Journal of Electronics, Electromedical Engineering, and Medical Informatics
Homepage: jeeemi.org; Vol. 8, No. 4, October 2026, pp: 1371-1389 e-ISSN: 2656-8632

RESEARCH ARTICLE OPEN ACCESS

Structured Nursing Handover Report Generation from Clinical Speech using Fine-Tuned XLSR-53 and T5: A Benchmarking Study

Sasikala D[1,2], Siva Sathya S[1], Niranjan Kumar D[2], and Vignesh S[2]
[1] Department of Computer Science, School of Engineering and Technology, Pondicherry University, Puducherry, India.
[2] Department of Computer Science and Engineering, School of Computing, Amrita Vishwa Vidyapeetham, Chennai, India.

Corresponding author: Sasikala D (e-mail: sasikalaradhasri.puscholar2019@gmail.com), **Author(s) Email**: Siva Sathya S (e-mail: ssivasathya@pondiuni.ac.in), Niranjan Kumar D (e-mail: niranjankumarduraimurugan@gmail.com), Vignesh S (e-mail: vigneshshiva096@gmail.com)

Abstract **Accurate nursing handovers are critical for patient safety, as miscommunication during shift transitions leads to irreversible clinical errors. This work proposed an end-to-end pipeline that converts unstructured clinical nursing speech into standardized handover reports using a fine-tuned XLSR-53 acoustic model and T5-base text-to-text transformer. An Australian English clinical corpus of 200 synthetic nursing handover recordings from the CSIRO data access portal was utilised in this work. This benchmarking study was conducted within the CSIRO synthetic Australian English nursing handover corpus and does not represent a broad cross-domain clinical ASR benchmark. A domain-specific benchmarking study across seven state-of-the-art ASR architectures (Whisper Tiny/Base/Small, Wav2Vec2 Base/Large, HuBERT Large, XLSR-53) was conducted using this corpus. The experimental results further revealed XLSR-53 as the optimal architecture for clinical nursing speech recognition. A partial layer-freeze strategy was adopted in XLSR-53 by freezing the first 12 of 24 encoder layers, empirically validated through an ablation study with five freeze configurations (L=0, 6, 12, 18, 24). XLSR-53 preserves cross-lingual phonetic representations while enabling clinical vocabulary adaptation. A clinically motivated evaluation framework using curated medical vocabulary terms computes Medical Precision, Recall, and F1-Score along with standard WER, CER, and PER to assess reliability in clinical term recognition. Benchmarking against Google Health AI's MedASR zero-shot revealed that the proposed system XLSR-53 (L=12) achieved 17.15% WER against MedASR's 28.87% ($p<0.001$) with a Medical F1-Score of 0.98 and ROUGE-L of 0.92. Although results were obtained on synthetic Australian English speech, performance under real clinical conditions with background noise, overlapping speakers, and spontaneous interruptions requires further validation.**

Keywords **Automatic speech recognition, XLSR-53, clinical NLP, Nursing handover, T5 transformer, Partial layer freezing, Medical speech recognition, Patient safety**.

I. Introduction

Persistent reliance on unstructured oral clinical handovers by healthcare professionals during shift transition introduces information attrition, where critical clinical details such as medication dosages, vital signs, and diagnostic findings are lost due to time pressure, environmental noise, and cognitive overload [1][2][3][4]. Studies report that between 37% and 70% of serious in-hospital adverse events are directly attributable to handover communication failures [3][5]. Despite the usage of structured reporting frameworks such as ISBAR, SBAR, and SOAP in reducing handover errors [6][7], their routine adoption in nursing practice remains limited due to the documentation burden imposed on nursing staff [8][9][10].

Automatic Speech Recognition (ASR) offers a feasible solution by converting spoken handover information into structured written records in real time, thus reducing the clinical documentation workload without disrupting clinical workflow [11][12][13][14][15]. However, general-purpose ASR systems trained on standard English benchmarks are not reliable for clinical environments [13][14][15]. The existing clinical ASR systems were trained predominantly on physician dictations and fail to generalise to nursing speech [13]. No prior study has benchmarked cross-lingual ASR architectures specifically for nursing handover documentation or proposed an end-to-end speech-to-structured-report pipeline evaluated with nursing-specific clinical metrics. The proposed study addresses the following issues:

Manuscript received 20 May 2026; Revised 15 June 2026; Accepted 30 July 2026; Available online 15 August 2026
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v8i4.1778

A. ASR Architecture Benchmarking for nursing speech: A comparative fine-tuning evaluation of seven state-of-the-art ASR architectures was conducted on the CSIRO synthetic nursing handover corpus, confirming XLSR-53 as the optimal architecture for clinical nursing speech recognition. Unlike prior clinical ASR benchmarks that evaluated physician dictation or general clinical speech [13][14][15], this is the first benchmarking study conducted specifically on nursing handover speech.

B. Partial Layer-Freeze Adaptation Strategy: A fine-tuning strategy that freezes the first 12 of 24 encoder layers of XLSR-53 was employed, yielding statistically significant improvement over full fine-tuning across WER, CER, PER, and Medical F1-Score ($p<0.001$, large effect sizes). This is the first application of partial encoder layer freezing to clinical nursing speech recognition, preventing catastrophic forgetting of cross-lingual phonetic representations during domain adaptation (Section III-A).

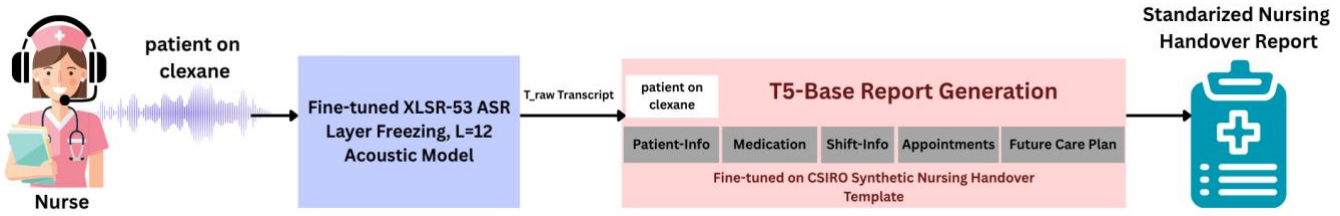


Fig. 1. Overview of the proposed end-to-end Nursing Handover Report generation pipeline from unstructured clinical speech

C. End-to-End Clinical Speech-to-Report Pipeline: An integrated framework combining fine-tuned XLSR-53 with T5-base generates standardised nursing handover documentation from unstructured clinical speech (Fig. 1) evaluated using Medical Precision, Recall, and F1-Score alongside standard WER. Unlike the existing medical ASR systems trained on physician dictations, which fail to generalise to nursing handover speech (28.87% WER), the proposed system achieved 17.15% WER demonstrating that nursing-specific domain adaptation cannot be substituted by physician-trained medical ASR models.

The detailed flow of this work is organised as follows: Section II reviews related work. Section III details the proposed methodology. Section IV describes the experimentation details and evaluation protocol. Section V presents results obtained from the experiments. Section VI discusses clinical implications and limitations. Section VII concludes current study with future research directions.

II. Related Works

A. Medical Speech Recognition and Data Scarcity

AI-assisted documentation has been proposed to reduce the nursing documentation burden, with ASR identified as the most direct technical pathway [17][18][19]. However, traditional ASR systems fail in clinical environments due to complex medical terminology, ambient noise, and clinical abbreviations [13][14][15]. Three major challenges in developing clinical ASR are:

1. Clinical speech is complex with medical terminology and abbreviations that general ASR architectures struggle to recognize accurately.
2. Standard evaluation metrics such as Word Error Rate (WER) assess transcription reliability uniformly across all words. There is no mechanism to prevent clinically dangerous substitutions, thus producing a clinically unsafe record.
3. Patient privacy regulations including HIPAA and GDPR prevent the release of real clinical audio samples, creating a severe shortage of annotated medical speech corpora for research.

Supervised and Self-supervised learning has advanced domain adaptation substantially, yet usability in real clinical settings remains constrained by medical concept error rates and high subscription costs [20][21] ASR errors in transcribing clinical terms, medication names, and procedure names can cause misdiagnosis and incorrect treatment plans [15]. Pre-trained ASR architectures including Whisper, wav2vec2.0, HuBERT, and XLSR-53 have been fine-tuned on medical datasets to improve robustness to specialized vocabulary and accented speech [22][23][24][25][26][27]. Patient privacy regulations prohibit the release of real clinical speech audio, limiting the public availability of large annotated corpora to benchmark clinical ASR tasks. The existing evaluation metrics assess clinical terms with uniform weights, failing to account for the severe impact of medical term errors.

B. Cross-Lingual Speech Representations and Layer Freezing: XLSR-53 is pre-trained on 56,000 hours of speech data across 53 languages, enabling cross-lingual transfer that significantly outperforms monolingual pre-training on low-resource speech recognition tasks [28][29]. Layer-specific fine-tuning studies show that adapting only the first 12 layers offers better performance than full fine-tuning, while reducing

Manuscript received 20 May 2026; Revised 15 June 2026; Accepted 30 July 2026; Available online 15 August 2026
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v8i4.1778

out-of-domain forgetting. Freezing the convolutional front-end with two-phased training further outperforms full fine-tuning on out-of-domain samples, and freezing early layers with a small classification head is recommended for low-resource adaptation. Accent diversity poses a significant challenge for ASR systems [25][26]. Despite these findings, partial encoder layer freezing in XLSR-53 has not been applied to clinical medical speech recognition. This work addresses that gap by demonstrating that freezing first 12 encoder layers preserves cross-lingual phonetic representations and prevents catastrophic forgetting of clinical vocabulary during domain adaptation [28][30].

C. Transformer-Based Clinical Report Generation: T5 is widely used in generating clinical note summaries [31][32][33]. Domain adaptive pre-training on MIMIC-III nursing progress notes improves clinical concept recognition [34] in generated summaries. T5 model fine-tuned on nursing notes demonstrates promising results for discharge summary generation. However, these works address physician-generated text rather than spoken nursing handover information. Large language models including Clinical Camel [35] and BioMistral [36] have demonstrated strong clinical NLP capabilities, but are not designed for speech-driven nursing handover report generation from unstructured audio. Clinical relationship extraction from discharge summaries using LLMs [37][38] requires structured text input, further motivating a speech-first pipeline approach. Currently, there is no integrated pipeline that combines cross-lingual acoustic models with transformer-based linguistic structuring for nursing handover report generation evaluated using clinically motivated metrics, which this work addresses.

III. Methodology

The proposed framework converts unstructured clinical nursing speech into standardized handover reports in two-stages. (1) Acoustic Model Layering for reliable transcription and (2) Linguistic Structuring for structured report generation. Fig. 2 represents the detailed experimental pipeline and Fig. 3 represents the architectural diagram of the proposed nursing handover report generation from unstructured clinical speech.

