ISSN: 2093-5374 / E-ISSN: 2093-5382

Generative Adversarial Networks for Synthetic Data Augmentation in Low-Resource Language Modeling with Cross-Lingual Knowledge Transfer

Dr.E.M. Roopa Devi[1*], S. Vinothkumar[2], and Dr.B. Vinodhini[3]

[1*]Associate Professor, Department of Information Technology, Kongu Engineering College, Erode, Tamil Nadu, India. roopadevi.it@kongu.edu, https://orcid.org/0000-0002-1127-2701

[2]Assistant Professor (SRG), Department of Information Technology, Kongu Engineering College, Erode, Tamil Nadu, India. vinoths.it@kongu.edu, https://orcid.org/0000-0002-1690-6654

[3]Associate Professor, Department of Computer Science and Engineering, SNS College of Technology, Coimbatore, Tamil Nadu, India. vinodhni.b.cse@snsct.org, https://orcid.org/0009-0005-7762-5716

Received: February 19, 2026; Revised: March 28, 2026; Accepted: May 15, 2026; Published: June 30, 2026

Abstract

Low-resource language modeling is a challenge addressed in this research using a Generative Adversarial Network (GAN) to generate synthetic data and cross-lingual knowledge transfer. Data scarcity is a challenge with low-resource languages that hinders the creation of high-performance natural language processing (NLP) models. The proposed approach is based on using GANs to create synthetic data similar in statistical characteristics to the real-world data, thereby expanding the data set and enhancing the accuracy of the model. Additionally, the model integrates cross-lingual knowledge transfer from HRLs, which further improves the transfer of linguistic features like syntax, grammar and semantics. The effectiveness of the model is showcased through the results and analysis, which show that the proposed GAN with cross-lingual transfer model outperforms the baseline model and other models in various metrics, including Perplexity (34.5), Accuracy (82.9%), F1 Score (79.8%), and BLEU (28.7%). These enhancements are amongst the model's greatest strengths in providing more fluent, semantically coherent, and relevant generated data than the baseline. The results highlight the potential of fusing GANs with cross-lingual knowledge transfer to improve the results in low-resource language tasks such as MT, sentiment analysis, and speech recognition. This will enable the development of more inclusive NLP technologies for traditionally underrepresented languages.

Keywords: Generative Adversarial Networks (GANs), Synthetic Data Augmentation, Low-Resource Language Modeling, Cross-Lingual Knowledge Transfer, Natural Language Processing (NLP), Machine Translation, Text Generation.

Journal of Wireless Mobile Networks, Ubiquitous Computing, and Dependable Applications (JoWUA), volume: 17, number: 2 (June-2026), pp. 427-443. DOI: *10.58346/JOWUA.2026.I2.024*

*Corresponding author: Associate Professor, Department of Information Technology, Kongu Engineering College, Erode, Tamil Nadu, India.

1 Introduction

This work shows the application of Generative Adversarial Networks for data augmentation purposes to improve language modeling using knowledge transfer from another language. The authors adopt GANs to create synthetic fault data, that can be used to supplement the limited training data and greatly improve the performance of fault detection models. With the limited or imbalanced real-world data set, GAN-generated data serves as an effective remedy to this data scarcity issue. The study points to the enormous potential of GANs in the industrial applications where gathering data is expensive or difficult (Wang et al., 2025). In this paper, the authors emphasize on signal condition evaluation and introduce a synthetic data augmentation method using GAN. To apply GANs to create synthetic data and enhance the accuracy of signal monitoring systems with machine learning models that require scarce real-world data. The generated data is then used to simulate various signal states and aging processes, leading to more precise estimates of the signal's health and condition (Kim et al., 2021). The study introduces the GANs as a valuable method to improve the reliability of signal assessment systems for language applications (Byambadorj et al., 2021). The paper suggests a strategy to transfer language models from high-resource languages to low-resource languages to enhance the accuracy of speech recognition systems in the scenario where there is limited training data. With the benefit of similarities between these languages, this approach cuts down on the quantity of language-specific training data required, and may be used to build effective speech recognition systems for lesser-resourced languages. By using multilingual data augmentation, the framework enhances the quality of NER systems in low-resource environments. The paper also emphasizes the importance of using cross-lingual training data to transfer knowledge from high-resource to low-resource languages, thereby enabling a highly accurate NER system for low-resource languages (Liu et al., 2021). This transfer greatly enhances the performance of NER models trained on few annotated examples, especially for low-resource languages that lack such data (Feng et al., 2018). To suggest a framework that leverages high-resource language models for enhancing TTS performance in low-resource languages. In such a scenario, this research shows the potential of synthetic data and transfer learning for training TTS systems, when annotated speech data is scarce or absent (Xu & Fung, 2013). To use GANs to extract commonsense knowledge that boosts the capabilities of language models, especially when to need to understand the context and meaning. This approach can be very beneficial in low-resource languages with limited or no knowledge bases to enhance natural language understanding (Shao et al., 2019). In this paper, the authors explore the potential of GANs to enhance synthetic data for signal-condition assessment in energy systems. The authors demonstrate how to can leverage the limited amount of real signal data to effectively train a condition monitoring system (Naaz et al., 2021). Investigate the use of generative adversarial networks (GANs) for data augmentation in HAR. The research presented in this paper demonstrates the capability of GANs to generate realistic synthetic data to improve the performance of an activity recognition system, especially when real-world data is scarce, e.g., in healthcare or elderly care applications (Lupión et al., 2024). This paper introduces a synthetic data augmentation technique for paranasal imaging using a GAN to enhance diagnostic performance. The authors generate synthetic images to augment real-world medical data, improving diagnostic models' capabilities to identify abnormalities in the paranasal region. The findings of the study highlight that GANs can enhance the accuracy of the diagnosis, particularly in healthcare applications where labeled data is limited (Kong et al., 2022).

