techniques: Early fusion (combining features before processing) [8][9] and late fusion (combining outputs) strategies show varying effectiveness depending on the task. Cross-modal attention mechanisms [8] enable models to selectively focus on relevant visual and textual information, significantly improving performance over simple concatenation approaches.

2.4.4 Performance Metrics

ROUGE Scores: Most studies [1][2][3][5][6][8][11] report ROUGE-1, ROUGE-2, and ROUGE-L scores ranging from 0.45 to 0.85, with recent multimodal approaches [8] achieving the highest scores. ROUGE-1 is particularly effective for capturing fine-grained details in medical summaries [6].

Clinical Validity: Human evaluation by medical professionals [3][5][8][9] consistently shows that while automated metrics are useful, clinical accuracy, actionability, and alignment with medical standards require expert assessment. Studies incorporating human evaluation [3][9] report that modular approaches produce higher yields of useful, factually correct sentences compared to purely abstractive baselines.

2.4.5 Computational Efficiency

Resource Requirements: Large-scale pre-trained models [7][9][11] require significant computational resources for training and inference. QLoRA-based approaches [2] offer a compromise by enabling efficient fine-tuning with reduced memory requirements. Prompt-based methods [7] provide the most resource-efficient deployment option, requiring no additional training.

Scalability: Systems designed for specific languages or medical specialties [10] may face challenges when scaling to diverse populations or medical domains. Models with strong generalization capabilities [7][12] trained on large, diverse datasets demonstrate better transferability across different healthcare contexts.

2.5 Research Gaps Identified

2.5.1 Limited Integration of Temporal and Historical Context

While RadChat [5] demonstrates the importance of incorporating previous medical examinations, most existing approaches [1][2][3] process clinical conversations or reports in isolation without considering longitudinal patient history. There is a need for frameworks that systematically integrate temporal information, disease progression patterns, and treatment response over time to generate more contextually aware summaries.

2.5.2 Insufficient Handling of Multimodal Medical Data

Although some studies [8][9] integrate visual and textual information, comprehensive frameworks that incorporate diverse data types - including laboratory results, vital signs, genomic data, and patient-reported outcomes - remain underdeveloped. The potential of multimodal learning for capturing the complete clinical picture has not been fully explored.

2.5.3 Lack of Real-Time and Interactive Summarization

Current approaches [1][2][3][6] primarily focus on batch processing of complete conversations or reports. There is limited research on real-time summarization systems that can generate incremental summaries during ongoing consultations, allowing physicians to review and correct information immediately. Interactive systems that incorporate physician

feedback during the documentation process could significantly improve accuracy and usability.

2.5.4 Insufficient Focus on Explainability and Trust

While some work addresses interpretability [7][12], there is a critical gap in developing summarization systems that provide transparent explanations for their outputs, including citations to source material and confidence scores for generated statements. Clinicians need to understand why specific information was included or excluded from summaries to trust and effectively use these tools in clinical decision-making.

2.5.5 Inadequate Evaluation of Clinical Impact

Most studies [1][2][3][6][8] rely primarily on automated metrics (ROUGE, BLEU) and limited human evaluation, without comprehensive assessment of clinical outcomes, workflow integration, or impact on patient care quality. Rigorous clinical trials evaluating the effectiveness of summarization systems in reducing physician burnout, improving documentation quality, and enhancing patient outcomes are notably absent from the literature.

2.5.6 Standardization and Interoperability

The lack of standardized evaluation protocols, benchmark datasets covering diverse medical specialties, and interoperability standards for integrating summarization tools with existing healthcare IT systems [12] hinders the widespread adoption and comparative evaluation of different approaches.

2.5.7 Addressing Hallucination and Factual Accuracy

While some studies [3][6] acknowledge the problem of generated information that is not supported by source data, comprehensive solutions for detecting, preventing, and correcting hallucinations in medical text generation remain limited. This is particularly critical in healthcare, where factual accuracy is paramount for patient safety.

3. DESIGN OF THE PROPOSED SYSTEM

3.1 System Architecture

The system architecture of *"From Findings to Final Notes: Hybrid AI for Clinical Report Summarization"* is designed to enable automated, structured, and contextually rich clinical documentation by integrating multimodal data sources. The architecture adopts a **hybrid AI approach**, combining traditional Natural Language Processing (NLP), transformer-based language models, and structured data integration mechanisms. This layered architecture ensures modularity, scalability, and secure handling of sensitive medical information.

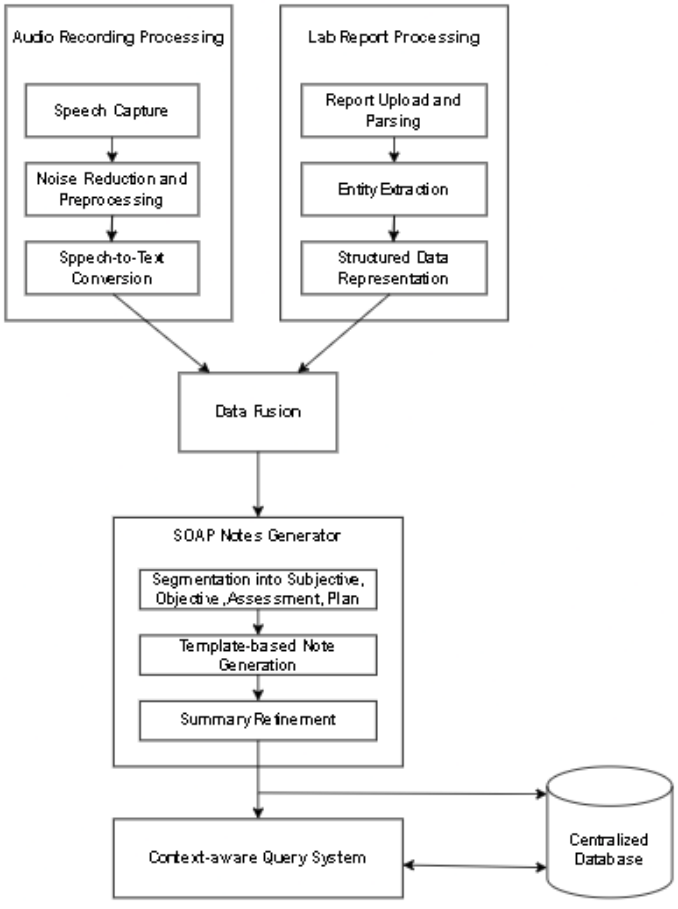

Figure 1: Architecture of the proposed system

Figure 1 illustrates the overall system architecture, beginning with two primary data streams: **Audio Recording Processing** and **Lab Report Processing**. These independent modules serve as the system's input channels, capturing conversational and diagnostic information respectively. The outputs from both are then processed and merged within the **Data Fusion Layer**, which plays a critical role in unifying unstructured and structured information.

The fused data is passed to the **SOAP Notes Generator**, the analytical core of the system, which applies advanced NLP models and entity-based segmentation to automatically produce standardized clinical documentation. These notes are generated following the widely recognized SOAP (Subjective, Objective, Assessment, Plan) format, ensuring interpretability and compatibility with clinical workflows.

The **Context-Aware Query System** operates as the user-interactive layer, allowing both doctors and patients to retrieve and interact with information efficiently. It supports semantic understanding of queries, enabling the system to provide relevant responses based on contextual relationships between data elements. To preserve confidentiality and data security, Role-Based Access Control (RBAC) is implemented at the database level, ensuring that access to sensitive patient information is restricted to authorized personnel only.

