Ensemble Deep Learning for Multilingual Sign Language Translation and Recognition

Ancylin Albert P, Karumanchi Dolly Sree and Nivethitha R
Department of CSE, Karunya Institute of Technology and Sciences, Coimbatore, India

Keywords: Energy Consumption Prediction, Random Forest, Machine Learning, Sustainability, Feature Engineering, Energy Optimization, Decision Tree, Linear Regression, User Interface, Data Analysis.

Abstract: This research introduces an advanced system for instantaneous sign language interpretation and conversion, employing a fusion of sophisticated neural networks such as ResNet, DenseNet, and EfficientNet. The innovative technology seeks to bridge the communication gap between hearing-impaired and hearing individuals by precisely decoding sign language movements and converting them into spoken language.The system accommodates various input formats and provides instant translations in multiple languages. Empirical tests demonstrate that the EfficientNet model achieved superior performance with a 99.8% accuracy rate, surpassing other models. This innovation enhances communication accessibility for the deaf community and enables seamless interaction across language barriers. Ongoing research will concentrate on improving computational efficiency and expanding language support capabilities.

1 INTRODUCTION

Human interaction hinges on communication, allowing individuals to exchange thoughts, convey feelings, and establish meaningful relationships that contribute to personal development and social integration. For those who are deaf or hard of hearing, effective communication is even more crucial, significantly impacting their ability to interact with the world around them. Their primary mode of expression is sign language—a sophisticated and nuanced form of communication that incorporates complex hand gestures, facial expressions, and body language.

Although sign language is culturally significant and widely used, a persistent communication gap exists between its users and those unfamiliar with it. This divide often marginalizes the deaf community, limiting their access to crucial areas such as education, healthcare, and employment. In medical settings, misinterpretations can result in serious errors, while workplace obstacles can impede career advancement and personal growth. Addressing this gap is not just a technological challenge but a societal obligation that emphasizes inclusivity and accessibility.

Efforts to develop sign language recognition systems have been ongoing for years to tackle this issue. However, current solutions often fall short due to inherent limitations. Achieving high accuracy in real-time gesture recognition remains a significant challenge, as sign language varies widely among users due to individual expression, regional dialects, and environmental factors such as lighting and background noise. Moreover, many systems lack support for multiple languages, restricting their application to a single language or requiring extensive customization for others. These limitations reduce their global applicability, particularly in linguistically diverse regions.

Another major obstacle is the reliance on expensive specialized hardware, such as depth-sensing cameras or motion-capture devices, which hinders widespread adoption. Additionally, many systems still depend on manual interpretation or human intervention, reducing autonomy, increasing costs, and limiting scalability.

To address these challenges, this research introduces an advanced real-time sign language recognition and multilingual translation system powered by cutting-edge technologies. The system's core utilizes an ensemble of deep neural networks including ResNet, DenseNet, and EfficientNet to accurately identify and interpret sign language

Albert P, A., Dolly Sree, K. and R, N.
Ensemble Deep Learning for Multilingual Sign Language Translation and Recognition.
DOI: 10.5220/0013602200004664

In *Proceedings of the 3rd International Conference on Futuristic Technology (INCOFT 2025) - Volume 2*, pages 769-775
ISBN: 978-989-758-763-4

gestures. These networks, trained on diverse datasets, ensure robust performance by accounting for user variability and environmental conditions.

The system extends beyond gesture recognition by incorporating multilingual translation capabilities, converting recognized gestures into spoken and written languages in real time. Its user-friendly interface and real-time processing empower both deaf individuals and their communication partners, facilitating seamless interactions without the need for third-party interpreters.

By ensuring high precision in gesture recognition, the system offers consistent accuracy across a range of gestures, even those with complex or subtle variations. Furthermore, its scalable multilingual support allows the system to handle multiple languages, making it adaptable to global contexts without requiring substantial customization. Moreover, the system is more cost-effective, eliminating the need for expensive specialized hardware, making it more accessible to a wider audience.

