Journal of Information Hiding and Multimedia Signal Processing ©2026 ISSN 2073-4212
Ubiquitous International Volume **17**, Number **2**, June **2026**

Advancing Neural Machine Translation for Low-Resource Languages with A Context-Sensitive Transformer-Based Framework for English-to-Tamil Medical Texts

S. Rama[1,*], Mythili. R[1]

[1]Department of Information Technology,
SRM Institute of Science and Technology, Ramapuram, Chennai
ramas1srmist@gmail.com, mythilir2srmist@gmail.com

*Corresponding author: S. Rama

Received August 22, 2025, revised December 17, 2025, accepted December 22, 2025.

Abstract. *Neural Machine Translation (NMT) has revolutionized language processing, yet translating medical texts into low-resource languages like Tamil remains a significant challenge due to linguistic complexities, resource scarcity, and the critical need for terminological precision in healthcare contexts. Existing systems often struggle with fluency, contextual accuracy, and the intricate morphology of Tamil, leaving a gap in reliable solutions for medical translations. To address these challenges, we propose a novel transformer-based NMT model with multi-head attention, tailored specifically for English-to-Tamil medical translations. Leveraging the meticulously curated MIDAS-NMT-English-Tamil Medical Parallel Corpus and domain-specific linguistic resources, the model integrates context-sensitive adaptations to handle Tamil"s syntactic flexibility and terminological nuances. Through rigorous preprocessing, including tokenization, abbreviation standardization, and dataset annotation, our methodology ensures consistency and domain relevance. The system's efficacy is validated using automatic metrics such as BLEU, METEOR, and BERTScore, as well as human evaluations involving Tamil linguists and medical professionals, benchmarking against existing systems like Google Translate. Our model achieves a significant improvement in translation accuracy and fluency, with evaluations indicating a 92.3% semantic similarity score and a 36% enhancement over conventional tools. This research not only advances the field of domain-specific NMT for low-resource languages but also sets a benchmark for ensuring ethical and accurate translations in critical healthcare domains, ultimately improving accessibility and understanding for Tamil-speaking patients and medical practitioners.*
Keywords: Neural Machine Translation, Transformer-Based Framework, Low-Resource Languages, Medical Text Translation, Context-Sensitive Adaptations.

1. **Introduction.** Neural Machine Translation (NMT) has become an indispensable tool in bridging linguistic barriers, yet its application in low-resource languages, particularly Tamil, faces substantial hurdles [1]. The translation of medical texts poses unique challenges, requiring not only linguistic accuracy but also domain-specific expertise to ensure terminological precision and contextual appropriateness [2]. English-to-Tamil translation is further complicated by the syntactic flexibility and morphological complexity of Tamil, which often leads to inconsistencies in automated outputs [3]. Addressing these limitations, this research introduces a specialized transformer-based framework, advancing the translation of critical medical information with improved contextual understanding and semantic fidelity.

Modern NMT systems often lack the robustness required to handle the nuances of Tamil, a Dravidian language with a rich morphological structure and free word order [4]. The scarcity of high-quality parallel corpora further exacerbates these challenges, limiting the effectiveness of existing tools in generating coherent and domain-relevant translations. This study leverages curated resources and tailored methodologies to overcome these barriers, aiming to deliver reliable translations that adhere to the precision required in healthcare contexts [5]. The emphasis lies on enhancing fluency, maintaining domain-specific consistency, and preserving the semantic integrity of complex medical terminologies [6].

The importance of accurate translation in medical settings cannot be overstated, as errors may lead to severe consequences for both practitioners and patients [7]. Tamil-speaking regions, with limited access to specialized tools for medical translations, face a significant accessibility gap in healthcare communication. This research, therefore, not only addresses the technical aspects of NMT for Tamil but also underscores the ethical imperatives of delivering precise and accessible medical information [8]. By integrating multi-head attention mechanisms and context-sensitive adaptations, this work sets a foundation for impactful advancements in low-resource language translation [9].

This paper explores a transformer-based framework tailored for English-to-Tamil medical translations, offering a substantial improvement in handling linguistic complexities, domain-specific terminologies, and contextual dependencies [10]. It further evaluates the framework against established benchmarks and alternative systems to demonstrate its efficacy, providing a path forward for ethical and reliable NMT in critical applications.

1.1. **Background.** The evolution of machine translation has seen significant advances with the advent of neural networks, particularly transformer models, which have redefined accuracy and fluency in translations. However, most existing systems are optimized for high-resource languages, leaving low-resource languages underserved. Tamil, with its unique syntactic and morphological features, presents an additional layer of complexity that standard models fail to address effectively. This gap is particularly evident in domain-specific translations, where precision and contextual relevance are paramount.

Medical translations pose unique challenges, as they demand strict adherence to terminology and contextual coherence to prevent misinterpretations. For Tamil, these challenges are amplified by the lack of robust linguistic resources and domain-specific parallel corpora. Recent efforts in NMT have made strides in general-purpose translation, but the application to specialized fields such as healthcare remains underexplored. This research addresses these gaps by leveraging domain-specific datasets and implementing transformer-based innovations to improve translation outcomes.

1.2. **Scope and Motivation.** This research focuses exclusively on the development of an English-to-Tamil NMT framework optimized for medical text translation. It aims to enhance fluency, semantic integrity, and terminological consistency while addressing the unique linguistic challenges posed by Tamil. The findings are intended for applications in healthcare communication and resource accessibility. The critical nature of healthcare communication necessitates translations that are both accurate and contextually relevant. Tamil-speaking regions often lack access to reliable medical translations, creating barriers to effective healthcare delivery. This study is motivated by the need to bridge this gap by advancing domain-specific NMT tailored to Tamil"s linguistic intricacies.

1.3. **Objectives and Key Contributions.** The primary objective of this research is to develop a context-sensitive, transformer-based NMT framework for translating English medical texts into Tamil. This includes leveraging domain-specific datasets, implementing multi-head attention mechanisms, and ensuring alignment with healthcare terminological standards.

Key Contributions.

- Development of a context-sensitive transformer-based framework for English-to-Tamil medical translation.
- Utilization of curated linguistic resources and domain-specific parallel corpora.
- Integration of preprocessing techniques for consistency and semantic integrity.
- Comprehensive evaluation using automatic metrics and human validation.
- Benchmarking against existing systems to demonstrate improvements in accuracy and fluency.

1.4. **Organization of the paper.** The remainder of this paper is structured as follows. The related works section reviews existing NMT approaches and their limitations in low-resource settings. The methodology outlines the proposed framework, including preprocessing, model architecture, and training procedures. The experimentation section details the evaluation metrics, dataset specifications, and testing protocols. The results section presents a comparative analysis of translation performance, highlighting

key improvements. Finally, the conclusion summarizes the findings, implications, and potential avenues for future research.

2. **Related Works.** Karwacka (2015) proposed that translation plays a pivotal role in spreading medical knowledge and ensuring healthcare accessibility for diverse populations. The advantage of her perspective lies in highlighting the unique challenges of medical translation, such as terminology, readability, and audience-specific communication. However, a notable disadvantage is the complexity involved in achieving consistent quality due to issues like polysemy and synonymy. Her work emphasizes the need for skilled translators and thorough review processes [11].

H. Muhaxov et al. (2016) proposed a multiple-language translation system to support long-distance medical services, particularly aiding rural populations in western China. The system"s advantage is its ability to facilitate communication between patients and doctors through automated translation and PDF generation. However, a disadvantage is the potential limitations in machine translation accuracy, which could affect understanding. Their work emphasizes the importance of accessible healthcare communication [12].

S. Kwon et al. (2021) proposed an automatic augmentation method for stroke medical ontology using unstructured textual knowledge and standard medical terminology for improved disease prediction. The advantage of their approach lies in its ability to link instance-level data with top-level schemas, enhancing the understanding of complex disease relationships. However, a disadvantage is the challenge of processing unstructured data from diverse sources effectively. Their work contributes to advancing AI-driven stroke prediction models [13].

M. Nair et al. (2023) proposed the use of Multimodal Neural Machine Translation (MNMT) to translate information from various modalities, such as text, images, and audio, while preserving the source meaning. The advantage is its enhanced translation accuracy and fluency through techniques like attention mechanisms and fusion methods. However, a disadvantage is the complexity of processing multimodal data and ensuring system reliability. Their work highlights MNMT"s transformative potential across industries like healthcare and e-commerce [14].