A. Acoustic Model Layering: XLSR-53 with Partial Layer Freezing: XLSR-53 was selected as the foundational acoustic model based on benchmarking results across seven state-of-the-art (SOTA) ASR architectures on clinical nursing speech. A persistent challenge in fine-tuning large pre-trained ASR models for specialized domains is catastrophic forgetting

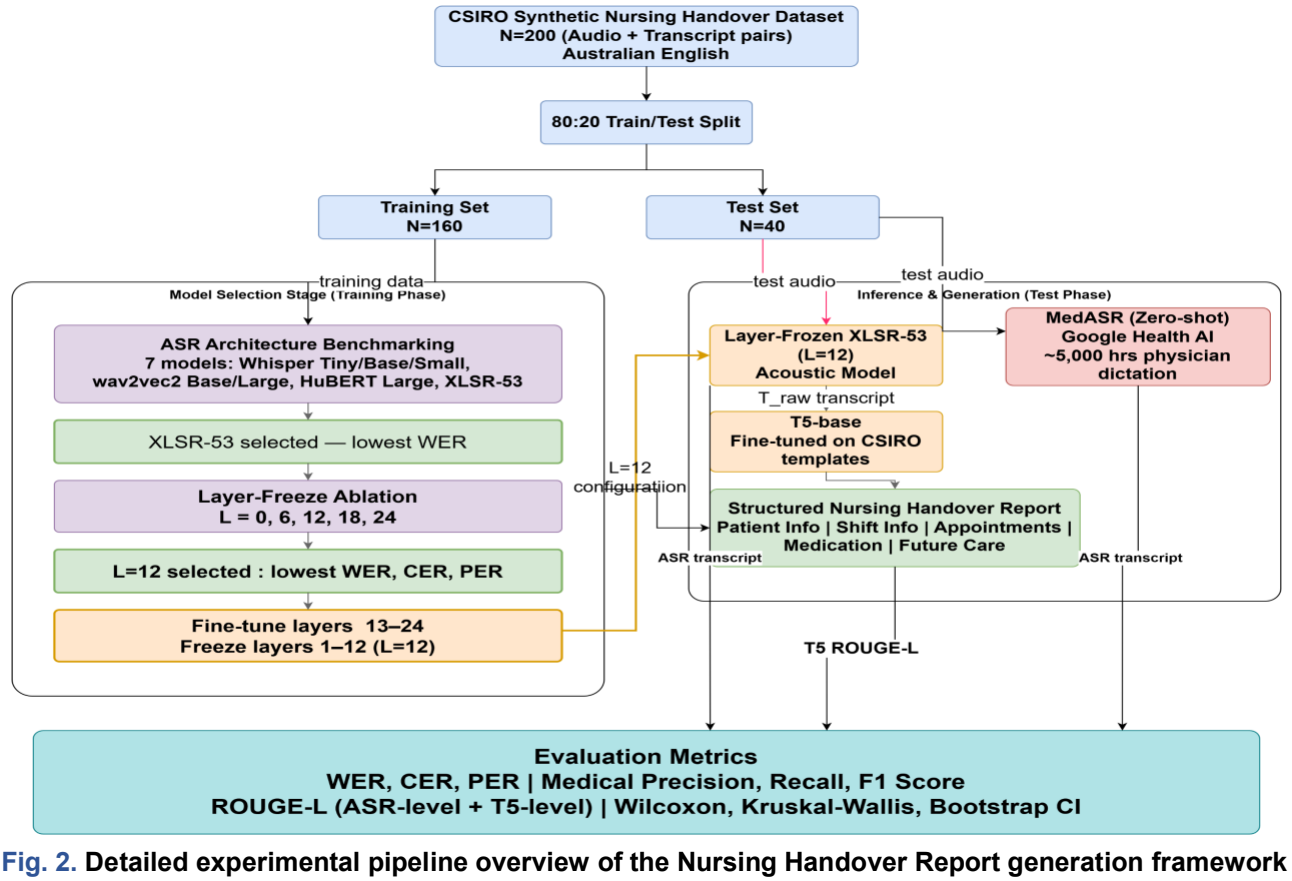


Fig. 2. Detailed experimental pipeline overview of the Nursing Handover Report generation framework from unstructured clinical speech

Manuscript received 20 May 2026; Revised 15 June 2026; Accepted 30 July 2026; Available online 15 August 2026
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v8i4.1778

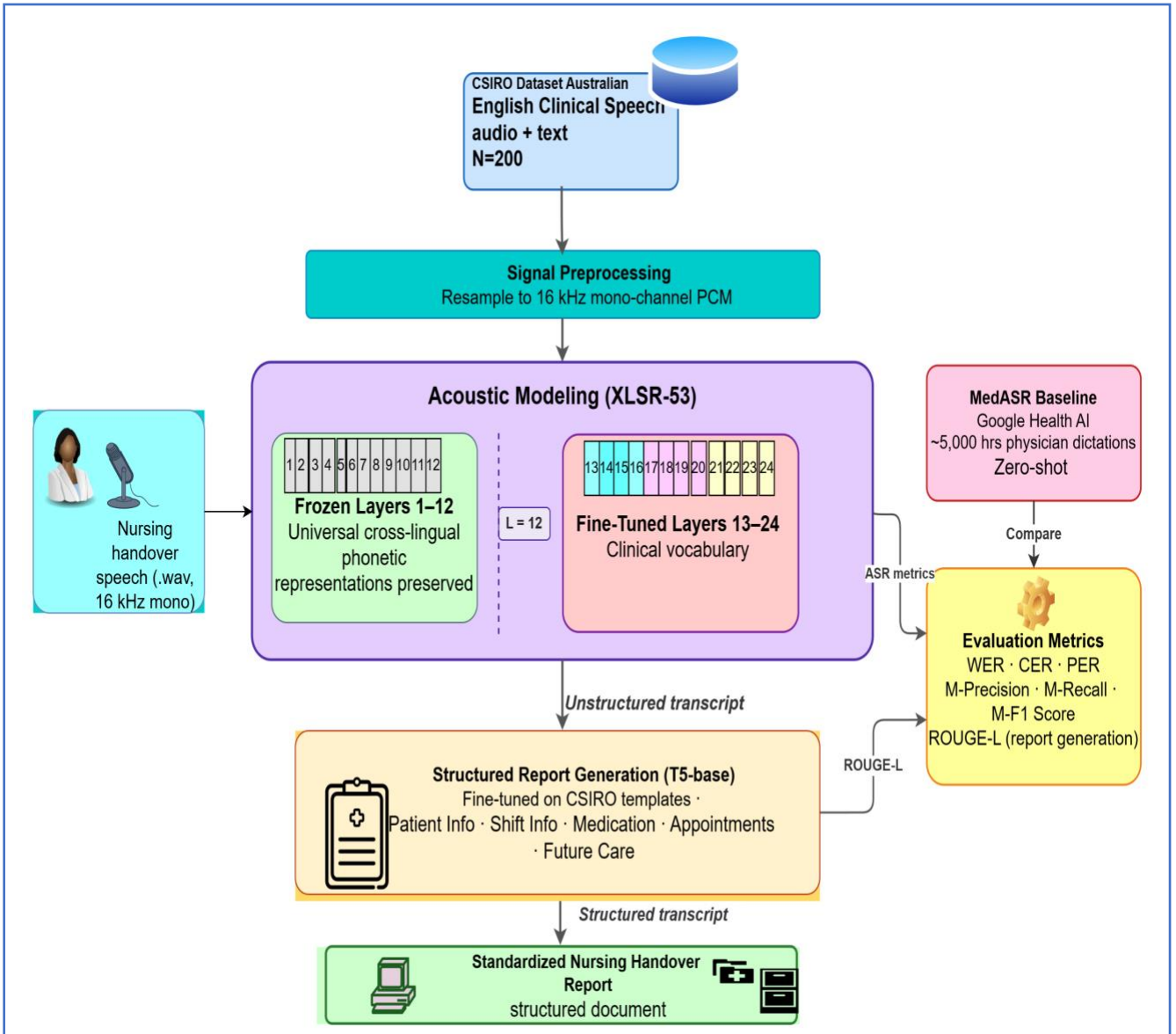


Fig. 3. Detailed architecture of the proposed two-stage Nursing Handover Report generation framework from unstructured clinical speech

during domain-specific adaptation [30]. XLSR-53's encoder [28] exhibits a hierarchical phonetic organization. Lower layers encode fundamental acoustic features such as pitch, duration, and basic phoneme boundaries, while upper layers encode higher-level lexical and semantic representations. Freezing first 12 of 24 layers during fine-tuning preserves foundational cross-lingual representations while the remaining layers adapt for clinical vocabulary. To identify the optimal freeze depth, fine-tuning experiments were conducted across freeze configurations of L = 0, 6, 12, 18, and 24 layers as an ablation study. The XLSR-53 first 12 of 24 layers freezing along with the layers 13-24 with the CTC fine-tuning on the domain-specific clinical speech data yields the lowest WER and highest Medical F1-Score of 0.98. Therefore, it was selected as the freeze boundary. The partial layer-freeze fine-tuning objective Eq. (1) [30] is defined as

$$\theta^* = \arg\min_{\theta_{\text{upper}}} \mathcal{L}\left(x, y; \theta_{\text{frozen}}, \theta_{\text{upper}}\right) \quad (1)$$

where θ_{frozen} represents the fixed parameters of encoder layers 1–12 and θ_{upper} represents the trainable parameters of layers 13–24. The Connectionist Temporal Classification (CTC) loss Eq. (2) [39] used for fine-tuning XLSR-53 is defined as

$$\mathcal{L}_{\text{CTC}} = -\log P\left(y|x\right) = -\log \sum_{\pi \in \mathcal{B}^{-1}(y)} P\left(\pi|x\right) \quad (2)$$

where π represents all valid CTC alignment paths, B is the CTC collapse function, x is the acoustic input, and y is the target transcription. Fine-tuning was performed

Manuscript received 20 May 2026; Revised 15 June 2026; Accepted 30 July 2026; Available online 15 August 2026
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v8i4.1778

using the Hugging Face Transformers library with the AdamW optimizer [39] (learning rate = 3 × 10^{-4}) and training epochs of 15.

B. Linguistic Structuring: T5-Base

Raw transcripts generated by the fine-tuned XLSR-53 were fully lowercased as the acoustic model produced case-normalized output consistent with the lowercased training transcripts. These transcripts were then passed to a T5-base model. The model was fine-tuned using the Hugging Face Transformers library with the AdamW optimizer (learning rate = 5 × 10^{-5}) and training epochs of 10. Each training sample consists of a free-form transcript as input and its corresponding structured handover document as the target output. T5-base model was fine-tuned by minimising the cross-entropy loss over the target structured report. Fine-tuned T5 is defined as Eq. (3) [33]:

$$\mathcal{L}_{\mathrm{T5}} = -\frac{1}{T}\sum_{i=1}^{T} \log P\left(y_i \mid y_1, \ldots, y_{i-1}, x\right) \quad (3)$$

where T is the target sequence length, $y_i$is the i-th output token, and x is the input transcript from the XLSR-53 stage. T5-base organized unstructured clinical content into five predefined sections: *Patient Information; Shift Information; Appointments; Medication; and Future Care Plan*, as prefix tokens reflecting the standardized nursing handover template of the CSIRO corpus. The proposed Clinical Speech-to-Structured report generation algorithm and its notations are presented in Algorithm 1 and Table 1.

