ISSN 2007-9737

GAN-based Incremental Learning for English to Khasi Machine Translation: A Low-Resource Language Model

Ibadonbok Syiemlieh[1], Zahid Ahmed[2], Arnab Kumar Maji[1], Sufal Das[1,*]

[1] North-Eastern Hill University,
Department of Information Technology,
India

[2] Vel Tech Rangarajan Dr. Sagunthala R & D Institute of Science and Technology,
Department of Information Technology,
India

{ibadonbok32, zahidprince786, arnab.maji, sufal.das}@gmail.com

Abstract. Design of a high-quality machine translation system for low-resource languages is still a big problem because there are not sufficient parallel corpora and the patterns of languages frequently shift. In order to address such a gap, this paper presents a modified generative adversarial network-based hybrid translation framework known as the Incremental Generative Adversarial Network Integrated Hybrid approach (I-GANH). It provides incremental learning and robust translation for the English-Khasi language pair. This framework consists of an incremental bilingual data processor that updates vocabulary and phrase alignments in real time. A hybrid translation core that integrates statistical alignment priors with a transformer-based encoder–decoder to improve syntactic and semantic grounding. A modified conditional GAN that generates high-quality pseudo-parallel sentences and learns new linguistic patterns over time and an adaptive optimization controller that balances adversarial learning signals and avoids catastrophic errors during model updates. It can update the newly generated representations and GAN components when new data is generated. This makes it more efficient and maintains performance consistency between training iterations. Experimental evaluations indicate that the suggested framework improves translation consistency, increases the coverage of rare language structures, and provides a scalable approach for constructing translation systems in low-resource settings. Overall, the proposed framework offers a flexible option to implement machine translation that remains effective in changing language settings.

Keywords. Machine translation, incremental learning, GAN, machine learning, Khasi language, low-resource language.

1 Introduction

Machine translation has come a long way because of the growth of massive bilingual corpora, neural architectures with a lot of processing power, and attention-driven sequence modeling [27, 32].

Undoubtedly, these improvements have largely helped languages that are having an extensive amount of resources. However, they are unable to endorse languages that are rich but poorly documented due to structural barriers [12, 38].

Like Khasi and other low-resource languages, they have complex morphological variants, flexible syntactic ordering, and semantic connections that depend on the context [14, 21, 28].

These properties make it more challenging for a typical Neural Machine Translation (NMT) framework [36]. Most of the time, NMT encountered difficulties with lexical sparsity, unstable cross-lingual alignments, and poor generalization when it comes to sparse or changing vocabulary, because it usually depends on dense parallel supervision and stable statistical distributions [6, 17, 21, 26].

Because of this, the outputs of the translation may not always be precisely accurate in meaning, or they may not always be consistent with the original meaning.

Additionally, they may not be able to demonstrate new language patterns that have formed within the speech community. The conventional approaches for improving low-resource MT, like back-translation, multilingual transfer, and hybrid Statistical Machine Translation (SMT) + NMT pipelines, work well but have a big limitation, as they work under static training base assumptions [8]. All of these models usually need to be completely retrained from scratch when new words, phrases, or domain-specific constructions come out. This not only costs a lot of computation power but also increases the certain risk of catastrophic errors that arise when new parameters are optimized and previously learned language knowledge is lost [20].

This gap is especially significant in the real world, as the language changes all the time with new expressions, adapted words, and adaptations for specific fields. Moreover, generative models like Generative Adversarial Networks (GANs) have shown potential for data augmentation, but traditional GAN-based MT frameworks do not possess mechanisms to train incrementally or keep meaning stable [19, 21, 37]. These limitations strongly emphasize the requirement for a translation system capable of dynamically updating its language representations as well as preserving the knowledge that it has previously gained.

To address these problems, this paper introduces a modified generative adversarial network-based hybrid translation framework known as the Incremental Generative Adversarial Network Integrated Hybrid approach (I-GANH). It is an end-to-end framework that is designed for incremental learning in low-resource translation contexts.

This framework includes an incremental bilingual data processor for expanding vocabulary in real time. A hybrid translation core that combines SMT-derived alignment priors with Transformer-based neural modeling. A modified conditional GAN that creates high-fidelity pseudo-parallel data with incremental updates and an adaptive optimization controller that can keep the training stable even when there is an adversarial imbalance or forgetting [1]. I-GANH aims to produce consistent translation quality, better flexibility, and larger linguistic generalization for under-resourced language pairings by allowing selected parameter refining and structured assimilation of new linguistic information.

The novelty of the proposed I-GANH lies in its ability to unify dynamic linguistic adaptation, hybrid translation modeling, and adversarial data generation within a single continuously evolving architecture. The proposed architecture includes an Incremental Bilingual Data Processor (IBDP) that can add new words and alignment tables in real time. With the help of this, the system could potentially be able to acquire new vocabulary without necessarily going through a thorough retraining process.

This framework is different from traditional GAN-based or hybrid SMT–NMT systems. Here, to ensure that the translation process remains on the right track, the hybrid translation core (HTC) incorporates alignment priors generated from SMT. A transformer encoder-decoder enhances contextual fluency, which provides an extremely solid foundation for progressive learning.

An innovative novel concept is the Modified Conditional GAN, which is also referred to as the MC-GAN. It operates with streaming updates through the production of pseudo-parallel English-Khasi pair combinations that accurately represent shifting linguistic distributions.

It uses automated conditioning and attention-regulated noise filtering to keep semantic coherence during incremental updates. The adaptive optimization controller (AOC) is the final component, and it introduces a novel method of controlling a system that strikes a balance between adversarial, reconstruction, and retention losses to prevent catastrophic failures.

Systematically integrating these interconnected components provides a novel, scalable, and rationally sound strategy for adaptive low-resource machine translation.

1.1 Contribution

The primary contributions of this study are:

— Introduced a novel incremental bilingual translation framework that can grow over time by incorporating new language data and maintaining what it has already learned.

— The proposed framework is a hybrid learning methods, which mix rule-based and statistical priors with neural and generative models, to improve the quality of translations and the consistency of their meanings in situations with few resources.

— The proposed framework can enhance the quality and adaptability of machine translation through the creation of synthetic bilingual data and the application of adaptive optimization techniques for stable and efficient model training.

The remaining parts of the paper are arranged like this: Section 2 addresses the background of machine translation and incremental learning. Section 3 discusses the motivations for the suggested approach. Section 4 discusses the details of the proposed framework. Section 5 has the details of the experiment, Section 6 includes the results and an observation, and Section 7 includes the conclusion of this article.

2 Related Studies

Existing studies in machine translation (MT) have shown significant improvements by implementing incremental learning, with a focus on enhancing adaptability to low-resource languages and minimizing fatal errors. However, the findings from this research uncover significant inconsistencies in methodological assumptions, dataset characteristics, and performance improvement processes. Early research by Kirkpatrick et al. [18] introduced elastic weight consolidation (EWC) to address catastrophic forgetting in incremental learning.

