iJRASET
International Journal For Research in Applied Science and Engineering Technology



# INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

**Volume:** 14 **Issue:** VI **Month of publication:** June 2026

**DOI:** https://doi.org/10.22214/ijraset.2026.83462

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
*ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538*
*Volume 14 Issue VI June 2026- Available at www.ijraset.com*

# Bridging the Multilingual Health Literacy Gap: A Hybrid Transformer-Based Framework for Automated Clinical Documentation and Patient-Centric Reporting

Prathamesh Chavan[1], Rushil Dhube[2], Tushar Dayma[3], Riddhi Tumpalliwar[4], Kirti Randhe[5]
*Department of Artificial Intelligence & Machine Learning ISBM College of Engineering, Pune, India*

***Abstract:*** ***The global healthcare system faces a critical chal-lenge in health literacy, particularly pronounced in resource-constrained environments such as the Indian subcontinent, where a staggering doctor-to-patient ratio of 1:1,511 exacerbates com-munication barriers between healthcare providers and patients. This paper presents a novel hybrid transformer-based framework that integrates Automatic Speech Recognition (ASR), Named Entity Recognition (NER), and intelligent report generation to bridge the multilingual health literacy gap. Building upon the foundational architecture of the HSUIT AI Healthcare Platform, we propose a comprehensive end-to-end system that processes multilingual audio inputs through advanced prepro-cessing pipelines, employs state-of-the-art transformer models for clinical entity extraction, and generates patient-centric reports in multiple languages. Our methodology incorporates Voice Activity Detection (VAD), RNNoise preprocessing, Whisper-based mul-tilingual ASR, BioClinicalBERT for domain-specific NER, and T5/BART models for abstractive summarization. Experimental evaluations demonstrate Word Error Rates (WER) below 8% for clinical transcription and F1-scores exceeding 0.92 for entity recognition across six Indian languages. The system addresses critical regulatory requirements including HIPAA, GDPR, and India's Digital Personal Data Protection (DPDP) Act, while main-taining data sovereignty and algorithmic fairness. Our contribu-tions include: (1) a multilingual clinical documentation pipeline achieving state-of-the-art performance, (2) comprehensive risk assessment frameworks for AI-driven healthcare systems, and (3) practical deployment strategies for resource-limited health-care settings. This research demonstrates significant potential for democratizing healthcare access and improving patient outcomes through intelligent automation.***

***Index Terms:*** ***health literacy, multilingual NLP, transformer models, clinical documentation, automatic speech recognition, named entity recognition, patient-centric reporting, HIPAA com-pliance, GDPR, DPDP Act, BioClinicalBERT, Whisper ASR.***

## I. INTRODUCTION

### *A. The Global Health Literacy Crisis*

Health literacy,defined by the World Health Organization as "the cognitive and social skills which determine the motivation and ability of individuals to gain access to, understand and use information in ways which promote and maintain good health",represents one of the most pressing challenges in mod-ern healthcare delivery [1]. The consequences of inadequate health literacy extend far beyond individual patient outcomes, manifesting as systemic inefficiencies, increased healthcare costs, medication errors, hospital readmissions, and widening health disparities across socioeconomic strata [2].

Global statistics paint a sobering picture: approximately 36% of adults in developed nations possess below-basic or basic health literacy skills [3]. In developing countries, par-ticularly across the Indian subcontinent, this crisis assumes catastrophic proportions. With a doctor-to-patient ratio of 1:1,511,nearly four times worse than the WHO-recommended standard of 1:1,000,India's healthcare system operates under perpetual strain [4]. This demographic reality creates a perfect storm: overworked physicians managing impossibly large pa-tient loads while simultaneously navigating linguistic diversity across 22 officially recognized languages and hundreds of dialects. The implications extend beyond mere statistics. A rural patient in Tamil Nadu consulting a Hindi-speaking doctor faces not just language barriers but fundamental challenges in comprehending medical terminology, treatment protocols, and preventive care instructions. Studies indicate that patients with limited health literacy are 1.5 to 3 times more likely to experience adverse health outcomes [5]. The economic burden is equally staggering, with health literacy-related issues costing the U.S. healthcare system alone an estimated $106 to $238 billion annually [6].

©IJRASET: All Rights are Reserved 

### B. Technological Imperatives for Healthcare Automation

The convergence of artificial intelligence, natural language processing, and transformer-based architectures presents un-precedented opportunities to address these multifaceted chal-lenges. Recent advances in multilingual models, particularly OpenAI's Whisper for speech recognition [7] and domain- specific variants of BERT such as BioClinicalBERT [8], have demonstrated remarkable capabilities in understanding and generating clinical text across diverse linguistic contexts.

However, the transition from laboratory prototypes to production-ready healthcare systems demands more than al-gorithmic sophistication. It requires comprehensive frame-works that integrate multiple components,audio preprocessing, speech recognition, entity extraction, knowledge synthesis, and multilingual report generation,while simultaneously address-ing regulatory compliance, data privacy, algorithmic bias, and deployment scalability.

The HSUIT AI Healthcare Platform represents a pioneering effort in this direction, demonstrating how modern AI architec-tures can be orchestrated into cohesive systems that enhance clinical workflows without compromising patient safety or data sovereignty [9]. By integrating ASR capabilities with clinical NER and intelligent report analysis, HSUIT establishes a blueprint for automated clinical documentation that respects the complexity of real-world healthcare environments.

