

Sign Language Recognition Using Residual Network Architectures for Alphabet And Digraph Classification

^{1*}Martins E. Irhebhude, ²Adeola O. Kolawole and ^{2,3}Wali M. Zubair

^{1,2}Department of Computer Science, Faculty of Military Science and Interdisciplinary Studies, Nigerian Defence Academy, Kaduna, Nigeria

³Department of Computer Science, College of Science Technology, Kaduna Polytechnic, Kaduna, Nigeria

email: ^{1*}mirhebhude@nda.edu.ng, ²adeolakolawole@nda.edu.ng, ³zubairwm@kadunapolytechnic.edu.ng

*Corresponding author

Received: 08 October 2024 | Accepted: 11 December 2024 | Early access: 19 December 2024

Abstract - Communication is crucial in human life, enabling the exchange of information through various methods beyond spoken language. Sign language translation is crucial for bridging communication gaps between hearing-impaired and hearing individuals, promoting effective interaction and understanding. This study presents a comprehensive model for identifying alphabet and digraph signs using feature extraction techniques from ResNet architectures, specifically ResNet18, ResNet50, and ResNet101. The system was designed to integrate both hand gestures and facial expressions, enhancing the accuracy of sign language recognition. Classification of sign language images into alphabet and digraph categories was assessed using Support Vector Machine (SVM). The resulting classification accuracies were 61.7% for ResNet18, 64.5% for ResNet50, and 66.5% for ResNet101. The research results emphasize how deeper ResNet models are effective in improving recognition accuracy. This proposed model has significant implications for educational applications as it addresses attention-related challenges and aims to enhance student engagement in learning processes, thereby contributing to developing more inclusive educational environments.

Keywords: Alphabet Sign, Digraph Sign, ResNet, Hearing-Impaired, Sign Language Recognition, Support Vector Machine.

1 Introduction

Hand gesture recognition is used to create systems for exchanging information among individuals with disabilities or controlling devices. (Sahoo et al., 2021). According to Al-Hammadi et al. (2020), a significant application of hand gesture recognition is to facilitate the translation of sign language. A gesture is a physical movement involving the hands, arms, face, or body, aimed at conveying information or meaning (Irhebhude et al., 2023). Gesture recognition involves tracking human movements and interpreting them as meaningful commands with semantic significance. In the field of Human-Computer Interaction (HCI), two primary methodologies are used to interpret gestures: data glove techniques and vision-based methods (Ma et al., 2021).

In the data glove method, hand gesture data is collected using motion sensors, gloves, and trackers (Côté Allard et al., 2019). However, this approach is costly due to the need for hardware and can be cumbersome as it restricts the movement of signers. The vision-based method uses cameras and imaging sensors to gather data, unlike the other method (Haria et al., 2017). HCI is essential in information technology, involving computer-human interaction. Hand gestures are a nonverbal and innate way to interact with computers. Recognizing hand gestures is crucial in HCI, especially for those who are hearing impaired (Haria et al., 2017).

Hearing-impaired individuals use sign languages, which are visual-based natural languages, for communication. Given that most hearing individuals do not understand sign language, sign language translation (SLT) has become essential for facilitating communication between these two groups (Gupta & Singh, 2024). In recent years,

researchers have increasingly explored deep learning models for neural SLT to address this communication challenge.

Effective communication is essential for individuals to collaborate, express emotions and ideas, and contribute to societal advancement. Individuals with hearing impairments naturally develop sign language as a means of communication tailored to their needs (Sharma & Singh, 2020). Sign language, as a primal and innate form of communication, predates the early stages of human evolution. The development of sign language aligns with early historical theories, emerging even prior to spoken languages. Throughout history, sign language has evolved and seamlessly integrated into everyday communication, becoming an essential component of human interaction (Olabanji & Ponnle, 2021).

Sign language recognition systems benefit from an understanding of the hearing-impaired culture. Hearing-impaired culture encompasses the shared experiences, values, and traditions of a community where sign language serves as the primary mode of communication (Wen et al., 2021). This cultural awareness can lead to more accurate, reliable, and inclusive speech recognition (SLR) systems that promote accessibility and empowerment for individuals and communities with hearing impairment. The Nigerian hearing-impaired community was first introduced to American Sign Language (ASL) in the 1960s. However, according to Asonye et al. (2018), the indigenous Nigerian Sign Language (NSL) has historically been marginalized and misrepresented. NSL, which developed organically within the Nigerian hearing-impaired community, remains the primary means of communication for many Nigerians with speech and hearing difficulties.