This architecture not only minimizes manual documentation effort but also enhances the precision and completeness of medical records. It facilitates seamless integration between conversational data and laboratory findings, thereby improving the overall efficiency and reliability of healthcare documentation systems.

3.2 Description of Modules

The proposed system consists of multiple interconnected modules that together facilitate the end-to-end automation of clinical documentation. Each module performs a distinct function but interacts harmoniously with the others to maintain workflow consistency and ensure accurate data handling.

I. Audio Recording Processing: This module initiates the system's workflow by capturing doctor–patient conversations through a secure and HIPAA-compliant interface. The recorded speech undergoes noise reduction, audio normalization, and preprocessing to enhance clarity and remove ambient distortions. A transformer-based speech-to-text conversion model is then applied to accurately transcribe spoken dialogue into text. This transcription forms the primary unstructured input, representing the subjective and observational aspects of a medical consultation. The emphasis on linguistic precision ensures that essential details such as symptoms, patient history, and medical advice are retained during conversion.

II. Lab Report Processing: The lab report processing module is designed to manage structured diagnostic data. Laboratory reports uploaded by healthcare professionals are parsed using optical character recognition (OCR) (if scanned) and text parsing algorithms (if digital). The extracted data undergoes entity extraction to identify key clinical elements such as test names, parameters, and values. The system then formats these results into a structured data representation, aligning them with standardized terminologies and ensuring interoperability with other modules. This structured data serves as the objective evidence in the final documentation.

III. Data Fusion: The data fusion module acts as the integrative core of the system. It combines transcribed conversational data and structured lab report information into a unified framework. The process involves semantic alignment and temporal synchronization, ensuring that each piece of diagnostic data is properly associated with its corresponding conversational context. This step creates a coherent patient record where symptoms, diagnoses, and test results are meaningfully connected. By leveraging hybrid AI models, this fusion not only enriches the contextual understanding of patient cases but also enables downstream modules to perform more accurate summarization and reasoning.

IV. SOAP Notes Generator: This module forms the analytical backbone of the system. Using NLP pipelines, transformer-based summarization models (like BERT or BART), and Named Entity Recognition (NER), it segments the fused data into the four standardized components of SOAP notes:

- **Subjective:** Patient's verbal reports and symptoms extracted from audio transcriptions.
- **Objective:** Quantitative data and clinical findings derived from lab reports.
- **Assessment:** Analysis and interpretation of the patient's condition.
- **Plan:** Recommended treatments, medications, or next steps suggested by the clinician.

The generator refines the output using template-based generation and context-aware summarization to ensure clarity, coherence, and clinical accuracy. This automated approach not only standardizes documentation but also reduces redundancy and manual workload for practitioners.

V. Database Management System: The database serves as the secure backbone of the entire architecture, storing both structured and unstructured data generated throughout the system. It maintains transcribed conversations, processed lab data, fused patient records, and generated SOAP notes in well-defined collections or relational tables. Implemented using scalable database technologies it supports efficient indexing and retrieval for real-time querying. The integration of Role-Based Access Control (RBAC) ensures that data access is restricted based on user privileges, maintaining patient confidentiality and compliance with healthcare privacy standards.

VI. Context-Aware Query System: The final module offers a user-centric interaction platform, allowing both doctors and patients to query the system intuitively. Built using a semantic query engine, it interprets natural language questions and retrieves relevant information based on contextual and relational understanding. For example, a doctor may query "show lab trends for the last visit," and the system responds with precise, contextually linked results. The underlying database, integrated through secure APIs, ensures that query results are fetched efficiently while maintaining data protection via RBAC. This mechanism enforces privacy, accountability, and traceability in all data transactions. Thus, each module contributes to a unified goal: transforming raw, multimodal clinical inputs into structured, meaningful, and secure medical documentation. The modular design also allows scalability for future enhancements, such as integrating

imaging data or predictive diagnostic models, ensuring the system remains adaptable to evolving healthcare needs.

3.3 Algorithms

The system employs a combination of advanced algorithms and preprocessing techniques to automate clinical documentation, generate structured SOAP notes, and enable context-aware querying of patient records. **Automatic Speech Recognition (ASR)** is performed using OpenAI's **Whisper transformer model (medium)**, which captures contextual relationships in audio sequences to accurately convert speech into text. **Summarization** is carried out with **BART-large transformer models** fine-tuned on conversational data, producing concise summaries of doctor-patient dialogues while retaining critical clinical information. **Named Entity Recognition (NER)** leverages **BERT-based models** to identify and classify clinical entities such as patient names, medical conditions, medications, and treatments, providing structured representations from unstructured text.

For future enhancements, **sequence-to-sequence transformers** will be used for automated SOAP note generation, structuring summaries into Subjective, Objective, Assessment, and Plan sections, while **Retrieval-Augmented Generation (RAG)** and vector similarity search algorithms will enable context-aware querying by retrieving relevant historical records. Additionally, **knowledge graph algorithms** may be integrated to model relationships among clinical entities, supporting semantic reasoning and interpretable query responses. Complementing these core algorithms, **preprocessing techniques** such as audio silence removal, normalization, transcript cleaning to remove filler words, and token-based chunking for managing long transcripts ensure high-quality inputs for the models. Together, these algorithms and techniques provide a robust framework to transform raw clinical audio into structured, actionable, and queryable patient information, forming the backbone of an intelligent clinical assistant system.

4. IMPLEMENTATION

4.1 Software and Hardware Requirements

To ensure smooth execution and optimal performance of the proposed clinical assistant system, specific software and hardware resources are required. These resources facilitate audio processing, natural language understanding, model inference, and storage of transcripts and patient records.

Software Requirements

- **Operating System:** Windows 10/11, Linux (Ubuntu 20.04 or later), or macOS.
- **Programming Language:** Python 3.9 or later.
- **Machine Learning & NLP Libraries:** Transformers, HuggingFace Hub, PyTorch, TensorFlow (optional), Whisper, BERT, BART.
- **Audio Processing Libraries:** Librosa, SoundFile, NumPy.
- **Data Handling Libraries:** Pandas, NumPy.
- **Database Management:** MongoDB or PostgreSQL for storing transcripts and patient records.
- **Development Tools:** VS Code, Jupyter Notebook, PyCharm, or Google Colab.
- **Additional Tools:** Kaggle API for dataset access, Git for version control, pip for package management.

Hardware Requirements

- **Processor:** Intel i5/i7 or AMD Ryzen 5/7 (8th generation or later).
- **RAM:** Minimum 16 GB (32 GB recommended).
- **GPU:** NVIDIA GPU with at least 6 GB VRAM (e.g., GTX 1660 or higher) for accelerated model inference.
- **Storage:** Minimum 500 GB SSD; additional external storage for large audio datasets.
- **Audio Devices:** Microphone and speakers for recording and playback.
- **Internet Connectivity:** Stable broadband for downloading datasets, accessing APIs, and cloud operations.

4.2 Dataset Description

The project utilizes the **"Audio Recording Whisper"** dataset, available on Kaggle, curated by najamahmed97. The dataset comprises **272 audio recordings** in WAV and MP3 formats, each accompanied by corresponding **clean transcripts stored as .txt files**, making it highly suitable for Automatic Speech Recognition (ASR) and transcription tasks. The recordings include a variety of speech patterns, speaker tones, and durations, providing a realistic corpus for training and evaluating ASR models.