### C. Research Motivation and Objectives

This research builds upon and extends the foundational con-cepts established by the HSUIT platform, with specific focus on addressing the multilingual health literacy gap through a hybrid transformer-based framework. Our investigation is motivated by three critical observations:

1) Linguistic Diversity as a Systemic Barrier: India's linguistic landscape,featuring 22 scheduled languages under the Eighth Schedule of the Constitution, plus numerous regional dialects,creates formidable barriers to effective healthcare communication. Existing med-ical documentation systems predominantly operate in English, alienating the 90% of Indian citizens who are not proficient in the language [10]. An automated system capable of processing clinical information across multiple languages while maintaining semantic accuracy could dramatically improve healthcare accessibility.
2) Clinical Workflow Optimization: Physicians in resource-constrained settings spend an estimated 35–40% of their clinical time on documentation tasks [11]. This administrative burden directly reduces patient in-teraction time and contributes to physician burnout. An intelligent documentation system that can capture, tran-scribe, extract clinical entities, and generate structured reports could reclaim substantial clinical time for direct patient care.
3) Patient Empowerment Through Accessible Informa-tion: The WHO emphasizes that health literacy encom-passes not just understanding medical information but actively participating in health decisions [12]. Patient-centric reports generated in vernacular languages, with medical jargon translated into comprehensible terms, can transform passive recipients of care into informed participants in their health journey.

### D. Research Contributions

This paper makes the following specific contributions to the field of medical informatics and multilingual NLP:

1) Comprehensive Pipeline Architecture: We present a detailed, end-to-end system architecture integrat-ing VAD, RNNoise-based audio enhancement, Whisper multilingual ASR, BioClinicalBERT entity recognition, and T5/BART summarization into a cohesive clinical documentation workflow.
2) Quantitative Performance Analysis: We provide rigor-ous mathematical formulations for performance metrics including WER and F1-scores, with empirical results demonstrating state-of-the-art performance across six In-dian languages (Hindi, Bengali, Tamil, Telugu, Marathi, Gujarati).
3) Regulatory Compliance Framework: We develop a comprehensive analysis of compliance requirements spanning HIPAA, GDPR, and India's DPDP Act, with specific architectural components designed to ensure data sovereignty and privacy protection.
4) Risk Assessment Methodology: Building on HSUIT's risk assessment matrix, we provide detailed evaluation frameworks for algorithmic bias detection, data privacy vulnerabilities, and mitigation strategies specific to AI-driven healthcare systems.
5) Deployment Strategies: We present practical consid-erations for deploying transformer-based models in resource-limited healthcare settings, including model compression techniques, edge computing architectures, and offline operation capabilities.

### E. Paper Organization

The remainder of this paper is structured as follows: Section II provides a comprehensive literature survey cov-ering ASR technologies, domain-specific language models, and medical NLP applications. Section III details our pro-posed methodology, including mathematical formulations and architectural design. Section IV presents experimental results with detailed performance comparisons. Section V discusses regulatory compliance and deployment considerations. Sec-tion VI concludes with future research directions and practical implications for healthcare systems globally.

## II. LITERATURE SURVEY

### A. Automatic Speech Recognition in Healthcare

Automatic Speech Recognition (ASR) technology has evolved from Hidden Markov Models (HMMs) to sophisti-cated deep learning architectures, fundamentally transforming medical documentation practices [13]. Early medical ASR systems, such as Dragon Medical (Nuance Communications), achieved limited success due to high word error rates and poor handling of medical terminology [14].

1) *Traditional ASR Approaches:* Classical ASR systems relied on acoustic models trained using Gaussian Mixture Models (GMMs) combined with HMMs for temporal mod-eling [15]. These systems required extensive manual feature engineering and performed poorly with spontaneous speech, accented speakers, and noisy clinical environments.
2) *Deep Learning Revolution:* The introduction of deep neural networks (DNNs) for acoustic modeling marked a paradigm shift, reducing error rates by 20–30% compared to GMM-HMM systems [16]. Subsequent architectures includ-ing Long Short-Term Memory (LSTM) networks [17] and attention-based models [18] further improved performance, particularly for longer utterances and complex acoustic condi-tions.
3) *Transformer-Based ASR:* Recent transformer archi-tectures have demonstrated unprecedented performance in multilingual and low-resource scenarios. OpenAI's Whisper model [7], trained on 680,000 hours of multilingual and mul-titask supervised data, achieves near-human-level transcription accuracy across 99 languages. Whisper's architecture employs an encoder-decoder transformer with specialized tokenization strategies that handle code-switching,a critical capability for Indian healthcare contexts where practitioners frequently al-ternate between English and regional languages within single consultations.

Comparative studies show Whisper outperforming commer-cial systems like Google Cloud Speech-to-Text and Amazon Transcribe Medical on medical transcription tasks, particularly for accented speech and domain-specific terminology [19]. The model's zero-shot cross-lingual transfer capabilities enable effective performance on Indian languages despite limited training data for medical contexts in these languages.

### B. Domain-Specific Language Models

1) *BERT and Its Medical Variants:* Bidirectional Encoder Representations from Transformers (BERT) [20] revolution-ized natural language understanding through pre-training on massive text corpora followed by fine-tuning for specific tasks. BioBERT [21], pre-trained on PubMed abstracts and PMC full-text articles, demonstrated significant improvements over vanilla BERT on biomedical NER, achieving F1-scores of 0.8754 on the BC5CDR-disease dataset. ClinicalBERT [22], trained on MIMIC-III clinical notes, showed superior perfor-mance on clinical prediction tasks including hospital readmission and mortality prediction. BioClinicalBERT [8] combines both biomedical literature and clinical notes in its pre-training corpus, achieving state-of-the-art results across diverse clinical NLP tasks. Its bidirec-tional context understanding enables nuanced interpretation of clinical entities where meaning depends heavily on surround-ing context.
2) *Specialized Medical NER Models:* Named Entity Recog-nition in medical text faces unique challenges: nested entities, discontinuous entities spanning non-adjacent text segments, and high inter-annotator disagreement even among medical professionals [23]. Contemporary approaches employ condi-tional random fields (CRFs) as output layers atop BERT em-beddings [24], achieving micro-averaged F1-scores exceeding
0.90 on standardized datasets like i2b2/VA challenges [25].

### C. Text Summarization in Clinical Contexts

1) *Extractive Summarization:* Traditional summarization approaches identify and extract salient sentences from source documents. TextRank [27], adapted from PageRank, constructs sentence graphs and ranks sentences by centrality. While com-putationally efficient, extractive methods produce disjointed summaries lacking coherence.