While the EWC helps to preserve previous knowledge, usually it encounters challenges to capture parameter uncertainty and is also computationally demanding in changing environments. In a similar vein, Gomes et al. [10] developed adaptive random forest (ARF) for streaming data. ARF is able to successfully maintain stability across dynamic datasets, but its performance may drop during sudden concept changes if no drift detection is used.

These initial techniques contributed to the formation of learning that continues throughout the lifespan. However, they did not adapt well to complex, high-dimensional, or multilingual translation tasks. Since 2019, most of the research in machine translation, including studies by Thabah et al. [34], Shah et al. [29], Kalimuthu et al. [15], and Singh et al. [31], has used neural machine translation (NMT), statistical machine translation (SMT), and incremental learning strategies.

In spite of the fact that these models contributed to progress for particular languages, they often relied on limited datasets like religious texts or small parallel corpora. Sometimes the SMT worked better for low-resource languages, but NMT models with attention mechanisms generally produced more fluent and adaptable translations.

However, these models typically faced challenges due to small datasets, limited domain coverage, and high training requirements. Transformer-based architectures proposed by Gillioz et al. and Donald et al. [9, 7] and convolutional neural network (CNN)-based models [5] proposed by Hujon et al. [10] have marked a significant shift in deep contextual representation learning.

Although transformer-based approaches possess a positive impact on domain generalization, their performance dropped when they trained on generic datasets. It clearly demonstrates that domain dependency is still an issue that is challenging to solve. The CNN-based approaches achieved an improvement in the quality of translation for the English-Khasi pair.

However, it frequently encountered difficulty with preserving phrase-level coherence and dealing with lengthy phrases. Transfer learning proposed by Hujon et al. [14], and knowledge transfer proposed by Huang et al. [13] have demonstrated their potential in low-resource contexts by assisting models in adapting to new language pairs.

Still, these methods led to parameter inflation due to the added layers or modules. Therefore, scalability is more difficult to achieve in incremental learning environments. Most studies rely on static datasets, which makes it hard for models to learn over time or adapt to new domains without having to start over.

As a result, there is an unavoidable requirement for translation frameworks that are able to keep their knowledge up to date and preserve it as they progress.

So far, the majority of the research that has been conducted on machine translation and incremental learning has been conducted independently. There exist very few works that have incorporated approaches for incremental learning like EWC or rehearsal with neural or transformer-based translation.

Another factor that hinders the ability of models to generalize is the use of limited parallel corpora, such as the Bible or Tatoeba. To avoid performance drops in different language settings, better ways to adapt and transfer across domains are essential.

Extending the number of parameters in a model space for the purpose of knowledge transfer is associated with a significant rise in the computing computational costs. Research should focus on building translation models that maintain accuracy without becoming overly large.

Although EWC and ARF have shown their stability under changing data, very few research studies have addressed how to evaluate translation uncertainty, which is crucial for practical use and which is very important for real-world use.

Overall the existing studies show that neural and transfer-based methods have improved translation, but how incremental learning works in low-resource machine translation remains unclear. Most of the studies rely on fixed datasets and do not adapt well to new domains or changing languages.

In addition, some issues such as computational demands, understanding how models make decisions, and the risk of forgetting prior knowledge remain obstacles to combining machine learning with incremental learning.

3 Motivation

The majority of traditional NMT models are fixed. They cannot adjust to new language patterns or vocabulary changes. Generative adversarial networks (GANs) [11, 2] use adversarial learning, which allows the model to adapt over time through feedback between the generator and discriminator.

This process helps the model to improve their translation accuracy as language inputs change, so it keeps learning without needing to be retrained from scratch. Both the SMT and NMT usually require substantial parallel data, which is not available for English-Khasi translation.

A modified conditional GAN uses monolingual Khasi data to generate synthetic English sentences and create pseudo-parallel English-Khasi sentence pairs to improve the English-to-Khasi translation system during training. The discriminator checks these pairs to add high-quality synthetic samples to the dataset which helps to overcome data scarcity and keep translations relevant.

Many translation systems do not measure uncertainty in their outputs. It can make them less reliable in low-resource settings.

In GANs, the discriminator helps to estimate translation confidence by distinguishing real from synthetic translations. This process brings in a way for the system to prefer more confident and accurate translations during retraining. Most of the current incremental machine translation systems rely only on NMT probability scores or SMT phrase alignments without connecting statistical structure to neural representation.

A new hybrid approach may be improved by using SMT-based bilingual priors in the early translation steps, and the NMT-GAN subsystem gradually learns more complex semantic relationships. The combination of the structure of SMT with the flexibility of NMT allows for refining translations more consistently. The combination of GAN adaptive learning with SMT-NMT methods enables a system to produce high-quality translations, adapt over time, scale well, handle data shortages, and estimate confidence.

4 Proposed Methodology

The proposed framework is an incremental, modified generative adversarial network-based hybrid translation known as the Incremental Generative Adversarial Network Integrated Hybrid approach (I-GANH), which is designed to build an English-Khasi translation system that gets progressively better over time and works well even when there are not sufficient resources. I-GANH is different from static translation systems because it has the potential to be refined incrementally.

This means that the model can change as new vocabulary, verbal, or semantic arrangements accumulate through incoming data. The framework combines the alignment accuracy of statistical machine translation (SMT) with the contextual generalization of neural machine translation (NMT) in a modified generative adversarial network (GAN) [11]. The GAN architecture has been modified so that it can handle incremental vocabulary expansion, context-sensitive pseudo-data creation, and adaptive adversarial optimization. This ensures that it will continue to get better without having to be retrained from scratch.

Here, the GAN has been modified in such a way that it can upgrade itself over time, and at the same time, it can maintain translation consistency.

Modification in Generator: The generator G_θ processes monolingual Khasi sentences x and alignment priors A as conditional inputs, then generates synthetic English translations $\hat{y}$:

$$\hat{y} = G_\theta(x, A). \tag{1}$$

Here x denotes a monolingual Khasi sentence, A represents the SMT-derived alignment priors, $\hat{y}$ denotes the synthetic English sentence generated by the proposed Modified Conditional GAN, and G_θ represents the generator parameterized by θ. The generated English sentence is paired with the original Khasi sentence to construct a pseudo-parallel English-Khasi corpus for training the Hybrid Translation Core (HTC).

When new words or grammatical structures are encountered, the integration layer adapts dynamically as follows:

$$E' = [E; f_{\text{ctx}}(x_{\text{new}})], \tag{2}$$

where E denotes the pre-existing embedding matrix, E' denotes the updated embedding matrix, and $f_{\text{ctx}}(\cdot)$ computes the contextual embedding of the newly encountered word x_{new} by averaging the representations of its neighboring tokens.