Researchers have documented various regional varieties of NSL. For instance, Morgan (2002) studied Hausa Sign Language, utilized by hearing-impaired communities in northern Nigeria. Bura Sign Language, another regional sign language, has also been identified and analyzed in academic literature (Blench et al., 2006). The regional sign languages have a similar linguistic basis but show different vocabulary and grammar based on local spoken languages and cultural contexts. The preference for ASL over NSL has restricted the ability of hearing-impaired Nigerians to receive education, employment, and social services in their own language. Efforts to promote and preserve NSL as a vital component of the culture and identity of hearing-impaired Nigerians are crucial for ensuring their linguistic rights and inclusion in all aspects of society (Asonye et al., 2018).

Sign language is a visual-gestural mode of communication used by the hearing-impaired community. However, it faces significant challenges in facilitating effective communication between signers and non-signers. Recent advancements in the fields of deep learning and computer vision have been explored as potential solutions to address these challenges. However, as identified by Bragg et al. (2019), current research often excludes input from hearing-impaired individuals, who have firsthand experience with the challenges of sign language recognition algorithms. Additionally, Bragg et al. (2019) noted that many of the datasets used to train sign language recognition algorithms do not accurately represent real-world scenarios. Irhebhude et al. (2023) in their work, extracted diagraph signs from a school for the hearing-impaired in Kaduna State, Nigeria, and developed diagraph sign language recognition using a Residual Network (ResNet18) as a feature extractor and Support Vector Machine (SVM) as the classifier. The study did not, however, include recognition of alphabet signs. To overcome these limitations, the paper suggests using a real-world locally obtained dataset of the 26 English alphabets and some selected 16 diagraph signs, and applying ResNet18, ResNet50, and ResNet101.

The remainder of the paper is as follows: section II describes the Residual network models. Related studies are reported in section III. The methodology is discussed in section IV. Section V discusses the experimental results. The summary of findings, conclusion, and recommendations, are presented in section VI.

2 Residual Network

Convolutional Neural Networks (CNNs) consist of a wide range of architectures, such as the Residual Network family, which show variations within a foundational framework, primarily based on the number of layers and the total parameters (Hasanah et al., 2023). Residual Network (ResNet), is a deep convolutional neural network architecture introduced by He et al. (2016) in their 2015 paper "Deep Residual Learning for Image Recognition" (He et al., 2016). ResNet tackles the issue of vanishing gradients in deep neural networks by introducing "skip connections." These connections enable layers to learn residual functions based on the layer inputs, rather than learning unreferenced functions (He et al., 2016). The main advantage of ResNet is that it simplifies the optimization of extremely deep convolutional neural networks (Hasanah et al., 2023). ResNet models are named based on the depth of the network, such as ResNet-50 or ResNet-101, with the number denoting the number of layers in the model. Each level within the ResNet framework serves a distinct purpose that aids in the machine learning process and helps in extracting crucial features for classification assignments. It is widely known that

increasing the number of layers in a CNN architecture promotes deeper learning, which often leads to improved performance. However, this advantage comes with the drawback of longer training times due to the significant increase in parameters associated with deeper architectures (Hasanah et al., 2023).

2.1 ResNet-18

ResNet-18 is a foundational model within the ResNet family, known for its simplicity. It comprises 18 layers, with 17 convolutional layers and one fully connected layer. The architecture makes use of residual blocks, each containing two convolutional layers with 3×3 kernels, separated by batch normalization and ReLU activations. These residual blocks are characterized by a shortcut connection that bypasses the convolutional layers, enabling the network to learn residual mappings (He et al., 2016).

2.2 ResNet-50

ResNet-50 introduces a more complex architecture with 50 layers. It includes the use of Bottleneck blocks, which consist of three convolutional layers: a 1×1 convolutional layer for dimensionality reduction, a 3×3 convolutional layer, and another 1×1 convolutional layer for dimensionality expansion. The Bottleneck blocks help in reducing the computational complexity while maintaining high performance (He et al., 2016).

2.3 ResNet-101

ResNet-101 is an extension of the ResNet family, featuring a deeper network with 101 layers. It employs the same Bottleneck block structure as ResNet-50 but with an increased number of layers. The expanded architecture enables ResNet-101 to capture more intricate patterns and features, but it also demands greater computational resources and memory (He et al., 2016).