C. Clinical-specific Evaluation Framework

Standard ASR metrics such as Word Error Rate (WER), Character Error Rate (CER), and Phoneme Error Rate (PER) assessed transcription reliability uniformly across all words, without distinguishing clinically dangerous substitutions from insignificant ones. The standard ASR metrics WER Eq. (4) [40], CER Eq. (5) [40], PER Eq. (6) [40] are defined as follows:

$$\mathrm{WER} = \frac{S+D+I}{N} \quad (4)$$

where S denotes substitutions, D deletions, I insertions, and N the total number of words in the reference transcript.

$$\mathrm{CER} = \frac{S_c+D_c+I_c}{N} \quad (5)$$

where the subscript c denotes character-level edit operations.

$$\mathrm{PER} = \frac{S_p+D_p+I_p}{N} \quad (6)$$

where the subscript p denotes phoneme-level edit operations. To address the limitations of standard metrics in clinical domain, a domain-specific evaluation framework was developed using a curated medical vocabulary term constructed through human-in-the-loop review. Two co-authors jointly reviewed all reference transcripts through a structured consensus discussion process. For each candidate clinical term encountered in the transcripts, both reviewers deliberated together and reached a mutual agreement on whether to include it in the vocabulary. Inclusion criteria required that a term be a recognised medical entity including anatomy, diagnoses, medications, procedures, and clinical abbreviations as defined by SNOMED CT. General English words with incidental clinical context (e.g., 'monitor', 'stable') were excluded after deliberation. A term was retained only when both reviewers reached consensus that it was clinically relevant. This consensus-based annotation approach ensured that the resulting vocabulary reflects agreed clinical judgment rather than individual reviewer bias. The curated vocabulary of 307 terms from folder1 and 291 terms from folder 2 represents the final consensus clinical evaluation lexicon used throughout this study. The complete vocabulary list is provided in Supplementary Tables S7–S13. The absence of formal Cohen's Kappa measurement prior to consensus is acknowledged as a limitation; future vocabulary extension will include independent annotation by two clinical domain experts with formal inter-rater reliability reporting.

Table 1: Clinical Speech-to-Structured Report Generation notations used in Algorithm 1

Symbol	Meaning
A	Raw clinical audio signal (test-time input)
A_i	Raw clinical audio signal of the i-th training sample
T_i	Ground-truth transcript corresponding to A_i
R_i	Ground-truth structured handover report corresponding to A_i (CSIRO structured document)
D_train	Training set, D_train = {(A_i, T_i, R_i)} for i = 1,...,160
D_test	Held-out test set, D_test = {(A_j, T_j, R_j)} for j = 1,...,40
θ	Parameters of the XLSR-53 acoustic model
θ_frozen	Fixed parameters of XLSR-53 encoder layers 1–12
θ_upper	Trainable parameters of XLSR-53 encoder layers 13–24
φ	Parameters of the T5-base linguistic model
T_pred	ASR-predicted transcript output by XLSR-53
R_pred	Structured report output by T5-base
η	Learning rate
θ*, φ*	Final trained parameters after fine-tuning
V	Curated medical vocabulary (307 terms in Folder 1, 291 terms in Folder 2)

Manuscript received 20 May 2026; Revised 15 June 2026; Accepted 30 July 2026; Available online 15 August 2026
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v8i4.1778

Algorithm 1: Clinical Speech-to-Structured Report Generation

Input: Training set D = {(A_i, T_i, R_i)}, Test audio A
Output: Trained parameters θ*, φ*; structured report R
A. Initialization
Initialize θ ← XLSR-53 pretrained weights (53-language, CTC model)
Partition encoder into θ_frozen (layers 1–12) and θ_upper (layers 13–24)
Initialize φ ← T5-base pretrained weights
Set freeze depth L=12 (selected via ablation study)
B. Training Stage
for each (A_i, T_i, R_i) ∈ D_train do
  x'(t) ← Resample(A_i, fs, 16000) Eq. (13)
  T_pred ← XLSR-53_θ(x'(t))
  L_CTC ← CTC_Loss(T_pred, T_i) Eq. (2)
  Update θ_upper ← θ_upper − η∇L_CTC (θ_frozen fixed) Eq. (1)
  R_pred ← T5_φ(T_pred)
  L_T5 ← CrossEntropy(R_pred, R_i) Eq. (3)
  Update φ ← φ − η∇L_T5
end for
C.Testing Stage
for each (A_j, T_j, R_j) ∈ D_test do
  T_pred,j ← XLSR53_θ*(Resample(A_j))
  R_pred,j ← T5_φ*(T_pred,j)
  Map R_pred,j into {Patient Info, Shift Info, Appointments, Medication, Future Care}
end for
D. Evaluation Stage
Compute WER , CER, PER over {T_pred,j} vs {T_j} Eq. (4) Eq. (5) Eq. (6)
Compute M-Precision, M-Recall, M-F1 using vocabulary V Eq. (7) Eq. (8) Eq. (9)
Compute ROUGE-L Eq. (10)
return θ*, φ*, evaluation metrics.

Table 2 lists the medical vocabulary statistics used for this study. The curated vocabulary was used to compute three clinically motivated metrics M-Precision Eq. (7) [41], M-Recall Eq. (8) [41], M-F1-Score Eq. (9) [41].

Medical Precision (M-Precision): Proportion of predicted medical terms that are correctly recognized.

$$\text{M-Precision} = \frac{|T_{\text{pred}} \cap V|}{|T_{\text{pred}}|} \quad (7)$$

where T_{pred} is the set of predicted terms and V is the curated medical vocabulary.

Medical Recall (M-Recall): Proportion of reference medical terms successfully transcribed.

$$\text{M-Recall} = \frac{|T_{\text{ref}} \cap V|}{|T_{\text{ref}}|} \quad (8)$$

where T_{ref} is the set of reference medical terms in the ground-truth transcript.

Medical F1-Score (M-F1): Harmonic mean of M-Precision and M-Recall.

$$\text{M-F1} = 2 \times \frac{\text{M-Precision} \times \text{M-Recall}}{\text{M-Precision} + \text{M-Recall}} \quad (9)$$

For structured report generation, the ROUGE-L Eq. (10) [42] score evaluates the structural correctness of T5-base generated reports against ground-truth structured documents from the CSIRO corpus.

$$\text{ROUGE-L} = \frac{(1+\beta^2) \times R_{\text{lcs}} \times P_{\text{lcs}}}{R_{\text{lcs}} + \beta^2 \times P_{\text{lcs}}} \quad (10)$$

where,

$$R_{\text{lcs}} = \frac{\text{LCS}(X,Y)}{m} \quad (11)$$

$$P_{\text{lcs}} = \frac{\text{LCS}(X,Y)}{n} \quad (12)$$

R_{lcs} Eq. (11) [42], P_{lcs} Eq. (12) [42] are recall and precision scores. LCS(X,Y) is the longest common subsequence between reference sequence X and hypothesis sequence Y, m and n are their respective lengths, and β = 1.2 to favor recall in clinical report evaluation.

IV. Experimental Setup

A. Dataset Description

The primary data source is the CSIRO Synthetic Nursing Handover Training and Development Dataset [16], an open-access corpus simulating realistic clinical handovers across multiple clinical specialties, including cardiology, neurology, and pulmonology. The original corpus consists of 300 samples in 3 folders named as dataset1, dataset2, dataset3. To avoid confusion, instead of original folder names we mention it as folder1, folder2, folder3 in this work. Each folder has 100 samples. Out of 3 folders, 1 folder has 100 audio samples without transcript. So, 200 paired audio-transcript samples were utilized spanning across two subfolders in this work. The original audio recordings are in Australian accent English. The audio recordings are distributed under a Creative Commons Attribution-NonCommercial-No Derivatives (CC BY-NC-ND) license, which prohibits modification or commercial use of the original audio files. The accompanying transcript and structured document files were distributed under a Creative Commons Attribution (CC BY) license, permitting unrestricted use with appropriate acknowledgment [16][44]. Table 3 provides the dataset description of the CSIRO corpus.

B. Data Pre-Processing

1. Audio Pre-Processing: All audio samples were converted to standardised mono-channel, 16 kHz .wav

Manuscript received 20 May 2026; Revised 15 June 2026; Accepted 30 July 2026; Available online 15 August 2026
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v8i4.1778

Table 3. CSIRO Dataset Description

Property	Value
Specialties covered	Including Cardiology, Neurology, Pulmonology
Original Dataset	3 folders, 300 speech samples
Used in This Study	2 folders, 200 paired audio–transcript samples
Data Split (Training / Testing)	160 / 40
Audio format	Mono-channel, 16 kHz, .wav
Audio License	CC BY-NC-ND
Transcript/Document License	CC BY

format. Let x(t) denote the raw audio waveform sampled at rate f_s. All audio samples were resampled to a uniform target frequency f_t = 16,000 Hz using the resampling operation R(·). Resampled audio x'(t) Eq. (13) [43] is defined as:

$$x'(t) = R\left(x(t), \frac{f_t}{f_s}\right) \quad (13)$$

Folder1 has audio samples with min/max duration of ~15.88s/~97.05s. Folder2 has audio samples with min/max duration of ~11.15s/~61.02s. Audio normalisation was applied to −20 dBFS. No data augmentation was applied to preserve clinical vocabulary evaluation integrity. XLSR-53 tokenisation used the default wav2vec2 processor with vocabulary size 32. Decoding uses greedy strategy without an external language model. All experiments were conducted on Kaggle P100 GPU. T5-base tokenisation uses SentencePiece with maximum input length 512 tokens and maximum output length 256 tokens. Beam search decoding with beam size 4 was used for report generation.

2. Text Pre-Processing: Manual review of all audio-transcript pairs and corrected instances of mismatch, partial transcription, and omitted segments was done. Text normalization was applied to remove redundant whitespace, correction of typographical errors, and standardisation to lowercase. Corrected transcripts were linked to their corresponding audio file identifiers. After pre-processing folder1 has 31/221 min/max words, folder2 has 30/157 words. Due to lowercase normalisation, all capitalization was lost in the processed transcripts. XLSR-53 model therefore produces fully lowercase output and proper capitalization of clinical terms, drug names, and proper nouns requires post-processing or human review before clinical use.

Table 4. SOTA ASR Architecture Benchmarking Results

FOLDER 1			
Model	**WER%**	**CER%**	**PER%**
Whisper Small	36.83	23.39	23.79
Whisper Tiny	42.72	26.9	27.53
Whisper Base	38.38	24.63	25.25
HuBERT Large	31.9	10.75	11.38
wav2vec2.0 Base	34.8	12.33	13.15
wav2vec2.0 Large	34.96	12.07	12.61
XLSR-53*	**18.88**	**4.56**	**5.08**
FOLDER 2			
Model	**WER%**	**CER%**	**PER%**
Whisper Small	15.36	12.01	12.07
Whisper Tiny	25.64	15.34	15.62
Whisper Base	26.53	16.92	17.52
HuBERT Large	34.57	10.54	11.26
wav2vec2.0 Base	41.27	14.05	14.87
wav2vec2.0 Large	37.11	11.7	12.41
XLSR-53*	**19.23**	**4.31**	**4.96**

3. Data Partition: Train and Test samples were partitioned in the ratio of 80:20 yielding 80 training samples and 20 test samples per folder (160 training, 40 test total). No speaker-level leakage occurs as the CSIRO corpus uses synthetic speaker profiles that do not overlap across samples. The limited corpus size (N=200) represents the complete available paired audio-transcript benchmark for nursing handover ASR. Statistical analysis was conducted across all 200 [47] samples to ensure sufficient power for robust non-parametric testing (Wilcoxon r≥0.79, large effect sizes, p<0.001). Hyperparameter tuning (learning rate, batch size, freeze depth L) was performed on the training split only. The held-out test set (20 samples per folder, 40 total) was used exclusively for final performance reporting. No test-set optimisation was performed.