The generator deploys an incremental fine-tuning method that only modifies the layers that are relevant to new tokens and keeps the previous constraints exactly the same:

$$\theta' = \theta + \eta \cdot \nabla_{\theta_{\text{new}}} L_{adv}^{inc}. \tag{3}$$

Here θ and θ' denote the model parameters before and after the incremental update, respectively; η is the learning rate; and $\nabla_{\theta_{\text{new}}} L_{adv}^{inc}$ represents the gradient of the incremental adversarial loss with respect to the model parameters associated with the newly introduced vocabulary. This makes sure that learning is flexible and avoids catastrophic forgetting.

Modification in Discriminator: The discriminator D_ϕ is redesigned into a dual-stream model that determines both semantic and syntactic fidelity:

— Semantic fidelity (D_{sem})—evaluates contextual similarity between y and $\hat{y}$,

— Syntactic fidelity (D_{syn})— assesses structure-level correspondence.

The overall output of the discriminator is calculated as:

$$D_\phi(x, y) = \sigma(W_1 D_{\text{sem}} + W_2 D_{\text{syn}}). \tag{4}$$

Here D_ϕ denotes the discriminator parameterized by ϕ, x and y denote the input sentence pair, D_{sem} and D_{syn} represent the semantic and syntactic fidelity scores, respectively, W_1 and W_2 are weighting coefficients that balance the contributions of the two fidelity measures, and $\sigma(\cdot)$ denotes the sigmoid activation function.

A memory-replay strategy has been implemented to support incremental learning. Specifically, a small memory buffer $M = \{(x_i, y_i)\}$ containing representative sentence pairs from previous training phases is reused during refinement to preserve previously learned knowledge and mitigate catastrophic forgetting.

The aggregate incremental adversarial loss is defined as:

$$L_{adv}^{inc} = \mathbb{E}_{x,y\sim D}\left[\log D_\phi(x,y)\right] + \mathbb{E}_{x_{\text{new}}}\left[\log(1 - D_\phi(x_{\text{new}}, G_\theta(x_{\text{new}}, A)))\right]. \quad (5)$$

Here L_{adv}^{inc} denotes the incremental adversarial loss, $\mathbb{E}[\cdot]$ denotes the expectation operator, D is the distribution of bilingual real sentence pairs, x_{new} is a newly encountered Khasi sentence, D_ϕ is discriminator parameterized by ϕ and G_θ is generator parameterized by θ. The first term maximizes confidence of discriminator for real bilingual pairs while the second term encourages generator to produce English sentences that are hard to distinguish from translations.

This incremental adversarial objective allows GAN to adapt to new data observed without reoptimizing previously learned parameters and thus keeps translation performance over time.

The architecture of the proposed framework is shown graphically in Figure 1. The proposed framework comprises four main components: (i) Incremental bilingual data processor (IBDP) constantly updates bilingual vocabulary and alignments; (ii) Hybrid translation core (HTC) integrates SMT-derived alignment priors with a Transformer encoder-decoder for English to Khasi neural machine translation; (iii) Modified conditional GAN generates synthetic English sentences corresponding to monolingual Khasi sentences and thus creates pseudo-parallel pairs for incremental HTC training; and (iv) Adaptive optimization controller (AOC) regulates adversarial learning, preserves prior knowledge and reduces forgetting when making incremental updates.

The translation process moves through a series of adaptive steps that make sure that learning never stops and the quality of the translation remains the same. The entire translation process is outlined phase by phase below:

Incremental Bilingual Data Processor (IBDP): IBDP first dynamically pulls new words from incoming Khasi phrases and updates both the bilingual phrase table P and the embedding matrix E. This phenomenon can be demonstrated formally as:

$$P' = P \cup \{(x_{\text{new}}, y_k, p(x_{\text{new}} \mid y_k))\}. \quad (6)$$

Consider P to denote an existing bilingual phrase table and P' to denote an updated table. x_{new} is a Khasi word or phrase encountered for the first time and y_k is a candidate English translation for it. The alignment probability between this newly encountered phrase pair is $p(x_{\text{new}} \mid y_k)$.

The embedding matrix is updated according to Eq. 2 so that newly encountered words can be included in existing semantic representations.

Hybrid Translation Core (HTC): In this phase, SMT-derived alignment priors are used as alignment constraints inside the Neural Machine Translation (NMT) model. This is called the Hybrid Translation Core (HTC). During translation, let x denote the English input sentence and y denote the corresponding Khasi reference sentence.

The Transformer-based encoder-decoder predicts the translated Khasi sentence as:

$$\hat{y} = T_\psi(x, A). \quad (7)$$

Here x denotes the English input sentence, $\hat{y}$ is the translated Khasi sentence, $T_\psi(\cdot)$ represents the Transformer-based encoder–decoder parameterized by ψ, and A denotes the SMT-derived alignment priors.

The alignment-guided translation loss is defined as:

$$L_{\text{trans}} = \|y - T_\psi(x, A)\|^2. \quad (8)$$

Here L_{trans} is the translation loss, y denotes the reference Khasi sentence, and $T_\psi(x, A)$ represents the predicted Khasi translation. Eq. 8 minimizes the discrepancy between the predicted and reference translations, thereby preserving bilingual alignment consistency during training.

Modified Conditional GAN-based for synthesization: The Modified Conditional GAN augments the bilingual corpus by generating synthetic English counterparts for monolingual Khasi sentences. The generated English sentences are paired with the original Khasi sentences to construct pseudo-parallel