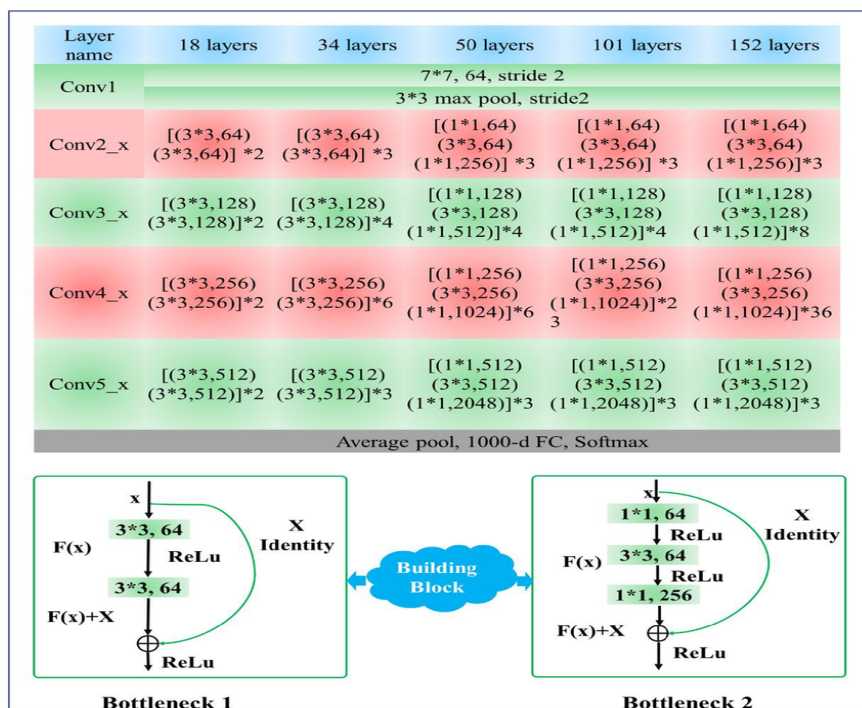


Figure 1: The structure of the ResNet: ResNet-18, ResNet-34, ResNet-50, ResNet-101, and ResNet-152 (Lin et al., 2022).

The ResNet architecture, shown in Figure 1, consists of two main building blocks: bottleneck 1 and bottleneck 2. ResNet-18 and ResNet-34 use bottleneck 1, while bottleneck 2 is used in ResNet-50, ResNet-101, and ResNet-152. The number in the model name, such as 18 or 152, indicates the total number of layers in the model (He et al., 2016). In comparison to bottleneck 1, bottleneck 2 incorporates three convolutional layers of sizes 1×1 , 3×3 , and 1×1 . The initial 1×1 convolutional layer functions to reduce the dimensionality of the input, making the 3×3 convolutional layer acts as a bottleneck with constrained input/output dimensions. The subsequent 1×1 convolutional layer then restores the dimensionality of the input to its original size (He et al., 2016).

The identity mapping in ResNet architectures is an important feature that helps mitigate the degradation problem. This issue arises when the training accuracy plateaus and then declines rapidly as the network becomes deeper. The identity mapping is achieved through skip connections, which allow the input of a layer to be added directly to its output, bypassing one or more layers. This allows the network to learn residual functions, which are easier to optimize than learning the original, unreferenced functions (He et al., 2016). The identity mapping in ResNet is defined as shown in equation 1:

$$Y = F(x, \llbracket wi \rrbracket) + X \quad (1)$$

where x is the input to the layer, $F(x, \llbracket wi \rrbracket)$ is the residual function learned by the layer, and Y is the output of the layer. The addition of X to $F(x, \llbracket wi \rrbracket)$ is the identity mapping, which ensures that the network can at least learn the identity mapping if the residual function is difficult to optimize.

ResNet18, ResNet50, and ResNet101 were chosen as pre-trained models and used as feature extracted to train an SVM classifier used for sign language recognition in this work.

3 Related Literature

This section discusses some of the related works on sign language recognition, the techniques, dataset, and results. Shi et al. (2022) proposed an innovative residual neural network architecture that integrates an enhanced residual module with a Bi-directional Long Short-Term Memory (BiLSTM) model to effectively classify 3D sign language gestures. The study evaluated this approach using challenging 3D sign language datasets from Chalearn and Sports-1M. The researchers devised a multi-path hybrid residual neural network architecture that combined improved residual modules for spatial feature extraction and BiLSTM for capturing temporal dynamics. The residual module incorporated motion excitation to enhance motion information captured and hierarchical-split blocks for extracting features across multiple scales. The hybrid neural network was trained end-to-end, assessed on benchmark datasets, and compared against state-of-the-art sign language recognition methods. Results showed that the model achieved a classification accuracy of 78.9% on the first dataset and 82.7% on the second dataset, surpassing existing algorithms and nearing human-level accuracy of 88.4%.