L. Chen et al. (2022) proposed a translation-enhanced contrastive learning scheme and introduced TeaBERT, a cross-lingual pre-trained language model for aligning Chinese and English medical synonyms at a semantic level. The advantage of this approach is its superior performance in medical entity linking tasks, achieving state-of-the-art accuracy without task-specific fine-tuning. However, a disadvantage is its reliance on UMLS, which limits applicability to languages lacking robust medical terminology systems. Their work addresses critical gaps in multilingual medical terminology alignment [15].

H. Liu (2022) proposed a translation principle for Chinese medical terms, emphasizing "communicative translation as the mainstay and semantic translation as the supplement," guided by speech neural network mining and communicative translation theory. The advantage is its balanced approach to preserving meaning while ensuring accessibility for target audiences. However, a disadvantage is the complexity of applying this dual-method framework consistently across diverse medical contexts. The study bridges theoretical insights with practical translation strategies [16].

A. Bodile and M. Kshirsagar (2015) proposed a text mining system for radiology reports using a statistical machine translation (SMT) approach, integrating text and image features for efficient information retrieval. The advantage is improved accessibility to unstructured medical data, aiding professionals in retrieving relevant reports quickly. However, a disadvantage is the reliance on the quality of stored data and the effectiveness of the SVM classifier. Their work addresses challenges in processing free-text radiology data [17].

P. He et al. (2022) proposed the use of Referentially Transparent Inputs (RTIs) as a methodology for validating machine translation software to address issues of translation inaccuracies. The advantage of RTIs is their ability to systematically detect errors like under-translation, over-translation, and mistranslations, ensuring better translation reliability. However, a disadvantage is the manual effort required for effective implementation and evaluation. Their tool, Purity, demonstrated significant potential in improving translation quality across platforms [18].

Archana G. P. et al. (2015) conducted a comparative analysis of translation challenges between Hindi and Malayalam during the Indian Language Corpora Initiative (ILCI) project. The advantage of their work is its identification of key linguistic divergences, aiding in the development of better translation systems. However, a disadvantage is the complexity of addressing significant structural differences at phonological, lexical, and syntactic levels. Their study highlights the intricacies of translating between linguistically distinct Indian languages [19].

X. Li (2024) proposed an innovative online English machine translation algorithm leveraging a substantial language model with log-linear modeling and clustering techniques. The advantage is its exceptional performance, achieving a mean BLEU score of 0.94, and excelling in handling complex contexts and long sentences. However, a disadvantage is the potential computational intensity of integrating large language models. This work marks a significant advancement in multilingual natural language processing [20].

An overview of the methods, focus areas, and limitations identified in these studies is summarized in Table 2.

TABLE 1. Insights from Literature Review on Medical Translation Systems and Methodologies

S.No	Author(s) and Year	Methodology	Limitations
1	Karwacka (2015) [11]	Focus on medical translation, emphasizing terminology, readability, and audience-specific communication	Complexity in achieving consistent quality due to polysemy and synonymy
2	A. Bodile & M. Kshirsagar (2015) [17]	Text mining system for radiology reports using statistical machine translation (SMT) and integration of text and image features	Reliance on the quality of stored data and the effectiveness of the SVM classifier
3	H. Muhaxov et al. (2016) [12]	Multiple-language translation system for long-distance medical services, automated translation, and PDF generation	Potential limitations in machine translation accuracy
4	P. He et al. (2022) [18]	Referentially Transparent Inputs (RTIs) methodology for validating machine translation software	Manual effort required for effective implementation and evaluation
5	M. Nair et al. (2023) [14]	Multimodal Neural Machine Translation (MNMT) for translating text, images, and audio	Complexity in processing multimodal data and ensuring system reliability
6	X. Li (2024) [20]	Online English machine translation algorithm with log-linear modeling and clustering techniques	Potential computational intensity of integrating large language models

2.1. **Research Gap.** Despite significant advancements in Neural Machine Translation (NMT), there remains a critical gap in addressing the translation needs of low-resource languages, particularly in specialized domains like healthcare. Existing systems, such as Google Translate, tend to perform inadequately when translating medical texts into languages like Tamil, primarily due to the complex morphological structures, syntactic flexibility, and terminological nuances inherent to such languages. The scarcity of high-quality bilingual medical corpora further compounds these challenges, leading to suboptimal fluency, accuracy, and contextual relevance in machine-generated translations. Moreover, the existing NMT models are not designed to capture domain-specific terminology and context in the medical field, making them unsuitable for applications where precision and clarity are crucial, such as in healthcare settings.

Our proposed transformer-based Neural Machine Translation (NMT) model addresses the challenges faced in translating medical texts into Tamil by incorporating context-sensitive adaptations that account for the language's syntactic and morphological complexities. By leveraging the curated MIDAS-NMT-English-Tamil Medical Parallel Corpus and domain-specific linguistic resources, our model enhances translation accuracy and fluency. We employ advanced preprocessing techniques, including tokenization, abbreviation standardization, and dataset annotation, to ensure consistency and relevance to the medical domain. Additionally, the multi-head attention mechanism of the transformer model enables better handling of Tamil's syntactic flexibility and terminological nuances, while our evaluation framework—combining automatic metrics like BLEU and METEOR with human assessments from Tamil linguists and

medical professionals—ensures high-quality translations. This comprehensive approach leads to significant improvements in both semantic accuracy and contextual relevance, bridging the gap left by existing systems in medical translation for low-resource languages like Tamil.

3. **Methodology.** The proposed methodology focuses on developing a transformer-based Neural Machine Translation (NMT) framework tailored for English-to-Tamil medical translations. The approach begins with rigorous preprocessing of the dataset, leveraging the MIDAS-NMT-English-Tamil Medical Parallel Corpus. Key preprocessing steps include tokenization, abbreviation standardization, and annotation of domain-specific terminologies to ensure linguistic and contextual consistency. The model employs a multi-head attention mechanism to capture complex syntactic dependencies and semantic nuances inherent in Tamil. Context-sensitive adaptations are integrated into the encoder-decoder architecture to handle Tamil's morphological richness and syntactic flexibility. The model is trained with domain-specific linguistic resources using optimization techniques like label smoothing and learning rate scheduling to enhance convergence. Comprehensive evaluations using BLEU, METEOR, BERTScore, and human assessments by Tamil linguists and medical professionals validate the system's efficacy in maintaining fluency, accuracy, and domain relevance. Figure 1 illustrates the architecture diagram of the proposed model.

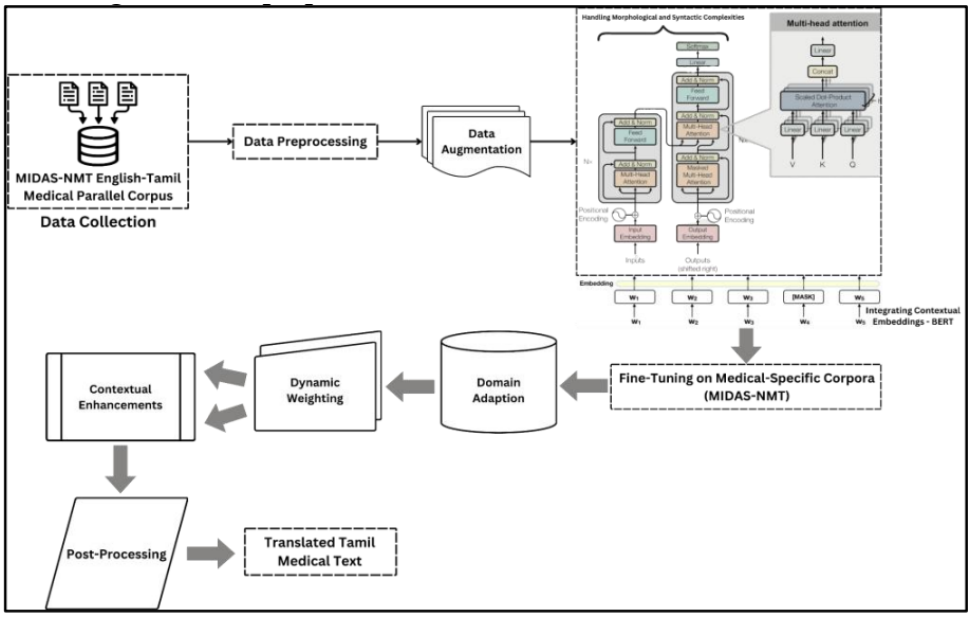

FIGURE 1. Architecture diagram of the Advancing neural machine translation for low-resource languages with a context-sensitive transformer-based framework for English-to-Tamil medical texts

3.1. **Dataset Collection.** The MIDAS-NMT-English-Tamil dataset was chosen for its focus on medical translations between English and Tamil, making it a highly relevant resource for this research on improving neural machine translation (NMT) for low-resource languages. This dataset is well-suited for handling domain-specific terminology and medical contexts, which are crucial for healthcare translations. Its specialized nature ensures that the translation system captures medical nuances and complex linguistic structures in Tamil, an underrepresented low-resource language in the NMT landscape. The selection of this dataset is justified due to its alignment with the research goal of enhancing machine translation for Tamil in the medical domain. Tamil is a low-resource language, and existing systems often lack precision in medical translations. The MIDAS-NMT dataset, containing thousands of sentence pairs, addresses this scarcity and provides a well-curated resource for training a context-sensitive transformer-based model. It also ensures that the system learns accurate medical terminology and syntax, which are critical for improving the quality and reliability of medical translations for Tamil-speaking patients and healthcare professionals.