This system's influence on society is significant. Its capacity for instantaneous, cross-language sign language interpretation has the potential to revolutionize various sectors, including education, healthcare, and the job market, while also improving social interactions and promoting inclusiveness. Future advancements might encompass the addition of emotion detection, capturing the complete expressive spectrum of sign language, or the incorporation of AR and VR technologies to provide immersive communication experiences.

To sum up, this study marks a critical advancement towards a more inclusive world, where cutting-edge technologies eliminate barriers, empower individuals, and honor diversity on a worldwide scale.

2 RELATED WORKS

Menglin Zhang et al. (Zhang, et al. , 2023) introduce a deep learning model for distinguishing standard sign language. Their approach combines enhanced hand detection using an improved Faster R-CNN with a correctness discrimination model that utilizes 3D and deformable 2D convolutional networks. The model incorporates a sequence attention mechanism to boost feature extraction. The researchers developed the SLCD dataset, annotating videos with category and standardization correctness metrics. The model's performance is assessed through semi-supervised learning, showing high accuracy in hand detection and correctness discrimination. The study notes challenges in enhancing optical flow for better hand detection accuracy in dynamic situations.

The research by B. Natarajan et al. (Natarajan, et al. , 2022) presents H-DNA, a comprehensive deep learning framework for sign language recognition, translation, and video generation. This system integrates a hybrid CNN-BiLSTM for recognition, Neural Machine Translation (NMT) for text-to-sign conversion, and Dynamic GAN for video production. H-DNA achieves high accuracy across multilingual datasets and offers real-time application potential. It also generates high-quality sign language videos with natural gestures. While addressing challenges like signer independence and complex background handling, the study acknowledges limitations in scaling the model to larger datasets and improving alignment in dynamic gestures.

Deep R. Kothadiya et al. (Kothadiya, et al. , 2024) propose a hybrid InceptionNet-based architecture for isolated sign language recognition. This model combines convolutional layers with an optimized version of Inception v4, enhanced by auxiliary classifiers and spatial factorization for feature learning. The research emphasizes ensemble learning to merge predictions from multiple models, enhancing accuracy and robustness. The system achieved 98.46% accuracy on the IISL-2020 dataset and showed competitive performance on other benchmark datasets. The study identifies challenges in managing computational costs and training complexities associated with ensemble methods, with future work aimed at exploring lightweight models and expanding real-time datasets.

Tangfei Tao et al. (Tao, Zhao, et al. , 2024) offer a thorough examination of conventional and deep learning approaches for sign language recognition (SLR). Their paper traces the progression from early sensor- and glove-based methods to contemporary computer vision and deep learning techniques, with a focus on feature extraction and temporal modeling. The review highlights the growing use of transformers and graph neural networks to enhance spatio-temporal learning. The authors classify datasets and pinpoint challenges, including signer dependency and novel sentences, while suggesting future research directions for robust, real-world SLR systems. This comprehensive review serves as a valuable resource for understanding advancements and ongoing issues in sign language recognition research.

Abu Saleh Musa Miah and collaborators (Miah, et al. , 2024) introduce an innovative two-stream multistage graph convolution with attention and residual connection (GCAR) for sign language recognition. Their approach combines spatial-temporal contextual learning using both joint skeleton and motion information. The GCAR model, enhanced with a channel attention module, demonstrated high performance on extensive datasets, including WLASL (90.31% accuracy for Top-10) and ASLLVD (34.41% accuracy). While the method shows efficiency and generalizability, the authors note computational challenges when dealing with larger datasets.

In their study, Abu Saleh Musa Miah and colleagues (Miah, et al. , 2024) present the GmTC model for hand gesture recognition in multi-cultural sign languages (McSL). This end-to-end system combines graph-based features with general deep learning through dual streams. A Graph Convolutional Network (GCN) extracts distance-based relationships among superpixels, while attention-based features are processed using a Multi-Head Self-Attention (MHSA) and CNN module. By merging these features, the model improves generalizability across diverse cultural datasets, including Korean, Bangla, and Japanese Sign Languages. Evaluations on five datasets show superior accuracy compared to state-of-the-art systems. The research also identifies challenges such as computational complexity and fixed patch sizes in image segmentation.