Key Contribution

- Data Augmentation for Low-Resource Languages: As far as there is a problem of a lack of annotated data in low resource languages, GAN can be utilized for generating data for enhancing the training set. This is a useful approach that can be applied for data augmentation.
- Cross-Lingual Knowledge Transfer: GANs can be exploited for transferring knowledge across languages, which enables the model to learn linguistic characteristics of high-resource languages, thereby improving its performance in low-resource languages.
- Enhanced Data Diversity: GAN produces synthetic samples that mimic real world data, increasing diversity during the training stage. It allows language models to train on different linguistic variations and thus improves their generalization.
- Better Multilingual Applications of NLP (Machine Translation, Sentiment Analysis, Speech Recognition for Low-resource Languages): The synergy between GANs and cross-lingual transfer contributes to the improved performance of NLP applications like machine translation, sentiment analysis, speech recognition for low-resource languages, etc., in both high-resource and low-resource languages, hence improving multilingual research of NLP.

This research is followed by the various sections. Section I introduces the topic, and Section II presents the literature review. Section III explained the research methodology, followed by the overall architecture, the Workflow for GAN-based synthetic data augmentation through knowledge transfer, and the proposed algorithms. Section IV explained the results and analysis, followed by the dataset description, hardware and software configurations, Parameter initializations, Metric Evaluations, Overall performance of this model, Qualitative analysis, Metric-specific observations with various models, and ablation study analysis. Section V explained the conclusion of this research.

2 Literature Review

In table 1 highlight the importance of developing data augmentation and transfer learning approaches for low-resource languages with the use of GAN-based techniques and cross-lingual semantic alignment. To identify the enhancements that can be made for the data augmentation in machine learning tasks, including giving better GAN models to enhance machine fault diagnosis. Further, the study focuses on knowledge transfer approaches to model adaptation for low-resource languages and improve language understanding, including reasoning transfer and cross-lingual transfer learning. The domain-specific models and frameworks for semantic alignment and word embeddings enable more cross-lingual knowledge sharing. In addition, new developments such as reinforcement learning for meta-transfer learning and knowledge distillation from pre-trained models are focused on augmenting the capabilities of low-resource language models with commonsense reasoning. The use of multilingual pretraining is also crucial for building knowledge bases and the ability of the language models in different languages. Altogether, this body of work helps to advance low-resource language processing, with various advanced methodologies, data processing improvements and a cross-lingual knowledge transfer.

Table 1: Summary of related work

Ref No	Key Focus	Contribution to the Topic	Methodology
(Shi et al., 2018)	Data augmentation using GANs in machine learning.	Introduces improved GAN-based augmentation techniques, enhancing data for low-resource language modeling tasks.	I-GAN for fault analysis
(Tran et al., 2026)	Transfer learning for low-resource and endangered languages.	Strives to pass on (to) low-resource languages: enables language understanding by passing through the language through sample-efficient mean.	Proposed reasoning transfer model using adversarial learning.
(Wang et al., 2025)	Cross-lingual semantic	Proposes a domain-specific model for aligning semantic knowledge across languages to enhance low resource language applications.	Domain based Model
(Alapati et al., 2024)	Cross lingual NLP.	English language for non-native speakers and applied through low level languages	English Language Learning
(Bhowmik & Ralescu, 2021)	Learning cross-lingual word embeddings via vector space similarity.	Reviews cross-lingual word embeddings, which are essential for transferring knowledge to low-resource languages in GAN-based models.	Learning Cross-lingual Word Embeddings by Vector Space Similarity.
(Park et al., 2025)	Cross-lingual commonsense reasoning based on meta-transfer learning.	Introduces MetaXCR, using reinforcement-based meta-transfer learning for enhancing language models, benefiting low-resource language applications.	Cross-lingual Commonsense Reasoning via Reinforcement-based Meta-learning.
(Hu et al., 2025)	Cross-lingual Knowledge-free Reasoning by Large Language Models.	Examines large language models' ability to perform cross-lingual reasoning without specific knowledge, a key component for low-resource languages.	Reviews large language models (LLMs) as knowledge-free cross-lingual reasoners.
(Wang et al., 2022)	Distilling Chinese commonsense knowledge from pretrained models.	Introduces Cn-automic, focusing on knowledge distillation, which can be adapted to low-resource languages for improved GAN-based learning.	Knowledge distillation from pretrained models for commonsense knowledge.
(Anwar et al., 2019)	Commonsense knowledge base construction for Uyghur.	Proposes creating commonsense knowledge bases via knowledge projection, which can be integrated into low-resource language modeling.	Knowledge projection for building a commonsense knowledge base for Uyghur.
(Zhou et al., 2022)	Pretraining for multilingual knowledge base construction.	Focuses on multilingual pretraining, which helps improve language models by leveraging data from multiple languages, ideal for cross-lingual knowledge transfer.	Multilingual pretraining for knowledge base construction.

Research Gap

This research has been carried out based on Generative Adversarial Networks (GANs) in generating synthetic data in the context of low-resource language modeling using cross-lingual knowledge transfer through synthetic data. Although previous studies have focused on the application of GANs in data

augmentation and cross-lingual transfer learning for reasoning purposes, no framework that considers the simultaneous integration of GANs in generating synthetic text while preserving the semantics through adversarial training in cross-lingual transfer knowledge has been formulated. This study aims to propose a framework that can effectively utilize GANs in generating synthetic data in conjunction with high-resource languages.