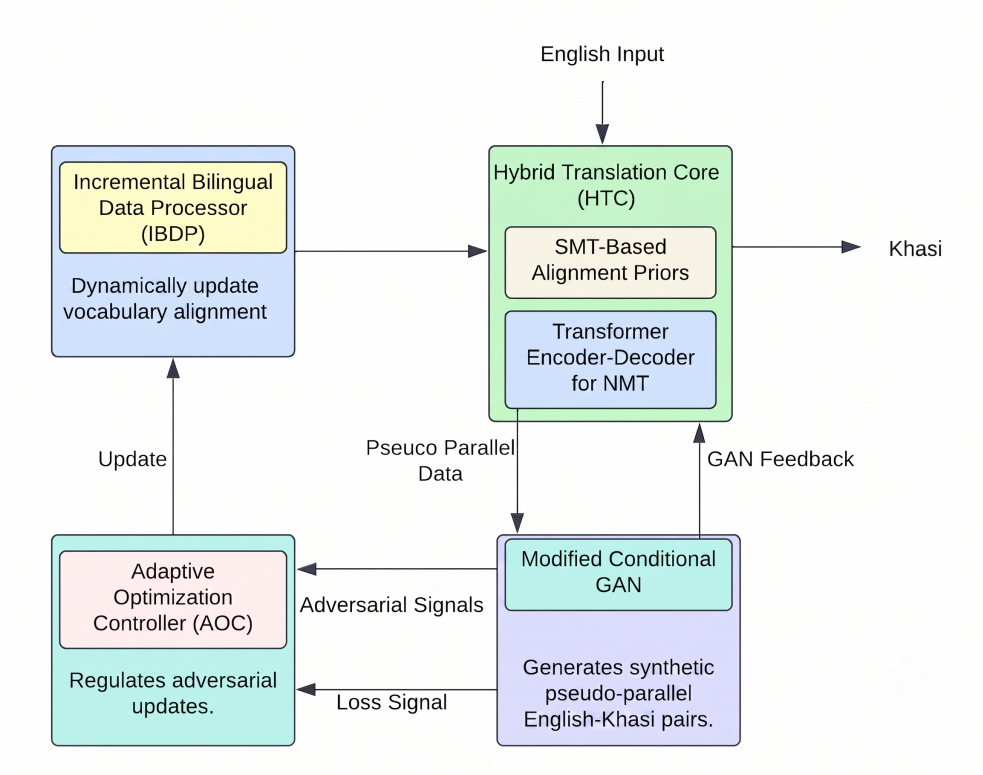

Fig. 1. Block Diagram of the Proposed Method

English–Khasi sentence pairs for incremental training. The generator G_θ generates synthetic English sentences $\hat{y} = G_\theta(x, A)$ from monolingual Khasi inputs and SMT-derived alignment priors. The discriminator D_ϕ checks for semantic (D_{sem}) and syntactic (D_{syn}) fidelity using Eq. 4.

The proposed incremental adversarial loss function stays flexible when it comes to new language structures that are defined in Eq. 5.

Adaptive Optimization Controller (AOC): Finally, the AOC makes sure that learning is stable through balancing adversarial updates and avoiding catastrophic error. This is achieved by using the Fisher Information to create a stability loss:

$$L_{\text{stability}} = \frac{1}{2} \sum_i F_i (\theta_i - \theta_i^*)^2. \tag{9}$$

Here $L_{\text{stability}}$ denotes the stability regularization loss, F_i represents the Fisher Information associated with parameter θ_i, θ_i denotes the current model parameter, and θ_i^* represents the corresponding parameter learned during previous training. Eq. 9 penalizes significant changes to previously important parameters, thereby reducing catastrophic forgetting during incremental learning.

The overall optimization objective is defined as:

$$L_{\text{total}} = \lambda_1 L_{\text{adv}}^{\text{inc}} + \lambda_2 L_{\text{trans}} + \lambda_3 L_{\text{stability}}. \tag{10}$$

Here L_{total} is the overall optimization objective, $L_{\text{adv}}^{\text{inc}}$ is the incremental adversarial loss, L_{trans} is the translation loss, $L_{\text{stability}}$ is the stability regularization loss, and λ_1, λ_2, and λ_3 are weighting coefficients that balance the contribution of each loss component.

By updating only the parameters associated with newly introduced vocabulary according to Eq. (3), the proposed framework preserves previously acquired knowledge while adapting efficiently to new language patterns.

The ability of the proposed system to incrementally optimize enables it to continue adapting at all times without requiring a full retraining process.

As such, rather than full training processes, each incremental update reduces overall computational costs (i.e., processing time) while maintaining semantic consistency, alignment quality, and translation fluency across successive updates. Across iterations, semantic fluency and alignment accuracy remain the same. Algorithm 1 summarizes the proposed I-GANH framework.

Algorithm 1 initializes the model parameters, the bilingual phrase table, and the embedding matrix at the beginning. The IBDP continually expands the vocabulary and phrase alignments based on the contextual embeddings of new bilingual Khasi words or sentence pairs; it does so in a way that it will never need to retrain the entire model. Next, the modified conditional GAN generates English counterparts of monolingual Khasi sentences using bilingual knowledge and contextual embeddings already developed to produce pseudo-parallel English-Khasi sentence pairs.

At the same time, the discriminator judges how well the generated sentence pair has preserved both its syntax and semantics and whether or not it distinguishes itself sufficiently from the real bilingual data. Both the generator and the discriminator are trained against each other in an adversarial manner to minimize the incremental adversarial loss while maintaining previously acquired knowledge by applying stability regularization to avoid catastrophic forgetting.

Subsequently, after the pseudo-parallel sentence pairs have been added to the bilingual corpus, they will be used by the HTC, which will use SMT-derived alignment priors, along with translation loss and adversarial consistency, to refine HTC. It is this continual feedback loop that allows our framework to integrate new vocabulary, improve translation quality and fluency, and adapt to changes in language over time, all without requiring full training, thus enabling us to develop a highly efficient and scalable low-resource English-Khasi machine translation system.

Algorithm 1 Incremental Generative Adversarial Network Integrated Hybrid Approach (I-GANH)

Require: Monolingual Khasi text X_{mono}, Limited bilingual Khasi–English pairs (X_b, Y_b).
Ensure: English-Khasi translation model with adaptive vocabulary and improved translation quality.
Initialize model parameters θ' using Eq. 3, generator G_θ using Eq. 1, discriminator $(D_\phi$ using Eq. 4 and 5, bilingual phrase table P' using Eq. 6, and embeddings E' using Eq. 2.
for each incoming Khasi sentence $x \in X_{mono}$ **do**
  Extract new tokens $x_{new} \leftarrow$ identify unseen words.
  Determine potential English alignments y_k using co-occurrence statistics.
  Update phrase table and embeddings using Eq. 6 and Eq. 2
end for
Input an English sentence x_{eng} and SMT-derived alignment priors A into the Hybrid Translation Core using Eq. 7 .
Generate the Khasi translation $\hat{y}_{HTC} \leftarrow Transformer(x_{eng}, A)$ Eq. 8.
Generate synthetic English counterparts for monolingual Khasi sentences using Eqs. 1 and 2.
if new vocabulary detected **then**
  Expand embeddings $E' \leftarrow [E; f_{ctx}(x_{new})]$.
  Fine-tune parameters for new tokens only using Eq. 3.
end if
Compute semantic and syntactic fidelity using Eq. 4.
Reuse memory buffer $M = \{(x_i, y_i)\}$ to preserve old knowledge.
Compute adversarial loss using Eq. 5.
Compute total loss using Eq. 10.
Apply stability constraint using Eq. 9.
for each incremental step t **do**
  Update parameters:

$$\theta_t = \theta_{t-1} - \eta \cdot \nabla_\theta (L_{total}(t) + \beta L_{stability}).$$

  Update only layers linked to new tokens to avoid forgetting.
end for
Set the final translation $Y_{trans} \leftarrow \hat{y}_{HTC}$.
Save updated parameters (θ_t, ϕ_t), phrase table P', and embeddings E'.

5 Experimental Details

This section discusses the arrangements made for the experiments, the datasets that were used,

the preprocessing pipeline, the training steps, the evaluation method, and the baselines that were used to compare how well the proposed I-GANH approach worked.

A standard CPU environment (Intel Xeon 2.20 GHz, 12 GB RAM) was used for all tests on Google Colab. Our goal is to determine if the system can improve the quality of translations between English and Khasi, deal with uncommon or previously unseen words, and keep working well with large amounts of bilingual data.