Sahoo et al. (2021) presented a hand gesture recognition system employing deep convolutional neural network (CNN) features integrated with machine learning techniques to develop a user-independent system capable of accurately recognizing hand gestures without requiring specific user training or calibration. The researchers' approach involved extracting deep CNN features from pre-trained models such as AlexNet and VGG-16, which were subsequently fed into a support vector machine (SVM) classifier. The authors investigated the utilization of CNN features from various layers independently and in combination to optimize hand gesture recognition accuracy. Evaluations conducted on a standardized American Sign Language (ASL) dataset demonstrated the system's effectiveness. Using features from the fully connected layers of AlexNet, the system achieved an accuracy of 92.8%. Combining features from multiple CNN layers further improved accuracy to 94.1%, surpassing benchmarks set by existing methodologies.

Jain et al. (2021) explored the use of SVM and CNN models to develop an effective system for recognizing ASL gestures. In the first phase, features from the dataset were extracted after applying various preprocessing techniques, using SVM with four different kernels i.e., 'poly', 'linear', 'rbf', and 'sigmoid'. CNN with single and double layers was applied to the training dataset to train the model. Experimental results showed that the hybrid system using SVM and CNN achieved an accuracy of 98.58% in recognizing ASL gestures. The optimal filter size was found to be 8×8 for both single and double-layer CNN.

Venugopalan & Reghunadhan (2023) aimed to create an advanced sign language recognition system that can assist in improving communication and accessibility for hearing-impaired COVID-19 patients, who faced additional challenges in expressing their needs and concerns during the pandemic. The videos of hand gestures for the most common Indian sign language (ISL) words used for urgent communication by COVID-19-positive hearing-impaired patients were included in the proposed dataset. The proposed SLR utilized a deep learning model designed as a combination of the VGG-16 and bidirectional long short-term memory (BiLSTM) sequence network. The classification of gestures was done with a hybrid model of VGG-16 and BiLSTM networks and achieved an average accuracy of 83.36%. The model was also tested on another ISL word dataset as well as the Cambridge hand gesture dataset to further assess its performance and achieved promising accuracies of 97% and 99.34% respectively.

Chowdhury et al. (2017) created a novel method for converting Bengali Sign Language into readable text using a combination of Artificial Neural Network (ANN) and SVM techniques. The researchers collected a dataset of Bengali sign language gestures and video data were preprocessed. Microsoft Kinect was used to take the input, which is the hand sign performed in front of the camera. The captured hand sign was eventually recognized, after joint and wrist detection and by assessing the contours. The contour feature was extracted and presented to the SVM for the classification of the sign. The contour finding algorithm utilized the convex hull method, and the features extracted after detection were passed through the SVM for recognition. To validate the performance of the proposed model, a dataset of both male and female hand gesture images was utilized. Experimental results demonstrated 84.11% classification accuracy.

Jiang & Zhu (2019) developed an effective method for identifying and recognizing Chinese Sign Language (CSL) using wavelet entropy and SVM techniques. The researchers collected a dataset of CSL gestures, the sign language gesture data was preprocessed, wavelet entropy was used for feature extraction of the CSL gestures and the extracted wavelet entropy features were used as input to an SVM classifier. The experiment was implemented on 10-fold cross-validation and it yielded an overall accuracy of $85.69 \pm 0.59\%$.

Sreemathy et al. (2023) introduced an effective system for continuous, word-level recognition of sign language through a blend of machine-learning approaches. The authors employed a proprietary image dataset comprising 80 static signs, encompassing a total of 676 images. The study proposed two distinct models: You Only Look Once version 4 (YOLOv4) and SVM integrated with MediaPipe. SVM utilized linear, polynomial, and Radial Basis Function (RBF) kernels. SVM with MediaPipe achieved an accuracy of 98.62%, while YOLOv4 achieved a higher accuracy of 98.8%, surpassing existing state-of-the-art methods in the field.

Al-Hammadi et al. (2020) developed an effective method for recognizing sign language gestures using deep learning techniques while focusing on efficient hand gesture representation. Two separate instances of 3D Convolutional Neural Networks (3DCNNs) were employed to capture distinct features: one focusing on detailed hand shapes and the other on broader body configurations. Multi-Layer Perceptron's (MLPs) and auto encoders were utilized to integrate and globalize these local features, followed by classification using the SoftMax function. Additionally, to mitigate the training expenses associated with the 3DCNN module, the researchers explored domain adaptation techniques and conducted extensive experiments to refine knowledge transfer efficiency. The efficacy of the proposed approach was evaluated on the King Saud University Saudi Sign Language (KSU-SSL) datasets. Experiments were carried out in two scenarios; signer dependent mode and signer-independent mode, with the MLP achieving accuracies of 98.62% and 87.69% respectively and the auto encoder achieving accuracies of 98.75% and 84.89%.