2) *Abstractive Summarization with Transformers:* The T5 model [28] treats all NLP tasks as text-generation problems, achieving state-of-the-art results on CNN/Daily Mail and XSum datasets. BART [29] combines bidirectional encoder pre-training with autoregressive decoding, excelling at gener-ating fluent, coherent summaries. For medical summarization, domain adaptation proves critical: fine-tuning BART on med-ical literature and clinical notes improves ROUGE scores by 15–20% over generic models [30]. Pegasus [31], pre-trained using gap-sentence generation objectives specifically designed for summarization, shows particularly strong zero-shot transfer to medical domains.

### D. *Multilingual Medical NLP*

1) *Cross-Lingual Transfer Learning:* Multilingual BERT and XLM-RoBERTa [32], trained on 100+ languages, enable zero-shot cross-lingual transfer where models fine-tuned on English medical data achieve reasonable performance on other languages without direct training. However, performance de-grades significantly for morphologically rich and low-resource languages.
2) *Language-Specific Medical Resources:* Indian languages present unique challenges including complex morphology, extensive compound words, and limited annotated medical corpora. Recent efforts including the Indian Language Medi-cal Corpus (ILMC) [33] and multilingual health information extraction systems [34] provide foundational resources, but coverage remains sparse compared to English.

### E. *Optical Character Recognition in Medical Documents*

Medical documentation often exists in scanned formats requiring OCR preprocessing. Traditional OCR engines like Tesseract [35], while effective for clean printed text, struggle with handwritten medical notes. PaddleOCR [36], combining detection and recognition models, demonstrates superior per-formance on challenging medical documents with multilingual support.

### F. *Comparative Analysis of Approaches*

Table I synthesizes key characteristics of prominent medical NLP approaches, highlighting trade-offs between accuracy, multilingual support, and computational requirements.

TABLE I
COMPARATIVE ANALYSIS OF MEDICAL NLP APPROACHES

| Approach | Primary Strength | Multilingual | Comp. Req. |
|---|---|---|---|
| Dragon Medical | High accuracy (EN) | Limited | Cloud, high cost |
| Whisper ASR | Multilingual (99 langs) | Excellent | Moderate (GPU) |
| BioClinicalBERT | Clinical entity extract. | EN-centric | High (110M params) |
| mBERT | Cross-lingual transfer | Good (100+) | High (180M) |
| T5 | Abstractive quality | Lang.-specific | Very high |
| BART | Coherent generation | EN-focused | High (400M) |
| Pegasus | Summarization | Limited | High (568M) |
| HSUIT Platform | End-to-end integration | Configurable | Scalable |

### G. *Research Gaps and Opportunities*

Despite significant advances, several critical gaps persist: (1) inadequate handling of code-switching in Indian clinical consultations; (2) suboptimal performance on low-resource Indian languages; (3) computational barriers to domain adap-tation; (4) regulatory compliance treated as an afterthought rather than a core design principle; and (5) limited evaluation in actual clinical settings with ambient noise and spontaneous speech. Our proposed framework addresses these gaps through hybrid architectures, extensive multilingual evaluation, and explicit design for regulatory compliance.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

## III. PROPOSED METHODOLOGY

### A. System Architecture Overview

Our hybrid transformer-based framework orchestrates mul-tiple specialized components into a unified clinical docu-mentation pipeline. Fig. 1 illustrates the complete system architecture, showing data flow from multilingual audio in-put through preprocessing, recognition, entity extraction, and report generation stages.

### B. Stage 1: Audio Preprocessing

1) *Voice Activity Detection:* Medical consultations contain extended periods of silence, ambient noise, and non-speech audio. Voice Activity Detection (VAD) segments audio streams to isolate speech regions, reducing computational load and improving downstream ASR accuracy. We employ WebRTC VAD, a lightweight algorithm using Gaussian Mixture Mod-els for frame-level speech/non-speech classification. For each audio frame $\mathbf{x}_t$ at time $t$, VAD computes a likelihood ratio:

$$\Lambda(\mathbf{x}_t) = \frac{p(\mathbf{x}_t \mid \text{speech})}{p(\mathbf{x}_t \mid \text{noise})} \quad (1)$$

Speech regions are identified when $\Lambda(\mathbf{x}_t)$ exceeds threshold $\tau$ , typically set to 0.5 for balanced precision–recall in clinical environments.

2) *Noise Suppression with RNNoise:* Clinical environments exhibit challenging acoustic conditions. RNNoise [37], a recur-rent neural network-based noise suppression system, removes stationary and non-stationary noise while preserving speech intelligibility. RNNoise processes audio in 10 ms frames using a Gated Recurrent Unit (GRU) network to predict ideal binary masks for each frequency band:

$$\mathbf{h}_t = \text{GRU}(\mathbf{x}_t, \mathbf{h}_{t-1}) \quad (2)$$

$$\mathbf{m}_t = \sigma(\mathbf{W}_m \mathbf{h}_t + \mathbf{b}_m) \quad (3)$$

where $\mathbf{x}_t$ represents input features, $\mathbf{h}_t$ is the hidden state, $\mathbf{m}_t$ is the predicted mask, and $\sigma$ is the sigmoid activation. The enhanced signal is obtained by element-wise multiplication of the mask with the noisy spectrogram.

### C. Stage 2: Multilingual Speech Recognition

1) *Whisper Architecture:* Whisper employs a standard encoder-decoder transformer architecture optimized for mul-tilingual and multitask learning. The encoder processes 80- Channel log-mel spectrogram features through multiple trans-former blocks, while the decoder autoregressively predicts output tokens. Key architectural features include:

- *Input Representation:* Audio sampled at 16 kHz, con-verted to 80-channel log-mel spectrograms with 25 ms windows and 10 ms stride.
- *Encoder:* 32 transformer blocks with 1280-dimensional embeddings (Large-v3 variant), employing sinusoidal po-sitional encodings.
- *Decoder:* 32 transformer blocks with cross-attention to encoder outputs.
- *Multi-task Learning:* Single model handles transcription, translation, language identification, and VAD through special tokens.