First, all the datasets were uploaded in CSV format. Then, a uniform cleaning and normalization workflow was used to get eliminated of punctuation noise, broken Unicode, and HTML markup.

To make the data clean, we have removed duplicates, handled missing values, and tokenized sentences. Khasi-specific characters and diacritics were also normalized; we employed the AutoTokenizer for the model Helsinki-NLP/opus-mt-en-mul, which can be fine-tuned for Khasi as it is multilingual. Next, SentencePiece BPE with 16k merge operations was used to tokenize the text. This made separate vocabularies for the English source (SRC) and Khasi target (TGT).

Through perplexity-based filtering, sentences with more than 80 tokens were filtered out, and synthetic examples that were too noisy were taken out.

The datasets were finally split into four sets: 80% were used for training, 10% were used for validation, and 10% were used for testing. It was made sure that each set had an equal number of features from the other two sets.

The I-GANH framework model training took place in three stages that were all related to each other. At that point, AdamW optimization, cross-entropy loss, and a warmup–decay learning rate schedule were used to train a small Transformer with two encoder and two decoder layers.

In Stage 2, incremental GAN-driven lexical expansion was introduced. A generator (2-layer MLP) and discriminator (3-layer MLP) were trained using the standard non-saturating GAN loss to create semantically meaningful embeddings for Khasi tokens that had not been seen before.

These generated embeddings were added to the decoder's embedding matrix and bilingual phrase table automatically.

The hybrid translation fusion mechanism was put in place in Stage 3. It combined transformer predictions, GAN-generated lexical candidates, and statistical phrase-table matches.

An embedding space with 512 dimensions is considered in the configuration of the proposed model. A six-layer Transformer encoder-decoder architecture, which included eight attention heads and a feed-forward layer with 2048 dimensions is considered.

To optimize generalization better, a loss rate is considered 0.1 and label smoothing is considered 0.1. To optimize the performance of the generator and discriminator, AdamW is applied. The learning rates were set at 2×10^{-4} and 1×10^{-4}, respectively.

The values of $\beta_1 = 0.9$ and $\beta_2 = 0.98$ were considered, and the weight decay was set to 10×4. Gradient clipping was set to 1.0. The training was carried out with a batch size of 32 for a maximum of 50 epochs. The discriminator itself consisted of 3 fully linked layers containing a hidden dimension of 512, LeakyReLU activation, and 0.2 dropout.

A replay buffer of 500 sentence pairs was preserved and updated every 100 incremental sentences, with an equal distribution of freshly observed and replayed occurrences.

The loss coefficients were regulated using $\lambda_1 = 0.50$, $\lambda_2 = 0.30$, and $\lambda_3 = 0.20$, and the stability coefficient was set to $\beta = 0.10$. Based on validation results, an early halting strategy of 5 epochs has been applied.

5.1 Dataset Description

The datasets used during the experiment are batch-wise to make the model adapt to new data when new batches arrive.

The first batch of data was downloaded from the Bapyn-EnKha-Parallel-Corpus [22], which is hosted on Hugging Face. According to the dataset description, the corpus contains 40,000 English-Khasi pairs. The corpus is compiled from various sources. A total of 26,000 pairs were

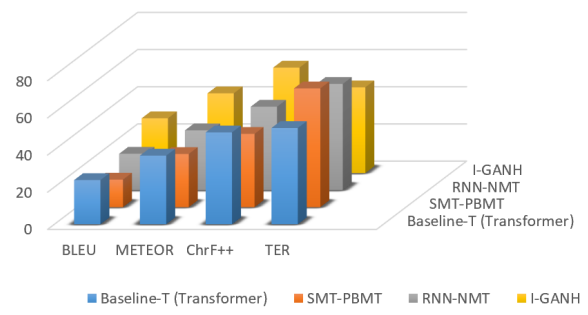

Fig. 2. Overall Model Translation Performance Analysis

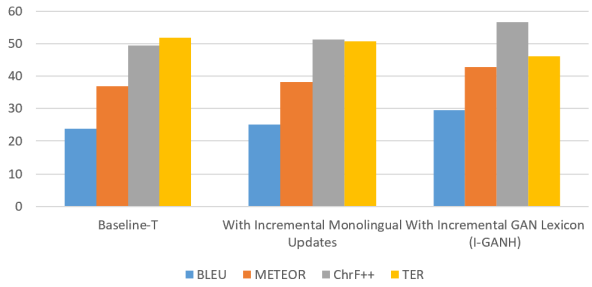

Fig. 3. Ablation Analysis

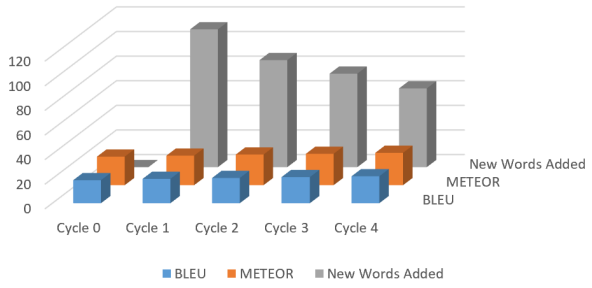

Fig. 4. Incremental Performance Improvements

compiled using sentences contained within the IndicNECorp 1.0 corpus [23].

Approximately 1,400 Khasi-English pairings were taken from the Tatoeba Project [16].

Sentence pairings that have been translated from other languages (such as Thai) and paired with an English version can be found in the Tham Annotated Corpus. Also, some data has been gathered by web scraping and manual translation. This research project uses all 40,000 pairs in the corpus for the initial training batch.

The next batch and the remaining batch were sourced from the English–Khasi Synthetic Parallel Corpus [35] hosted on Kaggle by user10383. The synthetic data contains 5 million pairs of sentences. Therefore, we have divided into 50 batches, i.e., 100000 pairs of sentences per batch. In tables 1 to 4, all the details of the dataset are listed.

5.2 Evaluation Metrics

Four distinct evaluation metrics were used to test how well I-GANH performed. These metrics were chosen because they can measure different parts of translation quality. BLEU served as a standard n-gram precision-based measure that is commonly used in machine translation [3, 24].

This ensured that the results could be compared to earlier studies that compared English to Indic and English to Khasi.

METEOR is being used because it is good at stemming, matching synonyms, and recall [3, 4]. This is important because Khasi often uses phrases with a lot of synonyms to get its point across.

Subsequently, ChrF++ is used, which may identify all character-level n-gram patterns that show morphological details like compounding and affixation, which BLEU might miss [4, 3]. Lastly, TER had been used to count how many changes had to be made to turn the result of the model into the reference translation [4, 33]. This gave a good idea of how much work was involved in the translation and made it possible to figure out where the mistakes were found.