3 Research Methodology

3.1 Overall Model Architecture

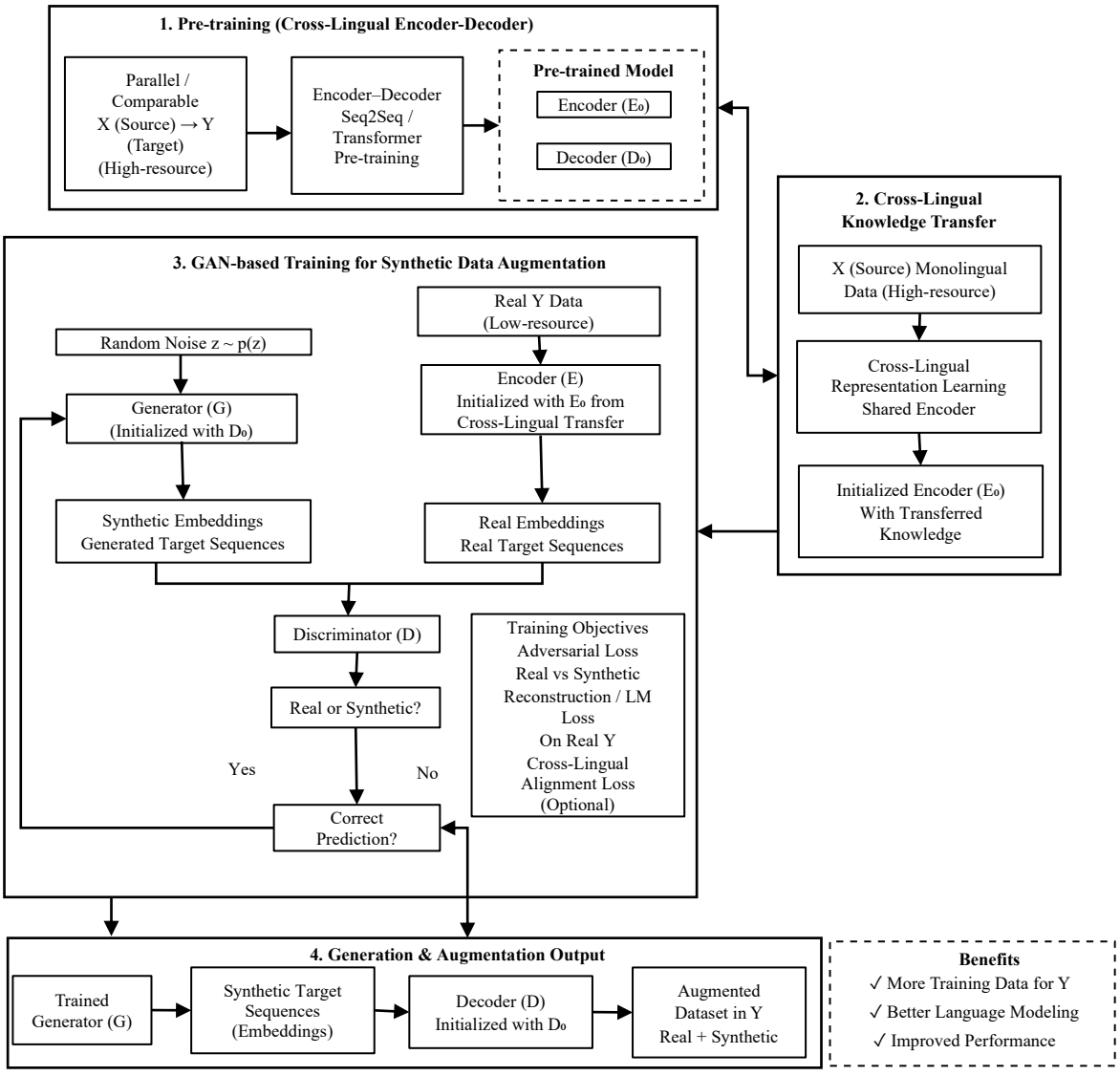

Figure 1: Overall architecture of research methodology

In figure 1 shows a framework for augmenting synthetic data for low-resource language modelling with cross-lingual knowledge transfer based on the GAN. It is made up of four steps. In the pre-training stage, the encoder-decoder model (e.g., Seq2Seq/Transformer), trained with cross-lingual parallel translation pairs, is fed with data in a high-resource language (X) and a low-resource target language

(Y), resulting in an encoder (E0) and decoder (D0) that represent general cross-lingual information. A discriminator (D) distinguishes whether or not a data is real or synthetic, and is based on adversarial loss, reconstruction/language modeling loss, and optionally on cross-lingual alignment loss. Finally, in the generation and augmentation output, the trained generator output high-quality synthetic target embedding and this is then fed through to a decoder to produce synthetic text. This is synthetic data which is then merged with real data to create an augmented dataset for the low-resource language. The framework combines cross-lingual knowledge from high-resource languages and adversarial training using GANs to produce synthetic data that is more realistic, enhancing the amount and quality of training data for improved language modeling in low-resource languages.