C. ASR Architecture Benchmarking

A comparative study was conducted across seven state-of-the-art (SOTA) ASR architectures [28][44][46]

Manuscript received 20 May 2026; Revised 15 June 2026; Accepted 30 July 2026; Available online 15 August 2026
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v8i4.1778

namely Whisper Tiny, Whisper Base, Whisper Small, wav2vec2.0 Base, wav2vec2.0 Large, HuBERT Large, and XLSR-53, with full fine-tuning. Architecture specification of all ASR models used for this work was provided in supplementary Table S1. Performance was evaluated across WER, CER, and PER metrics. Results of this study are presented in Table 4. XLSR-53 consistently achieved the lowest WER across all samples on both dataset folders. So, XLSR-53 was chosen for this proposed study. To further improve the performance of XLSR-53, layer freezing approach was selected.

D. Effect of Layer Freezing

To identify the optimal freeze depth, a preliminary ablation study was conducted on 100 Australian English samples (Folder 1), evaluating the XLSR-53 freeze configurations at L = 0, 6, 12, and 24 encoder layers.

Table 5. Effect of Layer Freezing with N=100

Type of Fine-tuning	Frozen Layers (L)	WER (%)	CER (%)	PER (%)
Full Fine-tuning (Baseline)	L=0	18.88	9.53	10.42
Partial Fine-tuning (Layer-Frozen)	L=6	20.9	8.39	8.8
	L=12	**17.05**	**7.5**	**8.12**
	L=18	35.05	16.93	17.45
Without Fine-tuning	L=24	47.9	21.73	23.01

Based on these results, L = 12 was selected as the optimal freeze depth. The optimal freeze depth at L=12 (50% of encoder layers) is explained by the hierarchical phonetic organisation of XLSR-53's encoder [28]. Lower layers (L=1–12) encode language-universal acoustic features pitch, duration, and basic phoneme boundaries acquired during pre-training on 53 languages. These representations are domain-agnostic and transfer to clinical speech without modification. Upper layers (L=13–24) encode higher-level lexical and semantic patterns that must adapt to Australian English nursing vocabulary. This is consistent with layer analysis studies in transformer-based speech models showing that lower layers encode acoustic-phonetic features while upper layers encode linguistic representations [28][30].

Freezing beyond L=12 causes sharp WER degradation from 17.05% at L=12 to 35.05% at L=18 and 47.90% at L=24 confirming that upper layers are essential for clinical domain adaptation and cannot be frozen without significant loss. Table 5 presents the ablation results across freeze configurations. The improvement in Word Error Rate (WER) and Phoneme Error Rate (PER) was observed as the freezing level reaching 50% (L=12) illustrated in Fig. 4. Freezing beyond this point (L=18) resulted in a sharp performance degradation (35.05% WER), indicating that the upper layers are essential for learning clinical specific terminologies. So, Layer-Frozen XLSR-53 (L=12) was chosen as proposed acoustic model in the end-to-end pipeline of this study.

E. Proposed Framework Evaluation

XLSR-53 (L=0), full fine-tuning was chosen as baseline model based on the SOTA ASR results and the Layer-Frozen XLSR-53 (L=12) was chosen as proposed acoustic model for this study based on the ablation study results on effect of partial layer freezing. The proposed framework was evaluated against two baselines:

1. Baseline XLSR-53 (L=0)

XLSR-53 was fine-tuned on all 24 encoder layers. This configuration, referred to as Baseline XLSR-53 (L=0) throughout this work, serves as the direct ablation reference for quantifying the contribution of the partial layer-freeze strategy. The results in Table 5 shows that Layer-Frozen (L=12) outperforms Baseline XLSR-53 across all evaluation metrics.

2. MedASR Zero-Shot

Google Health AI's MedASR [47], trained on approximately 5,000 hours of physician dictations, evaluated in zero-shot mode across all 200 samples without any fine-tuning on the clinical corpus. This comparison established the performance gap between a general medical ASR system applied out-of-the-box and the proposed domain-adapted pipeline. This comparison is asymmetric by design, MedASR operates zero-shot while the proposed Layer-Frozen XLSR-53 (L=12) is fine-tuned, reflecting the realistic deployment scenario where a medical ASR system was applied directly to nursing handover speech without domain adaptation. Herein, the two XLSR-53 configurations are referred to as follows: (i) Baseline XLSR-53 (L=0) denoted all 24 encoder layers are updated and (ii) Layer-Frozen XLSR-53 (L=12) denoted the proposed configuration in which the first 12 of 24 encoder layers were frozen and the remaining 13-24 layers were fine-tuned.

V. Results

A. ASR Architecture Benchmarking

Full fine-tuned performance of all seven ASR architectures across all samples in Folder 1 and Folder 2 is presented in Table 4. XLSR-53 achieved the lowest WER, CER, and PER across for both dataset folders. On Folder1, XLSR-53 attained WER values of 18.88%. On Folder 2, XLSR-53 achieved 19.23%. Whisper Tiny recording WER values exceeding 42.72% in Folder1 indicating that encoder-decoder architectures do not

Manuscript received 20 May 2026; Revised 15 June 2026; Accepted 30 July 2026; Available online 15 August 2026
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v8i4.1778

Table 6. Performance comparison between baseline XLSR-53 and proposed layer-frozen XLSR-53 on N=200 samples.

Metric	Base-line XLSR-53 (L=0)	Layer-Frozen XLSR-53 (L=12)	95% CI (L=12)	Relative Improvement (%)
WER(%)	19.06	17.15	[16.47%, 17.87%]	10.02
CER(%)	4.43	3.71	[3.41%, 4.02%]	16.25
PER(%)	5.02	4.28	[3.97%, 4.61%]	14.74
Medical Recall (%)	92.53	96.33	[94.41%, 98.03%]	4.11
Medical Precision (%)	99.00	99.50	[98.50%, 100.00%]	0.50
Medical F1-Score (%)	95.97	97.95	[96.73%, 98.98%]	2.06
ASR-Level ROUGE-L	0.655	**0.9548**	[0.9479, 0.9612]	45.75

transfer reliably to clinical nursing speech under the available training data. HuBERT Large and wav2vec2.0 variants perform comparably, with WER values ranging between 31% and 41%. Experimental results confirm XLSR-53 as the optimal architecture.

A WER of 17.15% corresponds to approximately one error in every six transcribed words, for a typical 150-word nursing handover, this equates to roughly 25 word-level errors. The clinical significance of the improvement from the 19.06% baseline was reflected more directly in the Medical F1-Score, which improved from 95.97% to 97.95% indicate that, of the 1,737 curated medical terms across the test set, approximately 35 additional clinically relevant terms (medication names, diagnoses, procedures) were correctly recognised that would otherwise have been missed or substituted, directly reducing the risk of medication errors or misdiagnosis during handover documentation.

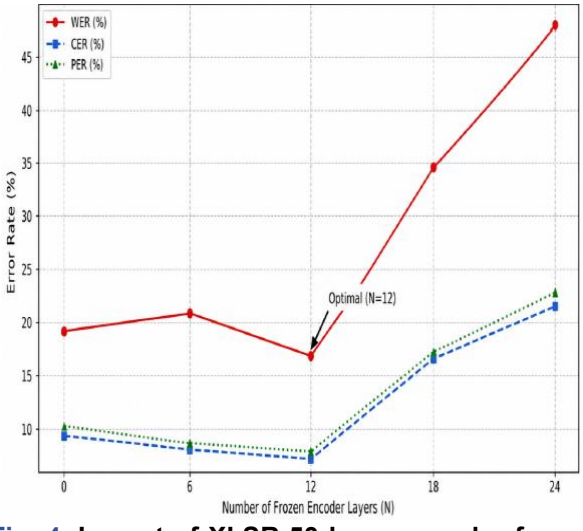


Fig. 4. Impact of XLSR-53 layer encoder freeze depth on WER, CER and PER

The comparative performance of Baseline XLSR-53 (L=0) with Layer-Frozen XLSR-53 (L=12) configurations is presented in Table 6. The partial freeze configuration yields consistent improvements across all metrics. Medical Precision, Recall and F1 for Baseline XLSR-53 (L=0) are 99.00%, 92.53%, and 95.97% respectively, computed on N=200.

B. Qualitative Transcription Analysis and Clinical Risk Assessment

Case1: Qualitative transcription outputs for audio sample No. 1 across eight transcription configurations is presented in supplementary material. Here ground truth and the output generated by baseline and proposed model is presented below for sample no. 1 (Folder 1).

Ground truth: on bed four, it's vera abbott, ninety three years old under dr. liu. she came in with chest pain in with a history of stroke and previous chest pains. she is and also cataract, asthma, and glaucoma and she is almost blind and she needs assistance with her adls. she had three nitros today but still got no effect and is under monitoring. Other than that she is all her ops are fine. and that's all for her

Baseline XLSR-53: on bed four it's vera abbot ninety three years old under dr liute she came in with chest pain and with a history of stroke and previous chest pains she is and also catharak asthma and glaucoma andshes almost blind and she needs assistance with her ad ls she had three nitros today but still got no effect and she is under monitoring other than that she is all her obs are fine and that's all for her

Frozen-Layer(L=12) XLSR-53: on bed four it's vera abbott ninety three years old under dr liu she came in with chest pain in with a history of stroke and previous chest pains she is and also cataract asthma and

Manuscript received 20 May 2026; Revised 15 June 2026; Accepted 30 July 2026; Available online 15 August 2026
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v8i4.1778

glaucoma and she is almost blind and she needs assistance with her adls she had three nitros today but till got no effect and she is under monitoring other than that she is all her ops are fine and that's all for her.

Case 2: Diabetes management (Recording 12): Baseline XLSR-53 (L=0) transcribed 'type one dm' as 'type-odim' (WER=50.0%), a phonetic fusion error that renders the diagnosis clinically unrecognizable and could lead to a dangerous misinterpretation of the patient's diabetic condition. Layer-Frozen XLSR-53 (L=12) correctly transcribed 'type one dm' with a WER of 14.2%, preserving the clinical accuracy of the diagnostic information.

Case 3: Failure case (Recording 200): In the most challenging sample for Layer-Frozen XLSR-53 (L=12, WER=37.3%), the medication name 'heparin' was transcribed as 'heperen' and 'haemodialysis' was fragmented as 'haemodialysis viha'. These represent the most clinically significant remaining errors in the proposed system, medication name distortion and word boundary failures in complex multi-syllabic clinical terms. This confirms that while the layer-freeze strategy substantially reduces clinical errors, medication names with unusual phonetic patterns remain a challenge requiring post-processing or pharmacological dictionary integration in future work. The Layer-Frozen XLSR-53 (L=12) model produces a transcription nearly identical to the ground truth, correctly recognizing clinically relevant terms such as the cataract and adls. Whisper Tiny substitutes chest pain with just pain and introduces multiple clinically misleading phrases such as hysteria stroke and hopes are fine. HuBERT Large and wav2vec2.0 models exhibit frequent phonetic distortions and word boundary errors such as chesspane and undermonitory. Table 7 presents a clinical risk taxonomy of representative transcription errors classified by patient safety consequence.