5.3 Baseline Frameworks

To determine the improvements made by I-GANH, its outcomes were compared to three state-of-the-art models typically employed in low- and medium-resource machine translation studies.

Baseline-T: is a standard transformer trained just on the unified parallel corpus, without incremental lexicon updates [39].

RNN-NMT: is a 2-layer bidirectional LSTM encoder-decoder model with attention

Table 1. Dataset Description for English–Khasi Machine Translation

Sl.No	Dataset Name	Sentence Pairs Used	Source
1	Bapyn-EnKha Parallel Corpus	40,000	Hugging Face [22]
2	English–Khasi Synthetic Parallel Corpus	1,000,000	Kaggle [35]
3	WMT 2023 IndicMT Corpus	24,000	WMT 2023 Shared Task [23]

Table 2. Train Sample

Index	English (en)	Khasi (kha)
13626	Then the spirit took me up , and I heard behind me...	Te U Mynsiem jong U Blei u la rah joit ïa nga ...
12104	Among whom are ye also the called of Jesus Christ...	Kane ruh ka kynthup ïa phi kiba ha Rom , ïa ki...
21191	And the Pharisees came to him , and asked him...	Katto katne ngut ki Pharisi ki la wan ha u bad...
3842	SKHEMBOK SUCHIANG (NPP) – 29423.	SKHEMBOK SUCHIANG (NPP) – 29423.
37691	If the contact was accidental, apologize and...	Lada ka jingïakynduh ka long kaba shu jia ryng...

Table 3. Validation Sample

Index	English (en)	Khasi (kha)
31475	Until I make thy foes thy footstool.	tad ynda nga buh ïa ki nongshun jong me kum ka...
28674	Now therefore so shalt thou say unto my servant...	Kumta ïathuh ha u shakri jong nga u Dabid ba n...
11056	missiles	ksangkti
21495	So the people took victuals in their hand, and...	Kumta u Gideon u phah noh ïa kita ki Israel ba...
30603	Nevertheless my brethren that went up with me...	Ki rangbah kiba ïaleit ryngkat bad nga , pynba...

Table 4. Test Sample

Index	English (en)	Khasi (kha)
19725	The wolf and the lamb shall feed together, and...	Ki suri bad ki khunlangbrot kin ïabam lang, k...
39482	The park does not ban picnics.	Ka park kam khang ïa ki jingleit piknik.
9604	superterrific	basngewphylla
23794	O Lord, I know that the way of man is not in...	A Trai, nga tip ba ym don uno uno uba long u...
17396	Now therefore write ye this song for you, and...	Mynta, thoh ïa kane ka jingrwai. Hikai ïa ka...

that simulates traditional neural translation systems [30].

SMT-PBMT (Phrase-Based Machine Translation) is a Moses-style statistical model that incorporates lexicalized reordering and Kneser-Ney smoothing [25].

These baselines establish significant contrast, which enables evaluation of improvements in handling rare terms, lexical innovation, and overall translation fluency.

6 Results Analysis and Discussion

This section discusses the quantitative and qualitative findings of the proposed I GANH framework on the combined English-Khasi dataset.

The evaluation looks at translation quality, lexical generalization, and the advantages of incremental GAN-driven adaptation. The performance is compared to three competitive baselines-Transformer (Baseline-T), RNN-NMT, and SMT-PBMT-with BLEU, METEOR, ChrF++, and TER.

Table 5 summarizes the aggregated performance of all models over the entire test set of mixed-domain data. The comparative performance of all models is also illustrated in Figure 2.

The impact of the Incremental GAN-Driven Lexicon Expansion was investigated using an ablation study to distinguish the unique contribution of GAN-based embedding creation.

Three configurations were compared: a solo Transformer baseline, a Transformer with incremental data updates but no GAN component, and the entire suggested model, which combines the Transformer with GAN-driven lexical synthesis.

This design allows for a clear examination of how dynamically created embeddings affect translation quality beyond what incremental data alone can provide. Table 6 summarizes the results of the ablation study. Figure 3 provides a graphical comparison of the ablation results.

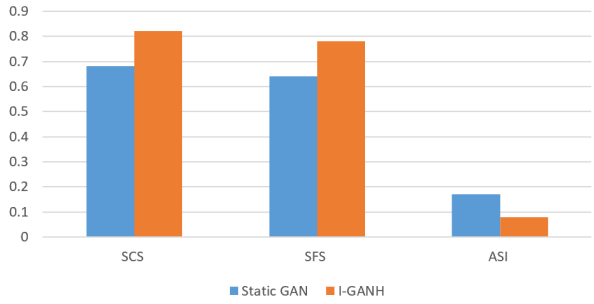

Fig. 5. Evaluation of Synthetic Data

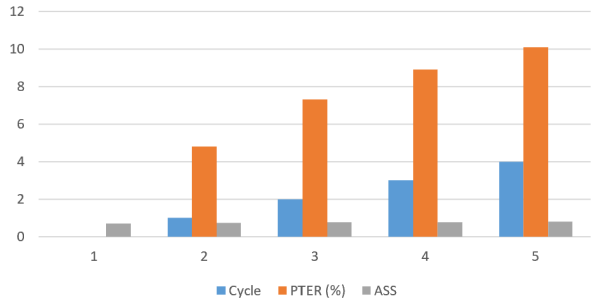

Fig. 6. Phrase Table Expansion and Alignment Improvement

6.1 Effects of Incremental Updates

The effect of incremental learning was evaluated over five update cycles, with each cycle introducing additional Khasi vocabulary and sentence structures. Performance improved gradually, as illustrated in Table 7 and Figure 4.

6.2 GAN-supported Synthetic Data Quality

The score for semantic coherence (SCS), score for syntactic fidelity (SFS), and index of adversarial stability (ASI) were used to assess the quality of the pseudo-parallel data produced by MC-GAN. The evaluation results are presented in Table 8 and illustrated in Figure 5.

6.3 Improvement of Phrase Table Expansion and Alignment

Incremental updates progressively improved SMT alignments. The phrase table expansion rate

Table 5. Overall Model Translation Performance

Model	BLEU	METEOR	ChrF++	TER
Baseline-T (Transformer)	23.78	36.92	49.41	51.78
SMT-PBMT	14.82	28.63	39.42	63.90
RNN-NMT	19.55	32.14	44.87	57.11
I-GANH	29.46	42.87	56.73	46.10

Table 6. Ablation Analysis for Progressive Lexical Growth

Model Variant	BLEU	METEOR	ChrF++	TER
Baseline-T	23.78	36.92	49.41	51.78
With Incremental Monolingual Updates	25.11	38.27	51.32	50.66
With Incremental GAN Lexicon (I-GANH)	29.46	42.87	56.73	46.10

Table 7. Incremental Performance Improvements

Update Cycle	BLEU	METEOR	New Words Added
Cycle 0	18.7	23.2	0
Cycle 1	19.8	24.0	112
Cycle 2	20.5	24.9	87
Cycle 3	21.2	25.5	76
Cycle 4	21.9	26.3	64

Table 8. Evaluation of Synthetic Data

Metric	Static GAN	I-GANH
SCS	0.68	0.82
SFS	0.64	0.78
ASI	0.17	0.08

Table 9. Phrase Table Expansion and Alignment Improvement

Cycle	PTER (%)	ASS
0	—	0.71
1	4.8	0.74
2	7.3	0.77
3	8.9	0.79
4	10.1	0.82

(PTER) and alignment stability score (ASS) are presented in Table 9 and illustrated in Figure 6.