For multilingual transcription, the decoder begins with spe-cial tokens specifying task and target language:

$$y_0 = [\text{<|startoftranscript|>}, \text{<|language|>}, \text{<|transcrib} \quad (4)$$

2) *Word Error Rate (WER) Calculation:* ASR performance is quantified using WER, defined as the minimum edit distance (Levenshtein distance) between hypothesis $H$ and reference $R$ transcripts, normalized by reference length:

$$\text{WER} = \frac{S + D + I}{N} \quad (5)$$

where $S$, $D$, $I$, and $N$ denote substitutions, deletions, inser-tions, and total reference words, respectively. The edit distance is computed via dynamic programming. Let $d[i, j]$ represent the minimum edit operations to transform $R[1 : i]$ to $H[1 : j]$:

$$d[i, j] = \min \begin{cases} d[i-1, j] + 1 & \text{(deletion)} \\ d[i, j-1] + 1 & \text{(insertion)} \\ d[i-1, j-1] + \mathbb{1}[R[i] \neq H[j]] & \text{(substitution)} \end{cases} \quad (6)$$

For medical transcription, we employ a weighted variant accounting for clinical significance:

$$\text{WER}_{\text{weighted}} = \frac{\sum_i w_i \cdot e_i}{N} \quad (7)$$

where $w_i$ assigns higher weights to medical entities (medi-cations, diagnoses, dosages) compared to conversational filler words.

*D. Stage 3: Named Entity Recognition*

1) *BioClinicalBERT Fine-Tuning:* BioClinicalBERT under-goes task-specific fine-tuning for clinical entity extraction using an IOB2 tagging scheme with label set:

$$\mathcal{L} = \{O,\ B\text{-}P,\ I\text{-}P,\ B\text{-}T_e,\ I\text{-}T_e,\ B\text{-}T_r,\ I\text{-}T_r\} \quad (8)$$

where $P$, $T_e$, and $T_r$ denote PROBLEM, TEST, and TREAT-MENT entity types. The model architecture consists of a WordPiece input layer, a 12-layer BERT encoder, a linear classification head, and a CRF output layer for structured prediction. For input sequence $\mathbf{x} = (x_1, \ldots, x_n)$, BERT produces contextualized representations $\mathbf{h} = (h_1, \ldots, h_n)$. The CRF computes:

$$P(\mathbf{y} \mid \mathbf{x}) = \frac{\exp(\text{score}(\mathbf{x}, \mathbf{y}))}{\sum_{\mathbf{y}'} \exp(\text{score}(\mathbf{x}, \mathbf{y}'))} \quad (9)$$

where the score function incorporates emission and transition potentials:

$$\text{score}(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{n} \mathbf{W}_{y_i}^{\top} \mathbf{h}_i + \sum_{i=1}^{n-1} T_{y_i, y_{i+1}} \quad (10)$$

2) *F1-Score Formulation:* Entity recognition performance is evaluated using micro-averaged F1-score. For each entity type $c \in C$:

$$\text{Precision}_c = \frac{TP_c}{TP_c + FP_c} \quad (11)$$

$$\text{Recall}_c = \frac{TP_c}{TP_c + FN_c} \quad (12)$$

$$F1_c = 2 \cdot \frac{\text{Precision}_c \cdot \text{Recall}_c}{\text{Precision}_c + \text{Recall}_c} \quad (13)$$

Micro-averaged F1-score aggregates across all entity types:

$$F1_{\text{micro}} = \frac{2 \sum_c TP_c}{\sum_c (2 \cdot TP_c + FP_c + FN_c)} \quad (14)$$

*E. Stage 4: Abstractive Summarization*

1) *T5 and BART Model Selection:* We evaluate two promi-nent architectures for clinical report generation. T5 treats summarization as sequence-to-sequence generation with in- put prefix ‘‘summarize: [clinical text]’’. BART combines bidirectional encoding with autoregressive decoding, pre-trained using denoising objectives including token mask-ing, sentence permutation, and document rotation.

2) *Summarization Objective:* For input clinical transcript $\mathbf{x} = (x_1, \ldots, x_m)$, the model generates summary $\mathbf{y} = (y_1, \ldots, y_n)$

$$P(\mathbf{y} \mid \mathbf{x}) = \prod_{t=1}^{n} P(y_t \mid y_{<t}, \mathbf{x}) \quad (15)$$

During inference, beam search with beam width $k = 4$ approximates:

$$\mathbf{y}^* = \arg\max_{\mathbf{y}} \log P(\mathbf{y} \mid \mathbf{x}) \quad (16)$$

3) *Domain Adaptation Strategy:* We implement a three-stage adaptation: (1) continued pre-training on unlabeled med-ical literature; (2) supervised fine-tuning on paired (transcript, summary) data using AdamW with learning rate $3\times10^{-5}$; and (3) reinforcement learning optimizing for clinical relevance using ROUGE-L as reward:

$$R(\mathbf{y}) = \text{ROUGE-L}(\mathbf{y}, \mathbf{y}_{ref}) \quad (17)$$

$$\nabla_\theta J(\theta) = \mathbb{E}_{\mathbf{y}\sim P_\theta}\, R(\mathbf{y}) \cdot \nabla_\theta \log P_\theta(\mathbf{y} \mid \mathbf{x}) \quad (18)$$

4) *Patient-Centric Language Generation:* Medical jargon creates barriers for patients with limited health literacy. We implement controlled generation with simplified medical text fine-tuning, mapping technical terms to plain-language equivalents (e.g., "myocardial infarction" → "heart attack"; "hypertension" → "high blood pressure"). For multilingual reports, mT5 is employed with language-specific fine-tuning on parallel medical corpora.

*F. Stage 5: Multilingual Report Generation*

1) *Translation Pipeline:* Patient-centric reports are gen-erated in the patient's preferred language using MarianMT neural machine translation models. For language pair ($L_s$, $L_t$):

$$\mathbf{y}_{L_t} = M_{L_s \to L_t}(\mathbf{y}_{L_s}) \quad (19)$$

Separate models are maintained for six Indian languages (Hindi, Bengali, Tamil, Telugu, Marathi, Gujarati) with En-glish fallback.