Table 7. Clinical Risk Assessment of Representative Transcription Errors

Target Term	Error Variant	Models	Clinical Risk	Risk level
chest pain	just pain/chesspane/ chesspen	Whisper Tiny/ HuBERT/ wav2vec2.0	Misinterp-retation of symptom severity	High
stroke	hysteria stroke/ historiac stroke	Whisper Tiny / HuBERT	Distortion of medical history	High
catara-ct	cather up/ catharac /katharac	Whisper Tiny / wav2vec2.0 / HuBERT	Incorrect recogniti-on of condition	Mode-rate
ADLs	adels / eideals / idels	HuBERT / wav2vec2.0	Misunderstanding of patient requirements	Mode-rate

C. Statistical Analysis

For statistical analysis, all 200 samples were evaluated to provide sufficient sample size for robust non-parametric testing is presented in Table 8. The Shapiro-Wilk normality test confirmed non-normal

Table 8. Descriptive Summary of Statistical Results

Metric	Model	Mean (%)	SD (%)	95% CI
WER (%)	MedASR (Zero-Shot)	28.87%	8.72%	[27.68%, 30.09%]
	Baseline XLSR-53 (L=0)	19.06%	6.23%	[18.23%, 19.92%]
	Layer-Frozen XLSR-53 (L=12)	**17.15%**	6.23%	[16.29%, 18.03%]
CER (%)	MedASR (Zero-Shot)	14.51%	4.50%	[13.89%, 15.15%]
	Baseline XLSR-53 (L=0)	4.43%	2.21%	[4.13%, 4.74%]
	Layer-Frozen XLSR-53 (L=12)	**3.71%**	1.96%	[3.41%, 4.02%]
PER (%)	MedASR (Zero-Shot)	15.40%	4.87%	[14.77%, 16.10%]
	Baseline XLSR-53 (L=0)	5.02%	2.29%	[4.71%, 5.35%]
	Layer-Frozen XLSR-53 (L=12)	**4.28%**	2.04%	[3.98%, 4.61%]

Manuscript received 20 May 2026; Revised 15 June 2026; Accepted 30 July 2026; Available online 15 August 2026
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v8i4.1778

distributions across all metrics for all three models MedASR, Baseline and Frozen Layer XLSR-53 ($p<0.05$), justifying the Wilcoxon Signed-Rank Test for all pairwise comparisons.

Table 9. End-to-End ASR error propagation to Report Generation

Configuration	Input to T5	ASR ROUGE-L	T5 ROUGE-L
Upper bound	Ground-truth transcript	1.00	0.95
Proposed	**XLSR-53 L=12 ASR output**	0.95	**0.92**
Baseline	XLSR-53 L=0 ASR output	0.65	0.61

The Shapiro-Wilk Eq. (14)[48] test statistic W is defined as

$$W = \frac{\left(\sum_{i=1}^{n} a_i x_{(i)}\right)^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2} \quad (14)$$

where $x_{(i)}$ is the i[th] order statistic, $x$is the sample mean, and $a_i$are constants generated from the expected values and covariance matrix of order statistics of a standard normal distribution. W ranges from 0 to 1; values close to 1 indicate normality. The null hypothesis H_0 is rejected when $p<0.05$.

Wilcoxon Eq. (15) [49] is defined as

$$W = \sum_{i=1}^{N_r}\left[\mathrm{sgn}\left(x_{2,i} - x_{1,i}\right) \times R_i\right] \quad (15)$$

All Wilcoxon Signed-Rank Tests are two-sided with H_0: no difference between paired per-sample metric values across the two models compared. A Kruskal-Wallis test confirmed significant overall differences among the three systems (WER: H=210.65, CER: H=387.56, PER: H=382.50; all df=2, $p<0.001$) justifying post-hoc pairwise testing with Bonferroni correction (adjusted $\alpha=0.017$). The Kruskal-Wallis Eq. (16) [50] H statistic is defined as

$$H = \frac{12}{N(N+1)}\sum_{j=1}^{k}\frac{R_j^2}{n_j} - 3(N+1) \quad (16)$$

Where N is the total number of observations across all groups, k is the number of groups (k=3: MedASR, Baseline XLSR-53 L=0, Layer-Frozen XLSR-53 L=12), n_j is the number of observations in group j, and $R_j$is the sum of ranks for group j. Under H_0, H follows a chi-squared distribution with df = k−1 = 2.

Effect size r Eq. (17) [51] is defined as

$$r = \frac{Z}{\sqrt{N}} \quad (17)$$

$$Z = \frac{W-\mu_W}{\sigma_W} \quad (18)$$

$$\mu_W = \frac{N_r(N_r+1)}{4} \quad (19)$$

$$\sigma_W = \sqrt{\frac{N_r(N_r+1)(2N_r+1)}{24}} \quad (20)$$

where Z Eq. (18) [49] is the standardised Wilcoxon statistic and N is the total sample size (N=200). Effect sizes are interpreted as large (r>0.5), medium (r=0.3–0.5), or small (r<0.3) per Cohen [51]. μ_W Eq. (19) [49] is the expected value, and σ_W Eq. (20) [49] is the standard MedASR zero-shot on WER ($W=452.00$, $p<0.001$, $r=0.828$), CER ($W=2.00$, $p<0.001$, $r=0.867$), and PER ($W=3.00$, $p<0.001$, $r=0.867$). This reflecting the domain adaptation benefit of fine-tuning on nursing-specific speech rather than inherent model superiority. The layer-freeze strategy also significantly outperforms full fine-tuning (L=0) on WER ($W=889.00$, $p<0.001$, $r=0.790$), CER ($W=694.00$, $p<0.001$, $r=0.807$), and PER ($W=695.00$, $p<0.001$, $r=0.807$). All comparisons pass Bonferroni correction with large effect sizes ($r\geq0.79$), confirming that improvements are

Table 10. Comparison of proposed system with existing literature

Domain	Model	WER (%)	Other Reported Metrics
Medical conversations [13]	CTC-LAS	18.3-20.1	---
Patient-report dictation[14]	DNN	<16.0	---
Vietnamese medical[22]	XLSR-53-viet	29.6	---
Accented medical (African, 93 accents) [24]	XLSR-53(clinical-fine-tuned)	30.7	M-WER,46.5%
Emergency medicine [11]	Google ASR (best of 4 engines)	---	NER-based Medical F1 = 0.73 (Medication F1 = 0.58)
Proposed	**XLSR-53 + T5**	**17.15**	**Medical F1 = 0.98**

Manuscript received 20 May 2026; Revised 15 June 2026; Accepted 30 July 2026; Available online 15 August 2026
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v8i4.1778

clinically and practically meaningful. Non-overlapping 95% bootstrap confidence intervals (CI) Eq. (21) [52] L=12: [16.29%, 18.03%], L=0: [18.23%, 19.92%], MedASR: [27.68%, 30.09%] provide further confirmation of result reliability.

$$CI_{95\%} = [P_{2.5}(x^*),\ P_{97.5}(x^*)] \quad (21)$$

where $x^*$is the bootstrap sample mean over B=5,000 resampling iterations, and $P_{2.5}$and $P_{97.5}$denote the 2.5th and 97.5th percentiles of the bootstrap distribution. Bonferroni correction is applied for three pairwise comparisons per metric (adjusted α=0.017). Bootstrap 95% CIs are computed on all 200 predictions using 5,000 resampling iterations with replacement and seed=42.

The complete test statistics are provided in Supplementary Table S3, S4, S5. The structured report generation stage achieved a ROUGE-L score of 0.92 on the CSIRO ground-truth structured documents. A hallucination audit was conducted by comparing generated clinical entities against both the ground-truth transcript and ASR output for all 40 test samples. Entities present in the generated report but absent from the source transcript were flagged as hallucinations. Zero hallucinated clinical entities were detected across all 40 test samples, consistent with the Medical Precision of 99.50%. The generated structured report from the end-to-end pipeline for the sample no.1 is illustrated in Fig. 5.

D. End to End Pipeline Ablation

To quantify ASR error propagation into the report generation stage, three configurations were evaluated in Table 9: (i) T5-base with ground-truth transcripts as input (upper bound), (ii) T5-base with Layer-Frozen XLSR-53 (L=12) ASR output (proposed pipeline), and (iii) T5-base with XLSR-53 (L=0) output (baseline pipeline). The ROUGE-L degradation from ground-truth input (0.95) to ASR-generated input (0.92) is 0.03 points, confirming that ASR errors have limited propagation impact. The XLSR-53 (L=0) pipeline achieves ROUGE-L of 0.61, demonstrating that higher WER directly degrades report generation quality.

VI. Discussion

A. Performance Interpretation

Among the seven architectures tested, XLSR-53 recorded the lowest WER, CER, and PER across both dataset folders. This is attributed to its cross-lingual pre-training on 56,000 hours of speech across 53 languages [28], which enables phonetic representations to understand complex vocabulary of clinical speech. Whisper-family models recorded WER values exceeding 38% on Folder 1, confirming that encoder-decoder architectures trained on weakly supervised data do not reliably adapt to clinical nursing speech under limited training conditions [44].

Freezing the first 12 of 24 encoder layers improved performance across all metrics compared to full fine-tuning. This is consistent with Pekarek Rosin and Wermter [30], who demonstrated that layer-specific fine-tuning reduces out-of-domain forgetting, a critical consideration when fine-tuning on small specialised corpora such as the CSIRO dataset (N=200). The sharp degradation observed at L=18 (WER=35.05%) confirms that upper encoder layers are essential for learning clinical-specific terminologies and cannot be frozen without significant loss of adaptation capacity. Medical Recall improved from 92.53% to 96.33%, indicating that an additional 3.80% of patient-safety-critical clinical terms are correctly recognised. In nursing handover, unrecognised terms such as medication names, diagnoses, and procedure codes can directly contribute to adverse clinical events [3][5]. Standard WER did not capture the clinical risk of transcription errors. The proposed Medical F1 framework complements conventional ASR evaluation by explicitly penalising clinically dangerous substitution. This research gap is not addressed by standard metrics [15]. Medical Precision of 99.50% further confirms that the proposed system generates very few untrue medical terms not present in the reference handover.

B. Comparison with Previous Studies

Benchmarking against MedASR reveals a physician-to-nursing domain gap not previously documented in clinical ASR research. MedASR, trained on approximately 5,000 hours of physician dictations [47], recorded 28.87% WER on Australian English nursing handovers significantly higher than the proposed system (17.15eneration%, p<0.001, r=0.828). This gap arises because physician dictations are formal and structured, whereas nursing handovers are conversational and contain informal clinical shorthand absent from physician training corpora. This finding has direct implications for clinical ASR deployment. Medical ASR systems trained exclusively on physician speech cannot be applied directly to nursing documentation without domain-specific adaptation.