6.4 Key Observations

From Figures 2 and 3 it can be observed that the suggested I-GANH achieves significant improvements in all evaluation metrics.

Improvements in BLEU and METEOR represent improved n-gram coherence and semantic adequacy, whereas higher ChrF++ scores suggest stronger handling of subword-level morphology. The reduction in TER confirms that post-editing requirements have been lowered. The suggested approach increases BLEU by +5.68 points over the Transformer baseline, demonstrating better fluency, while the +5.95 improvement in METEOR demonstrates greater handling of synonyms and morphological variants.

A significant +7.32 improvement in ChrF++ reveals stronger subword-level generalization, while a 5.68-point decrease in TER indicates that fewer human modifications are required for proper translation. The GAN module contributes the most, particularly in METEOR and ChrF++, demonstrating its success in creating meaningful embeddings for new Khasi words and phrases, with the synthetic corpus showing the biggest benefits due to its high lexical variety.

Khasi's low-frequency verbs, classifiers, and noun compounds lead to significant improvements in OOV handling when sentences include at least one rare token (frequency ¡ 5). This is due to the combined strengths of GAN-generated

embeddings and the hybrid statistical fusion mechanism. From Figures 4,5 and 6 it can be observed that the BLEU score increased by 3.2 points between Cycles 0 and 4, indicating that the system successfully acquires fresh data without compromising previously learned representations.

The I-GANH definitely produces more intelligible and structurally consistent sentences than a traditional GAN because of its attention-gated conditioning and replay-regularized discriminator. The improvement in ASS shows that incremental learning makes SMT alignments more robust and semantically solid, which supports the hybrid translation core.

6.5 Qualitative Improvements

In addition to computer-based analysis, 200 English-Khasi translated sentences were evaluated manually.

The translated versions were evaluated by two independent bilingual annotators who were fluent in both English and Khasi. The criteria for the evaluation of the translations were idiomatic accuracy, correct usage of Khasi classifier components such as ka, u, and ki, morphological segmentation and compounding, hallucinated or unsupported words and contextual consistency across multiple sentence translations.

The quality of each parameter was rated on a scale of 1 (poor/incorrect) to 5 (excellent/correct), with higher ratings indicating better translation quality. The translated sentences were independently inspected by each of the annotators, and the outputs produced by the proposed framework were compared to the results achieved by the Transformer baseline.

The manual evaluation showed that the proposed framework is more accurate in capturing Khasi idioms and appropriate in processing classifier components.

The suggested method also produced fewer hallucinated words than the Transformer baseline and used more consistent wording across multi-sentence settings. These qualitative findings are consistent with the quantitative benefits shown in the ChrF++ and METEOR metrics.

7 Conclusion

The paper introduced the Incremental Generative Adversarial Network Integrated Hybrid approach (I-GANH), a unified framework that focuses on improving machine translation for low-resource languages by implementing adaptive learning and hybrid modeling. This framework effectively combines incremental bilingual data processing, SMT-guided alignment priors, transformer-based neural translation, and a modified conditional GAN that can provide linguistically coherent pseudo-parallel data.

The experimental results demonstrate how the framework can easily add fresh terminology, adjust to changing language structures, and keep translation quality consistent without reconstructing the entire process.

The incremental design has less processing overhead, better resistance to data sparsity, and better consistency across translation outputs. Even though this strategy has certain advantages, it also has certain limitations. Incremental updates can sometimes add noise if there are not plenty of fresh terms or if they are not utilized in the right way. Moreover, balancing adversarial optimization is influenced by hyperparameter selections.

Furthermore, using pseudo-generated data could possibly have an effect on unconventional syntactic formulations. Future studies can explore uncertainty-aware filtering of incremental data, enhanced regularization techniques for increasing adversarial stability, and cross-lingual transfer mechanisms to facilitate support for related low-resource languages. Extending the system to multimodal or speech-based translation could make it even more useful. In general, I-GANH points the way toward low-resource machine translation that can adapt and grow.

Competing Interests On behalf of all authors, the corresponding author states that there is no conflict of interest.

Data Availability Statement The data supporting this research are publicly available in the Hugging Face and Kaggle repositories. Link to Dataset in the footnote[1].

Ethical Standards The research meets all ethical guidelines, including adherence to the legal requirements of the study country.

Author Contributions All authors contributed equally to this manuscript. All authors approved the final submitted draft.

References

1. **Ahmed, Z., Das, S. (2023).** A comparative analysis on recent methods for addressing imbalance classification. SN Computer Science, Vol. 5, No. 1, pp. 30.
2. **Ahmed, Z., Das, S. (2024).** A novel hybrid resampling approach to address class-imbalanced issues. SN Computer Science, Vol. 5, No. 7, pp. 865.
3. **Chatzikoumi, E. (2020).** How to evaluate machine translation: A review of automated and human metrics. Natural Language Engineering, Vol. 26, No. 2, pp. 137–161.
4. **Chauhan, S., Daniel, P. (2023).** A comprehensive survey on various fully automatic machine translation evaluation metrics. Neural Processing Letters, Vol. 55, No. 9, pp. 12663–12717.
5. **Chua, L. O., Roska, T. (2002).** The CNN paradigm. IEEE Transactions on Circuits and Systems I: Fundamental Theory and Applications, Vol. 40, No. 3, pp. 147–156.
6. **Devi, M. I., Purkayastha, B. S. (2023).** An exploratory study of SMT versus NMT for the resource constraint English to Manipuri translation. International Conference on Information and Communication Technology for Intelligent Systems, Springer, pp. 329–338.
7. **Donald Jefferson Thabah, N., Purkayastha, B. S. (2021).** Low resource neural machine translation from English to Khasi: A transformer-based approach. Proceedings of the international conference on computing and communication systems: I3CS 2020, NEHU, Shillong, India, Springer, pp. 3–13.
8. **Dwivedi, R. K., Nand, P., Pal, O. (2025).** Hybrid NMT model and comparison with existing machine translation approaches. Multidisciplinary Science Journal, Vol. 7, No. 4, pp. 2025146–2025146.
9. **Gillioz, A., Casas, J., Mugellini, E., Abou Khaled, O. (2020).** Overview of the transformer-based models for NLP tasks. 2020 15th Conference on computer science and information systems (FedCSIS), IEEE, pp. 179–183.
10. **Gomes, H. M., Bifet, A., Read, J., Barddal, J. P., Enembreck, F., Pfharinger, B., Holmes, G., Abdessalem, T. (2017).** Adaptive random forests for evolving data stream classification. Machine Learning, Vol. 106, pp. 1469–1495.
11. **Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y. (2014).** Generative adversarial nets. Advances in neural information processing systems, Vol. 27.
12. **Haddow, B., Bawden, R., Miceli-Barone, A. V., Helcl, J., Birch, A. (2022).** Survey of low-resource machine translation. Computational Linguistics, Vol. 48, No. 3, pp. 673–732.
13. **Huang, K., Li, P., Ma, J., Yao, T., Liu, Y. (2023).** Knowledge transfer in incremental learning for multilingual neural machine translation. Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 15286–15304.
14. **Hujon, A. V., Singh, T. D., Amitab, K. (2023).** Transfer learning based neural machine translation of English-Khasi on

[1] www.kaggle.com/datasets, huggingface.co/docs/hub/en/repositories

low-resource settings. Procedia Computer Science, Vol. 218, pp. 1–8.