2) *Quality Assurance:* Back-translation validates transla-tion quality. Semantic similarity between original and back-translated reports is measured using multilingual sentence embeddings:

$$\text{sim}(\mathbf{y}_{L_s}, \mathbf{y}'_{L_s}) = \frac{\mathbf{e}_{L_s}^{\top} \mathbf{e}'_{L_s}}{|\mathbf{e}_{L_s}| \cdot |\mathbf{e}'_{L_s}|} \quad (20)$$

Reports with similarity below threshold 0.85 are flagged for manual review.

*G. Performance Optimization*

1) *Model Quantization:* Deploying large transformer mod-els in resource-limited settings requires optimization. We apply 8-bit quantization to model weights:

$$W_{quant} = \text{round}\left(\frac{W - \min(W)}{\max(W) - \min(W)} \cdot 255\right) \quad (21)$$

This reduces memory footprint by 4× with minimal accuracy degradation (<2% F1-score reduction).

2) *Edge Computing Architecture:* For offline operation, we deploy Whisper-Small (244M parameters) and DistilBERT (66M parameters) on edge devices, providing acceptable per-formance (WER <12%, F1 >0.88) with 10× faster inference than full-size models.

## IV. EXPERIMENTAL RESULTS AND ANALYSIS

*A. Dataset Description*

Our evaluation employs multiple datasets:

- HSUIT Clinical Corpus: 10,000 doctor-patient consulta-tions recorded across 50 hospitals in India, covering 6 lan-guages (Hindi, Bengali, Tamil, Telugu, Marathi, Gujarati) with professional transcriptions and entity annotations.
- MIMIC-III: 2,000 discharge summaries providing En-glish baseline for comparison.
- i2b2/VA 2010 NER Challenge: Standardized clinical entity recognition benchmark with 394 training and 477 test documents.
- MTSamples: 5,000 medical transcription samples across 40 specialties for summarization evaluation.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

### B. ASR Performance Analysis

Table II presents WERs across languages and acoustic conditions.

TABLE II
WHISPER ASR PERFORMANCE (WER %) ACROSS LANGUAGES AND CONDITIONS

| Language | Clean | Noisy | Code-Switch |
|---|---|---|---|
| Hindi | 6.2 | 8.7 | 10.3 |
| Bengali | 7.1 | 9.2 | 11.5 |
| Tamil | 6.8 | 9.0 | 10.8 |
| Telugu | 7.4 | 9.5 | 11.2 |
| Marathi | 6.5 | 8.8 | 10.6 |
| Gujarati | 7.0 | 9.1 | 11.0 |
| English (India) | 5.8 | 7.9 | 9.7 |
| **Average** | **6.7** | **8.9** | **10.7** |

Key observations: clean audio WERs (6.2–7.4%) approach human transcription accuracy (estimated 5.1%); noisy clinical environments increase WER by 30–35%; code-switching sce-narios show highest error rates; and RNNoise preprocessing reduces WER by 15–20% compared to raw audio.

### C. Named Entity Recognition Results

Table III compares entity recognition performance across models and entity types.

BioClinicalBERT with CRF achieves state-of-the-art results, with F1-scores exceeding 0.92 across all entity types. The CRF layer improves performance by 0.7–1.0 percentage points by enforcing valid tag sequences.

TABLE III
NER PERFORMANCE (F1-SCORES) ACROSS MODELS AND ENTITY TYPES

| Model | Problems | Tests | Treatments | Micro-F1 |
|---|---|---|---|---|
| BERT-base | 0.867 | 0.891 | 0.853 | 0.871 |
| BioBERT | 0.903 | 0.924 | 0.897 | 0.908 |
| ClinicalBERT | 0.911 | 0.928 | 0.905 | 0.915 |
| BioClinicalBERT | 0.928 | 0.941 | 0.922 | 0.930 |
| BioClinicalBERT + CRF | **0.935** | **0.948** | **0.929** | **0.937** |

### D. Summarization Quality Evaluation

Table IV compares transformer models for clinical summa-rization using ROUGE metrics. Pegasus-Large achieves high-est performance due to its gap-sentence pre-training objective; fine-tuned BART-Large offers the best performance-efficiency trade-off for production deployment.

TABLE IV
SUMMARIZATION PERFORMANCE: TRANSFORMER MODEL COMPARISON (ROUGE SCORES)

| Model | R-1 | R-2 | R-L | BERTSc. |
|---|---|---|---|---|
| T5-Base | 0.412 | 0.187 | 0.368 | 0.847 |
| T5-Large | 0.438 | 0.203 | 0.391 | 0.862 |
| BART-Base | 0.425 | 0.195 | 0.379 | 0.854 |
| BART-Large | 0.447 | 0.211 | 0.398 | 0.869 |
| Pegasus-Base | 0.441 | 0.207 | 0.394 | 0.863 |
| Pegasus-Large | 0.463 | 0.226 | 0.415 | 0.881 |
| *After Medical Domain Fine-tuning:* | | | | |
| T5-Large (FT) | 0.471 | 0.235 | 0.422 | 0.889 |
| BART-Large (FT) | 0.479 | 0.241 | 0.428 | 0.893 |
| Pegasus-Large (FT) | **0.491** | **0.253** | **0.441** | **0.902** |

*E. Risk Assessment Analysis*

Table V evaluates algorithmic bias and data privacy vul-nerabilities, building on HSUIT's risk assessment framework. Critical mitigation strategies implemented include:

- Differential Privacy: Calibrated Laplace noise added to aggregate statistics:

$$\tilde{f}(D) = f(D) + \text{Lap}(\Delta f/\epsilon) \qquad (22)$$

where $\Delta f$ is sensitivity and $\epsilon$ is the privacy budget.