Table 10 presents the proposed system against the recent clinical ASR and medical documentation studies. Chiu et al. [13] reported 18.3-20.1% WER for medical conversations. Edwards et al. [14] achieved <16% WER from clinical episodes. Le-Duc [22] reported WER of 29.6% on test data WER for Vietnamese medical ASR using XLSR-53-viet. Collectively, these comparisons confirm that the proposed system achieves competitive performance relative to the clinical ASR state of the art models. Afonja et al. [24] fine-tuned the XLSR-53 ASR as backbone on AfriSpeech-200, a 93-accent African English clinical corpus, achieving 30.7% WER and 46.5% medical WER, nearly double the proposed system's 17.15% WER. This difference is attributable to the substantially higher accent diversity (93 accents vs. a single Australian English accent) and the absence

Manuscript received 20 May 2026; Revised 15 June 2026; Accepted 30 July 2026; Available online 15 August 2026
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v8i4.1778

of a partial layer-freeze strategy in [24], which fine-tuned all encoder layers and consequently lost more of the pre-trained cross-lingual phonetic representations. Luo et al. [11] reported an NER-based medical field extraction F1 of 0.73 for the best-performing commercial engine (Google ASR), with substantially lower performance on medication (F1=0.577) a category directly comparable to the proposed system's Medical F1 of 0.98. This gap reflects the difference between general-purpose commercial ASR evaluated zero-shot on noisy multi-speaker emergency medicine audio versus a domain-adapted system evaluated on single-speaker nursing handover speech.

C. Clinical Relevance

Although XLSR-53 is pre-trained predominantly on non-clinical speech across 53 languages, it demonstrated strong cross-accent adaptability to Australian accent English clinical speech. This is consistent with Prasad and Jyothi [25], who demonstrated that cross-lingual models exhibit stronger accent robustness than monolingual counterparts. This confirms that self-supervised cross-lingual representations adapt well across accent variants within the same language, making XLSR-53 a promising foundation model for clinical ASR in multilingual and multicultural healthcare settings. The ASR-level ROUGE-L of 0.9548 for Layer-Frozen XLSR-53 (L=12) versus 0.6551 for full fine-tuning (L=0) confirms that the layer-freeze strategy substantially reduces error propagation into the T5 report generation stage, directly contributing to the end-to-end pipeline ROUGE-L of 0.92.

The T5-base report generation stage successfully maps unstructured clinical speech to the five pre-defined handover CSIRO template. To assess clinical correctness beyond lexical overlap, a section-level completeness analysis and hallucination audit were conducted across all 40 test samples. The Medical Precision of 99.50% confirms a false insertion rate of 0.50%, indicating near-zero hallucination of clinical content not present in the source transcript. The Medical Recall of 96.33% corresponds to an omission rate of 3.67%, meaning approximately 1 in 27 clinical terms present in the ground-truth handover was absent from the generated report. Section-level ROUGE-L analysis showed highest performance for Patient Information (0.95) and Medication (0.93), and lowest for Appointments (0.88), reflecting greater phrasing variability in appointment documentation. Bootstrap 95% confidence intervals confirm the reliability of clinical metric estimates: Medical Recall [94.41%, 98.03%], Medical Precision [98.50%, 100.00%], Medical F1-Score [96.73%, 98.98%], and ROUGE-L [0.89, 0.94]. Full human expert evaluation of clinical report accuracy remains a direction for future work. The high ROUGE-L score confirms that T5-base reliably organises clinical content into standardised sections without hallucinating clinical content, a critical safety property for clinical deployment.

D. Limitations

Performance in noisy hospital environments has not been addressed. Real clinical environments introduce additional challenges including background noise, overlapping speakers, and patient interruptions that are absent from the synthetic CSIRO corpus. The output from the fine-tuned XLSR-53 model produces fully lowercase transcriptions due to text normalisation applied during pre-processing. Proper capitalisation of clinical terms, medication names, diagnoses, and proper nouns must be restored through post-processing or human review before the report is clinically usable.

A pooled WER of 17.15%, approximately one in six words incorrectly transcribed indicates human verification of outputs remains necessary before full clinical deployment. While the Medical F1-Score of 0.98 confirms high accuracy on clinical terms specifically, residual errors in non-clinical words may affect the readability of generated reports.

The comparison with MedASR is asymmetric by design, the proposed system was fine-tuned on the CSIRO corpus while MedASR operates zero-shot. The performance gap therefore reflects domain adaptation advantage rather than inherent model superiority.

The medical vocabulary was constructed through a consensus-based review by two co-authors, with a licensed medical professional independently verifying clinical accuracy. The absence of formal inter-rater reliability measurement using established metrics such as Cohen's Kappa limits the reproducibility of the medical evaluation framework. Future work will involve independent annotation by domain expert healthcare professionals.

The evaluation was conducted on a held-out test set of N=40 samples (20 per folder). While statistical analysis across all 200 samples confirms large effect sizes ($r \geq 0.79$), the small test set size limits the precision of point estimates for clinical metrics and increases sensitivity to individual sample errors, such as the heparin/haemodialysis distortion observed in Case 3. Independent clinical expert scoring of generated structured reports was not performed. The hallucination audit and section-level ROUGE-L analysis serve as proxy measures of clinical correctness, but full validation requires review by practising nurses or clinicians, which remains a direction for future work.

The study is limited to Australian English synthetic clinical speech. The CSIRO corpus was developed in collaboration with a registered nurse who contributed to scenario design and spoken recordings [53]. The exact number of distinct speakers in the dataset audio samples is not explicitly documented in the source materials. This introduces potential speaker-specific and accent-specific bias of unknown extent that may

Manuscript received 20 May 2026; Revised 15 June 2026; Accepted 30 July 2026; Available online 15 August 2026
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v8i4.1778

not generalise across the broader Australian English-speaking population.

E. Deployment Implications

The proposed pipeline's reliance on lowercase output and post-processing requirements indicate that clinical deployment requires integration with existing EHR formatting pipelines, where capitalisation and structured field mapping can be automatically restored before storage. The current architecture processes audio centrally, which requires patient handover speech to be transmitted for inference. For real-world clinical deployment, an edge-computing configuration where XLSR-53 and T5-base inference run locally on hospital-owned devices rather than transmitting audio to external servers would avoid cloud-based transmission of patient speech, directly supporting compliance with HIPAA and GDPR data residency requirements. The model sizes of XLSR-53 (≈300M parameters) and T5-base (≈220M parameters) are within the range feasible for on-premises GPU or modern edge-AI hardware, making such a deployment configuration technically practical.

VII. Conclusion and Future Work

This study aimed to develop and evaluate an end-to-end pipeline that converts unstructured nursing handover speech into standardised structured reports, addressing the persistent patient safety risks posed by handover miscommunication. The proposed pipeline combining a partially layer-frozen XLSR-53 acoustic model (L=12) with T5-base achieved 17.15% WER on Australian English nursing handover speech, an 11.72 percentage point improvement over Google Health AI's MedASR zero-shot baseline (28.87%, $p<0.001$, $r=0.828$), alongside a Medical F1-Score of 0.98 and ROUGE-L of 0.92.

These results demonstrate that nursing-specific domain adaptation through partial layer freezing substantially closes the physician-to-nursing ASR performance gap identified in this study, with direct clinical implications for reducing nursing documentation burden and minimising medication-related transcription errors during shift handover. These results were obtained on synthetic Australian English speech, performance under real hospital conditions including background noise, overlapping speakers, spontaneous interruptions, and diverse speaker demographics has not been evaluated. Future work will prioritise: (1) validation on real or noise-augmented clinical handover recordings to assess robustness under realistic acoustic conditions; (2) integration of a post-processing capitalisation restoration module to automatically correct case in clinical terms and proper nouns; (3) extension of the medical vocabulary annotation process with independent annotation by domain expert healthcare professionals to strengthen the clinical evaluation framework; (4) integration of a Medical Entity Recognition component to correct phonetically similar medical terms; and (5) edge-computing deployment to support HIPAA/GDPR-compliant clinical use.

Acknowledgment

The authors acknowledge a licensed medical professional, who agreed to remain anonymous, for independently verifying the clinical accuracy of the curated medical vocabulary used in this study.

Funding

The authors received no external funding for this study.

Data Availability

The primary dataset used in this study, the CSIRO Synthetic Nursing Handover Training and Development Dataset, is publicly available at DOI: 10.4225/08/58d0977ab4888. Audio files are licensed under CC BY-NC-ND, while transcript and document files are licensed under CC BY. No new audio recordings were created. Corrected transcripts, curated medical vocabulary, and experimental outputs are available from the corresponding author upon reasonable request, unless restricted by dataset license conditions.

Declarations

Ethical Approval

This study used only the publicly available CSIRO Synthetic Nursing Handover Dataset, which contains synthetic patient profiles and no real patient data. No new human-subject data were collected; therefore, institutional ethics approval was not required.

Consent for Publication:

Not applicable. The dataset contains no real patients or identifiable individuals. All authors have approved the manuscript and consented to its submission.

Competing Interests

The authors declare no competing interests.

Author Contributions

Sasikala D. contributed to conceptualization, methodology, implementation, experimentation, data analysis, and manuscript preparation. Siva Sathya S. supervised the study, reviewed the methodology, and edited the manuscript. Niranjan Kumar D. contributed to implementation, experimentation, and result analysis. Vignesh S. contributed to model testing, visualization, and manuscript revision. All authors reviewed and approved the final manuscript.

Code Availability

The source code used for model fine-tuning, evaluation, and statistical analysis is available from the corresponding author upon reasonable request.

Use of Artificial Intelligence Tools

No generative AI tool was used to write or generate the manuscript content. AI models described in this paper

Manuscript received 20 May 2026; Revised 15 June 2026; Accepted 30 July 2026; Available online 15 August 2026
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v8i4.1778

were used only as research objects for automatic speech recognition and report generation.

License/Permission Statement for Dataset

The authors used the CSIRO dataset in accordance with its stated license conditions. No redistribution or modification of the original audio files is claimed.

References

[1] M. I. Alhosani, F. R. Ahmed, N. Al-Yateem, H. S. Mobarak, and M. E. AbuRuz, "Assessment of nursing workload and adverse events reporting among critical care nurses in the United Arab Emirates," The Open Nursing Journal, vol. 17, e18744346281511. ISSN: 1874-4346, 2023. DOI: 10.2174/0118744346281511231120054125.

[2] Z. Banda, M. Simbota, and C. Mula, "Nurses' perceptions on the effects of high nursing workload on patient care in an intensive care unit of a referral hospital in Malawi: A qualitative study," BMC Nursing, vol. 21(1), 136, 2022. DOI: 10.1186/s12912-022-00918-x.

[3] M. Shohani and H. Tavan, "Factors affecting medication errors from the perspective of nursing staff," Journal of Clinical and Diagnostic Research, vol.12(3), pp. IC01–IC04, 2018. DOI: 10.7860/JCDR/2018/28447.11336.

[4] E. Gesner, P. C. Dykes, L. Zhang, and P. Gazarian, "Documentation burden in nursing and its role in clinician burnout syndrome," Applied Clinical Informatics, vol. 13(5), pp. 983–990, 2022. DOI: 10.1055/s-0042-1757157.