The dataset is particularly compatible with **Whisper-based models**, enabling experimentation with transformer-based speech-to-text algorithms. It serves as the primary input for generating textual transcripts, cleaning them, and subsequently producing structured documentation such as summaries or SOAP notes. Its diversity in audio quality, speaker variation, and readily available textual transcripts ensures the development of robust models capable of handling real-world audio scenarios.

Key attributes of the dataset include:

- **Number of Files:** 272 audio recordings with corresponding clean transcripts in .txt format.
- **File Formats:** WAV and MP3 audio files, .txt transcript files.
- **Content Type:** Human speech, suitable for transcription tasks.

This dataset provides realistic audio data and ready-to-use transcripts to evaluate transcription accuracy, summarization quality, and the extraction of meaningful information.

4.3 Initial Results

The initial implementation focused on building a functional prototype that could process audio data, generate accurate transcripts, and produce concise summaries. Using a dataset of 272 audio files with corresponding cleaned transcripts covering varied speech patterns and durations, the workflow progressed through stages of audio preprocessing, chunk

segmentation, transcription, evaluation, and summarization, with both quantitative and visual analyses at each stage.

I. Audio Preprocessing

The first step involved preparing the raw recordings for further processing. Each audio file underwent **noise reduction**, **silence trimming**, and **normalization** using Python libraries such as librosa and soundfile. These operations aimed to enhance clarity, maintain consistent volume levels, and remove irrelevant background portions. This not only improved the overall signal quality but also significantly reduced processing time for the subsequent stages.

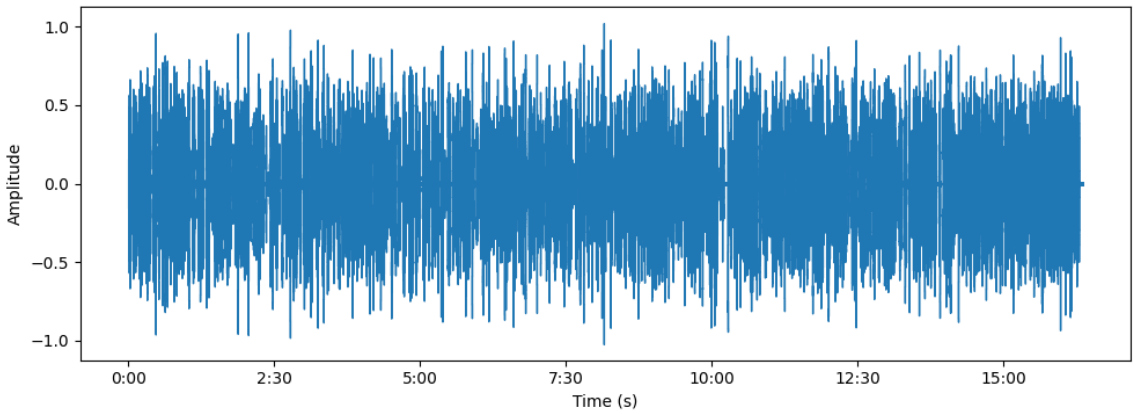

Figure 2(a): Raw Audio Waveform Before Preprocessing

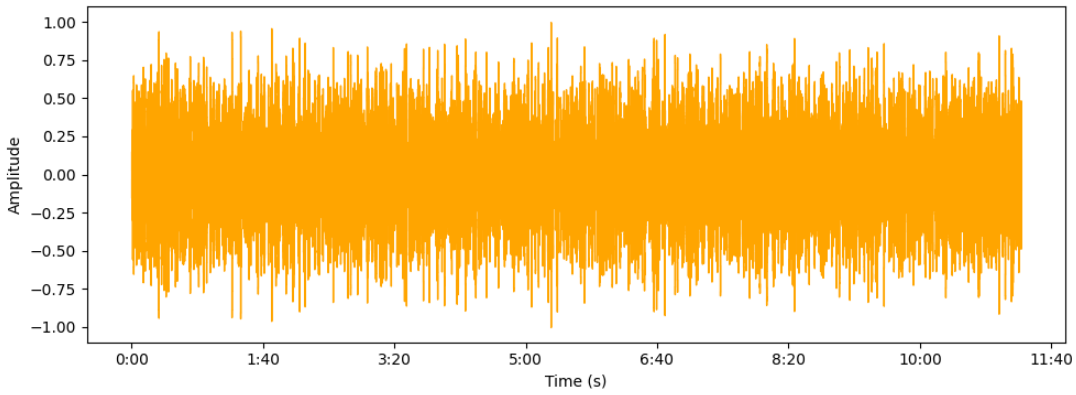

Figure 2(b): Enhanced Audio Waveform After Preprocessing

In Figure 2(a), the **raw audio waveform** shows frequent spikes in amplitude due to noise and extended flat regions corresponding to silent or inactive segments. These contribute little to actual speech recognition and increase computational load. After preprocessing, as illustrated in Figure 2(b), the waveform becomes more compact and focused on relevant speech segments. The amplitude variation stabilizes, and silent intervals are largely removed, confirming the efficiency of the preprocessing pipeline.

Quantitatively, the **average length reduction across the dataset was 29.17%**. Original audio lengths ranged from **350 to 1,100 seconds**, which were reduced to approximately **250 to 850 seconds** after preprocessing.

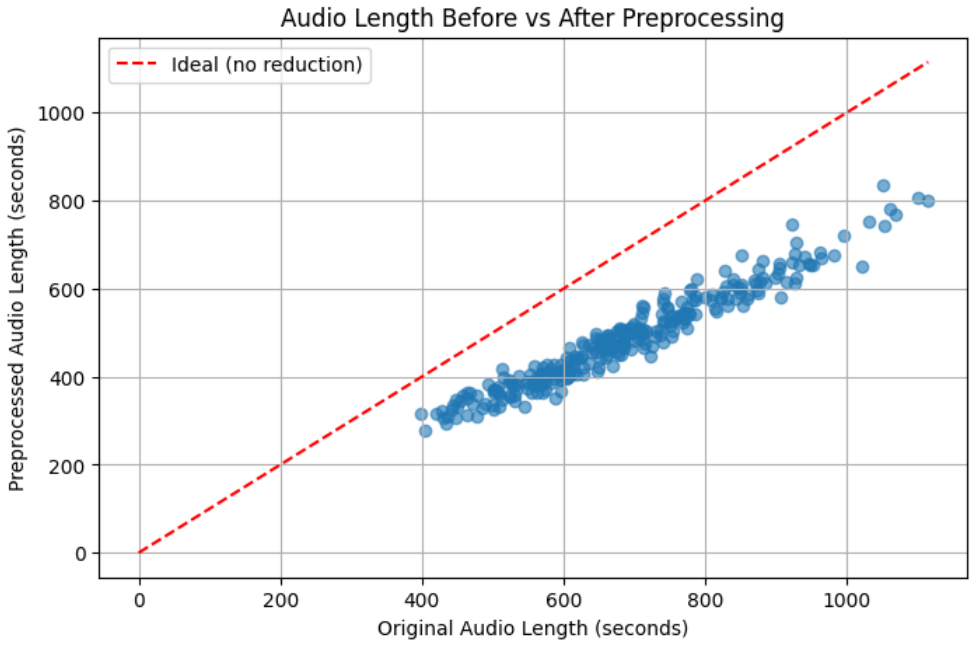

Figure 3: Audio Length Before vs After Preprocessing (Scatter Plot)

As seen in Figure 3, most points lie below the red dashed "Ideal (no reduction)" line, indicating successful shortening of recordings. Figure 4 shows that the majority of files experienced a **25–35% reduction**, demonstrating the consistent removal of redundant or non-speech segments while preserving meaningful content.