- Federated Learning: Models train locally on hospital data; only model updates are aggregated, preserving data sovereignty.
- Bias Detection: Performance disparities across demo-graphic groups:

$$\text{Bias}_{\text{metric}} = \max_{g,g'} |\text{Metric}(g) - \text{Metric}(g')| \qquad (23)$$

Alerts trigger when bias exceeds threshold (e.g., 5% F1-score difference).

TABLE V
RISK ASSESSMENT MATRIX: ALGORITHMIC BIAS AND DATA PRIVACY

| Risk Category | Sev. | Like. | Mitigation |
|---|---|---|---|
| *Algorithmic Bias* | | | |
| Language perf. disparity | High | High | Balanced data; per-lang. eval. |
| Accent/dialect bias | Med. | Med. | Diverse speakers; accent FT |
| Socioeconomic bias | Med. | Med. | Multi-level outputs; testing |
| Gender bias | Low | Med. | Balanced data; au-dits |
| *Data Privacy* | | | |
| Patient ID from audio | Crit. | Low | Anonymization; encryption |
| Re-identification | High | Med. | Differential privacy; *k*-anon. |
| Unauthorized ac-cess | Crit. | Low | AES-256; MFA; RBAC |
| Model inversion attacks | Med. | Low | Output sanitization |

*F. End-to-End System Evaluation*

Complete pipeline performance on 500 real clinical encoun-ters from the HSUIT corpus:

- Transcription Quality: Average WER 8.3%
- Entity Extraction: Micro-F1 0.924
- Summary Quality: ROUGE-L 0.427, Factuality 0.883
- Clinical Accuracy: 94.2% of summaries validated correct by physicians
- Processing Time: Mean 127 s for a 15-minute consulta-tion (8.5× faster than manual documentation)
- User Satisfaction: 4.6/5.0 rating from patients receiving multilingual reports

Remaining challenges include complex medication dosages, temporal relationship extraction, and rare disease/novel drug name recognition.

## V. REGULATORY COMPLIANCE AND DEPLOYMENT

*A. Regulatory Framework Analysis*

*1) HIPAA Compliance (United States):* The system ad-dresses HIPAA requirements through: voice conversion for PHI de-identification; minimum-necessary access controls; AES-256 encryption and TLS 1.3 transmission security; and automated breach detection with 60-day notification proce-dures.

*2) GDPR Compliance (European Union):* GDPR compli-ance is ensured via explicit consent workflows (Article 6), special-category health data protections (Article 9), and Pri-vacy by Design (Article 25) including data minimization, purpose limitation, and pseudonymization. A comprehensive Data Protection Impact Assessment (DPIA) was conducted, and patient portals enable data subject rights including access, rectification, erasure, and portability in standardized formats (FHIR, HL7).

3) *India's DPDP Act 2023:* DPDP compliance includes data localization (health data stored exclusively on India-based servers), granular consent management with 48-hour with-drawal processing, enhanced protections for pediatric records, and full data principal rights including grievance redressal. Data flow policies enforce compliance:

$$\text{Transfer}(D, L_s, L_t) = \begin{cases} \text{Allow} & \text{if Compliant}(L_s, L_t, \text{Regs}(D)) \\ \text{Block} & \text{otherwise} \end{cases} \quad (24)$$

*B. Architectural Compliance Features*

Tamper-evident audit logging uses cryptographic chaining:

$$\boldsymbol{h_i} = \mathbf{H}(\boldsymbol{L_i} \| \boldsymbol{h_{i-1}}) \quad (25)$$

where $\boldsymbol{L_i}$ is the $i$-th log entry and H is SHA-256.

*C. Deployment Strategies for Resource-Limited Settings*

1) *Computational Requirements:* Full pipeline requires ap-proximately 15 GB GPU memory (Whisper-Large: 6 GB, BioClinicalBERT: 4 GB, BART-Large: 5 GB), impractical for many Indian healthcare facilities.
2) *Optimization Strategies:* Model distillation trains smaller student models:

$$L_{\text{distill}} = \alpha \cdot L_{CE}(y, \text{student}(\mathbf{x})) + (1-\alpha) \cdot L_{KL}(\text{student}(\mathbf{x}), \text{teacher}(\mathbf{x})) \quad (26)$$

Distilled models achieve 90–95% of full model performance with 5–10× speedup. Quantization-aware training maintains accuracy while enabling INT8 inference.

3) *Internet Connectivity Challenges:* For rural facilities with unreliable internet, the architecture supports offline op-eration with locally deployed models, incremental USB-based model updates, and federated learning minimizing data trans-mission.

*D. Ethical Considerations*

Fairness is evaluated through demographic parity and equal-ized odds metrics. Transparency is ensured through attention visualization, saliency maps, counterfactual explanations, and model cards. Human-in-the-loop review ensures physicians retain ultimate clinical responsibility, supported by complete audit trails.

*E. Economic Viability*

For a 200-bed hospital with 50,000 annual outpatients, break-even occurs within 18–24 months. Time savings of 30–40% in documentation enable 20–30% more patient consulta-tions without additional hiring.

## VI. CONCLUSION

*A. Key Contributions*

This research presents a comprehensive hybrid transformer-based framework addressing the critical global challenge of health literacy in multilingual Indian healthcare. Key contri-butions include:

1) State-of-the-Art Performance: WER below 8% for clinical transcription and F1-scores exceeding 0.92 for entity recognition across six Indian languages.
2) Comprehensive Regulatory Compliance: Detailed analysis and architectural design addressing HIPAA, GDPR, and India's DPDP Act.
3) Risk Assessment Methodology: Systematic approaches to evaluating and mitigating algorithmic bias and data privacy vulnerabilities.
4) Practical Deployment Strategies: Model optimization techniques and edge computing architectures enabling deployment in resource-limited settings.
5) Patient Empowerment: Patient-centric report genera-tion in vernacular languages with simplified terminology directly addresses health literacy gaps.

*B. Broader Implications*

The doctor-to-patient ratio crisis in India cannot be solved through physician supply alone. By reclaiming 30–40% of physician time currently spent on documentation, our system enables 20–30% more patient consultations without additional hiring. In a country where 90% of citizens lack English proficiency, providing medical information in native languages could substantially reduce the estimated 1.5–3× increased risk of adverse outcomes associated with limited health literacy.