Kothadiya et al. (2022) introduced a deep-learning model designed to detect and interpret gestures as words. Specifically, Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) models, which are feedback-based learning architectures, were employed to recognize signs from individual frames of Indian Sign Language (ISL) videos. The proposed model explores four distinct sequential configurations combining LSTM and GRU layers, leveraging the custom dataset, IISL2020. Among these configurations, the model featuring a single LSTM layer followed by a GRU layer achieved 97% accuracy across 11 different signs.

Aksoy et al. (2021) employed deep learning and image processing techniques to detect Turkish Sign Language gestures. The authors curated a dataset consisting of 10,223 images covering 29 letters from the Turkish Sign Language alphabet. The images were enhanced using various image processing methods to optimize them for educational purposes. The study culminated in the classification of these images using a range of architectures including CapsNet, AlexNet, ResNet-50, DenseNet, VGG16, Xception, InceptionV3, NasNet, EfficientNet, HitNet, SqueezeNet, and a specially designed TSLNet for this research. Among these models, CapsNet and TSLNet demonstrated the highest accuracy rates, achieving 99.7% and 99.6%, respectively, in their classification performance.

Olabanji & Ponnle (2021) introduced a system for interpreting the native sign language of Nigeria. The methodology consists of three main stages: dataset generation, application of computer vision methodologies, and the development of a deep learning model. A multi-class CNN was specifically crafted to train and interpret the indigenous signs. The evaluation was conducted using a custom dataset containing 15,000 images of selected native words. The experimental results demonstrated a strong performance from the interpretation system, achieving an accuracy of 95.67%.

Liao et al. (2019) presented a multimodal dynamic sign language recognition method based on a deep 3-dimensional residual ConvNet and BiLSTM networks, which was named as BiLSTM-3D residual network (B3D ResNet). This approach comprised three primary components. Initially, it localized the hand object within video

frames to reduce the computational complexity of network operations. Subsequently, the B3D ResNet automatically extracted spatiotemporal features from video sequences, analyzing these features to establish intermediate scores corresponding to each action within the sequences. Finally, through video sequence classification, it accurately identified dynamic sign language expressions. Experimental validation was conducted on test datasets, including DEVISIGN_D and SLR_Dataset. Results demonstrated that the proposed approach achieved state-of-the-art recognition accuracy (89.8% on DEVISIGN_D and 86.9% on SLR_Dataset). Moreover, the B3D ResNet effectively recognized intricate hand gestures across extensive video sequences, achieving high accuracy in identifying 500 Chinese hand sign language vocabularies.

Gupta & Singh (2024) proposed an automated method for recognizing Indian sign language (ISL) gestures in English. Initially, the hand region was isolated using the Grasshopper optimization algorithm (GOA) based on a skin color model. The effectiveness of segmentation was evaluated using three approaches: GOA-based skin color detection algorithm (SCDA), particle swarm optimization-based SCDA (PSO-SCDA), and artificial bee colony-based SCDA (ABC-SCDA). Subsequently, a database was established containing gestures representing individual English alphabets. For gesture recognition, the system was trained using a template-based matching approach. The classification was performed using both SVM and convolutional neural network (CNN) techniques. The proposed recognition method achieved the highest accuracy of 97.85% with GOA-SCDA, compared to 89.29% and 93.96% with PSO-SCDA and ABC-SCDA, respectively. Additionally, CNN surpassed SVM in classification performance, achieving an accuracy of 99.2% and a precision of 81.8%.

The study in AkanshaTyagi (2022) developed an extraction technique that used the Fast Accelerated Segment Test (FAST), and Scale-Invariant Feature Transformation (SIFT). FAST and SIFT were hybridized and used to detect and extract features with CNN reserved classification. Results obtained showed that excellent recognition accuracies on four different datasets.

Irhebhude et al. (2023) devised a diagraph sign language recognition system by employing a ResNet18 as a feature extractor and SVM as the classifier. The researchers utilized a proprietary dataset comprising 796 images depicting students expressing 16 diagraph sounds, which include two and three-letter words conveyed through sign language. The system achieved an accuracy of 79.3%.