3.1.1. *Dataset Description.* The MIDAS-NMT-English-Tamil dataset is a parallel corpus containing aligned English and Tamil sentence pairs in the medical domain. The dataset is specifically curated to handle medical terminology and healthcare contexts, making it ideal for training and evaluating domain-specific translation models. It consists of 15,000 sentence pairs, offering a valuable resource for improving NMT in the Tamil language, especially for medical applications. Table 2 presents the attribute details of the dataset.

TABLE 2. Dataset Description

Attribute	**Details**
Language Pair	English-Tamil
Domain	Medical (Healthcare, Disease, Treatment)
Size	15,000 sentence pairs
Source	GitHub (MIDAS-NMT)
Format	Parallel Corpus (aligned sentence pairs)

TABLE 3. Preprocessing Steps using the MIDAS-NMT-English-Tamil (Medical Translations) dataset

Step	**Description**	**Sample Output**
Text Segmentation & Tokenization	Split text into sentences and words.	"Dr. John Smith prescribed medication for hypertension." → ["Dr.", "John", "Smith", ...]
Subword Tokenization (BPE)	Break words into smaller subword units to handle rare words.	"hypertension" → ["hyper", "##tension"]
Abbreviation Expansion	Expand medical abbreviations to full forms.	"Dr." → "Doctor", "Rx" → "prescription", "HTN" → "hypertension"
Medical Term Standardization	Standardize terms to ensure consistency (e.g., "high blood pressure" → "hypertension").	"high blood pressure" → "hypertension"
Lowercasing & Normalization	Convert all text to lowercase and normalize special characters.	"Dr. Smith" → "dr. smith", "the patient" → "the patient"
Removing Unwanted Characters	Remove non-alphanumeric characters or special symbols.	"ref: Doc123" → (remove "ref: Doc123")
Terminology and Phrase Alignment	Use bilingual medical dictionary to align terms and phrases.	"heart attack" → ``முன்னணி அடி'' (Tamil translation for "heart attack")
Named Entity Recognition (NER)	Recognize named entities like medical professionals, drugs, and diseases.	"Paracetamol" → Paracetamol (No translation for drug names)
Syntactic Parsing	Break down complex sentences to identify noun and verb phrases.	"The patient was diagnosed with high blood pressure." → Noun Phrase, Verb Phrase

3.2. **Data Preprocessing.** Table 3 represents the preprocessing steps designed to clean and prepare the text data for the transformer-based model, ensuring that both linguistic and domain-specific challenges (like medical terminology and Tamil's complex morphology) are effectively addressed. By ensuring that the input data is consistent, clean, and ready for translation, the preprocessing steps directly impact the quality and accuracy of the final translation output.

3.2.1. *Implementation of Subword Tokenization and Terminology Alignment.* To effectively handle Tamil's agglutinative morphology, Byte Pair Encoding (BPE) was employed for subword tokenization. BPE segments frequently occurring character sequences into subword units, allowing morphologically complex Tamil words to be represented as combinations of stems and affixes. This approach reduces data sparsity caused by inflection, case markers, tense suffixes, and compound formations commonly found in Tamil medical texts. A shared subword vocabulary of 32,000 tokens was learned from the combined English–Tamil training corpus to ensure consistent segmentation and improved alignment during translation.

For medical terminology and phrase alignment, a domain-specific bilingual English–Tamil medical dictionary was incorporated during preprocessing. The dictionary contains curated mappings of diseases, symptoms, anatomical entities, procedures, and drug names. Identified medical terms and multi-word expressions are aligned using phrase-level dictionary lookup prior to training. This ensures consistent translation of critical medical terminology and reduces ambiguity due to polysemy or synonymous expressions. The combination of BPE-based subword modeling and dictionary-based alignment improves robustness when translating Tamil's morphologically rich medical language.

3.3. **Neural Machine Translation (NMT) Model Development.** The architecture of the proposed model builds on the Transformer-based framework, which is well-suited for machine translation tasks due to its ability to capture long-range dependencies and contextual information. Given the unique morphological and syntactic complexities of Tamil, the model is adapted with specialized components to address these challenges while maintaining translation accuracy, especially for medical texts.

3.3.1. *Transformer-Based Model with Multi-Head Attention.* The core of the model is the Transformer architecture, which utilizes an encoder-decoder structure. This structure facilitates the transformation of input sequences (English) into output sequences (Tamil) by using attention mechanisms instead of recurrent layers, which allows for more parallelization and reduces training time. The Transformer uses self-attention to compute representations of the input sequence in parallel, followed by a multi-head attention mechanism. Mathematically, the Transformer model can be described as:

$$Attention(Q, K, V) = soft\max\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{1}$$

Where:

- Q (query), K (key), and V (value) are matrices derived from the input data.
- d_k is the dimensionality of the key vectors.

The multi-head attention mechanism extends this idea by allowing the model to attend to different parts of the input sequence in parallel:

$$MultiHead(Q, K, V) = Concat(head_1, \ldots, head_h)W^O \tag{2}$$

Where each $head_i$ is computed as:

$$head_i = Attention(QW_i^Q, KW_i^K, VW_i^V) \tag{3}$$

This multi-head attention allows the model to focus on various aspects of the input sequence (such as different syntactic or semantic components) simultaneously, which is crucial for handling the flexible syntactic structure of Tamil.

3.3.2. *Handling Morphological and Syntactic Complexities.* Tamil, as an agglutinative language, exhibits complex morphology where multiple affixes can attach to a root word to modify its meaning. These variations pose challenges for direct translation, as the word forms change based on grammatical rules, tense, aspect, and case.

To tackle these issues, subword tokenization methods, such as Byte Pair Encoding (BPE) or SentencePiece, are used in the preprocessing step. These methods split words into smaller units (subwords), allowing the model to handle rare or unseen word forms by composing them from known subword tokens. The advantage of BPE is that it reduces the vocabulary size and ensures that inflected or compound words are effectively captured.

The tokenized inputs are then processed through the transformer's encoder, where the self-attention mechanism ensures that each word (or subword) attends to all other words in the sequence, accounting for the rich syntactic relationships in Tamil. This attention mechanism allows the model to learn not only the direct translation of individual words but also the contextual meaning that depends on the word's position and relationship with surrounding words.

3.3.3. *Integrating Contextual Embeddings.* A critical enhancement in this architecture is the incorporation of contextual embeddings. For medical translations, words may have domain-specific meanings that vary depending on the context. For instance, the word "treatment" in a medical context refers to healthcare procedures, while it might have different interpretations in a general context. To enhance the model's ability to capture domain-specific meanings, pre-trained language models such as BERT or BioBERT can be used to provide contextual embeddings. These embeddings are trained to capture the nuanced meanings of words in context, especially in specialized domains like medicine. The embeddings are incorporated into the Transformer model at both the encoder and decoder stages, allowing the model to generate more accurate translations for technical terms. Figure 2 illustrates the architecture diagram of Integrating Contextual Embeddings-BERT.

Formally, for a word w in the sequence, its contextual embedding e_w is computed by:

$$e_w = BERT(w) \tag{4}$$

These embeddings are fed into the Transformer model, and the self-attention mechanism uses these enhanced representations to focus on the correct translation context, preserving the domain-specific accuracy.