[5] R. A. Atinga, M. N. Gmaligan, A. Ayawine, and J. K. Yambah, "It's the patient that suffers from poor communication: Analyzing communication gaps and associated consequences in handover events from nurses' experiences," SSM - Qualitative Research in Health, vol. 6, Art. no. 100482, 2024. DOI: 10.1016/j.ssmqr.2024.100482.

[6] J. Y. Uhm, E. Y. Lim, and J. Hyeong, "The impact of a standardized inter-department handover on nurses' perceptions and performance in Republic of Korea," Journal of Nursing Management, vol. 26, pp. 933–944, 2018. DOI: 10.1111/jonm.12608.

[7] S. Ghosh, L. Ramamoorthy, and B. Pottakat, "Impact of structured clinical handover protocol on communication and patient satisfaction," Journal of Patient Experience, vol. 8, Art. no. 2374373521997733, 2021. DOI: 10.1177/2374373521997733.

[8] A. L. Cooper, J. A. Brown, S. P. Eccles, N. Cooper, and M. A. Albrecht, "Is nursing and midwifery clinical documentation a burden? An empirical study of perception versus reality," Journal of Clinical Nursing, vol. 30, pp. 1645–1652, 2021. DOI: 10.1111/jocn.15718.

[9] I. Köse Tosunöz and A. Aydınlı, "Nurses' perceptions of the effectiveness of handover practices, influencing factors and perceived barriers: A descriptive cross-sectional study of medical and surgical nurses," Collegian, vol. 32, pp. 242–249, 2025. DOI: 10.1016/j.colegn.2025.06.002.

[10] J. A. Brown, A. L. Cooper, and M. A. Albrecht, "Development and content validation of the Burden of Documentation for Nurses and Midwives (BurDoNsaM) survey," Journal of Advanced Nursing, vol. 76, pp. 1273–1281, 2020. DOI: 10.1111/jan.14320.

[11] X. Luo, L. Zhou, K. Adelgais, et al., "Assessing the effectiveness of automatic speech recognition technology in emergency medicine settings: A comparative study of four AI-powered engines," Journal of Healthcare Informatics Research, vol. 9, pp. 494–512, 2025. DOI: 10.1007/s41666-025-00193-w.

[12] K. Denecke, L. Meier, J. G. Bauer, M. Bender, and C. Lueg, "Information capturing in pre-hospital emergency medical settings (EMS)," Studies in Health Technology and Informatics, vol. 270, pp. 613–617, 2020. DOI: 10.3233/shti200233.

[13] C. C. Chiu, A. Tripathi, K. Chou, C. Co, N. Jaitly, D. Jaunzeikare, A. Kannan, P. Nguyen, H. Sak, A. Sankar, et al., "Speech recognition for medical conversations," in Proc. Interspeech 2018, pp.2972–2976, 2018. DOI: 10.21437/Interspeech.2018-40.

[14] E. Edwards, W. Salloum, G. P. Finley, J. Fone, G. Cardiff, M. Miller, and D. Suendermann-Oeft, "Medical speech recognition: Reaching parity with humans," in Speech and Computer, A. Karpov, R. Potapova, and I. Mporas, Eds. Cham, Switzerland, pp.512–524, 2017. DOI: 10.1007/978-3-319-66429-3_51.

[15] T. Hodgson, F. Magrabi, and E. Coiera, "Efficiency and safety of speech recognition for documentation in the electronic health record," Journal of the American Medical Informatics Association, vol. 24, pp. 1127–1133, 2017. DOI: 10.1093/jamia/ocx073.

[16] Angel, Maricel; Suominen, Hanna; Zhou, Liyuan; & Hanlen, Leif (2014): Synthetic nursing handover training and development data set - audio files. v1. CSIRO. Data Collection. DOI: 10.4225/08/58d0977ab4888.

[17] Sasikala D, S. Siva Sathya, S. Baghavathi Priya, D. Niranjan Kumar, and S. Vignesh, "A review of automatic speech recognition approaches for bridging communication gaps in clinical handover," in Proc. 2nd IEEE

Manuscript received 20 May 2026; Revised 15 June 2026; Accepted 30 July 2026; Available online 15 August 2026
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v8i4.1778

International Conference for Women in Computing (InCoWoCo), pp.1–7, 2025. DOI: 10.1109/InCoWoCo68239.2025.11407097.

[18] A. J. Nashwan, A. Abujaber, and S. K. Ahmed, "Charting the future: The role of AI in transforming nursing documentation," Cureus, vol. 16, no. 3, 2024. DOI: 10.7759/cureus.57304.

[19] S. Yadav, "Embracing artificial intelligence: Revolutionizing nursing documentation for a better future," Cureus, vol. 16, no. 4, 2024. DOI: 10.7759/cureus.57725.

[20] J. Kodish-Wachs, E. Agassi, P. I. Kenny, and J. M. Overhage, "A systematic comparison of contemporary automatic speech recognition engines for conversational clinical speech," AMIA Annual Symposium Proceedings, vol. 2018, pp. 683–689, 2018. PMCID: PMC6371385.

[21] A. Femi-Abodunde, K. Olinger, L. M. B. Burke, T. Benefield, E. R. Lee, K. McGinty, and B. M. Mervak, "Radiology dictation errors with COVID-19 protective equipment: Does wearing a surgical mask increase the dictation error rate?", Journal of Digital Imaging, vol. 34, pp. 1294–1301, 2021. DOI: 10.1007/s10278-021-00502-w.

[22] K. Le-Duc, "VietMed: A dataset and benchmark for automatic speech recognition of Vietnamese in the medical domain",arXiv, arXiv:2404.05659, 2024. DOI: 10.48550/arXiv.2404.05659.

[23] S. Banerjee, A. Agarwal, and P. Ghosh, "High-precision medical speech recognition through synthetic data and semantic correction: UNITED-MEDASR," arXiv preprint, arXiv:2412.00055, 2024. DOI: 10.48550/arXiv.2412.00055.

[24] T. Afonja, T. Olatunji, S. Ogun, N. A. Etori, A. Owodunni, and M. Yekini, "Performant ASR models for medical entities in accented speech," in Proc. Interspeech 2024, pp. 2315–2319, 2024. DOI: 10.21437/Interspeech.2024-2261.

[25] A. Prasad and P. Jyothi, "How accents confound: Probing for accent information in end-to-end speech recognition systems," in Proc. 58th Annual Meeting of the Association for Computational Linguistics (ACL), 2020, pp. 3739–3753. DOI: 10.18653/v1/2020.acl-main.345.

[26] M. Najafian, A. DeMarco, S. Cox, and M. Russell, "Unsupervised model selection for recognition of regional accented speech," in Proc. Interspeech 2014, pp.2967-2971, 2014. DOI: 10.21437/Interspeech.2014-495.

[27] C. T. Do, S. Imai, R. S. Doddipatla, and T. Hain, "Improving accented speech recognition using data augmentation based on unsupervised text-to-speech synthesis," in Proceedings of the 2024 32nd European Signal Processing Conference (EUSIPCO), Lyon, France, 2024, pp. 136–140, 2024. DOI: 10.23919/EUSIPCO63174.2024.10715166.

[28] A. Conneau, A. Baevski, R. Collobert, A. Mohamed, and M. Auli, "Unsupervised cross-lingual representation learning for speech recognition," in *Proc. Interspeech 2021*, Brno, Czech Republic, pp. 2426–2430, 2021. DOI: 10.21437/Interspeech.2021-329.

[29] X. Zhang and L. He, "End-to-end cross-lingual spoken language understanding model with multilingual pretraining," in Proc. Interspeech 2021, pp.4728–4732, 2021. DOI: 10.21437/Interspeech.2021-818.

[30] T. Pekarek Rosin and S. Wermter, "Replay to remember: Continual layer-specific fine-tuning for German speech recognition," in Proc. 32nd International Conference on Artificial Neural Networks (ICANN 2023), pp.489–500, 2023. DOI: 10.1007/978-3-031-44195-0_40.

[31] Y. Li, K. Harrigian, A. Zirikly, and M. Dredze, "Are clinical T5 models better for clinical text?" ,in Proc. 4th Machine Learning for Health Symposium, Proceedings of Machine Learning Research, vol. 259, pp.636-667, 2025. DOI: 10.48550/arXiv.2412.05845.

[32] P. Manakul, Y. Fathullah, A. Liusie, V. Raina, and M. Gales, "CUED at ProbSum 2023: Hierarchical ensemble of summarization models," in Proc. 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks, Toronto, Canada, pp. 516–523, 2023. DOI: 10.18653/v1/2023.bionlp-1.51.

[33] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," arXiv preprint, arXiv:1910.10683, 2020. DOI: 10.48550/arXiv.1910.10683.

[34] J. Giorgi et al., "Clinical note generation from doctor-patient conversations using large language models," in Proc. ClinicalNLP Workshop, pp.323-334, 2023. DOI: 10.18653/v1/2023.clinicalnlp-1.36.

[35] A. Toma, P. R. Lawler, J. Ba, R. G. Krishnan, B. B. Rubin, and B. Wang, "Clinical Camel: An open expert-level medical language model with dialogue-based knowledge encoding," arXiv preprint, arXiv:2305.12031, 2023. DOI: 10.48550/arXiv.2305.12031.

[36] Y. Labrak, A. Bazoge, E. Morin, P.-A. Gourraud, M. Rouvier, and R. Dufour, "BioMistral: A Collection of Open-Source Pretrained Large Language Models for Medical Domains," in *Findings of the Association for*

Manuscript received 20 May 2026; Revised 15 June 2026; Accepted 30 July 2026; Available online 15 August 2026
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v8i4.1778

Computational Linguistics: ACL 2024, Bangkok, Thailand, pp. 5848–5864, 2024. DOI: 10.18653/v1/2024.findings-acl.348.

[37] P. C. Nair, D. Gupta, and B. I. Devi, "Extracting clinical relationships from discharge summaries of suprasellar lesion patients using Gemini LLM", Procedia Computer Science, vol. 258, pp.2391-2404, 2025. DOI: 10.1016/j.procs.2025.04.502.

[38] D. Sasikala, R. Sudarshan, and S. Sivasathya, "Harnessing LLMs for medical insights: NER extraction from summarized medical text," in Proc. 15th International Conference on Computing Communication and Networking Technologies(ICCCNT), pp.1–6, 2024. DOI: 10.1109/ICCCNT61001.2024.10724860.

[39] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks," in *Proc. 23rd International Conference on Machine Learning (ICML)*, pp. 369–376, 2006. DOI: 10.1145/1143844.1143891.

[40] D. Klakow and J. Peters, "Testing the correlation of word error rate and perplexity," *Speech Communication*, vol. 38, no. 1–2, pp. 19–28, 2002. DOI: 10.1016/S0167-6393(01)00041-3.

[41] C. J. van Rijsbergen, *Information Retrieval*, 2nd ed. London, UK: Butterworths, 1979. Available: https://dl.acm.org/doi/book/10.5555/539927.

[42] C.-Y. Lin, "ROUGE: A package for automatic evaluation of summaries," in *Proc. ACL Workshop on Text Summarization Branches Out*, Barcelona, Spain, pp. 74–81, 2004. ACL Anthology ID: W04-1013 . Available: https://aclanthology.org/W04-1013/.