Chao et al. (2019) introduced a behavior recognition approach aimed at addressing sign language recognition challenges. Drawing inspiration from Multi-Fiber Networks, the authors proposed a Convolutional Block Attention Module CBAM-ResNet neural network that extends the ResNet architecture using 3D convolutions and integrates a convolutional block attention module. To enhance channel information fusion, the fifth layer incorporated the 3D-ResNet structure, leveraging the strengths of Multi-Fiber Networks while compensating for their limitations. The proposed method was compared with models incorporating convolutional block attention modules, Convolutional Long Short-Term Memory (ConvLSTM), optical flow, and other methodologies. Achieving an accuracy of 83.3% on the Chinese Sign Language Recognition Dataset without optical flow. The proposed approach demonstrated a performance improvement of approximately 9% over Multi-Fiber Networks.

As such, the main contribution of this work is to present a new alphabet and diagraph sign language recognition system by employing Residual Network and SVM. A previous work that devised a diagraph sign language recognition system by employing a Residual Network (Irhebhude et al., 2023) has already shown that the proposed model could be implemented within educational curricula to address issues stemming from attention deficits and enhance students' engagement in learning activities. This paper expands upon the aforementioned conference paper incorporating the 26 English alphabets into diagraph signs for better classification and improved performance. Additionally, three Residual network architectures (ResNet-18, ResNet-50, and ResNet-101) specifically designed for the robust and efficient classification of sign language are presented.

4 Proposed Methodology

This study employed ResNet for the feature extraction and SVM for the classification. Irhebhude et al. (2023) employed the ResNet architecture for feature extraction from the image dataset and classified the extracted features using SVM. The proposed model by Irhebhude et al. (2023) (shown in Figure 2) was adopted by evaluating three variants of the ResNet architectures (ResNet-18, ResNet-50, and ResNet-101).