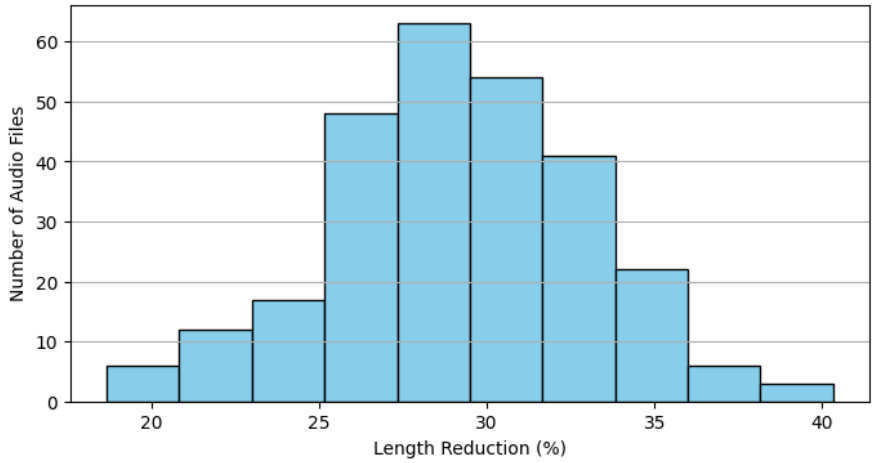

Figure 4: Distribution of Length Reduction (%) Across Dataset

Following this, as seen in Figure 4, the preprocessing not only reduces the overall audio length but also contributes to improved efficiency in downstream tasks such as chunking and transcription. Shorter, cleaner audio ensures that the segmentation into chunks is more balanced, minimizes memory usage, and allows the transcription model to focus on relevant speech segments, ultimately enhancing the accuracy and consistency of the generated transcripts.

II. Chunking and Segmentation

After cleaning, long audio files were **divided into smaller segments (chunks)** to optimize both memory usage and transcription accuracy. Each chunk represented a manageable portion of the recording (typically 30–60 seconds). This strategy ensured that even lengthy conversations could be processed without exhausting system resources and minimized the risk of transcription errors arising from lengthy context windows. Chunking also allowed **parallel or sequential processing** of segments, enabling smoother handling by the transcription model. Importantly, it ensured that silence or abrupt shifts in tone did not interfere with recognition accuracy. This segmentation proved highly beneficial during the transcription phase, leading to faster inference times and more coherent outputs. On average, each audio file was split into **6–10 smaller chunks**, depending on the total length after preprocessing.

III. Transcription Using Whisper Model

Following segmentation, the preprocessed and chunked audio files were transcribed using **OpenAI's Whisper model (medium variant)**. The model automatically converted spoken content into text while handling variations in accent, tone, and speech rate. Each chunk was transcribed separately, and the partial results were concatenated to form the final transcript for each file.

The output transcripts were then cleaned by removing filler expressions (e.g., "uh," "um," "you know") and correcting punctuation for improved readability. This stage established the foundation for further structured representation of information in later stages (see Figure 5 for an example of the Whisper model's audio-to-text conversion).

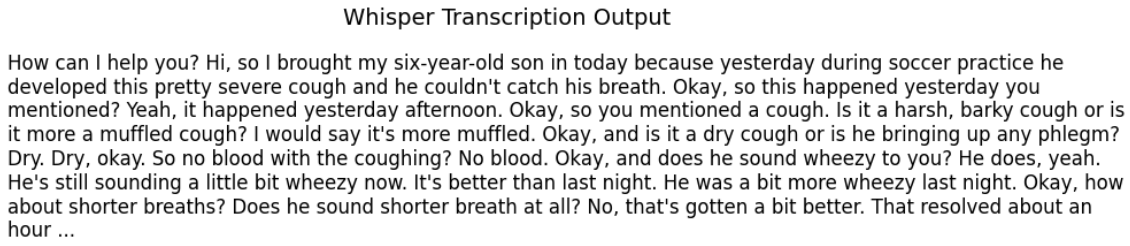

Figure 5: Example output showing Whisper model's conversion from speech to text.

IV. Transcription Accuracy Analysis (WER Evaluation)

To quantitatively measure transcription quality, the **Word Error Rate (WER)** was calculated for each transcript using the formula:

$$WER = (S + D + I)/N \times 100$$

where

S = Substitutions,

D = Deletions,

I = Insertions,

N = Total words in the reference transcript.

Across all **272 audio files**, the following results were observed:

- **Average (Mean) WER:** 0.2947 (≈29.47%)
- **Minimum WER:** 0.2168 (≈21.68%)
- **Maximum WER:** 1.0000 (100%)

This indicates that, on average, **70.5% of the words were recognized correctly**, while certain files exhibited higher errors due to noise or unclear speech.

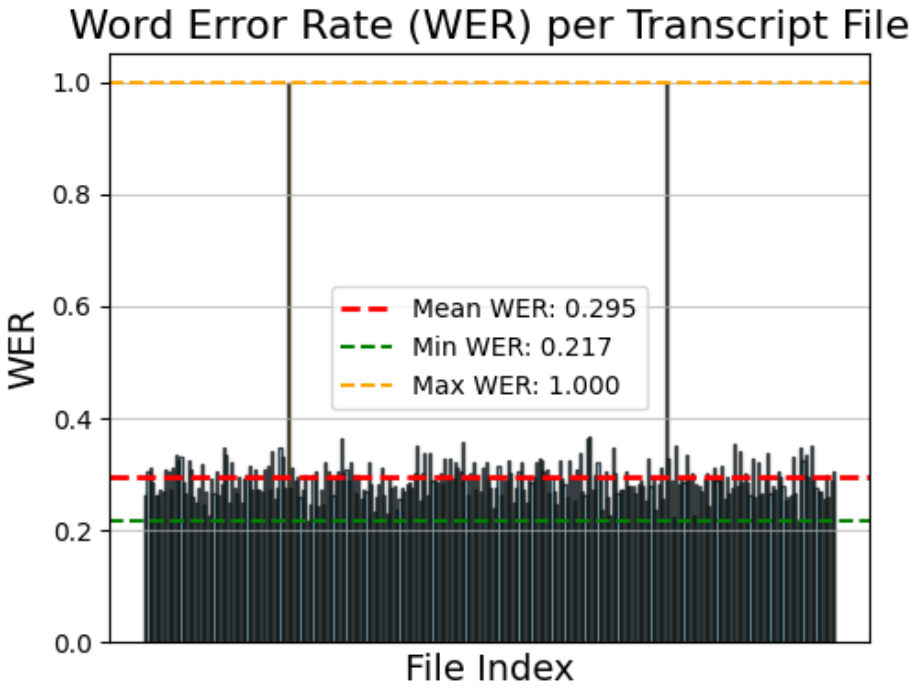

Figure 6: WER distribution across all files showing mean, minimum, and maximum values.

The bar graph in Figure 6 illustrates the performance variation across individual files. The **red dashed line** denotes the average WER (≈0.295), serving as a benchmark. The **green dashed line** represents the best case (≈0.217), and the **orange dashed line** marks the worst-performing file (WER = 1.0). Most results fall within the 0.25–0.35 range, demonstrating stable performance across diverse inputs. The few outliers correspond to heavily distorted or incomplete recordings.