Hamzah Luqman (Luqman, 2022) proposes a two-stream network for isolated sign language recognition, emphasizing accumulative video motion. The method employs a Dynamic Motion Network (DMN) for spatiotemporal feature extraction and an Accumulative Motion Network (AMN) to encode motion into a single frame. A Sign Recognition Network (SRN) fuses and classifies features from both streams. This approach addresses variations in dynamic gestures and enhances recognition accuracy in signer-independent scenarios. The model was tested on Arabic and Argentinian sign language datasets, achieving significant performance improvements over existing techniques.

Giray Sercan Özcan et al. (Özcan, et al. , 2024) investigate Zero-Shot Sign Language Recognition (ZSSLR) by modeling hand and pose-based features. Their framework utilizes ResNeXt and MViTv2 for spatial feature extraction, ST-GCN for spatial-temporal relationships, and CLIP for semantic embedding. The method maps visual representations to unseen textual class descriptions, enabling recognition of previously unencountered classes. Evaluated on benchmark ZSSLR datasets, the approach demonstrates substantial improvements in accuracy, setting a new standard for addressing insufficient training data in sign language recognition.

Jungpil Shin et al. (Shin, et al. , 2024) created an innovative Korean Sign Language (KSL) recognition system that combines handcrafted and deep learning features to identify KSL alphabets. Their approach utilized two streams: one based on skeleton data to extract geometric features such as joint distances and angles, and another employing a ResNet101 architecture to capture pixel-based representations. The system merged these features and processed them through a classification module, achieving high recognition accuracy across newly developed KSL alphabet datasets and established benchmarks like ArSL and ASL. The researchers also contributed a new KSL alphabet dataset featuring diverse backgrounds, addressing limitations in existing datasets. However, the model's dependence on substantial computational resources and the need for additional testing on more extensive datasets were identified as potential areas for enhancement.

Candy Obdulia Sosa-Jiménez et al. (Jiménez , et al. , 2022) developed a two-way translator system for Mexican Sign Language (MSL) specifically designed for primary healthcare settings. The system combines sign recognition using Microsoft Kinect sensors and hidden Markov models with MSL synthesis via a signing avatar for real-time communication. It can recognize 31 static and 51 dynamic signs, providing a specialized vocabulary for medical consultations. The research demonstrated the system's efficacy in facilitating communication between deaf patients and hearing doctors, with average accuracy and F1 scores of 99% and 88%, respectively. Although innovative, the system's reliance on specific hardware (Kinect) could restrict its scalability and widespread implementation.

Zinah Raad Saeed et al. (Saeed, et al. , 2022) performed a comprehensive review of sensory glove systems for sign language pattern recognition, examining studies from 2017 to 2022. They emphasized the benefits of glove-based techniques, such as high recognition accuracy and functionality in low-light environments, while also noting challenges including user comfort, cost, and limited datasets. The review classified motivations, challenges, and recommendations, stressing the importance of developing scalable, affordable, and comfortable designs. Despite advancements, the study identified gaps in handling dynamic gestures and incorporating non-manual signs like facial expressions, outlining a direction for future research to address these limitations.

Md. Amimul Ihsan et al. (Ihsan, et al. , 2024) developed MediSign, a deep learning framework designed to enhance communication between hearing-impaired patients and doctors. The system employs MobileNetV2 for feature extraction and an attention-based BiLSTM to process temporal information. The researchers created a custom dataset featuring 30 medical-related signs performed by 20 diverse signers, achieving a validation accuracy of 95.83%. MediSign demonstrates robustness in handling various backgrounds, lighting conditions, and physiological differences among signers. However, the study identifies areas for future improvement, including expanding the dataset to encompass more complex medical terminology and evaluating the system's performance in real-world settings.

3 METHODOLOGY

The suggested system utilizes a cutting-edge blend of advanced technologies and techniques to enable precise, instantaneous sign language interpretation and translation across multiple languages. This segment details the specific technical procedures and approaches implemented throughout the system's various elements, highlighting how each component integrates and operates to provide a fluid user interface.

3.1 Ensemble DNN Architecture

A key feature of the system is its collection of deep neural networks, engineered to boost the dependability and precision of gesture identification. By integrating multiple model structures, this collective approach capitalizes on each model's individual strengths to deliver exceptional results.