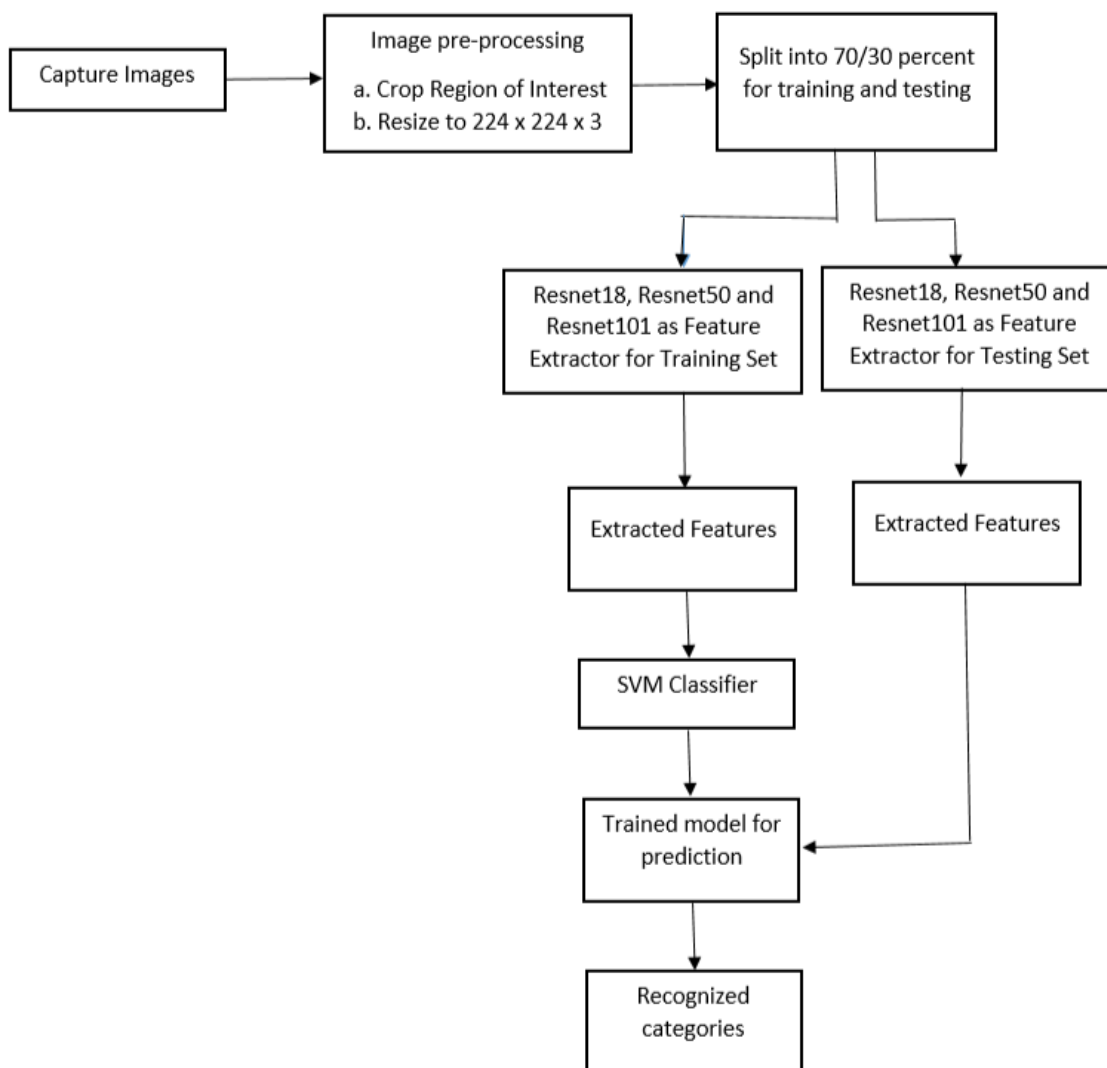












































Figure 2: Proposed Methodology

The experiment used a dataset captured at LGEA Kagoro Road Primary School in Kaduna State, Nigeria. LGEA Kagoro Road Primary School was established in 1977, the school follows the basic 7 policies and primarily uses English and Hausa languages alongside British Sign Language (BSL). The school serves 1028 students aged 7 years and older, focusing on the English alphabet (a - z) and digraph sounds (like 'sh', 'ie', 'ch', 'ng', etc.). Despite demonstrations, students struggle to comprehend the material due to slow learning processes. Therefore, a vision-based technique was proposed (Figure 2) by Irhebhude et al. (2023) to interpret and demonstrate sign languages, aiming to benefit both hearing impaired and hearing individuals alike.

4.1 Image Capture

To compile the dataset used in the study, 26 English alphabets and 16 specific diagraph signs were photographed, each comprising a variety of images. Diagraphs are pairs of letters that work as a team to create a unique sound, different from the individual sounds of the letters (Daisie, 2023). The dataset comprises 2,106 images of students demonstrating the alphabet and diagraph expressed in sign language. These images were captured using a camera to capture both facial expressions and hand gestures of male and female students, with each image corresponding to a distinct alphabet and diagraph sign. Table 1 shows example images from each category, including a breakdown of the alphabet, diagraphs, and the number of images captured for each class, with the highest class having 51 images and the lowest class having 43 images, which is a fairly balanced distribution dataset. The variation in the number of images was as a result of wrong display of sign following ground truth information from the instructors.

Table 1: List of alphabets and diagraph sounds with image sample

Alphabet/ diagraphs	Sample	Number of Images
A		51
AI		51
AR		43
B		51
C		51
CH		51
D		50
E		51
EE		51
ER		49
F		51
G		51
H		50
I		50
IE		51
J		51
K		51
L		49
M		51
N		49
NG		50
O		51
OA		51
OI		49
OO		51
OOO		49
OR		51
OU		50
P		51
Q		51
R		51
S		50
SH		50
T		51
TH		48
U		51
UE		51
V		51
W		51
X		50
Y		46
Z		49

The primary objective is to identify and classify the sampled alphabet and diagraph signs. Figure 2 depicts the proposed methodology for the sign language recognition system showing all the steps involved while Table 1 describes the various alphabets/diagraphs signs dataset and the number of images captured in each category.

4.2 Image Pre-processing

Before data splitting and testing, it is crucial to preprocess the images. Initially, all captured images were of varying sizes and subsequently cropped and resized to standardized dimensions of 224 by 224 pixels, ensuring uniformity in their format. The dataset was then partitioned into a training set comprising 70% of the data and a testing set comprising the remaining 30%.

4.3 Feature Extraction

In this stage, the ResNet algorithm served as the feature extractor to derive sign language recognition features from the dataset. The ResNet used 18-layer, 50-layer, and 101-layer plain network architectures, detailed in Irrehbude et al. (2023), to extract deep learned features. The features were automatically extracted in the layer before the fully connected layer before the subsequent input to the SVM classifier for classification. Skip connections in ResNet improve deep neural network training and performance by maintaining gradient flow, reducing information loss, and increasing optimization and generalization, enabling training of networks to appropriate layers (Oyedotun et al. 2021; Zhang et al. 2020). ResNet as a top-performing feature extraction model due to its automatic, reliable, and versatile nature across multiple applications (Xu et al., 2022).