V. Text Summarization

The final stage of the initial implementation involved summarizing the generated transcripts using a **BART-based transformer model**. This model was designed to compress lengthy doctor–patient conversations into concise, meaningful summaries while retaining the essential clinical information such as symptoms, observations, and relevant patient history.

Table I : Example of Original Transcript vs Generated Summary

Original Transcript (First Few Lines)	Generated Summary
D: What brought you in today? P: Sure, I'm I'm just having a lot of chest pain and and so I thought I should get it checked out. D: OK, before we start, could you remind me of your gender and age? P: Sure 39, I'm a male. D: OK, and so when did this chest pain start? P: It started last night, but it's becoming sharper. D: OK, and where is this pain located? P: It's located on the left side of my chest.	P is 39 and has chest pain. It started last night and has been going on for a couple of hours. It's sharp and on the left side of the chest. He feels lightheaded and has trouble breathing. He has no prior medical conditions or recent hospitalizations. P smokes about a pack of cigarettes a day and drinks alcohol regularly. His father had a heart attack at 45.

For testing, a single transcript was selected, and the summarizer successfully captured the core content of the conversation. It effectively extracted key details such as the patient's age, primary complaint, symptom onset and characteristics, relevant medical history, lifestyle factors (e.g., smoking and alcohol use), and pertinent family history. The generated summary demonstrated improved **coherence**, reduced **redundancy**, and

presented the information in a structured and readable manner, thereby confirming the feasibility of using automated summarization for structured documentation tasks like SOAP note preparation.

As shown in **Table I**, the summarization model distilled the first few lines of the original transcript into a compact, clinically relevant summary while preserving all key information. This demonstrates how the system can transform long dialogues into concise summaries suitable for downstream applications such as SOAP note generation and context-aware queries.

5.CONCLUSIONS AND FUTURE WORK

5.1 Conclusions

This project really shows how feasible and effective an AI-powered clinical assistant can be. It automates clinical documentation by pulling together conversational and diagnostic data in a smart way. The system called From Findings to Final Notes uses hybrid AI for clinical report summarization. It tackles key problems in healthcare documentation. Those include cutting down administrative work, boosting accuracy, and making workflows smoother for professionals in the field.

The first version of the system hit important goals in various parts of the documentation process. The audio preprocessing part cut down audio length by about 29.17% on average. It did this with noise reduction, silence trimming, and normalization. That approach removed non-speech parts efficiently. At the same time, it kept the important content intact. The chunking method helped with memory use and better transcription accuracy. It broke long recordings into chunks of 30 to 60 seconds that were easy to handle.

The transcription part relied on OpenAI's Whisper model in its medium version. It reached an average Word Error Rate of 29.47%. That means roughly 70.5% of words got recognized right across 272 audio files in the dataset. Some files had higher errors because of sound issues. Still, most transcriptions stayed in the 25 to 35% WER range. This showed steady performance with different speech styles and recording setups.

The summarization module used BART and did a good job condensing long doctor-patient talks. It turned them into short summaries that stayed clinically useful. Essential details like patient complaints, symptom details, medical history, and lifestyle factors all came through. The summaries had better flow, less repetition, and solid clinical accuracy. All this points to how well transformer models work for organized medical documentation.

The hybrid AI approach in the system mixes several advanced methods nicely. Those include Automatic Speech Recognition, Named Entity Recognition, and transformer-based summarization. Together, they form a full solution for clinical documentation. The

modular design makes it scalable and easy to maintain. It also allows for simple updates down the line. Plans to add Role-Based Access Control will handle privacy and security issues that come up in healthcare apps.

In the end, the system makes a strong case for intelligent automation in clinical work. It streamlines workflows quite a bit, cuts documentation time, lowers human mistakes, and lets professionals focus more on patient care directly. Getting the core modules to work validates the overall architecture. It also highlights the promise of AI tools to change how healthcare documentation happens.

5.2 Future Work

The starting implementation brought some good outcomes already. Even so, a few upgrades and additions could make the system stronger, more reliable, and more useful in clinical settings.

5.2.1 Complete SOAP Note Generation

Right now, the setup handles transcription and summarization mainly. Next steps will build a full module for generating SOAP notes. That means organizing summaries into the standard four parts. Those are Subjective, Objective, Assessment, and Plan. To do this, sequence-to-sequence models need training on medical datasets with labels for sections. Named Entity Recognition will spot and sort clinical items like symptoms, diagnoses, medications, and procedures. Template-based frameworks will help ensure accuracy and standard formatting in clinical terms.

5.2.2 Laboratory Report Integration

Bringing in structured lab data with conversation transcripts is a key area to tackle soon. This calls for OCR tools to handle scanned lab reports. Parsers will pull out key-value pairs from different lab formats. Entity extraction algorithms will identify test names, parameters, values, and reference ranges. Data fusion methods will then link lab results to consultation talks in time and meaning. The goal is a complete view of the patient.

5.2.3 Context-Aware Query System Implementation

An smart querying setup will let doctors and patients find information quickly and easily. This involves a Retrieval-Augmented Generation architecture for answers that fit the context. Vector similarity searches with embeddings will pull up relevant past records. Natural language understanding modules will parse user queries right. Knowledge graphs will map connections between clinical entities and aid semantic thinking.

REFERENCES

[1] A. A. Kumar *et al.*, "Improving Discharge Summary Generation through Clinical Text Summarization with QLoRA and LLMs," *Proc. 2025 6th Int. Conf. Control, Communication and Computing (ICCC)*, IEEE, pp. 1–5, 2025.

[2] B.-N. Nguyen *et al.*, "Enhancing Clinical Note Generation from Doctor-Patient Conversations through Semantic Partition-Oriented Summarization," *Proc. 2023 15th Int. Conf. Knowledge and Systems Engineering (KSE)*, IEEE, pp. 1–6, 2023.

[3] K. Krishna *et al.*, "Generating SOAP Notes from Doctor-Patient Conversations Using Modular Summarization Techniques," *arXiv preprint arXiv:2005.01795*, 2021.

[4] B. Chen and S.-H. Lin, "A Risk-Aware Modeling Framework for Speech Summarization," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 20, no. 1, pp. 211–222, 2012.

[5] X. Mei *et al.*, "RadChat: A Radiology Chatbot Incorporating Clinical Context for Radiological Reports Summarization," *Proc. 2024 IEEE Int. Conf. Bioinformatics and Biomedicine (BIBM)*, IEEE, pp. 2297–2302, 2024.

[6] C. Ma *et al.*, "An Iterative Optimizing Framework for Radiology Report Summarization With ChatGPT," *IEEE Trans. Artif. Intell.*, vol. 5, no. 8, pp. 4163–4175, 2024.

[7] B. Palanisamy *et al.*, "From Information Overload to Lucidity: A Survey on Leveraging GPTs for Systematic Summarization of Medical and Biomedical Artifacts," *IEEE Access*, vol. 13, pp. 7902–7922, 2025.

[8] T. Sultan *et al.*, "SumGPT: A Multimodal Framework for Radiology Report Summarization to Improve Clinical Performance," *IEEE Access*, vol. 13, pp. 15929–15945, 2025.

[9] M. S. Sabariram *et al.*, “An End-to-End Pipeline Integrating Segmentation, Anomaly Detection, and Report Generation for Medical Imaging Analysis,” *Proc. 2024 1st Int. Conf. Data, Computation and Communication (ICDCC)*, IEEE, pp. 65–70, 2024.

[10] P. Zelina *et al.*, “Extraction, Labeling, Clustering, and Semantic Mapping of Segments From Clinical Notes,” *IEEE Trans. NanoBioscience*, vol. 22, no. 4, pp. 781–788, 2023.