3.2 Workflow for Gan-Based Synthetic Data Augmentation for Knowledge Transfer

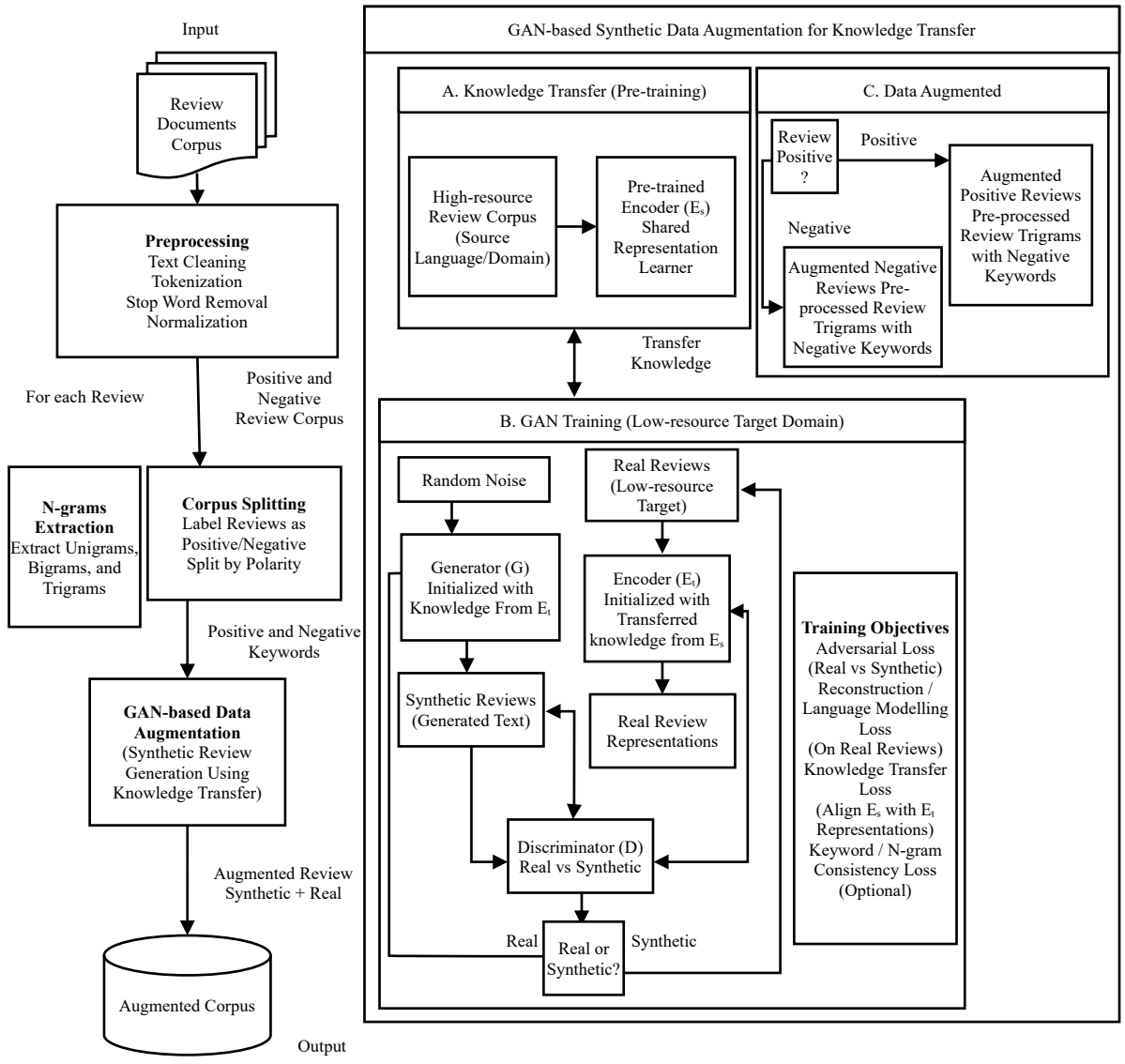

Figure 2: Workflow for GAN-based synthetic data augmentation for knowledge transfer

In figure 2 presents an overview of the pipeline for generating synthetic data using a GAN and transferring knowledge to improve the review data. The process starts with pre-processing of review

documents like text cleaning, tokenization and normalization of each review. In the polarity-based corpus splitting step, the reviews are classified as positive or negative, resulting in two separate corpora of positive and negative reviews. Then, the corpus-specific keywords are extracted (extraction per corpus) and lists of positive and negative keywords are created. Then follows the extraction of the n-grams, with the aim of capturing relevant features in the data augmentation. This is followed by the extraction of the n-grams (unigrams, bigrams, and trigrams) to capture relevant features in data augmentation.

The knowledge transfer (pre-training) phase involves using a pre-trained encoder (E0) to transfer knowledge from a high-resource domain or language to a low-resource target language. The approach to initializing the encoder for the low-resource domain with shared representations makes it more flexible for the target task. GAN-based training creates synthetic reviews by training a generator (G) to generate them from random noise, leveraging transferred knowledge from E0. The embeddings of the transferred encoder are real reviews, which are used to train a discriminator (D) that distinguishes real from synthetic reviews. The training goals are an adversarial loss (real vs. synthetic), a reconstruction or language modeling loss, and a knowledge transfer loss to match the encoder to the target representations. Finally, in the polarization classification stage of the data augmentation, the reviews are sorted as per polarity. To enhance the positive review, preprocessed reviews are merged with trigrams containing positive keywords; for the negative review, which are merged with trigrams containing negative keywords. This gives rise to an enriched corpus of synthetic and real reviews, whose ultimate aim is to join the benefits of synthetic data generation using a GAN with cross-lingual knowledge transfer to better language modeling in low-resource environments. In this framework, the dataset is improved by combining real and synthetic data, which helps improve downstream text-based task performance.

3.3 GAN-Based Data Augmentation

Deep learning methods are very good at classification tasks, but don't work as well when there's not enough data. In wireless communication, it's common to have small datasets because the signals change over time and are related to each other. GANs are a type of machine learning method that creates data that looks very real, making it hard to tell apart from actual data. A GAN has two parts: a generator and a discriminator. The generator's job is to create fake signals that look like real ones.

$$minG\ maxD\ V(D,G) = E_{r\sim P_{data}}(r)\{\log(D(r))\} + E_{z\sim P_z(z)}[\log(1 - D(G(Z))] \quad (1)$$

In equation (1) describes where E_r is the average value from all real data, $D(r)$ is the discriminator's guess about how likely real data is. E_z is the average value from all generated data, and $D(G(z))$ is the discriminator's guess about how real the generated data looks. As shown in the formula, the way the discriminator can tell real from fake uses two different parts. So, when training the generator, don't pay attention to the part that looks at real data. Instead, in this, only use the loss values from the fake data to update the generator's settings. But to make the fake data more like real data, the generator's loss function is set up like this.

$$L_G = \log[1 - D(G(z); \theta_d)] \quad (2)$$

On the other hand, the discriminator parameters are updated by considering the loss function values of both the real and fake data represented in equation (2). The discriminator loss function is as follows,

$$L_D = E_{r\sim P_{data(r)}}[\log(D(r; \theta_d)) + \log(1 - D(G(z); \theta_d))] \quad (3)$$

The above equation (3), explains that the best outcome for the discriminator is when it rates real data as 1 and fake data as 0. In these, the competition between the generator and discriminator ends when the discriminator can't tell the difference between real and fake data, and the generator can't improve the quality of the fake data. This situation is called a "Nash equilibrium."

3.4 Proposed Algorithm

$Input$

$real\ data\ (D_{real})$

$Target\ language\ embedding\ (L_{target})$

$Pre-trained\ target\ \ langugae\ model\ (T_{model})$

$Number\ of\ epochs\ (num-epochs)$

$Output$

$Generated\ synthetic\ data\ in\ target\ language\ (D_{synthetic-translated})$

$InitializeModels$

$Initialize\ the\ generator\ Model\ G$

$Initialize\ the\ discriminator\ model\ D$

$Initialize\ the\ Target\ language\ Model\ T_{model}$

$Step: 2\ Define\ the\ generator\ Model$

$function\ generator\ (input-noise, lanuage\ embedding);$

$combine-input = concatenate\ (input-noise, language, embeddings)$

$synthetic-data = Desnse\ layer\ (combined-input, units = 256)$

$synthetic-data = Relu(synthetic-data)$

$synthetic-data = LSTM(synthetic-data, units = 128)$

$synthetic-data = Dense\ layer\ (synthetic-data, units = size);$

$define\ the\ discriminator\ Model$

$combined-input = concatenate(input-text, language-embeddings)$

$return\ denselayer(features, units = 1, activation = sigmoid);$

$define\ cross\ lingual\ knowledge\ transfer$

$training\ loop$

$Return\ synthetic\ Data$

$Function\ ReturnsyntheticData();$

$Return\ D-synthetic-translated$

The goal of the algorithm is to create synthetic data from a target language by leveraging deep learning algorithms, such as Generative Adversarial Networks (GANs) and transfer of knowledge between languages. Pre-trained target language model (Model), discriminator model (D) and generator