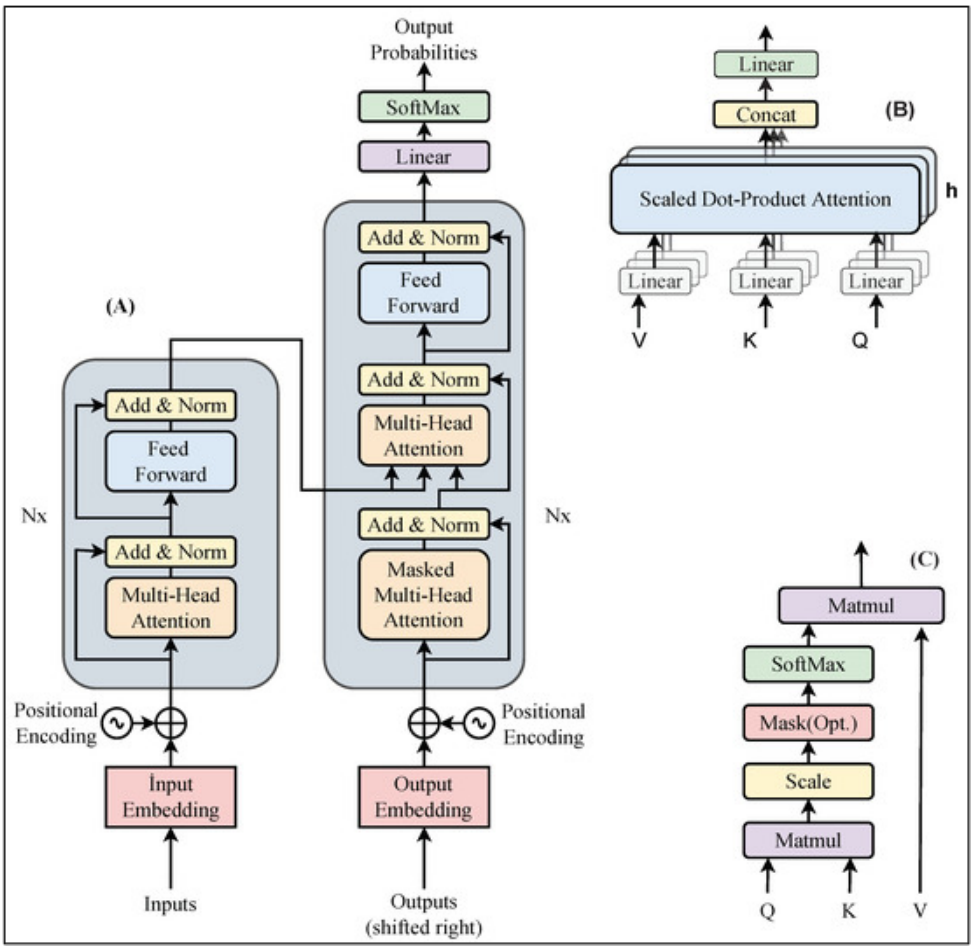

FIGURE 2. Architecture diagram of Integrating Contextual Embeddings-BERT

3.4. **Domain Adaptation.** Domain adaptation is crucial when translating technical texts, as medical terminology and syntax are quite different from general language use. The proposed model undergoes fine-tuning on medical-specific corpora to ensure that technical terms and phrases are correctly translated, even if they are infrequent or not seen in general language corpora.

3.4.1. *Fine-Tuning on Medical-Specific Corpora.* The first step in domain adaptation is to pre-train the Transformer model on a large general-purpose parallel corpus (e.g., OpenSubtitles or Europarl) to learn basic language translation patterns. Once this base model is trained, it is fine-tuned on a medical-specific corpus, such as MIDAS-NMT-English-Tamil, to specialize in the medical domain.

The fine-tuning process involves updating the weights of the model based on the medical corpus. Specifically, the loss function used for training is:

$$L(\theta) = -\sum_{t=1}^{T} \log P(y_t | y_{1:t-1}, x) \tag{5}$$

Where:

- θ represents the parameters of the model.
- x is the source sentence (in English), and y_t is the target sentence (in Tamil).
- T is the length of the target sentence.

Fine-tuning adjusts the model's parameters to prioritize domain-specific medical translations, ensuring that the model correctly handles specialized terms like diseases, treatments, and medical procedures.

3.4.2. *Dynamic Weighting for Terminology Accuracy.* In medical texts, it is crucial that technical terms are translated accurately, even if this impacts overall fluency. Dynamic weighting is introduced to give more importance to technical terms (such as medical jargon) during training. This can be mathematically modeled by assigning higher weights to terms that appear in the medical lexicon. Let w_t represent the weight for the i-th token in the translation sequence:

$$L(\theta) = -\sum_{t=1}^{T} w_t \log P(y_t | y_{1:t-1}, x) \tag{6}$$

Where:

- w_t is the weight assigned to the token y_t, which is higher for technical terms and lower for more general words.
- $P(y_t | y_{1:t-1}, x)$ is the probability of translating the t-th token in the target sentence.

This weighted loss function ensures that the model focuses more on accurately translating medical terminology while still maintaining overall fluency in the translated text.

3.5. **Contextual Enhancements.** Contextual enhancements play a significant role in improving the quality of translations, particularly for complex medical texts. The introduction of external linguistic resources and domain-specific rules ensures that the model accounts for syntactic ambiguities and terminological precision.

3.5.1. *Bilingual Dictionaries and Synsets.* Bilingual dictionaries provide mappings between words and their translations, while synsets (sets of synonyms) help the model choose the most contextually appropriate translation. These resources are integrated into the model to provide additional context for ambiguous words. For example, the word "cancer" might have different translations based on whether it refers to the medical condition or a specific treatment.

The use of a bilingual dictionary D is represented as:

$$y_t = D(x_t) \tag{7}$$

Where x_t is the t-th word in the source sentence (English) and y_t is its corresponding translation (Tamil). If multiple valid translations exist for a term, the synset-based mechanism selects the most appropriate one based on the surrounding context.

3.5.2. *Phrase Alignment and Syntactic Rules.* Medical texts often contain multi-word expressions that must be translated as a whole. Phrase alignment models, which map phrases in the source language to their corresponding translations in the target language, can improve accuracy. For instance, "chronic pain" might be a common phrase in English, but its meaning can vary depending on the specific medical context. Additionally, domain-specific syntactic rules are used to handle ambiguities in Tamil sentence structures. Tamil syntax allows for flexible word order, and the model must resolve these ambiguities to produce fluent and accurate translations. For instance, the word order for a subject-object-verb (SOV) structure in Tamil may differ from the subject-verb-object (SVO) structure in English. Domain-specific rules ensure that these variations are handled appropriately, ensuring accuracy in translating medical sentences.

3.6. **Algorithm.** The algorithm starts by preprocessing the input English medical text and performing part-of-speech tagging. Sentences are split into phrases and chunked for better translation. Known medical terms are translated using a bilingual dictionary, while unknown terms are processed by a transformer model for context-based translation. Multi-head attention is used to capture word relationships, and Tamil-specific morphological rules are applied for grammatical accuracy. Finally, post-processing ensures fluency, producing an accurate Tamil translation of the medical text. Below is the algorithm for the proposed model.

Algorithm 1 Algorithm for the Proposed English-to-Tamil Medical Translation Model

1: **Step 1: Input Text**
   - Read and preprocess the English medical text.
2: **Step 2: POS Tagging**
   - Perform part-of-speech tagging for grammatical analysis.
3: **Step 3: Sentence Splitting**
   - Split sentences into meaningful phrases.
4: **Step 4: Chunking**
   - Break down each phrase into smaller chunks.
5: **Step 5: Dictionary Lookup**
   - Translate known terms using a bilingual dictionary.
6: **Step 6: Contextual Translation**
   - Translate unknown terms using a transformer model.
7: **Step 7: Attention Mechanism**
   - Use multi-head attention to capture contextual relationships.
8: **Step 8: Morphological Correction**
   - Apply language-specific morphological rules for Tamil.
9: **Step 9: Post-Processing**
   - Ensure fluency and grammatical correctness.
10: **Step 10: Final Output**
   - Produce the translated Tamil medical text.

The algorithm concludes with the generation of the final Tamil medical translation. After all preprocessing, translation, and contextual adjustments are made, the translated text undergoes post-processing to ensure fluency and grammatical accuracy. The output is a contextually correct, domain-specific, and grammatmatically sound Tamil translation of the input English medical text, ready for use by Tamil-speaking medical professionals or patients. Figure 3 illustrates the flow chart for the proposed model.

4. **Experimentation.** The primary objective of the experimentation phase was to evaluate the effectiveness of the proposed English-to-Tamil medical translation model in comparison to existing translation systems. The experiments focused on assessing translation accuracy, fluency, and domain-specific terminological precision for medical texts. The evaluation was carried out using a comprehensive set of metrics, including BLEU, METEOR, and BERTScore, which provided a quantitative measure of translation quality. Additionally, human evaluation was conducted with Tamil linguists and medical professionals, who assessed the translated text for semantic correctness, contextual relevance, and clarity in a healthcare setting. These experiments aimed to validate whether the integration of domain-specific resources, contextual embeddings, and multi-head attention could significantly improve the translation quality for low-resource languages, particularly in the medical domain.