[11] G. Michalopoulos *et al.*, “MedicalSum: A Guided Clinical Abstractive Summarization Model for Generating Medical Reports from Patient-Doctor Conversations,” *Nuance Communications*, 2021.

[12] S. Sai *et al.*, “Generative AI for Transformative Healthcare: A Comprehensive Study of Emerging Models, Applications, Case Studies, and Limitations,” *IEEE Access*, vol. 12, pp. 31078–31106, 2024.

[13] M. Devarakonda *et al.*, “Problem-oriented patient record summary: An early report on a Watson application,” *Proc. 2014 IEEE 16th Int. Conf. e-Health Networking, Applications and Services (Healthcom)*, IEEE, pp. 281–286, 2014.

[14] F. Liu *et al.*, “A Supervised Framework for Keyword Extraction From Meeting Transcripts,” *IEEE Trans. Audio, Speech, and Language Processing*, vol. 19, no. 3, pp. 538–548, 2011.

From Findings to Final Notes: Hybrid AI for Clinical Report Summarization

Kotagiriwar Sriya
Department of CSE
Chaitanya Bharathi
Institute of Technology
Gandipet
Hyderabad-500075
Telangana,India
sriyakotagiriwar@gmail.com

M.S.L.Aashritha
Department of CSE
Chaitanya Bharathi
Institute of Technology
Gandipet
Hyderabad-500075
Telangana,India
mslaashritha7@gmail.com

Smt. Ch. Madhavi Sudha
Department of CSE
Chaitanya Bharathi
Institute of Technology
Gandipet
Hyderabad-500075
Telangana,India
madhavisudha_cse@cbit.ac.in

***Abstract*—Healthcare professionals waste a significant amount of time on paper work, and this reduces the time they have to interact with patients and contributes to burnout. Even the latest systems such as Electronic Health Records (EHR) and the off-the-shelf speech-to-text systems cannot literally generate structured and context-sensitive notes and cannot manipulate multimodal data audio and lab results. Thus, in this paper, an AI-aided Clinical Assistant based on the Generative Artificial Intelligence (GenAI) and Large Language Models (LLMS) was implemented to automate clinical documentation. Essentially, the system converts the verbal communication between the doctor and patient to text, lab reports to normalization, and generates neatly formatted SOAP (Subjective, Objective, Assessment, Plan) notes. To generate summaries, we are employing the models of transformers such as BERT and BART, and to extract clinical terminology, we are employing the model of Named Entity Recognition (NER). And top of all the interface allows one to query patient data in real-time. With chat data and diagnostic information, more accurate documentation is provided, and more work is facilitated in clinical practices, which must enhance health outcomes.**

***Index Terms*—Generative AI, Large Language Models, Clinical Documentation, SOAP Notes, Multimodal Data Fusion, Named Entity Recognition, BERT, BART, Context-Aware Queries, Healthcare Automation**

I. Introduction

The artificial intelligence boom in the field of medicine is beginning to turn the paradigm of how physicians approach digital technologies, indicating that we absolutely require intelligent devices to automate all those finer-grained health care processes. This, as a result, has led to clinicians spending the majority of their time on paper-work rather than spending time attending to patients, thereby only contributing to their inefficiencies, burnout and even costs to the quality of care provided patients and the quality of the records. Available documentation systems, such as believe EHRs and elementary speech-to-text, have been predominantly manual typing and a limited speech-to-text. They are out of context when it comes to doctor-patient conversations, missing important clinical observations and unable to combine multimodal data, such as laboratory results. This means that the records on patients remain discontinuous and their ability to make informed choices and care over the long run is effectively imprisoned.

In order to address this, we will be proposing a Clinical Assistant powered by AI which integrates in GenAI and LLMs to automate and enhance clinical documentation. The tool logs and transcripts physician-patient conversations, extracts vital data in lab reports and glues all the entries together into well-formatted SOAP reports. The assistant provides both contextual and semantic accuracy by drawing in NER to supply medical vocab and using transformer summarizers such as BERT and BART. In addition, the site has a query interface that makes use of a context-sensitive query to enable clinicians to dig into patient data, pose follow-up questions, and retrieve useful information in real time. Such combination of conversational intelligence, multimodal processing of data, and intelligent retrieval have established a new standard in documentation, which reduces the amount of manual effort, enhances the completeness of the data, and provides decision-makers with the timely and evidence-based support.

The paper will describe the ability of the system to create accurate and structured documentation and the manner in which it used various sources of data to weave the information. When compared with the conventional EHR systems, it can be seen that the integration of the GenAI-based summarization, entity detection system, and context-driven retrieval actually improves the quality of documentation and workflow efficiency.

II. Related Work

The growing administrative load on healthcare professionals has sparked a lot of research interest in automating the creation of clinical documentation. Managing multimodal data sources, preserving clinical accuracy, and following standardized documentation formats while transforming highly variable, conversational, and occasionally noisy patient-provider interactions into cohesive, structured medical records is the main challenge.

A. Semantic Structuring and Section-Aware Processing

Creating high-quality automated notes requires that clinical data be arranged in accordance with accepted medical documentation standards. It has been demonstrated that using semantic-

based partitioning to preprocess clinical dialogues greatly enhances the extracted information's coherence and relevance [1]. Sentences are assigned to the appropriate SOAP sections using allocation algorithms after semantic representations are created using Sentence-BERT and other sentence embedding techniques.

Postprocessing techniques improve outputs to make them look more like real clinical writing styles by making disease names the same and changing spoken numbers into written numbers. Experiments have shown that semantic-based allocation works much better than manually made query-based methods on benchmark datasets like MEDIQA-Sum 2023, which leads to big improvements in ROUGE scores.

The CLUSTER2SENT algorithm is much better than these extractive methods because it groups related statements together, pulls out important utterances that are relevant to each SOAP section, and makes one summary sentence for each group [7]. When it comes to making sentences that are coherent and true, this method works much better than purely abstract models. When using pre-trained models like T5-base, human evaluators often found that CLUSTER2SENT-based methods made sentences that were more useful than fully abstractive baselines.But there are still problems, such as the possibility of making up information, problems with pronoun resolution, and contradictions that come up when people summarize localized conversation.

Complementary research in unsupervised clinical text processing has focused on organizing and extracting information from unstructured medical records, especially in computationally underrepresented languages [2]. These methods use algorithms that look for patterns in the structure of clinical notes to break up records into digestible sections. It has been tested whether different classification models, such as transformer architectures like RoBERTa and LSA-based methods, can infer segment titles from content. These systems support semantic mapping to standardized ontologies, allow complex extraction pipelines, and aid in the analysis of patient similarity by category in addition to being useful for classification. Clinical experts claim that these methods significantly integrate clinical note segments and expedite information retrieval.

B. Domain Knowledge Integration and Medical Correctness

Ensuring that generated clinical summaries accurately reflect medical terminology remains fundamental, as errors can have serious consequences for patient care [5]. The MedicalSum model demonstrates how knowledge augmentation can be achieved through three mechanisms: incorporating medical words as guidance signals, leveraging UMLS semantic type knowledge to create clinically meaningful embeddings, and employing weighted loss functions that prioritize correct prediction of medical terminology.

This domain-aware training strategy ensures models prioritize clinical accuracy. Empirical results demonstrate substantial improvements, with MedicalSum achieving 0.8-2.1 point gains in ROUGE scores and a 6.2% error reduction in the Physical Examination section. These gains are significant given that medical summarization must balance generating fluent text while precisely capturing technical medical content. However, the authors emphasize that while such models can help reduce physician burnout, they should only assist, not replace, trained clinical professionals due to risks of omitting key information or hallucinating unsupported clinical facts.