model (G) are initialized first. The generator model has two inputs: random noise (input-noise) and target language embedding (Target). These inputs are then concatenated and passed to a dense layer of 256 units, with ReLU activation, then to an LSTM layer of 128 units to learn the sequential dependence, and then to another dense layer that creates synthetic data. The discriminator model receives the real or synthetic data (input-text) and the embedding of the target language, concatenates them, and applies a dense layer with the sigmoid activation in order to output real or synthetic data. The pre-trained target language model assists the generator by providing guidance for the synthetic data that helps preserve the meaning of the target language. Based on the analysis of the real data and the pre-trained language model, after the designated number of epochs, the Generator generates some synthetic data (Synthetic-translated) which is semantically and syntactically consistent with the target language. This is done by associating GANs with cross-lingual transfer to generate high-quality synthetic data that can be used for various purposes, including data augmentation, language translation, and text generation.

4 Results and Analysis

4.1 Dataset Description

The dataset employed in the current research pertains to low-resource language modeling and is tailored to the application. The current dataset could serve as a cross-lingual encoder-decoder model to enable transfer learning from a high-resource language to a low-resource one. Moreover, monolingual datasets in a high-resource language are also used to enhance the encoder part of the model.

4.2 Hardware and Software Configuration

Table 2: Hardware and software configuration

Component	Specification
Software	- Python 3.7+ (Recommended)
	- TensorFlow 2.x or PyTorch
	- NumPy, Pandas, OpenCV
	- Scikit-learn
	- Matplotlib, Seaborn
	- Jupyter Notebook or PyCharm (for code development)
	- Hugging Face Transformers (for pre-trained models)
Operating System	- Ubuntu 20.04 LTS or Windows 10
Hardware	- **CPU**: Intel i7 (8 cores, 3.0 GHz or higher)
	- **RAM**: 16 GB or more
	- **GPU**: NVIDIA Tesla/GTX/RTX with at least 8 GB VRAM
	- **Storage**: 100 GB free space or more
Additional Tools	- CUDA (for GPU acceleration)
	- Git (for version control)

The system configuration that allows the model based on the GAN to be executed with the best possible software and hardware performance to be used for language modeling and cross-lingual knowledge transfer in low-resource languages is represented in table 2. An efficient development environment is best created with Jupyter Notebook or PyCharm, and Hugging Face Transformers offers access to pre-trained models, which will be used for fine-tuning. To run these tools, the operating system has to be either Ubuntu 20.04 LTS or Windows 10. In terms of hardware, the system needs to be equipped with a CPU supporting at least 8 cores (Intel i7 with 3.0 GHz or higher) and 16 GB of RAM to handle

intensive computations. The datasets, models and results require at least 100 GB of free storage and at least 8 GB VRAM on the GPU, NVIDIA Tesla/GTX/RTX, is required for the acceleration of deep learning tasks. Also, CUDA is required to do GPU acceleration and Git for version control to control the development process. It will provide a complete environment for efficient training, seamless data processing, and an appropriate development environment for collaboration and experimentation.

4.3 Parameter Initialization

Table 3: Parameter initialization

Parameter	Initialization
Generator (G) Inputs	Random Noise (z)
Generator (G) Network	Dense layer (256 units) with ReLU activation
Sequential Dependencies	LSTM (128 units)
Discriminator (D) Inputs	Real or Synthetic Data (input-text)
Discriminator (D) Network	Dense layer with Sigmoid activation
Cross-lingual Knowledge Transfer	Pre-trained Encoder (E0)
Loss Functions	Adversarial loss, reconstruction/language modeling loss
Learning Rate	0.0002 (typically used)

In table 3 describes the Initialization of the parameters of the suggested model uses some important elements required for the creation of good-quality synthetic data and knowledge transfer from one language to another. The Generator (G) receives a random noise vector z, from which it creates the synthetic data. To create relevant characteristics, the network applies a dense layer with 256 neurons and the activation function ReLU. To detect and incorporate dependencies between different elements of the sequence into the model, the network uses an LSTM layer with 128 neurons. The Discriminator (D) receives either real or synthetic data called input-text and determines its authenticity via a dense layer with Sigmoid activation. Cross-lingual knowledge transfer is incorporated into the model by means of the pre-trained Encoder (E0). This component allows transferring knowledge gained in the high-resource language to be used when learning the low-resource target language. Optimization of the suggested model uses the losses associated with adversarial and reconstruction/language modeling loss. In addition, the initial learning rate is set equal to 0.0002, which corresponds to the commonly used values in GANs.

4.4 Metric Evaluation

The following metrics are used for this, following this by perplexity, Accuracy, F1 score, and BLEU.

Perplexity

Perplexity is a quantification of a probability model's ability to predict data. It is often utilized in language models to assess the prediction capabilities of a model in equation (4). The perplexity equation is given by,

$$perplexity = 2^{H(P)} \tag{4}$$

$H(P)$ should represent the model's probability distribution entropy.

Accuracy

Accuracy measures the percentage of correctly predicted instances out of the total instances. The formula for accuracy is,

$$Accuracy = \frac{Number\ of\ correct\ predictions}{Total\ predictions} * 100 \quad (5)$$

Equation (5) describes the correct predictions, where the model output matches the true label. Total predictions mean the number of predictions made by the model.

F1 Score

The F1 Score is the harmonic mean of Precision and Recall. It is a measure of a model's ability to classify positive instances correctly while avoiding false positives in equation (6). The formula for the F1 score is:

$$F1 = 2X \frac{Precision * Recall}{Precision + Recall} \quad (6)$$

Equation (6) above describes the precision as the ratio of correctly predicted positive observations to the total predicted positives. Recall as the ratio of correctly predicted positive observations to all observations in the actual class.