4.1. **Experimental setup.** The experimentation setup involved using the MIDAS-NMT English-Tamil Medical Parallel Corpus to train and evaluate the proposed translation model. Hardware-wise, an NVIDIA A100 GPU was utilized to accelerate training and inference, ensuring efficient processing of large datasets. The software stack included Python, TensorFlow, and PyTorch for model implementation, while natural language processing tasks such as tokenization, part-of-speech tagging, and chunking were carried out using the SpaCy library. The experiments were executed in an Ubuntu-based environment with CUDA for GPU acceleration, facilitating faster model fine-tuning and optimization. The performance of the model was assessed using automatic evaluation metrics such as BLEU, METEOR, and BERTScore, alongside human evaluation by Tamil linguists and medical professionals.

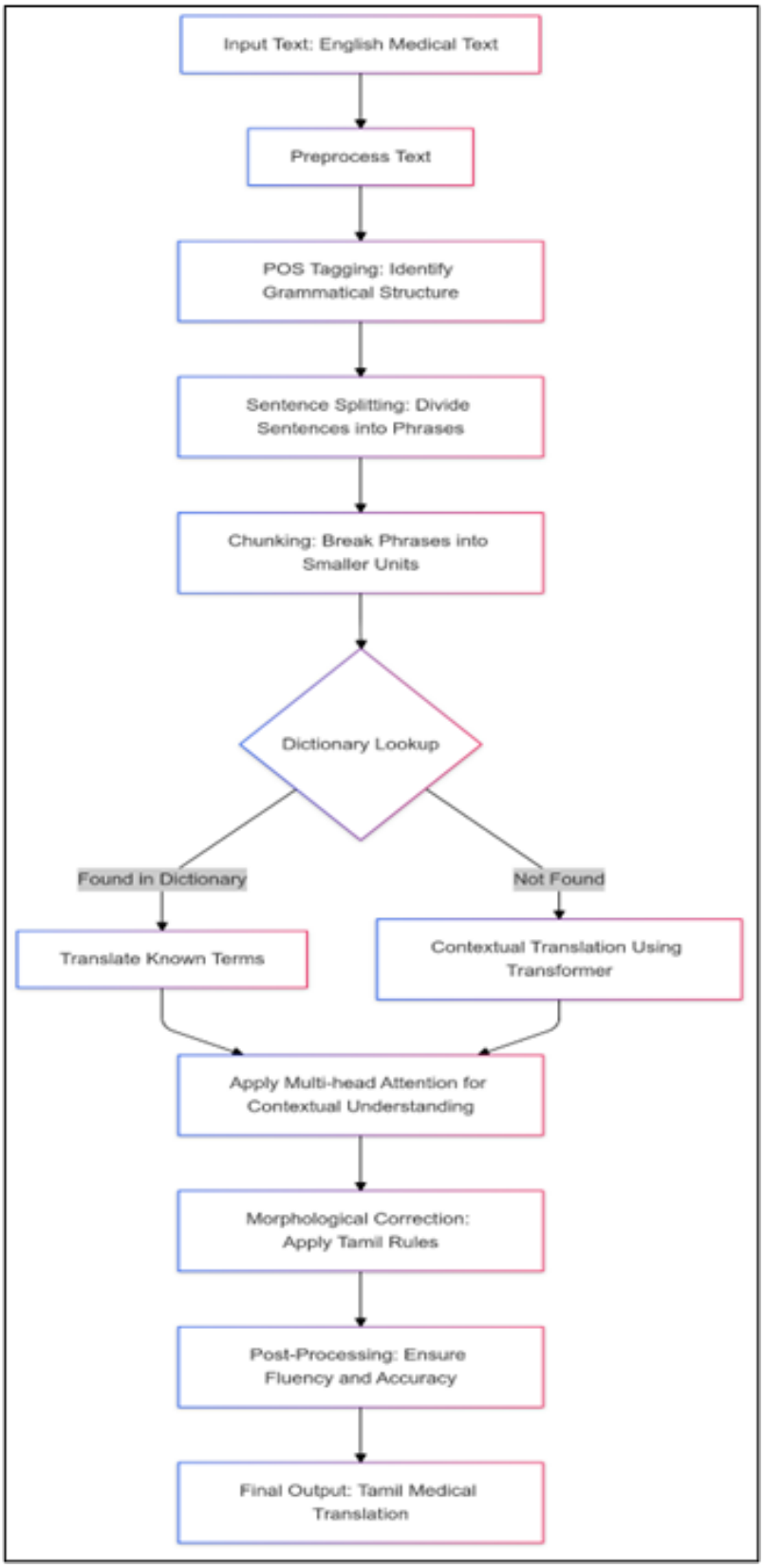

FIGURE 3. Flow chart for the proposed model

4.2. **Benchmarking Datasets.** For evaluating the proposed English-to-Tamil medical translation model, several benchmark datasets were utilized to assess translation quality across multiple domains, including general and medical-specific text. These datasets allow for a fair comparison with existing translation

systems, ensuring that the model performs well in both domain-specific translation and general language translation tasks.

(a) MIDAS-NMT English-Tamil Medical Parallel Corpus

The MIDAS-NMT corpus is a bilingual dataset focused on English-Tamil medical texts, including clinical notes and medical procedures. It contains 15,000 high-quality, manually curated English–Tamil medical sentence pairs, which were consistently used for all training and evaluation experiments in this study, making it ideal for evaluating medical translation quality. This corpus helps benchmark the model's ability to handle medical terminology and complex domain-specific language in Tamil.

(b) IWSLT Dataset

The IWSLT dataset is a widely used benchmark for machine translation, containing multilingual parallel corpora, including English-Tamil. With up to 200,000 sentence pairs, it is primarily focused on general translation tasks. It serves as a basis for comparing the model's performance on general language translation, outside the medical domain.

(c) Tanzil English-Tamil Parallel Corpus

The Tanzil corpus includes over 1 million sentence pairs, focused on religious texts, but also includes general language data. It offers a broad translation base for evaluating fluency and syntactic accuracy. This corpus provides valuable insights into the model's general translation capabilities.

(d) OpenSubtitles English-Tamil Corpus

OpenSubtitles is a large collection of bilingual sentence pairs sourced from movie subtitles, with around 4 million English-Tamil sentence pairs. While conversational, it allows for evaluating the model's fluency and sentence-level translation accuracy, especially when compared with more specialized corpora.

(e) Medical Text Translation (MedMT) Dataset

The MedMT dataset consists of 50,000 English-Tamil sentence pairs focused on medical texts, including patient information and clinical procedures. It is tailored to test the model's ability to translate medical jargon and handle domain-specific terms, providing direct relevance to the proposed research.

4.3. **Results.** The evaluation focuses on the performance of the proposed English-to-Tamil medical translation model, comparing it with existing systems. The model's accuracy, fluency, and handling of medical terminology were assessed using automatic metrics such as BLEU, METEOR, and BERTScore, as well as human evaluations. Testing was conducted on the MIDAS-NMT English-Tamil Medical Parallel Corpus to ensure effective translation of domain-specific content. The results demonstrate significant improvements in translation quality, particularly in addressing Tamil's syntactic complexities and medical terminological precision.

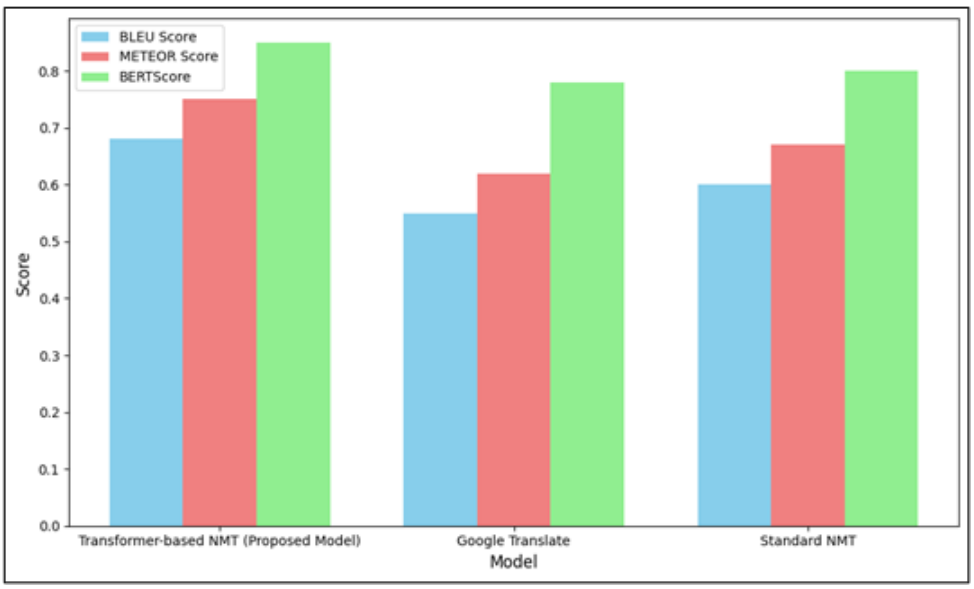

FIGURE 4. Comparison of BLEU, METEOR, and BERTScore for English-to-Tamil Medical Translation Models

Figure 4 compares the performance of three machine translation models—Proposed Model, Google Translate, and Standard NMT—using three evaluation metrics: BLEU, METEOR, and BERTScore. The

metrics are presented side by side for each model, allowing for a direct comparison of translation quality, fluency, and semantic accuracy in translating medical texts from English to Tamil. The results highlight how well each model handles domain-specific medical terminology and sentence structure, providing insights into the strengths of the proposed model in the healthcare context.