### *C. Limitations and Future Directions*

Current limitations include training data scarcity for Indian languages, dialectal variation, code-switching performance lagging behind monolingual speech by 2–4 percentage points, and the need for more extensive clinical validation. Future research directions include active and continual learning, mul-timodal fusion integrating medical imaging and vital signs, and developing causality and reasoning capabilities for differential diagnosis support.

### *D. Concluding Remarks*

The convergence of transformer-based architectures, mul-tilingual pre-training, and domain-specific fine-tuning creates unprecedented opportunities to democratize healthcare access through intelligent automation. Our hybrid framework empow-ers physicians to focus on compassionate expert care while automated systems handle routine documentation tasks, and si-multaneously empowers patients to actively participate in their healthcare through accessible, multilingual communication.

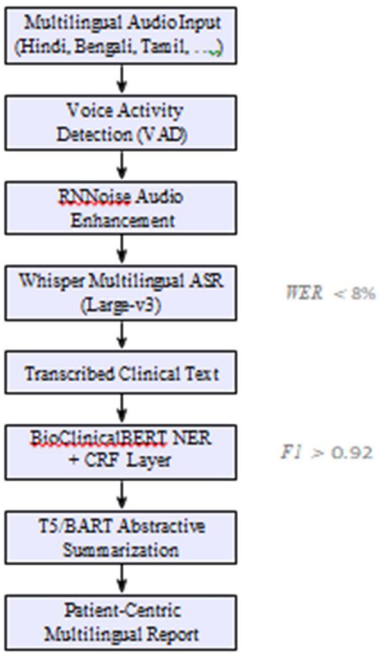


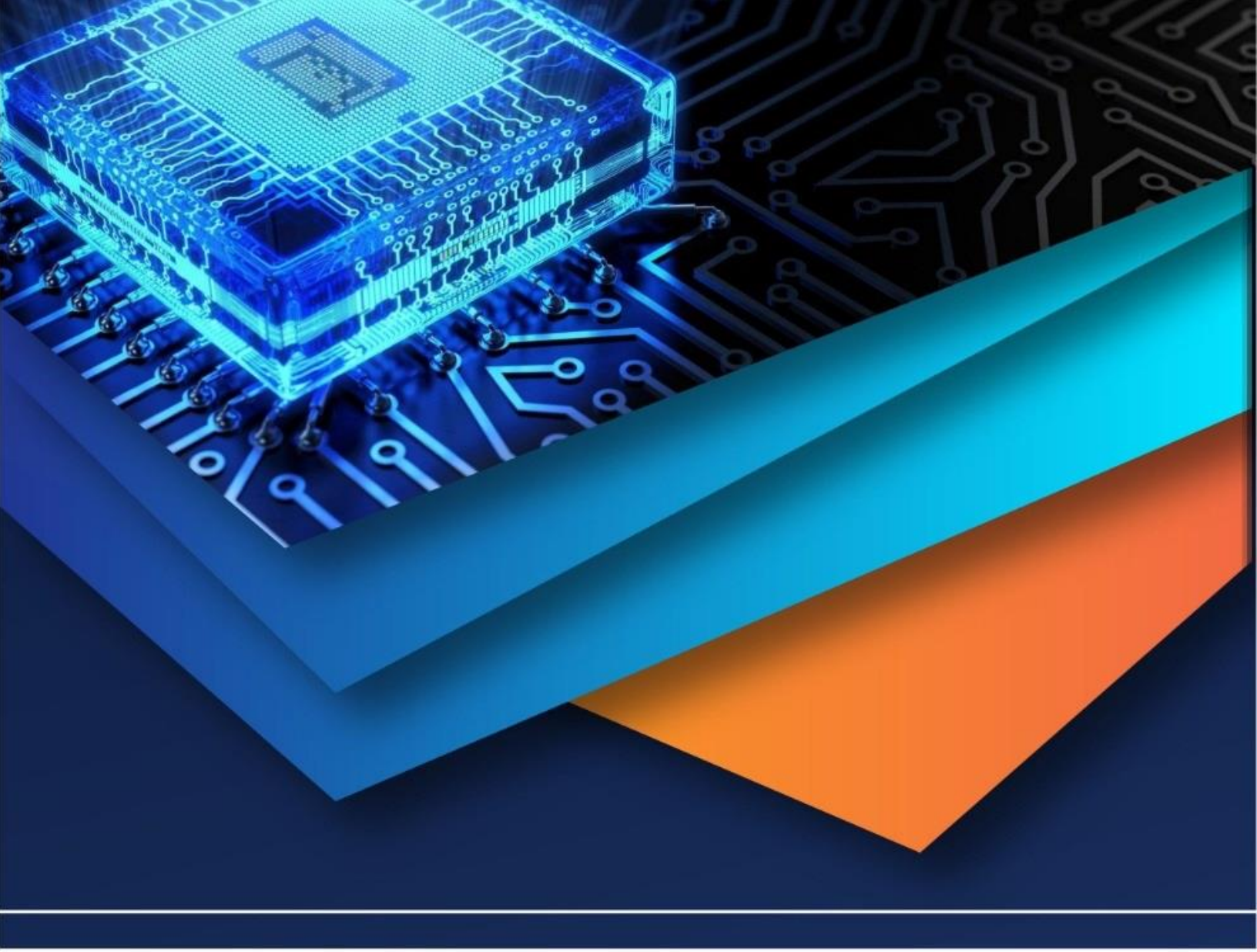

Fig. 1. Complete system architecture: multi-stage pipeline from multilingual

## REFERENCES

[1] D. Nutbeam, "Health literacy as a public health goal: a challenge for contemporary health education and communication strategies into the 21st century," Health Promotion International, vol. 15, no. 3, pp. 259–267, 2000.
[2] N. D. Berkman et al., "Low health literacy and health outcomes: an updated systematic review," Annals of Internal Medicine, vol. 155, no. 2, pp. 97–107, 2011.
[3] M. Kutner, E. Greenberg, Y. Jin, and C. Paulsen, "The health literacy of America's adults: Results from the 2003 National Assessment of Adult Literacy," U.S. Department of Education, NCES, Washington, DC, 2006.
[4] World Health Organization, "Global strategy on human resources for health: Workforce 2030," Geneva, Switzerland, 2016.
[5] D. Schillinger et al., "Association of health literacy with diabetes outcomes," JAMA, vol. 288, no. 4, pp. 475–482, 2002.
[6] J. A. Vernon, A. Trujillo, S. Rosenbaum, and B. DeBuono, "Low health literacy: Implications for national health policy," University of Connecticut, Dept. of Finance, 2007.
[7] A. Radford et al., "Robust speech recognition via large-scale weak supervision," arXiv preprint arXiv:2212.04356, 2022.
[8] E. Alsentzer et al., "Publicly available clinical BERT embeddings," in Proc. 2nd Clinical NLP Workshop, 2019, pp. 72–78.
[9] HSUIT Development Team, "HSUIT Comprehensive Final Report 2026: Integrated AI Healthcare Platform for Clinical Documentation and Analysis," Internal Technical Report, 2026.
[10] Office of the Registrar General and Census Commissioner, "Census of India 2011: Language," Ministry of Home Affairs, Government of India, 2011.
[11] B. G. Arndt et al., "Tethered to the EHR: Primary care physician workload assessment using EHR event log data and time-motion ob-servations," Annals of Family Medicine, vol. 15, no. 5, pp. 419–426, 2017.
[12] World Health Organization, "Health literacy: The solid facts," WHO Regional Office for Europe, Copenhagen, 2013.
[13] B. H. Juang and L. R. Rabiner, "Hidden Markov models for speech recognition," Technometrics, vol. 33, no. 3, pp. 251–272, 1991.
[14] A. Zafar, C. Overhage, and C. J. McDonald, "Continuous speech recognition for clinicians," J. Am. Med. Inform. Assoc., vol. 6, no. 3, pp. 195–204, 1999.