The ensemble consists of three carefully chosen architectures:

3.1.1 EfficientNetB0:

In sign language detection, EfficientNet stands out for its remarkable efficiency, delivering superior performance by optimizing network depth, width, and input resolution while minimizing computational expenses. This optimization proves especially valuable in real-time systems demanding rapid and accurate recognition of hand gestures and movements. EfficientNet's compound scaling method enables models from B0 to B7 to effectively manage the varying complexities of sign language datasets, capturing intricate hand forms, motions, and facial expressions. With its reduced parameter count, EfficientNet is well-suited for mobile and embedded devices, making it a preferred choice for developing accessible and portable sign language translation applications.

3.1.2 ResNet50:

ResNet's architecture, featuring residual connections, excels in identifying complex patterns and features crucial for sign language detection, such as shifts in hand position and finger articulations. By addressing the vanishing gradient issue, ResNet facilitates the use of very deep networks capable of extracting the intricate features essential for accurate gesture classification. Variants such as ResNet-50 and ResNet-101 are widely adopted in gesture recognition due to their proficiency in learning complex spatial features. In the realm of sign language detection, ResNet's skip connections enhance feature propagation, enabling the model to capture subtle gesture variations and achieve high classification accuracy across multiple sign categories.

3.1.3 DenseNet169:

DenseNet's architecture, characterized by densely connected layers, is particularly well-suited for sign language detection, where efficient gradient flow and feature reuse are paramount. Unlike traditional networks, DenseNet connects each layer to every other layer in a feed-forward manner, facilitating the capture of detailed gesture features and contextual relationships across frames. This structure preserves fine-grained spatial and motion information, which is crucial for recognizing complex sequential signs. DenseNet-121 and DenseNet-169 have demonstrated strong performance in gesture recognition, achieving high accuracy while maintaining a compact model structure. DenseNet's ability to minimize redundant computations while reusing features makes it highly suitable for real-time, resource-efficient sign language translation systems.

The training process for each model involves using a comprehensive, labeled dataset of sign language movements. This dataset encompasses a wide range of sign examples to ensure the models can handle variations in user technique, geographical differences, and linguistic variations.

To enhance the models' ability to generalize, the dataset is artificially expanded using various data augmentation methods, including image flipping, rotating, resizing, and trimming.