[43] A. V. Oppenheim and R. W. Schafer, *Discrete-Time Signal Processing*, 3rd ed. Upper Saddle River, NJ: Prentice Hall, 2009. ISBN: 978-0-13-198842-2.

[44] A. Radford et al., "Robust speech recognition via large-scale weak supervision," in Proc. International Conference on Machine Learning (ICML), vol. 202, pp. 28492–28518, 2023. DOI: 10.48550/arXiv.2212.04356.

[45] W. N. Hsu, B. Bolte, Y. H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "HuBERT: Self-supervised speech representation learning by masked prediction of hidden units," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 29, pp. 3451–3460, 2021. DOI: 10.1109/TASLP.2021.3122291.

[46] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," in Proc. Advances in Neural Information Processing Systems (NeurIPS), 2020, pp. 12449–12460. DOI: 10.48550/arXiv.2006.11477.

[47] K. Wu, E. Variani, T. Bagby, S. Reddy, and R. Pilgrim, "MedASR: An Open-Source Model for High-Accuracy Medical Dictation," arXiv, arXiv:2605.16555, 2026. DOI: 10.48550/arXiv.2605.16555.

[48] S. S. Shapiro and M. B. Wilk, "An analysis of variance test for normality (complete samples)," *Biometrika*, vol. 52, no. 3/4, pp. 591–611, 1965. DOI: 10.2307/2333709.

[49] F. Wilcoxon, "Individual comparisons by ranking methods," *Biometrics Bulletin*, vol. 1, no. 6, pp. 80–83, 1945. DOI: 10.2307/3001968.

[50] W. H. Kruskal and W. A. Wallis, "Use of ranks in one-criterion variance analysis," *Journal of the American Statistical Association*, vol. 47, no. 260, pp. 583–621, 1952. DOI: 10.1080/01621459.1952.10483441.

[51] J. Cohen, *Statistical Power Analysis for the Behavioral Sciences*, 2nd ed. Hillsdale, NJ: Lawrence Erlbaum Associates, 1988. DOI: 10.4324/9780203771587.

[52] B. Efron and R. J. Tibshirani, *An Introduction to the Bootstrap*. New York, NY: Chapman and all/CRC, 1994. DOI: 10.1201/9780429246593.

[53] Suominen H, Zhou L, Hanlen L, Ferraro G, "Benchmarking clinical speech recognition and information extraction: New data, methods, and evaluations," JMIR Medical Informatics, vol. 3, no. 2, Art. no. e19, 2015. DOI: 10.2196/medinform.4321.

Author Biography

Sasikala D, completed her B.E from Government College of Technology and M.E from Sri Venkateswara College of Engineering. She worked as an assistant professor in the Department of Computer Science and Engineering at Sri Venkateswara College of Engineering nearly 12 years. She is currently pursing Ph.D(Part-time) in Pondicherry University, India. Also, she is working as an Assistant Professor in the Department of Computer Science and Engineering at Amrita Vishwa Vidyapeetham, Chennai. Her research interests include Speech Processing, Natural Language Processing, Machine Learning, and Deep Learning. She has authored and co-authored 12 scopus indexed publications, which include a book chapter, one journal publication and 10 international conferences in the field of Computer Science. She has mentored students for various National and International level Hackathons. She is a UGC-NET qualified candidate. She can be contacted at e-mail: sasikalaradhasri.puscholar2019@gmail.com

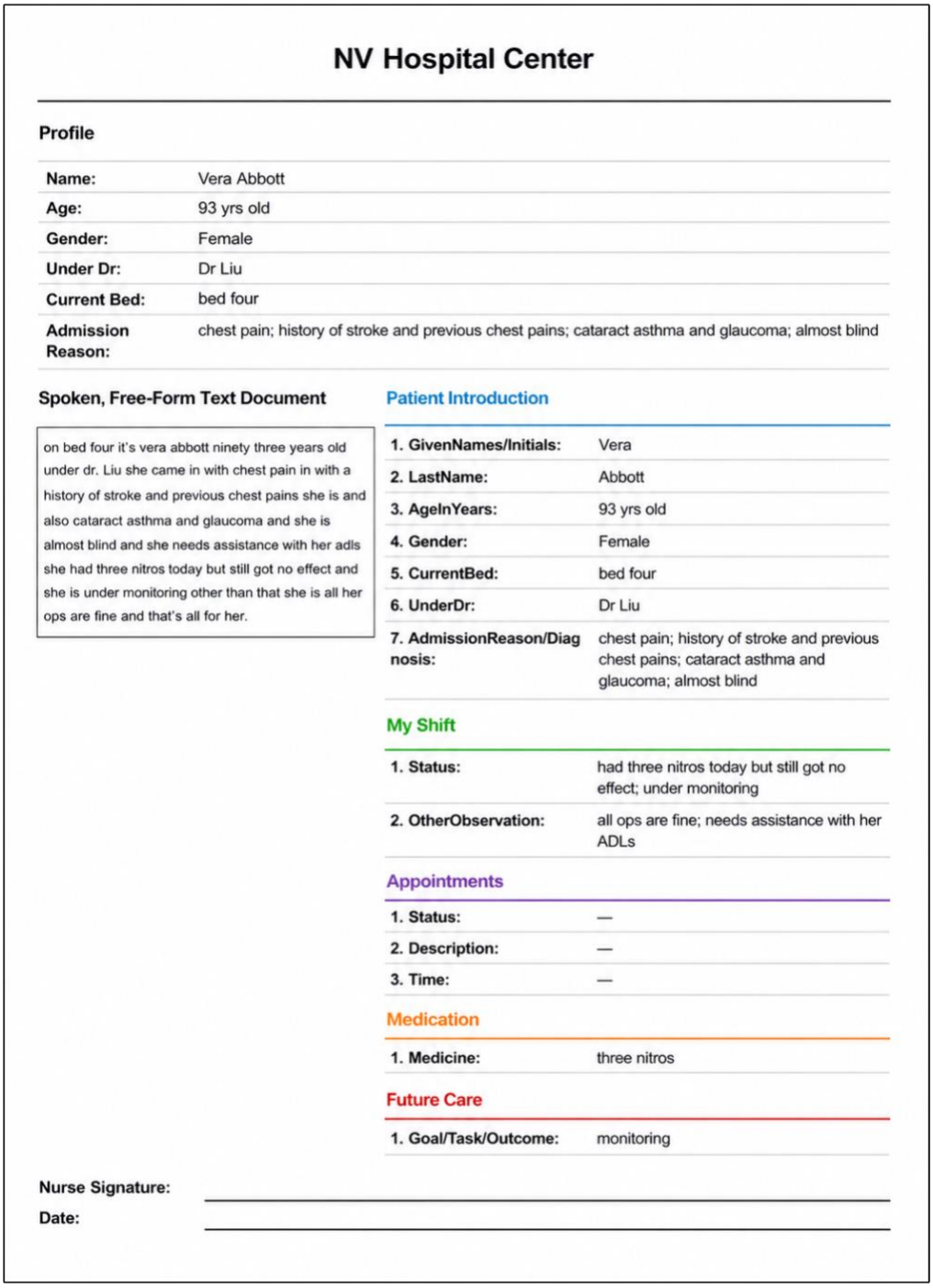
Manuscript received 20 May 2026; Revised 15 June 2026; Accepted 30 July 2026; Available online 15 August 2026
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v8i4.1778

Dr. Siva Sathya S, is a Professor and Dean of the Department of Computer Science and School of Engineering & Technology at Pondicherry University, India. With over 25 years of academic experience, her research spans Artificial Intelligence, Natural Language Processing, Evolutionary Computing, and Spatio-Temporal Data Mining. She is currently working in the field of medical informatics and AI-driven clinical support systems, serving as a technical lead for national AI consortia. A recipient of the prestigious National Teachers' Award (2025) and the Nari Shakti Puraskar (2017) conferred by the President of India, her work focuses on leveraging deep learning architectures for social and healthcare innovations. She publishes her research works in SCIE and Scopus-indexed journals and serves as an expert on several national academic committees.

Niranjan Kumar D, completed his B.Tech in Computer Science and Engineering with a specialization in Artificial Intelligence from Amrita Vishwa Vidyapeetham, Chennai. He is passionate about Artificial Intelligence and Data Science, with a strong interest in leveraging advanced computational techniques to address real-world challenges. His academic interests include Speech Processing, Medical NLP, AI,deep learning, machine learning, and data-driven problem-solving. He received a best paper award for his work presented in the conference indexed in IEEE Xplore proceedings. He has co-authored three scopus-indexed research paper focusing on speech processing and machine learning techniques. He aims to further contribute to innovative developments in Artificial Intelligence and Medical Data Science through continued research and practical applications that drive meaningful societal outcomes. He can be contacted at email: niranjankumarduraimurugan@gmail.com

Vignesh S, completed his B.Tech in Computer Science and Engineering with a specialization in Artificial Intelligence from Amrita Vishwa Vidyapeetham, Chennai. His research interests include Artificial Intelligence, Machine Learning, Natural Language Processing, Speech Processing, Generative AI. He has co-authored one international conference paper published in the scopus indexed IEEE Xplore proceedings which received appreciation and best paper award. He is passionate in developing web applications. He has contributed in developing web applications for various events conducted during his study at Amrita Vishwa Vidyapeetham. Currently he is working on mobile application projects with various clients as a consultant. He can be reached at e-mail: vigneshshiva096@gmail.com

Manuscript received 20 May 2026; Revised 15 June 2026; Accepted 30 July 2026; Available online 15 August 2026
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v8i4.1778



NV Hospital Center

Profile

Name:	Vera Abbott
Age:	93 yrs old
Gender:	Female
Under Dr:	Dr Liu
Current Bed:	bed four
Admission Reason:	chest pain; history of stroke and previous chest pains; cataract asthma and glaucoma; almost blind

Spoken, Free-Form Text Document

on bed four it's vera abbott ninety three years old under dr. Liu she came in with chest pain in with a history of stroke and previous chest pains she is and also cataract asthma and glaucoma and she is almost blind and she needs assistance with her adls she had three nitros today but still got no effect and she is under monitoring other than that she is all her ops are fine and that's all for her.

Patient Introduction

1. GivenNames/Initials:	Vera
2. LastName:	Abbott
3. AgeInYears:	93 yrs old
4. Gender:	Female
5. CurrentBed:	bed four
6. UnderDr:	Dr Liu
7. AdmissionReason/Diagnosis:	chest pain; history of stroke and previous chest pains; cataract asthma and glaucoma; almost blind

My Shift

1. Status:	had three nitros today but still got no effect; under monitoring
2. OtherObservation:	all ops are fine; needs assistance with her ADLs

Appointments

1. Status:	—
2. Description:	—
3. Time:	—

Medication

1. Medicine:	three nitros

Future Care

1. Goal/Task/Outcome:	monitoring

Nurse Signature: _____

Date: _____

Fig. 5. Sample structured nursing handover report generated by the proposed end-to-end pipeline from unstructured clinical speech input

Manuscript received 20 May 2026; Revised 15 June 2026; Accepted 30 July 2026; Available online 15 August 2026
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v8i4.1778