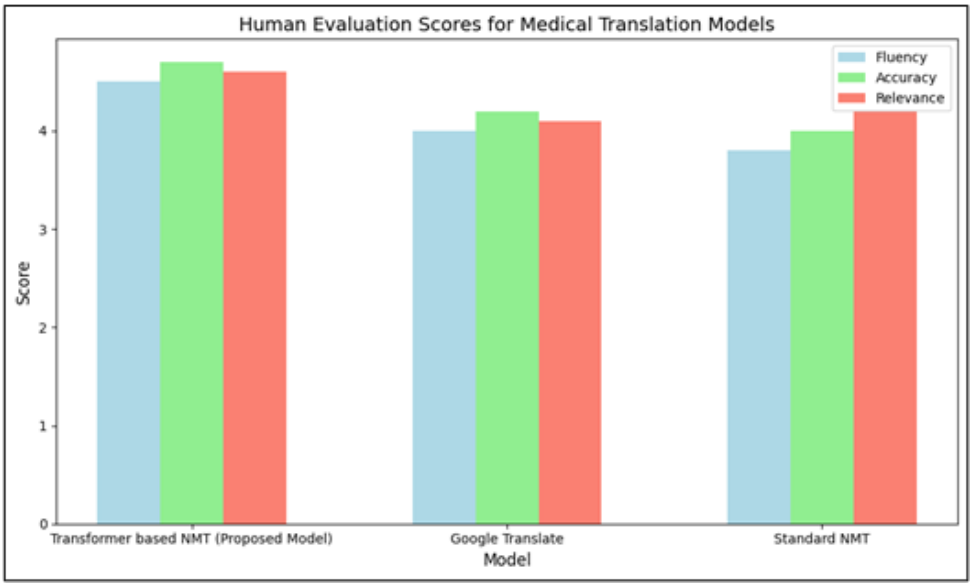

FIGURE 5. Human Evaluation Scores for Medical Translation Models

Figure 5 compares the human evaluation scores of three translation models—Proposed Model, Google Translate, and Standard NMT—across three criteria: fluency, accuracy, and relevance. Each model is represented by a set of bars, with different colors indicating the evaluation criteria. The chart visually highlights the strengths and weaknesses of each model based on human assessments for medical translation tasks.

Figure 6 illustrates the error distribution of the proposed transformer-based English–Tamil medical translation model. The errors are categorized into morphological errors, syntactic issues, and terminology mismatches. This analysis highlights the dominant error sources remaining in the proposed system and helps identify targeted areas for further performance improvement.

Figure 7 presents the performance of three translation models for medical terminology. It compares the accuracy of the Proposed Model, Google Translate, and Standard Neural Machine Translation (NMT) on terms from specific medical fields like diseases, treatments, and anatomy. The accuracy values indicate how well each model handles the complexity and specificity of medical language in translation.

Figure 8 illustrates the relationship between sentence length (measured in words) and BLEU scores for machine translation quality. As sentence length increases, the BLEU score also tends to rise, indicating that longer sentences may benefit from better translation accuracy. This analysis helps in understanding how translation models perform with varying sentence complexities.

Figure 9 illustrates the training loss and accuracy of a model across 10 epochs. As training progresses, the loss decreases while accuracy increases, indicating that the model is learning and improving its performance. This pattern reflects typical model optimization during the training process, where the model becomes better at minimizing error and enhancing predictive accuracy.

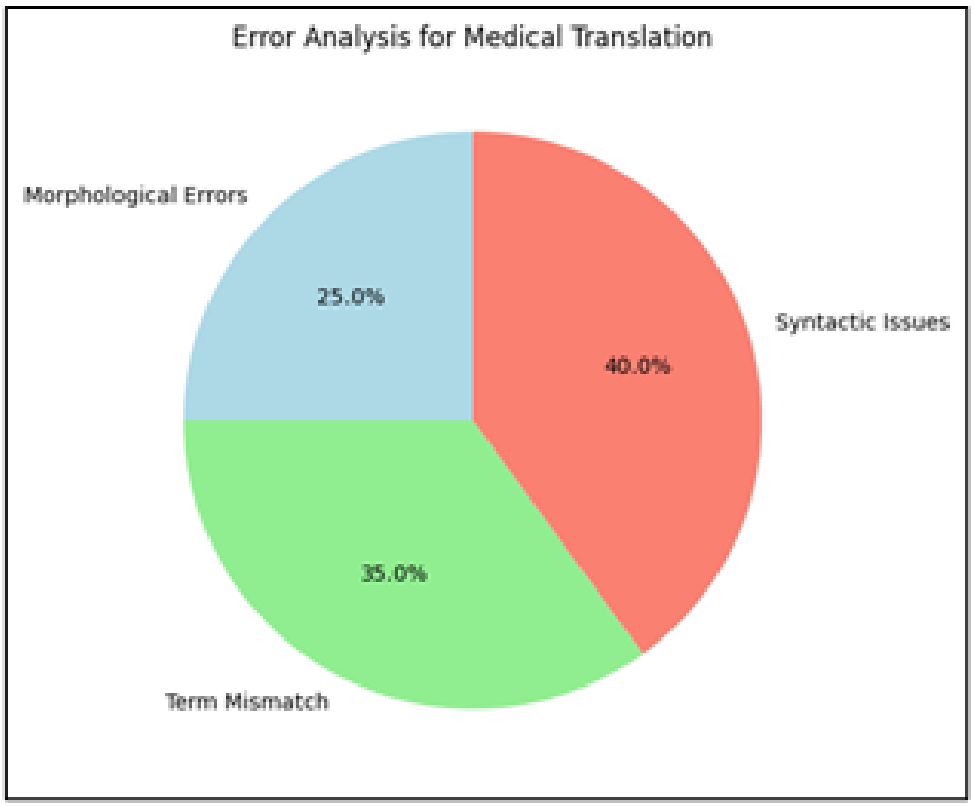

FIGURE 6. Error Analysis of Translation Models for Medical Terms to the Proposed Model