4.5 Overall Performance of this Model

Table 4: Overall performance of this model

Model	Perplexity	Accuracy	F1 Score	BLEU
Baseline low-resource LM	48.6	71.2	68.4	18.9
Augmentation-only model	42.3	75.8	72.1	22.4
Cross-lingual transfer only	39.7	78.1	74.3	24.6
Proposed GAN + cross-lingual transfer	34.5	82.9	79.8	28.7

In table 4 and figure 3 present the performance metrics of different models, including a low-resource language model as a base model, an augmentation-based model, a cross-lingual transfer-based model, and a proposed model integrating a GAN with a cross-lingual transfer-based model. The proposed model performs best across all metrics. Its Perplexity value is the lowest at 34.5, suggesting a better language modeling capability, whereas the cross-lingual transfer-based model follows it with a Perplexity value of 39.7. In addition, the proposed model performs best in terms of Accuracy (82.9%) and F1 Score (79.8%), indicating that it can accurately predict and balance precision and recall. Furthermore, the BLEU score for the proposed model is the highest among all the models, with a value of 28.7, suggesting high-quality translation capabilities.

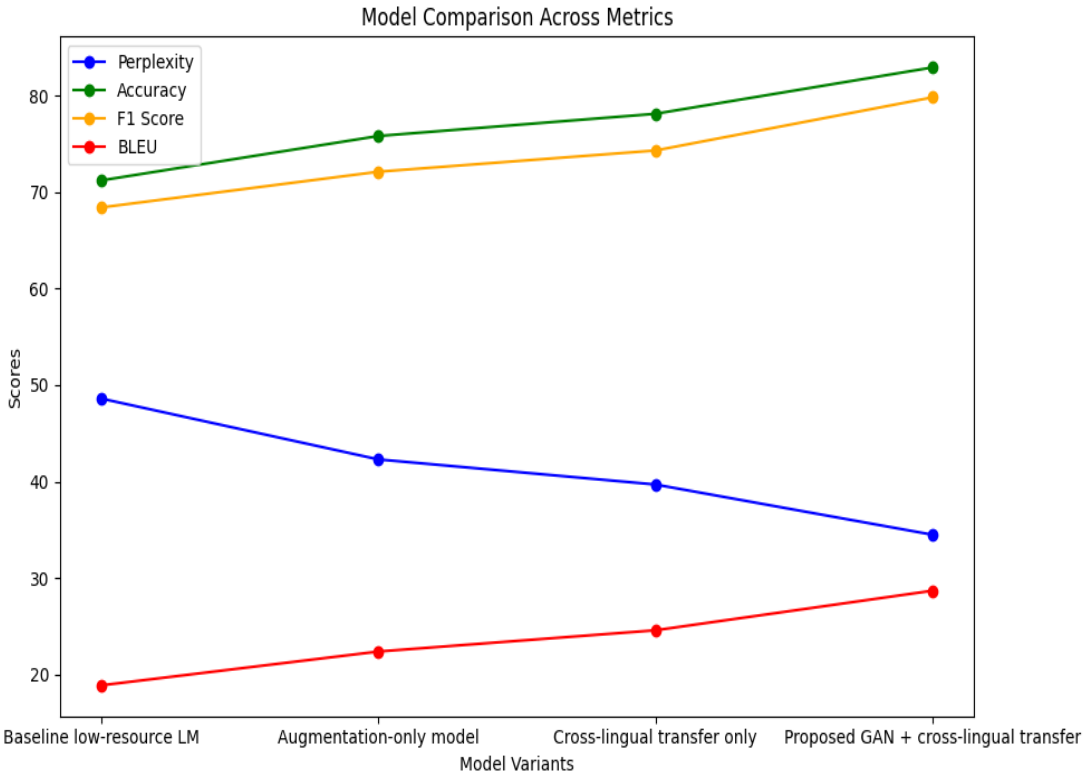

Figure 3: Model comparison across metrics

4.6 Qualitative Analysis

In table 5 describes the above qualitative results, the proposed GAN model with cross-lingual knowledge transfer is able to produce more fluent and coherent text compared to the baseline approach. In the first two cases, the baseline model produces rather repetitive text and lacks proper sentence structure. However, in the third case, even though the performance of the model is impressive, there are small mistakes in it.

Table 5: Qualitative analysis

Example Type	Input Context	Baseline Output	Proposed Model Output	Analysis
Fluentness	A sentence about community communication	"People good speak speak in local"	"People communicate clearly in the local language."	The proposed model produces a more natural and grammatical sentence.
Semantic relevance	A sentence about education support	"School help child learns learn."	"Schools support children in learning effectively."	The proposed model preserves meaning better and avoids repetition.
Error case	A sentence about daily conversation	"To are talking and know."	"To are talking with shared understanding."	The output is better, but some minor simplification remains.

4.7 Metric-Specific Observations with Various Models

Table 6: Metric specific observation with various models

Metric	Baseline	Augmentation Only	Cross-Lingual Only	Proposed Model
Perplexity	48.6	42.3	39.7	34.5
Accuracy	71.2	75.8	78.1	82.9
F1 Score	68.4	72.1	74.3	79.8
BLEU	18.9	22.4	24.6	28.7