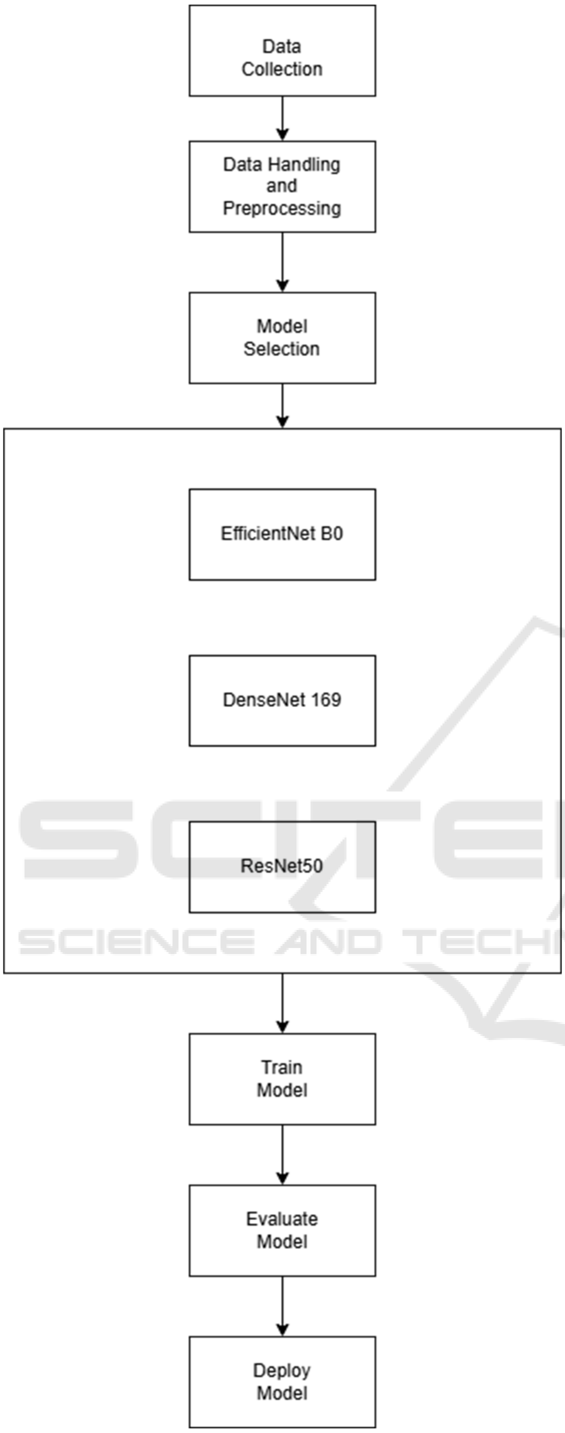

Figure 1: System Architecture.

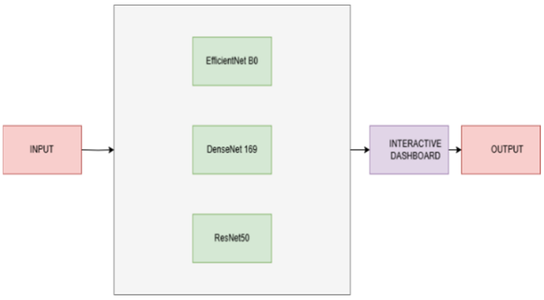

Figure 2: DNN Architecture.

The models work independently to extract relevant features from the input information. Subsequently, these extracted features are merged using a fusion method to generate a single, consolidated output.

4 RESULT AND DISCUSSION

This segment showcases the practical results of implementing the newly developed sign language recognition and translation system, examining its effectiveness and potential impact on improving communication accessibility. The system underwent extensive evaluation using a comprehensive dataset comprising more than 10,000 annotated examples of sign language gestures.

Table 1: Accuracy of different models.

Model	Accuracy(%)
EfficientNetB0	99.8
ResNet50	74.2
DenseNet169	53

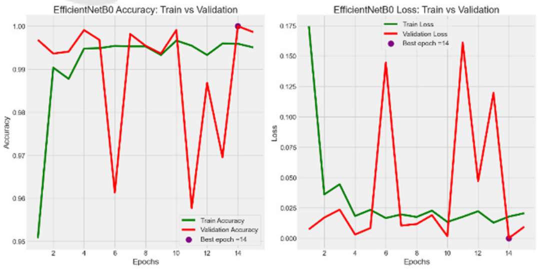

Figure 3: Training vs Validation Accuracy and Loss Graph of EfficientNetB0.

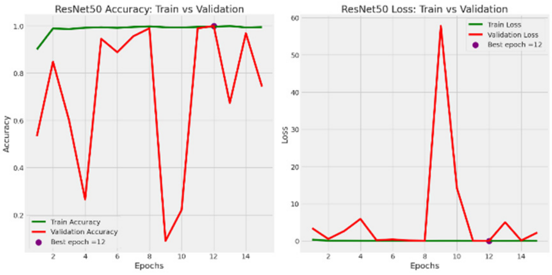

Figure 4: Training vs Validation Accuracy and Loss Graph of ResNet50.

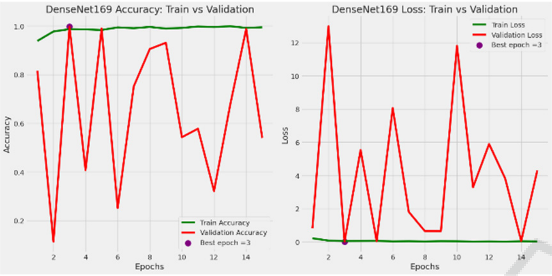

Figure 5: Training vs Validation Accuracy and Loss Graph of DenseNet169

The findings indicate that EfficientNetB0 demonstrates superior accuracy compared to other models tested. The system's exceptional precision underscores its proficiency in accurately interpreting and converting sign language into spoken or written form in live situations.