C. Multimodal Data Integration and End-to-End Pipelines

Modern healthcare generates data across multiple modalities, including spoken consultations, handwritten notes, laboratory results, and radiological images. Recent work has demonstrated that comprehensive pipelines combining OCR technology, speech recognition APIs, and fine-tuned language models can effectively process heterogeneous clinical inputs [3]. These systems employ parameter-efficient fine-tuning techniques such as Quantized Low-Rank Adaptation (QLoRA) to adapt large language models for medical domains. The approach integrates image-to-text modules utilizing PyTesseract OCR with t5-base models, speech-to-text modules employing AssemblyAI API, and radiology report processing components to generate structured discharge summaries.

Evaluation using ROUGE scores indicates effectiveness in capturing key information, with structured outputs stored in secure databases for efficient patient information management. However, significant limitations persist, particularly in handling handwritten notes and managing long or noisy text sequences. Despite these advances, multimodal integration remains an active research challenge requiring continued refinement to achieve clinical-grade reliability.

D. Contextual Understanding and Conversational Reasoning

Effective clinical reasoning requires understanding not just individual utterances but their relationships, temporal ordering, and connections to patient history [4]. RadChat exemplifies this by reframing radiology report summarization as a conversational task, generating "temporal radiological conversations" where findings function as questions and impressions serve as answers. This paradigm allows systematic integration of information from previous medical examinations.

The chatbot is developed by fine-tuning the Vicuna large chat model using these constructed conversations. Experimental validation on the MIMIC-CXR dataset demonstrates that RadChat outperforms state-of-the-art baselines across multiple dimensions. Human evaluation confirms RadChat's superior correctness, fluency, and diversity in generated impressions, remarkably even surpassing human reference summaries in diagnostic diversity. Ablation studies show substantial performance degradation when historical information is excluded, confirming that temporal context is essential for accurate clinical summarization. Future work aims to integrate actual radiological images to further enhance diagnostic accuracy.

E. Generative Pretrained Transformers in Medical Summarization

The broader application of Generative Pretrained Transformers (GPTs) to medical document summarization has

been extensively surveyed, revealing both substantial promise and significant challenges [6]. The rapid proliferation of medical research makes it increasingly difficult for healthcare professionals to stay current. GPTs offer potential solutions by distilling complex information into concise summaries, with applications across information extraction, automated summarization, clinical decision support, and research trend analysis.

Notable examples include GPT3-D3 for biomedical research summarization and GPT-4 for comprehensive medical record summarization. However, critical challenges persist. Ensuring faithfulness and accuracy remains paramount, as models frequently exhibit factual inconsistencies and may generate misleading information—particularly dangerous in healthcare contexts. The ”black-box” nature of transformer models raises interpretability concerns, especially given regulatory requirements for explainable AI in clinical settings. Data privacy and ethical considerations pose additional barriers, as medical data is highly sensitive and subject to strict regulatory frameworks.

Future directions include better integration into research database query systems, domain-specific refinement through continued pre-training on medical corpora, enhanced contextual comprehension for long clinical documents, and multimodal functionalities combining text with imaging and genomic data.

F. Evaluation and Research Gaps

Assessing automatically generated clinical documentation requires metrics capturing both linguistic fidelity and clinical appropriateness. Standard metrics like ROUGE provide useful benchmarks but may not fully reflect clinical utility [5, 7]. Research has shown that focusing exclusively on surface-level similarity can lead to systems that reproduce template-like text without capturing nuanced clinical reasoning.

Human evaluation by clinical professionals remains the gold standard, as clinicians can assess correctness, completeness, and clinical reasoning in ways that automated metrics cannot [4, 7]. Studies employing physician reviewers reveal that models excelling in ROUGE scores do not always produce the most clinically useful outputs. Evaluation frameworks combining automated metrics with human assessment provide more comprehensive quality measures.

Despite significant progress, multiple challenges remain limiting real-world deployment. Current approaches often excel in isolated aspects but struggle to simultaneously optimize across all dimensions [1, 3, 5]. Temporal reasoning and longitudinal patient understanding require deeper investigation [4]. Trustworthiness concerns pose perhaps the most critical barrier—healthcare professionals hesitate to rely on systems that may hallucinate symptoms, omit critical findings, or introduce subtle errors [5, 6]. Most existing work focuses on English-language clinical text, leaving substantial gaps in supporting global healthcare systems [2]. Scalability and computational resource constraints also limit deployment in resource-constrained settings [6].

These limitations collectively motivate comprehensive systems that holistically integrate semantic organization, multimodal data fusion, contextual reasoning, domain knowledge, and robust evaluation frameworks. The goal is to produce clinical documentation tools that not only reduce administrative burden but actively support clinical decision-making through accurate, complete, and verifiable automated note generation that clinicians can trust and integrate seamlessly into existing care delivery workflows.

III. Proposed Method

The architecture suggested will be used to automate clinical documentation using multimodal data about patients (audio-consultation and lab reports). The application uses speech recognition, natural language processing (NLP), and structured data processing to create standardized SOAP (Subjective, Objective, Assessment, and Plan) notes. Such notes are safely stored in a central database and can be accessed via a context-sensitive query system to a clinical decision support system.

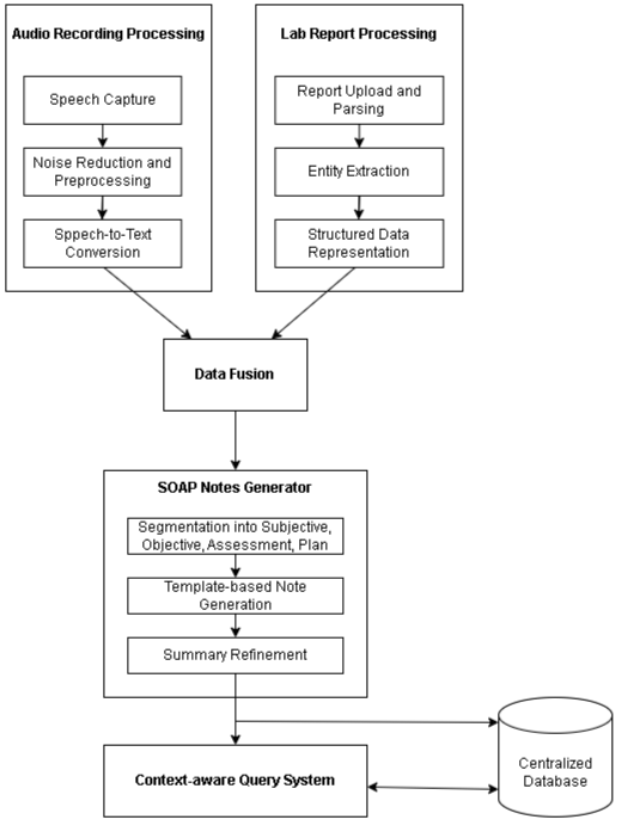

Fig. 1: Overview of the proposed system.

A. Audio Recording Processing

This module records and analyses doctor patient conversations. It uses libraries like Librosa and PyDub to do noise reduction, normalization and segmenting with the goal of increasing clarity. The audio will then be processed by way of a Whisper, an ultra-modern model of automatic speech recognition (ASR). There is post-processing that involves

removal of filler words as well as punctuation marks that restore punctuations to produce coherent medical transcripts.