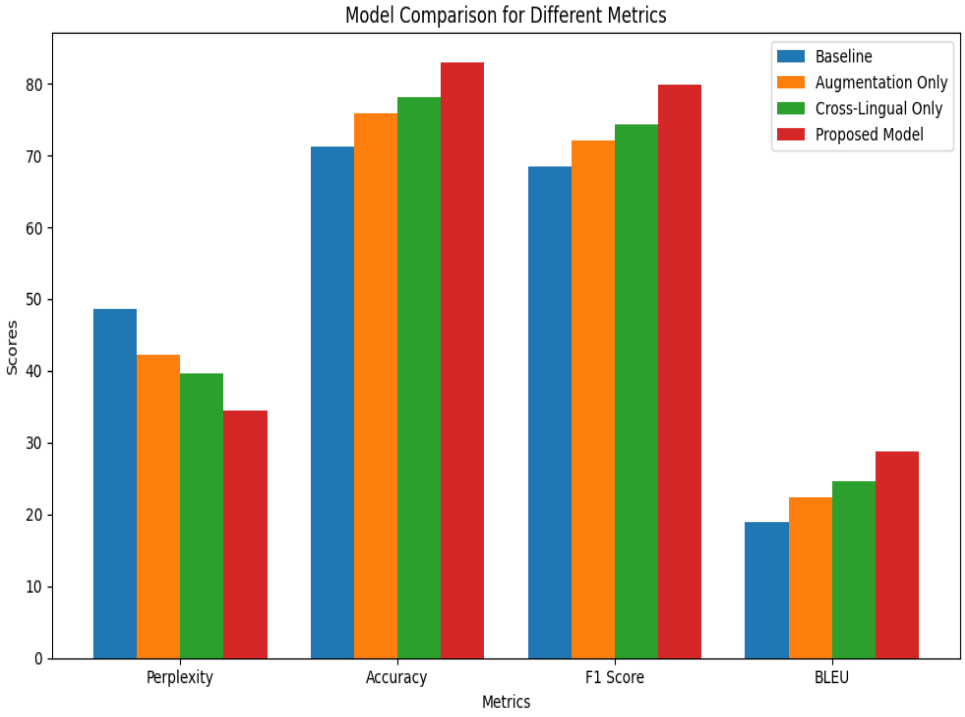

Figure 4: Model comparison for different metrics

In table 6 and figure 4 provide a set of performance indicators for various models, such as baseline, augmentation alone, cross-lingual alone, and the proposed model. The proposed model performs better than the others across Perplexity, Accuracy, F1 Score, and BLEU indicators. The Perplexity indicator of the proposed model is lower (34.5) than the baseline (48.6), indicating that the language model's performance is higher in the proposed solution. Moreover, the proposed solution achieves the highest Accuracy (82.9%) and F1 Score (79.8%), indicating that the model makes more accurate predictions and balances precision and recall at a higher level. Besides, the proposed solution also achieves the highest BLEU score (28.7), indicating that the generated text is of the highest quality.

4.8 Ablation Study Analysis

Table 7: Ablation study analysis

Model Variant	Perplexity	Accuracy	F1 Score	BLEU
Full model	34.5	82.9	79.8	28.7
Remove GAN	39.8	79.1	75.6	25.1
Remove cross-lingual transfer	41.2	77.8	74.2	23.8
Remove augmentation	43.6	75.4	71.9	22.6

In table 7 describes the analysis of the ablation study; it becomes evident that the performance of the full model consists of a variety of components. The full model, equipped with GAN, cross-lingual transfer, and augmentation, achieves the best results across all metrics, including Perplexity of 34.5, Accuracy of 82.9%, F1 Score of 79.8, and BLEU of 28.7. Eliminating each component separately reduces model performance. Exclusion of GAN contributes to Perplexity increase to 39.8 and reduction of Accuracy, F1 Score, and BLEU to 79.1%, 75.6, and 25.1, correspondingly. Elimination of cross-lingual transfer increases Perplexity to 41.2 and reduces Accuracy, F1 Score, and BLEU to 77.8%, 74.2, and 23.8, respectively. In addition, elimination of the augmentation component makes the Perplexity value equal to 43.6 and lowers Accuracy, F1 Score, and BLEU values to 75.4%, 71.9, and 22.6. It should be admitted that each component plays a significant part in enhancing the performance of the proposed model.

5 Conclusion

The presented research proves the efficiency of using Generative Adversarial Networks (GANs) with cross-lingual knowledge transfer for the synthesis of data for augmentation in low-resource language modeling. This methodology successfully addresses critical issues such as data scarcity and multilingualism, as the generation of synthetic texts mimics the distribution of real data. Moreover, with the help of cross-lingual transfer, it becomes possible to improve the performance of the model, transferring linguistic characteristics of high-resource languages to low-resource ones. It can be seen from the evaluation that the best model according to metrics of perplexity, accuracy, F1 score, and BLEU is GAN with cross-lingual transfer, with results of 34.5, 82.9%, 79.8%, and 28.7, respectively. That means that this hybrid approach successfully boosts both the language modeling and the translation capabilities even under the circumstances when there is a lack of data. As for qualitative analysis, the model proposed by researchers outperformed the baseline one in terms of fluency and semantic accuracy. Thus, the recommendations for future research may include work on improving the quality and coherence of texts created in the process of synthesis in the conditions of low-resource languages. Future research efforts may also be directed toward studying the scalability of the method in dealing with a much wider range of linguistic structures using larger data sets.

Declaration

Author Contribution

Funding Statement: The authors declare that no financial support was received for this research.

Conflict of Interest: The authors declare that to have no conflict of interest.

Data Availability Statement: https://github.com/elexis-eu/MWSA

References

[1] Alapati, P. R., Lawrance, J. C., Sambath, P., Murugan, R., Rengarajan, M., Raj, I. I., & Bala, B. K. (2024, April). Cross-Lingual Transfer Learning in NLP: Enhancing English Language Learning for Non-Native Speakers. In *2024 10th International Conference on Communication and Signal Processing (ICCSP)* (pp. 1042-1047). IEEE. https://doi.org/10.1109/ICCSP60870.2024.10544031

[2] Anwar, A., Li, X., Yang, Y., & Wang, Y. (2019). Constructing Uyghur Commonsense Knowledge Base by Knowledge Projection. *Applied Sciences*, *9*(16), 3318. https://doi.org/10.3390/app9163318

[3] Bhowmik, K., & Ralescu, A. (2021). Leveraging vector space similarity for learning cross-lingual word embeddings: A systematic review. *Digital*, *1*(3), 145-161. https://doi.org/10.3390/digital1030011

[4] Byambadorj, Z., Nishimura, R., Ayush, A., Ohta, K., & Kitaoka, N. (2021). Text-to-speech system for low-resource language using cross-lingual transfer learning and data augmentation. *EURASIP Journal on Audio, Speech, and Music Processing*, *2021*(1), 42. https://doi.org/10.1186/s13636-021-00225-4

[5] Feng, X., Feng, X., Qin, B., Feng, Z., & Liu, T. (2018, July). Improving Low Resource Named Entity Recognition using Cross-lingual Knowledge Transfer. In *IJCAI* (Vol. 1, pp. 4071-4077). https://doi.org/10.24963/ijcai.2018/566