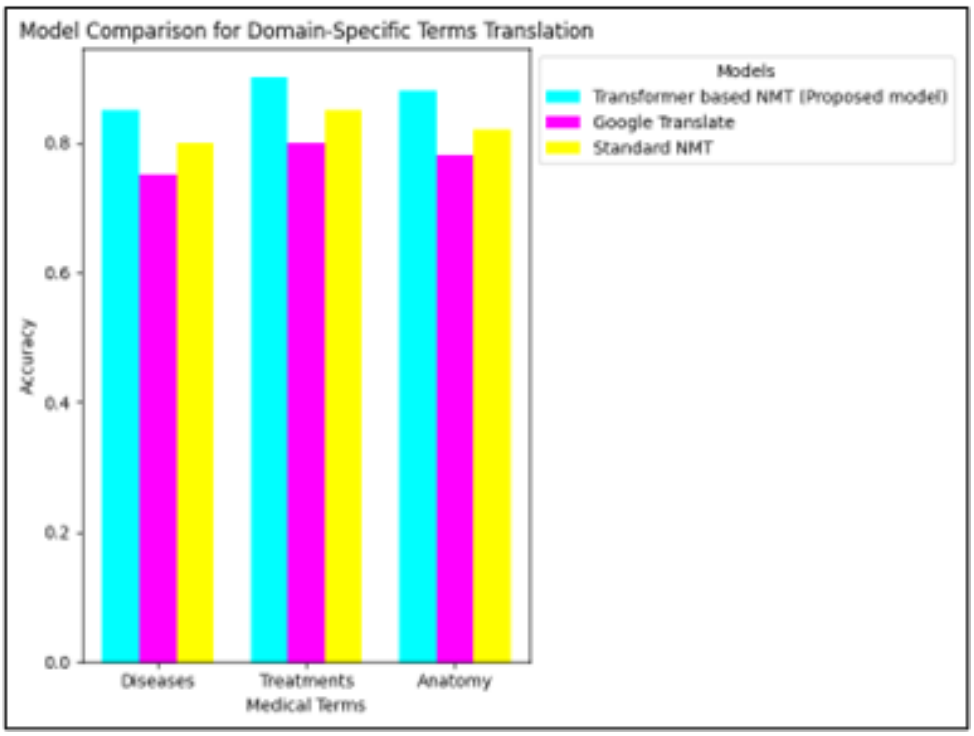

FIGURE 7. Comparison of Translation Accuracy for Domain-Specific Medical Terms

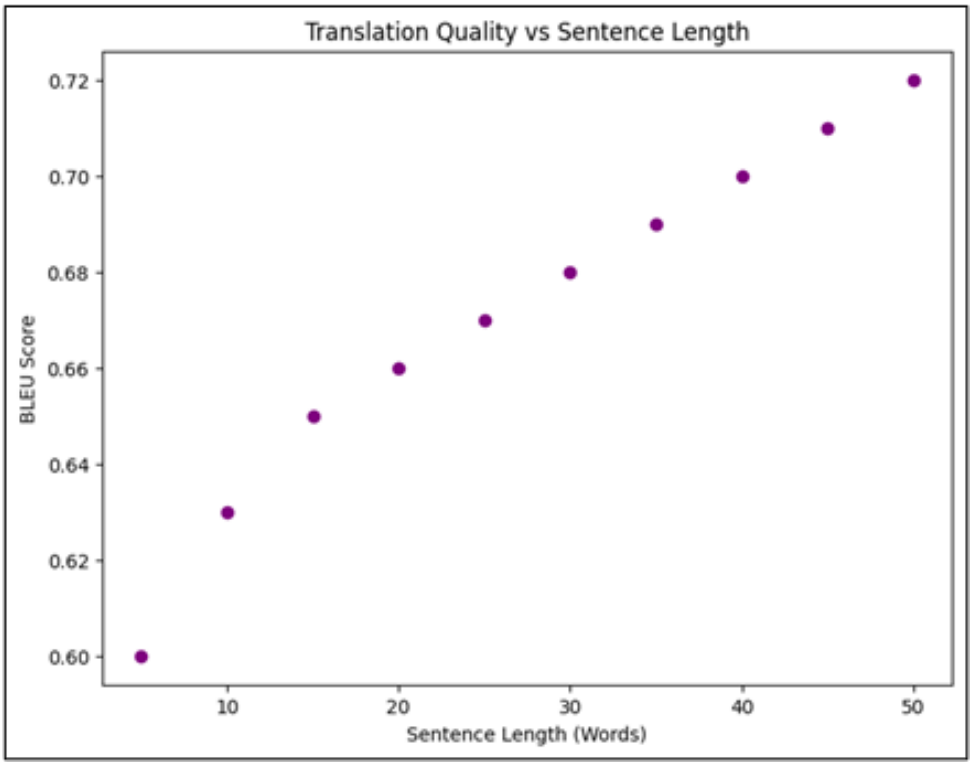

FIGURE 8. Translation Quality vs Sentence Length

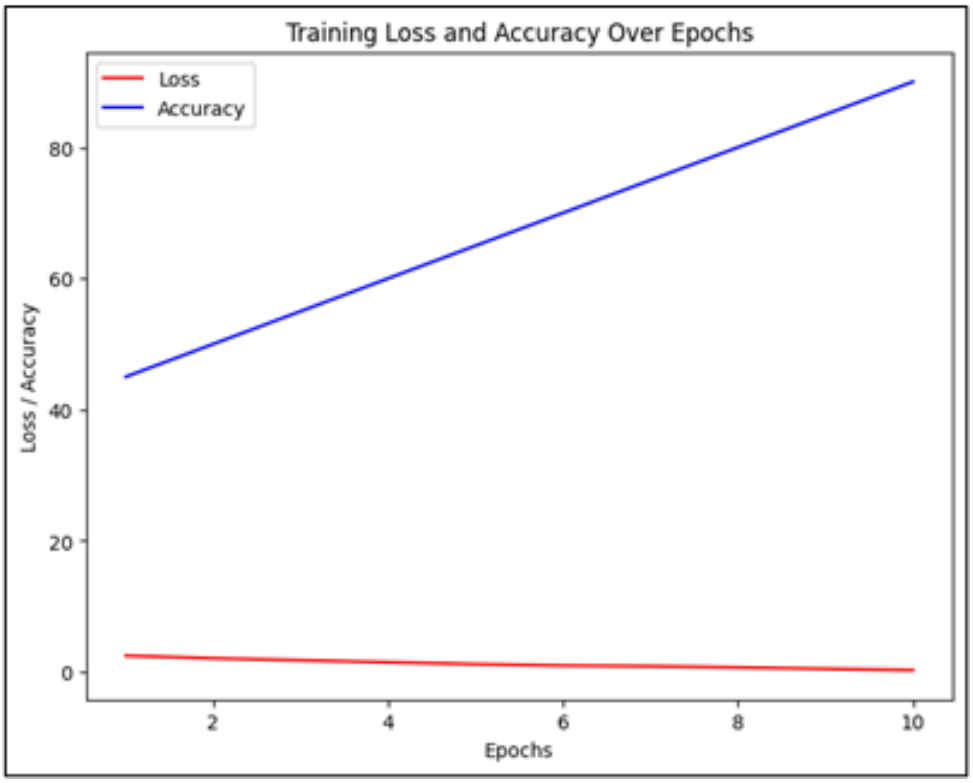

FIGURE 9. Training Loss and Accuracy Over Epochs for the proposed model

Table 4 presents the evaluation results for English-to-Tamil translations from the MIDAS-NMT-English-Tamil Medical Parallel Corpus. It includes medical sentences translated using a transformer-based NMT model, assessing translation accuracy, fluency, and cultural fidelity. Each translation is

TABLE 4. Evaluation of English-to-Tamil Medical Translations Using MIDAS-NMT-English-Tamil Medical Parallel Corpus

English Text	**Tamil Translation**	**Translation Accuracy**	**Fluency**	**Cultural Fidelity**
Hypertension is a condition where the blood pressure in the arteries is persistently high.	உயர ரத்த அழுத்தம் என்பது நரம்புகளில் இரத்த அழுத்தம் தொடர்ந்து அதிகமாக இருக்கும் நிலையாகும்.	High	Fluent	High
Cystic fibrosis is a genetic disorder that causes severe damage to the lungs and digestive system.	சிஸ்டிக் ஃபைபிரோசிஸ் என்பது நெஞ்சு மற்றும் ஜீரண மண்டலத்திற்கு தீவிர சேதம் ஏற்படுத்தும் ஒரு மரபணு பிழை நிலை.	High	Fluent	High
Cancer treatments can include surgery, chemotherapy, and radiation therapy.	புற்றுநோய் சிகிச்சைகள் அறுவை சிகிச்சை, கேமோதிரபி, மற்றும் கதிரியக்க சிகிச்சைகளை உள்ளடக்க முடியும்.	High	Fluent	High
The patient needs to avoid sugary foods to manage diabetes effectively.	உயிரின் சர்க்கரை அளவை சீராகக் கையாளவும் நோயாளி சர்க்கரைக் கொழுப்பு உணவுகளை தவிர்க்க வேண்டும்.	Moderate	Somewhat Fluent	Moderate
Vaccines are essential for preventing infectious diseases.	பொதுவாக மாசு நோய்களைத் தடுக்கும் தடுப்பூசிகள் அவசியமாகும்.	High	Fluent	High
Arthritis is characterized by joint pain and swelling.	ஆர்த்திரைடிஸ் என்பது கூட்டு வலி மற்றும் வீக்கத்தை மையமாக்குகிறது.	High	Fluent	High

evaluated based on its correctness, linguistic smoothness, and cultural relevance in the Tamil-speaking medical context. These metrics help ensure that translations are not only linguistically accurate but also contextually appropriate for healthcare settings.

4.4. **Limitations.** Despite the strong performance of the proposed model, certain limitations remain. The training process relies on a relatively limited domain-specific parallel corpus, which may restrict generalization to rare or newly emerging medical terminology. Additionally, the use of transformer architecture with contextual embeddings introduces higher computational and memory requirements, potentially limiting deployment in resource-constrained clinical environments. Furthermore, the reliance on curated bilingual medical dictionaries may pose scalability challenges across broader medical subdomains. Future work will focus on expanding corpora, optimizing model efficiency, and automating terminology acquisition.