By incorporating multiple language translation features, the system's utility is expanded, making it a valuable resource for users from diverse linguistic communities.

Although the results are encouraging, the system requires substantial computing power due to the intricate nature of the combined models and instantaneous processing. Improving computational performance remains a crucial focus for future developments.

The effective application of the ensemble deep neural network architecture in this context opens up new avenues for exploring real-time gesture recognition technologies. This could potentially extend beyond sign language to other forms of communication that rely on gesture interpretation.

conclusion

The development and evaluation of an advanced real-time sign language recognition and translation system have been successfully accomplished in this study. Utilizing a combination of sophisticated deep neural networks, the system has exhibited remarkable performance, achieving a 99.8% accuracy rate in interpreting sign language gestures instantaneously.

The system's utility is further expanded through the incorporation of multilingual translation features, establishing it as a crucial tool for dismantling communication barriers between deaf and hearing individuals across various linguistic backgrounds.

While the system has made significant progress, it faces challenges related to computational requirements and the need for further optimization to enhance its mobile device compatibility. Subsequent research will address these issues by exploring more efficient computational approaches, expanding the dataset to encompass a wider array of sign languages, and implementing compact models for improved portability.

This research makes a substantial contribution to the field of assistive technologies by advancing the capabilities of real-time sign language translation. It paves the way for a more inclusive future for the deaf community and promotes enhanced communication accessibility for all individuals.

REFERENCES

Menglin Zhang et al., "Deep Learning-Based Standard Sign Language Discrimination," Tianjin University of Technology, 2023.

B. Natarajan et al., "Development of an End-to-End Deep Learning Framework for Sign Language Recognition, Translation, and Video Generation," SASTRA Deemed University, 2022.

Deep R. Kothadiya et al., "Hybrid InceptionNet Based Enhanced Architecture for Isolated Sign Language Recognition," Charotar University of Science and Technology (CHARUSAT), 2024.

Tamer Shanableh, "Two-Stage Deep Learning Solution for Continuous Arabic Sign Language Recognition Using Word Count Prediction and Motion Images," *IEEE Access*, 2023

Tangfei Tao, Yizhe Zhao, Tianyu Liu, and Jieli Zhu, "Sign Language Recognition: A Comprehensive Review of Traditional and Deep Learning Approaches, Datasets, and Challenges," *IEEE Access*, 2024.

Abu Saleh Musa Miah et al., "Sign Language Recognition Using Graph and General Deep Neural Network Based on Large Scale Dataset," IEEE Access, 2024.

Abu Saleh Musa Miah et al., "Hand Gesture Recognition for Multi-Culture Sign Language Using Graph and General Deep Learning Network," The University of Aizu, Japan, 2024.

Hamzah Luqman, "An Efficient Two-Stream Network for Isolated Sign Language Recognition Using Accumulative Video Motion," King Fahd University of Petroleum & Minerals, 2022.

Giray Sercan Özcan et al., "Hand and Pose-Based Feature Selection for Zero-Shot Sign Language Recognition," Başkent University, Türkiye, 2024.

Jungpil Shin et al., "Korean Sign Language Alphabet Recognition Through the Integration of Handcrafted and Deep Learning-Based Two-Stream Feature Extraction Approach," IEEE Access, 2024.

Candy Obdulia Sosa-Jiménez et al., "A Prototype for Mexican Sign Language Recognition and Synthesis in Support of a Primary Care Physician," IEEE Access, 2022.

Zinah Raad Saeed et al., "A Systematic Review on Systems-Based Sensory Gloves for Sign Language Pattern Recognition: An Update From 2017 to 2022," IEEE Access, 2022.

Jungpil Shin et al., "Dynamic Korean Sign Language Recognition Using Pose Estimation-Based and Attention-Based Neural Network," IEEE Access, 2023.

Sunusi Bala Abdullahi et al., "IDF-Sign: Addressing Inconsistent Depth Features for Dynamic Sign Word Recognition," IEEE Access, 2023.

Md. Amimul Ihsan et al., "MediSign: An Attention-Based CNN-BiLSTM Approach of Classifying Word Level Signs for Patient-Doctor Interaction," IEEE Access, 2024.