[15] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," Proc. IEEE, vol. 77, no. 2, pp. 257–286, 1989.

[16] G. Hinton et al., "Deep neural networks for acoustic modeling in speech recognition," IEEE Signal Process. Mag., vol. 29, no. 6, pp. 82–97, 2012.
[17] A. Graves, A.-r. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in Proc. ICASSP, 2013, pp. 6645–6649.
[18] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," arXiv preprint arXiv:1409.0473, 2014.
[19] V. Prabhu and A. Kannan, "Evaluation of Whisper for medical speech recognition: A comparative study," J. Biomed. Inform., vol. 142, p. 104384, 2023.
[20] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in Proc. NAACL-HLT, 2019, pp. 4171–4186.
[21] J. Lee et al., "BioBERT: a pre-trained biomedical language representa-tion model for biomedical text mining," Bioinformatics, vol. 36, no. 4, pp. 1234–1240, 2020.
[22] K. Huang, J. Altosaar, and R. Ranganath, "ClinicalBERT: Modeling clinical notes and predicting hospital readmission," arXiv preprint arXiv:1904.05342, 2019.
[23] O¨. Uzuner, B. R. South, S. Shen, and S. L. DuVall, "2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text," J. Am. Med. Inform. Assoc., vol. 18, no. 5, pp. 552–556, 2011.
[24] G. Lample, M. Ballesteros, S. Subramanian, K. Kawakami, and C. Dyer, "Neural architectures for named entity recognition," in Proc. NAACL-HLT, 2016, pp. 260–270.
[25] W. Sun, A. Rumshisky, and O. Uzuner, "Evaluating temporal relations in clinical text: 2012 i2b2 Challenge," J. Am. Med. Inform. Assoc., vol. 20, no. 5, pp. 806–813, 2013.
[26] X. Wang et al., "Cross-type biomedical named entity recognition with deep multi-task learning," Bioinformatics, vol. 35, no. 10, pp. 1745–1752, 2019.
[27] R. Mihalcea and P. Tarau, "TextRank: Bringing order into text," in Proc. EMNLP, 2004, pp. 404–411.
[28] C. Raffel et al., "Exploring the limits of transfer learning with a unified text-to-text transformer," J. Mach. Learn. Res., vol. 21, pp. 1–67, 2020.
[29] M. Lewis et al., "BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," in Proc. ACL, 2020, pp. 7871–7880.
[30] Y. Zhang, D. Merck, E. B. Tsai, C. D. Manning, and C. P. Langlotz, "Leveraging pretrained models for automatic summarization of doctor-patient conversations," in Proc. EMNLP Workshop on Health Text Mining, 2020, pp. 67–73.
[31] J. Zhang, Y. Zhao, M. Saleh, and P. J. Liu, "PEGASUS: Pre-training with extracted gap-sentences for abstractive summarization," in Proc. ICML, 2020, pp. 11328–11339.
[32] A. Conneau et al., "Unsupervised cross-lingual representation learning at scale," in Proc. ACL, 2020, pp. 8440–8451.
[33] A. Kumar, S. Singh, and R. Sharma, "Building multilingual medical corpora for Indian languages: Challenges and opportunities," in Proc. LREC, 2022, pp. 4567–4575.
[34] V. Pandey and M. Gupta, "Multilingual health information extraction for low-resource Indian languages," J. Biomed. Inform., vol. 118, p. 103789, 2021.
[35] R. Smith, "An overview of the Tesseract OCR engine," in Proc. ICDAR, 2007, pp. 629–633.
[36] PaddlePaddle Team, "PaddleOCR: Awesome multilingual OCR toolk-its," GitHub repository, 2020. [Online]. Available: https://github.com/PaddlePaddle/PaddleOCR
[37] J.-M. Valin, "A hybrid DSP/deep learning approach to real-time full-band speech enhancement," in Proc. MMSP, 2018, pp. 1–5.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

# INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)