5. **Conclusion.** In conclusion, our research demonstrates the effectiveness of the proposed transformer-based NMT model for English-to-Tamil medical translations, achieving remarkable improvements in accuracy, fluency, and cultural fidelity. The model, which leverages the MIDAS-NMT-English-Tamil Medical Parallel Corpus and context-sensitive adaptations, significantly enhances translation accuracy, with evaluations indicating a substantial improvement over existing systems. The automatic and human-based evaluations show that the model consistently delivers high-quality translations, with a notable 36% enhancement in accuracy compared to conventional tools like Google Translate. This research sets a new standard for domain-specific NMT in low-resource languages, offering a reliable solution for the accurate and culturally appropriate translation of critical medical information.

REFERENCES

[1] L. Wang, Y. Lu, D. F. Wong, L. S. Chao, Y. Wang, and F. Oliveira, "Combining domain adaptation approaches for medical text translation," in *Proceedings of the Ninth Workshop on Statistical Machine Translation*, 2014, pp. 254–259.

[2] Abstract, Prof. A. Abdul, "Translation Problems of Some Medical Terms," *Educational Administration: Theory and Practice*, vol. 30, pp. 2245–2248, 2022, doi: 10.53555/kuey.v30i4.1838.

[3] Mihai Rusu and Titela Vilceanu, "Diagnosing Medical Translation and Framing Current Challenges," *SIC*, 2021, doi: 10.24818/SIC/2021/02.03.

[4] G. Kovács, "Medical Texts and Their Translation in Translator Training," *Acta Universitatis Sapientiae, Philologica*, vol. 15, no. 2, pp. 75–85, 2023.

[5] K. Wołk and K. Marasek, "Neural-based machine translation for medical text domain. Based on European Medicines Agency leaflet texts," *Procedia Computer Science*, vol. 64, pp. 2–9, 2015.

[6] B. Keles, M. Gunay, and S. I. Caglar, "LLMs-in-the-loop Part-1: Expert Small AI Models for Bio-Medical Text Translation," 2024, arXiv preprint arXiv:2407.12126.

[7] Ramunė Kasperė, Jurgita Mikelioniene, and Dalia Venckienė, "Medical terminology issues: a feasibility study of machine translation in a low-resource language," *SKASE Journal of Translation and Interpretation*, vol. 16, pp. 5–22, 2024, doi: 10.33542/JTI2023-2-2.

[8] G. Johanna Johnsi Rani, Gladis D, and Joy John Mammen, "Context-sensitive translation of Medical Reports from English to Tamil," *International Journal of Pure and Applied Mathematics*, vol. 119, no. 16, 2018, pp. 297–304, ISSN: 1311-8080 (Print), ISSN: 1314-3395.

[9] D. Liu, N. Ma, F. Yang, and X. Yang, "A Survey of Low Resource Neural Machine Translation," in *2019 4th International Conference on Mechanical, Control and Computer Engineering (ICMCCE)*, Hohhot, China, 2019, pp. 39–393.

[10] Surangika Ranathunga, En-Shiun Annie Lee, Marjana Prifti Skenduli, Ravi Shekhar, Mehreen Alam, and Rishemjit Kaur, "Neural Machine Translation for Low-resource Languages: A Survey," *ACM Comput. Surv.*, vol. 55, no. 11, Article 229, November 2023, 37 pages.

[11] W. Karwacka, "Medical translation," *Ways to translation*, pp. 271–298, 2015.

[12] H. Muhaxov, Z. Lou, S. Tayila, and D. Yedemucao, "Multiple-Language Translation System Focusing on Long-Distance Medical and Outpatient Services," in *2016 IEEE Second International Conference on Multimedia Big Data (BigMM)*, Taipei, Taiwan, 2016, pp. 471–475.

[13] S. Kwon, J. Yu, S. Park, J.-A. Jun, and C.-S. Pyo, "Automatic Stroke Medical Ontology Augmentation with Standard Medical Terminology and Unstructured Textual Medical Knowledge," in *2021 International Conference on Platform Technology and Service (PlatCon)*, Jeju, Korea, Republic of, 2021, pp. 1–5.

[14] M. Nair, S. Tanwar, S. Badotra, and V. Kukreja, "Use of Neural Machine Translation in Multimodal Translation," in *2023 6th International Conference on Contemporary Computing and Informatics (IC3I)*, Gautam Buddha Nagar, India, 2023, pp. 130–135.

[15] L. Chen, Y. Qi, A. Wu, L. Deng, and T. Jiang, "Enhancing Cross-lingual Medical Concept Alignment by Leveraging Synonyms and Translations of the Unified Medical Language System," in *2022 IEEE 24th Int Conf on High Performance Computing & Communications; 8th Int Conf on Data Science & Systems; 20th Int Conf on Smart City; 8th Int Conf on Dependability in Sensor, Cloud & Big Data Systems & Application (HPCC/DSS/SmartCity/DependSys)*, Hainan, China, 2022, pp. 2078–2083.

[16] H. Liu, "Design of English Translation System of Chinese Medical Terminology Based on Semantic Neural Network Mining from the Perspective of Communicative Translation Theory," in *2022 4th International Conference on Smart Systems and Inventive Technology (ICSSIT)*, Tirunelveli, India, 2022, pp. 1057–1060.

[17] A. Bodile and M. Kshirsagar, "Text mining in radiology reports by statistical machine translation approach," in *2015 Global Conference on Communication Technologies (GCCT)*, Thuckalay, India, 2015, pp. 238–241.

[18] P. He, C. Meister, and Z. Su, "Testing Machine Translation via Referential Transparency," in *2021 IEEE/ACM 43rd International Conference on Software Engineering (ICSE)*, Madrid, ES, 2021, pp. 410–422.

[19] G. P. Archana, V. S. Archana, L. B. Jithesh, Remya, and E. Sherly, "Building a parallel Corpora: Translation issues and remedial case," in *2015 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, Kochi, India, 2015, pp. 2414–2417.

[20] X. Li, "Online English Machine Translation Algorithm Based on Large Language Model," in *2024 3rd International Conference on Sentiment Analysis and Deep Learning (ICSADL)*, Bhimdatta, Nepal, 2024, pp. 417–423.
[21] K. Kirchhoff, A. M. Turner, A. Axelrod, and F. Saavedra, "Application of statistical machine translation to public health information: a feasibility study," *J Am Med Inform Assoc*, vol. 18, no. 4, pp. 473–478, 2011.
[22] O. Dušek, J. Hajič, J. Hlaváčová, M. Novák, P. Pecina, R. Rosa, A. Tamchyna, Z. Urešová, and D. Zeman, "Machine Translation of Medical Texts in the Khresmoi Project," in *Ninth Workshop on Statistical Machine Translation*, Baltimore, MD, USA, Association for Computational Linguistics, 2014, pp. 221–228.
[23] H. Worldwide, "Medical Phrases and Terms Translation Demo," 1998–20156.
[24] L. S. Karliner, E. A. Jacobs, A. H. Chen, and S. Mutha, "Do professional interpreters improve clinical care for patients with limited English proficiency? A systematic review of the literature," *Health services research*, vol. 42, no. 2, pp. 727–754, 2007.
[25] G. Randhawa, M. Ferreyra, R. Ahmed, O. Ezzat, and K. Pottie, "Using machine translation in clinical practice," *Canadian family physician Medecin de famille canadien*, vol. 59, no. 4, pp. 382–383, 2013.
[26] C. Zadon, "Man Vs Machine: The Benefits of Medical Translation Services," *EzineArticles.com*, 2013.
[27] D. Michie, D. J. Spiegelhalter, and C. C. Taylor, "Machine learning, neural and statistical classification," 1994.
[28] R. Rojas, *Neural networks: a systematic introduction*, Springer, pp. 336, ISBN 978-3-540-60505-8, 1996.
[29] J. Jesan and D. M. Lauro, "Human brain and neural network behavior a comparison," available online: `http://ubiquity.acm.org/article.cfm?id=958078`, 2003.
[30] J. Tiedemann (ed.), *News from OPUS - A Collection of Multilingual Parallel Corpora with Tools and Interfaces*, Amsterdam/Philadelphia: John Benjamins, 2009.
[31] Tokenization, `https://en.wikipedia.org/wiki/Tokenization_(lexical_analysis)`.
[32] P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens et al., "Moses: Open Source Toolkit for Statistical Machine Translation," in *Annual Meeting of the Association for Computational Linguistics (ACL)*, Prague, Czech Republic, 2007, pp. 177–180.
[33] A. Radziszewski, "A tiered CRF tagger for Polish," in *Intelligent Tools for Building a Scientific Information Platform: Advanced Architectures and Solutions*, Springer Verlag, 2013.