15. **Kalimuthu, M., Barz, M., Sonntag, D. (2019).** Incremental domain adaptation for neural machine translation in low-resource settings. Proceedings of the fourth Arabic natural language processing workshop, pp. 1–10.

16. **Kelly, C. (2014).** English–Khasi bilingual sentence pairs. `https://www.manythings.org/bilingual/kha`.

17. **Khazin, K. M., Sanjaya, D., Siregar, M., Meisuri, Adisaputra, A. (2023).** Comparison of machine translations (MT) technology; statistical (SMT) vs. neural (NMT). 2nd international conference on innovative technology & science 2021: Forging A Smart Technological World, AIP Publishing LLC, Vol. 2732, No. 1, pp. 040006.

18. **Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A. A., Milan, K., Quan, J., Ramalho, T., Grabska-Barwinska, A., et al. (2017).** Overcoming catastrophic forgetting in neural networks. Proceedings of the national academy of sciences, Vol. 114, No. 13, pp. 3521–3526.

19. **Koç, C., Özyurt, F., Iantovics, L. B. (2025).** Survey on latest advances in natural language processing applications of generative adversarial networks. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, Vol. 15, No. 1, pp. e70004.

20. **Mohamed, Y. A., Khanan, A., Bashir, M., Mohamed, A. H. H., Adiel, M. A., Elsadig, M. A. (2024).** The impact of artificial intelligence on language translation: a review. IEEE Access, Vol. 12, pp. 25553–25579.

21. **Mukhim, B., Kumar Maji, A., Das, S. (2025).** Sentiment analysis with khasi low-resource language through generation of sentiment words using machine learning. Computación y Sistemas, Vol. 29, No. 3.

22. **Nongkynrih, B. (2025).** Bapyn-EnKha parallel corpus: English–Khasi parallel dataset. Hugging Face Dataset. Accessed: 2026-08-03.

23. **Pal, S., Pakray, P., Laskar, S. R., Laitonjam, L., Khenglawt, V., Warjri, S., Dadure, P. K., Dash, S. K. (2023).** Findings of the WMT 2023 shared task on low-resource Indic language translation. Proceedings of the Eighth Conference on Machine Translation, pp. 682–694.

24. **Papineni, K., Roukos, S., Ward, T., Zhu, W.-J. (2002).** Bleu: a method for automatic evaluation of machine translation. Proceedings of the 40th annual meeting of the Association for Computational Linguistics, pp. 311–318.

25. **Popović, M. (2018).** Language-related issues for NMT and PBMT for English–German and English–Serbian. Machine Translation, Vol. 32, No. 3, pp. 237–253.

26. **Ranathunga, S., Lee, E.-S. A., Prifti Skenduli, M., Shekhar, R., Alam, M., Kaur, R. (2023).** Neural machine translation for low-resource languages: A survey. ACM Computing Surveys, Vol. 55, No. 11, pp. 1–37.

27. **Rivera-Trigueros, I. (2022).** Machine translation systems and quality assessment: a systematic review. Language Resources and Evaluation, Vol. 56, No. 2, pp. 593–619.

28. **Roy, A. K., Purkayastha, B. S. (2023).** Machine translation systems for official languages of North-Eastern India: A review. International Conference on Computational Intelligence in Communications and Business Analytics, Springer, pp. 301–315.

29. **Shah, P., Bakrola, V. (2019).** Neural machine translation system of indic languages-an attention based approach. 2019 Second International Conference on Advanced Computational and Communication Paradigms (ICACCP), IEEE, pp. 1–5.

30. **Sharma, S., Diwakar, M., Singh, P., Tripathi, A., Arya, C., Singh, S. (2021).** A review of neural machine translation based on deep learning techniques. 2021 IEEE 8th Uttar Pradesh Section International Conference

on Electrical, Electronics and Computer Engineering (UPCON), IEEE, pp. 1–5.

31. **Singh, T. D., Hujon, A. V. (2020).** Low resource and domain specific English to Khasi SMT and NMT systems. 2020 International Conference on Computational Performance Evaluation (ComPE), IEEE, pp. 733–737.

32. **Sitender, Bawa, S., Kumar, M., Sangeeta (2023).** A comprehensive survey on machine translation for English, Hindi and Sanskrit languages. Journal of Ambient Intelligence and Humanized Computing, Vol. 14, No. 4, pp. 3441–3474.

33. **Snover, M., Dorr, B., Schwartz, R., Micciulla, L., Makhoul, J. (2006).** A study of translation edit rate with targeted human annotation. Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers, pp. 223–231.

34. **Thabah, N. D. J., Purkayastha, B. S. (2019).** Khasi to English neural machine translation: an implementation perspective. International Journal of Engineering and Advanced Technology (IJEAT), Vol. 9, No. 2.

35. **User10383 (2025).** English–Khasi synthetic parallel corpus. Kaggle Dataset. Accessed: 2026-08-03.

36. **Wang, H., Wu, H., He, Z., Huang, L., Church, K. W. (2022).** Progress in machine translation. Engineering, Vol. 18, pp. 143–153.

37. **Xia, W., Zhang, Y., Yang, Y., Xue, J.-H., Zhou, B., Yang, M.-H. (2022).** Gan inversion: A survey. IEEE transactions on pattern analysis and machine intelligence, Vol. 45, No. 3, pp. 3121–3138.

38. **Xiao, Y., Wu, L., Guo, J., Li, J., Zhang, M., Qin, T., Liu, T.-y. (2023).** A survey on non-autoregressive generation for neural machine translation and beyond. IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 45, No. 10, pp. 11407–11427.

39. **Zhang, P., Chen, B., Ge, N., Fan, K. (2020).** Long-short term masking transformer: A simple but effective baseline for document-level neural machine translation. arXiv preprint arXiv:2009.09127.

Article received on 21/03/2026; accepted on 20/06/2026.
**Corresponding author is Sufal Das.*