B. Lab Report Processing

This element deals with Scanned and semi-structured information in lab reports, such as PDF and CSV. Methods like Optical Character Recognition (OCR) and Named Entity Recognition (NER) are used to identify clinical entities (e.g., test results, biomarkers). Controlled vocabularies like LOINC and SNOMED CT are used to normalize the data to allow interoperability with textual data.

C. Data Fusion

Data Fusion module brings together two heterogeneous inputs, i.e. transcription of dialogues, and structured reports into an integrated patient record. It carries out the time match, de-duplication, and map schema in order to ensure consistency of the data. The merged dataset is the basis of creating detailed and clinically-relevant summaries.

D. SOAP Notes Generator

It employs NLP pipelines, as well as certain slick simplistic summarizing using templates. It simply tags all over and divides it into SOAP fragments, and retrieves contextual summaries using Transformers such as BERT or T5. The final outcome is an un-polluted, legible files which are ready to drop directly into the EHR, no effort.

E. Centralized Database

All these SOAP notes are pasted in either postgreSQL or MongoDB that suits the data. It has indexing making lookups fast, encryption to ensure everything is locked and audit logs to ensure you will always know who looked into what.

F. Context-Aware Query System

It's like a search engine that sniffs with the help of embeddings on how to find the correct records. Once a clinician hits a button with a question, the Natural Language Understanding (NLU) springs into action, determines the intent and retrieves the most in the database as SOAP notes- it is so efficient and easy to use when clinicians need quick information.

IV. Initial Implementation

The implementation phase in this project converts the proposed architecture to a functioning prototype that automates the clinical documentation. It deals with a corpus of audio doctor-patient records and clean transcripts. Systematically sorting out the data to preprocess and evaluate, we use the libraries of Python, namely the os and pathlib libraries. Audio files are improved with noise cancellation, amplitude equalization, and cutting silences using librosa and soundfile, which facilitate better sound and ensure the quality of audio. To make transcription more intelligent and memory consumption smarter, the purged recordings are divided into smaller chunks.

To perform the transcription, we apply the Whisper model to transparent medical conversations and convert the information into text produced by the OpenAI Whisper model. This is smoothened out by the elimination of filler words and the incorporation of punctuation to facilitate easier reading.

Such a preliminary installation puts in place the main former functionality of the system, where audio-to-text conversion can be made dependable, quantitative evaluation through WER, and content summarization can be performed thus providing the basis to further stages of integration, where multimodal data can be added to the system, notes can be structured, and queries must be smarter.The quality of transcription is assessed using the Word Error Rate (WER) metric, calculated as:

$$\text{WER} = \frac{S + D + I}{N} \times 100 \tag{1}$$

where S is the number of substitutions, D is the number of deletions, I is the number of insertions, and N is the total number of words in the reference transcript. WER is computed per transcript file and then averaged across the dataset to obtain the mean WER, providing a quantitative measure of model performance. Fig. 2 illustrates the WER for each transcript file, highlighting file-wise variation along with mean, minimum, and maximum WER values.

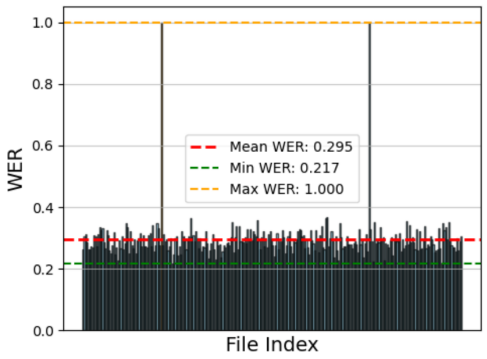

Fig. 2: Word Error Rate (WER) per transcript file showing individual file performance and summary statistics (mean, min, max WER).

The system also provides brief summaries of the transcripts based on summarizing models like BART and extracts important information in the text being transcribed. These summaries should serve as the initial stage in organizing data to be used in subsequent stages of the workflow, such as SOAP note creation.

This initial implementation establishes the core functionality of the system enabling accurate audio-to-text extraction, quantitative evaluation through WER, and content summarization, forming the foundation for later phases involving multimodal data integration, structured note generation, and intelligent querying.

V. Conclusion

The system offers a powerful and adaptable solution capable of automating clinical documentation using a mixture of speech recognition and NLP-based post-processing.

This addresses disfluencies, filler words, and medical jargon, thus the system creates the readable, clear transcripts and formal reports that are accurate to the context. The evaluation of its performance is done using such measures as Word Error Rate (WER) and semantic correctness that reveal that the method proves to be reliable and effective.

This solution will facilitate more effective clinical workflows, enhance accuracy of records and facilitate context-aware decision-making. It also preconditions the possibility of the further extensions such as multilingual transcription, real-time summarization, and seamless connection with electronic health record (EHR) systems.

ACKNOWLEDGMENT

We would like to thank and acknowledge Smt Ch.Madhavi Sudha of CSE Dept, Chaitanya Bharathi Institute of Technology for providing continuous support.

REFERENCES

[1] B. -N. Nguyen, H. -Q. Le and D. -C. Can, "Enhancing Clinical Note Generation from Doctor-Patient Conversations through Semantic Partition-Oriented Summarization," 2023 15th International Conference on Knowledge and Systems Engineering (KSE), Hanoi, Vietnam, 2023, pp. 1-6, doi: 10.1109/KSE59128.2023.10299512.

[2] P. Zelina, J. Halámková and V. Nováček, "Extraction, Labeling, Clustering, and Semantic Mapping of Segments From Clinical Notes," in IEEE Transactions on NanoBioscience, vol. 22, no. 4, pp. 781-788, Oct. 2023, doi: 10.1109/TNB.2023.3275195.

[3] A. A. Kumar, A. Abraham, P. Niranjan, S. Soman and A. Varghese, "Improving Discharge Summary Generation through Clinical Text Summarization with QLoRA and LLMs," 2025 6th International Conference on Control, Communication and Computing (ICCC), Thiruvanathapuram, India, 2025, pp. 1-5, doi: 10.1109/ICCC64910.2025.11077215.

[4] X. Mei, L. Yang, D. Gao, X. Cai, T. Liu and J. Han, "RadChat: A Radiology Chatbot Incorporating Clinical Context for Radiological Reports Summarization," 2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), Lisbon, Portugal, 2024, pp. 2297-2302, doi: 10.1109/BIBM62325.2024.10822010.

[5] G. Michalopoulos, K. Williams, G. Singh and T. Lin, "MedicalSum: A Guided Clinical Abstractive Summarization Model for Generating Medical Reports from Patient-Doctor Conversations," in *Findings of the Association for Computational Linguistics: EMNLP 2022*, Abu Dhabi, UAE, 2022, pp. 4741–4749, doi: 10.18653/v1/2022.findings-emnlp.349.

[6] B. Palanisamy, A. Chakrabarti, A. Singh, V. Hassija, G. S. S. Chalapathi and A. Singh, "From Information Overload to Lucidity: A Survey on Leveraging GPTs for Systematic Summarization of Medical and Biomedical Artifacts," in *IEEE Access*, vol. 13, pp. 7902–7922, 2025, doi: 10.1109/ACCESS.2024.3521596.

[7] K. Krishna, S. Khosla, J. P. Bigham and Z. C. Lipton, "Generating SOAP Notes from Doctor-Patient Conversations Using Modular Summarization Techniques," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics (ACL 2021)*, 2021, arXiv:2005.01795 [cs.CL].