[6] Hu, P., Liu, S., Gao, C., Huang, X., Han, X., Feng, J., ... & Huang, S. (2025, April). Large language models are cross-lingual knowledge-free reasoners. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)* (pp. 1525-1542). https://doi.org/10.18653/v1/2025.naacl-long.72

[7] Kim, B., Kim, J., Ko, Y., & Seo, J. (2021, May). Commonsense knowledge augmentation for low-resource languages via adversarial learning. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 35, No. 7, pp. 6393-6401). https://doi.org/10.1609/aaai.v35i7.16793

[8] Kong, H. J., Kim, J. Y., Moon, H. M., Park, H. C., Kim, J. W., Lim, R., ... & Kim, S. (2022). Automation of generative adversarial network-based synthetic data-augmentation for maximizing the diagnostic performance with paranasal imaging. *Scientific Reports*, *12*(1), 18118. https://doi.org/10.1038/s41598-022-22222-z

[9] Liu, L., Ding, B., Bing, L., Joty, S., Si, L., & Miao, C. (2021, August). MulDA: A multilingual data augmentation framework for low-resource cross-lingual NER. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)* (pp. 5834-5846). https://doi.org/10.18653/v1/2021.acl-long.453

[10] Lupión, M., Cruciani, F., Cleland, I., Nugent, C., & Ortigosa, P. M. (2024). Data augmentation for human activity recognition with generative adversarial networks. *IEEE Journal of Biomedical and Health Informatics*, *28*(4), 2350-2361. https://doi.org/10.1109/JBHI.2024.3364910

[11] Naaz, F., Herle, A., Channegowda, J., Raj, A., & Lakshminarayanan, M. (2021). A generative adversarial network-based synthetic data augmentation technique for battery condition evaluation. *International Journal of Energy Research*, *45*(13), 19120-19135. https://doi.org/10.1002/er.7013

[12] Park, G., Park, J., & Lee, H. (2025). Cross-Lingual Summarization for Low-Resource Languages Using Multilingual Retrieval-Based In-Context Learning. *Applied Sciences*, *15*(14), 7800. https://doi.org/10.3390/app15147800

[13] Shao, S., Wang, P., & Yan, R. (2019). Generative adversarial networks for data augmentation in machine fault diagnosis. *Computers in Industry*, *106*, 85-93. https://doi.org/10.1016/j.compind.2019.01.001

[14] Shi, H., Wang, L., Ding, G., Yang, F., & Li, X. (2018, August). Data augmentation with improved generative adversarial networks. In *2018 24th international conference on pattern recognition (ICPR)* (pp. 73-78). IEEE. https://doi.org/10.1109/ICPR.2018.8545894

[15] Tran, K. T., O'Sullivan, B., & Nguyen, H. D. (2026, March). Reasoning transfer for an extremely low-resource and endangered language: Bridging languages through sample-efficient language understanding. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 40, No. 39, pp. 33277-33286). https://doi.org/10.1609/aaai. v40i39.40613

[16] Wang, C., Li, J., Chen, Y., Liu, K., & Zhao, J. (2022, December). Cn-automic: Distilling chinese commonsense knowledge from pretrained language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing* (pp. 9253-9265). https://doi.org/10.18653/v1/2022.emnlp-main.628

[17] Wang, Y., Lin, M., Hu, Q., Bai, S., Li, Y., & Bao, L. (2025). A domain-specific cross-lingual semantic alignment learning model for low-resource languages. *Neural Networks*, 108114. https://doi.org/10.1016/j.neunet.2025.108114

[18] Wang, Y., Lin, M., Hu, Q., Bao, L., Bai, S., & Li, Y. (2025). Large and Small models for collaborative cross-lingual data augmentation in entity relationship extraction for low-resource languages. *Journal of King Saud University Computer and Information Sciences*, *37*(4), 56. https://doi.org/10.1007/s44443-025-00055-w

[19] Xu, P., & Fung, P. (2013). Cross-lingual language modeling for low-resource speech recognition. *IEEE transactions on audio, speech, and language processing*, *21*(6), 1134-1144. https://doi.org/10.1109/TASL.2013.2244088

[20] Zhou, W., Liu, F., Vulić, I., Collier, N., & Chen, M. (2022, May). Prix-LM: Pretraining for multilingual knowledge base construction. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 5412-5424). https://doi.org/10.18653/v1/2022.acl-long.371

Authors Biography

Dr.E.M. Roopa Devi is received the B. E (Electrical and Electronics) degree from Anna University and also M.E (Computer Science and Engineering) degree from Anna university and PhD (Information and Commuincation Engineering) from Anna university. Her research interest is in the area of Network Security and Data Analytics. Currently she is working as an Associate Professor in Kongu Engineering College. She took up academic pursuit and has 16 years' experience on Academic and research. Published three books on Artificial Intelligence, Block chain Technologies, Agile Methodlogies and Machine Learning. She has 20 published papers in various international journals and conferences.

S. Vinothkumar obtained his Bachelor's degree in Electronics and Communication Engineering from Institute of Road and Transport Technology, Erode and Master's Degree in Computer and Communication Engineering from Kongu Engineering College, Perundurai. He is doing Ph. D in Internet of Things under Anna University, Chennai. He has 14 years of teaching experience. He is currently working as an Assistant Professor (Senior Grade) in the department of Information Technology, Kongu Engineering College, Erode. He Received "Best Faculty" award for the year 2021 – 2022 also received "Best Placement Coordinator" award for the year 2024-2025. He conducted 1 sponsored seminar and completed 16 consultancy activities. He published more than 28 research articles in reputed International Journals and Completed AICTE sponsored PG certification course on Cyber Security and Digital Forensics at IIIT Kottayam.

Dr.B. Vinodhini, graduated from Bharathidasan University in the Department of Computer Science and Engineering. She received her Master's degree in Computer Science and Engineering from Avinashilingam University and obtained her Ph.D. degree from Anna University in Information Communication. She has 18 years of teaching experience and is currently working as an Associate Professor in the Department of Computer Science and Engineering at SNS College of Technology, affiliated with Anna University. Her research interests include Data Science, Machine Learning, and Wireless Communications. She has filed 7 patents and published research papers in reputed Conferences and Journals.