4.4 Classification

By learning key features for individual classes in the dataset, the SVM completes classification tasks for the alphabet and diagraph sign language. The alphabets and diagraphs were represented by a hand sign with facial expression which were recognized and the correct sign was identified by the models. SVM classifier is chosen for its efficiency, accuracy, robustness, and ability to utilize extracted features in image classification, particularly for large, high-dimensional datasets (Kashef, 2021). SVM efficiency and accuracy are influenced by exceptional feature extracted and optimal parameters, which can be improved through innovative feature selection methods (Wang et al., 2023).

5 Experimental Results and Discussions

The results of the analysis are presented in this section. The model was trained using 70% and 30% of the dataset for training and testing respectively. To experiment, the dataset consists of 26 alphabets and 16 classes of diagraph images. The words require both hand gestures and other parts of the face as shown in validation results shown in Figures 9-11. The classification model attained an accuracy of 61.7% for ResNet18, 64.5% accuracy for ResNet50, and 66.5% accuracy for ResNet101. The confusion matrix and Area Under Curve (AUC) of the Receiver Operating Curve (ROC) were used in evaluating the performance of the model. The hyperparameter tuning of the training is shown in Table 2.

Table 2: SVM Model Hyperparameters

Parameters	Value
Kernel Function	Linear
Box Constraint Level	1
Kernel Scale Mode	Auto
Multiclass Coding	One-vs-one
Standardise Data	Yes

The ROC curves depicted in Figures 3 to 5 illustrate the performance of the various models. Notably, ResNet50 exhibited the highest Area Under the Curve (AUC) with an impressive overall performance of 99.9%. Following closely, ResNet18 achieved an AUC of 97.4%. In contrast, ResNet101 showed the lowest AUC performance at

96.4%. These findings underscore ResNet50's exceptional classification capability, demonstrating its effectiveness in distinguishing between the various alphabets and diagraph sounds compared to the other models evaluated.

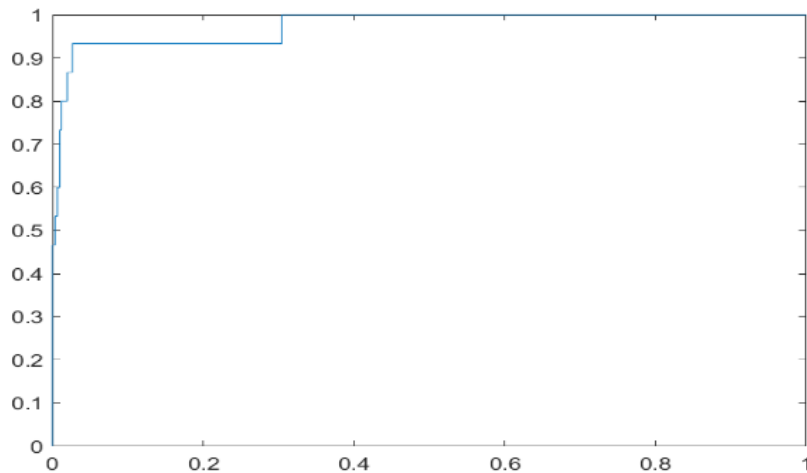


Figure 3: ROC Showing Classification Performance of ResNet18

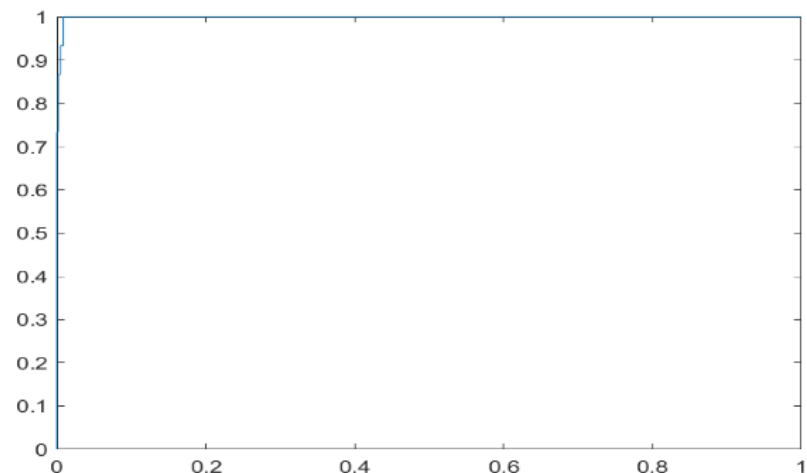


Figure 4: ROC Showing Classification Performance of ResNet50

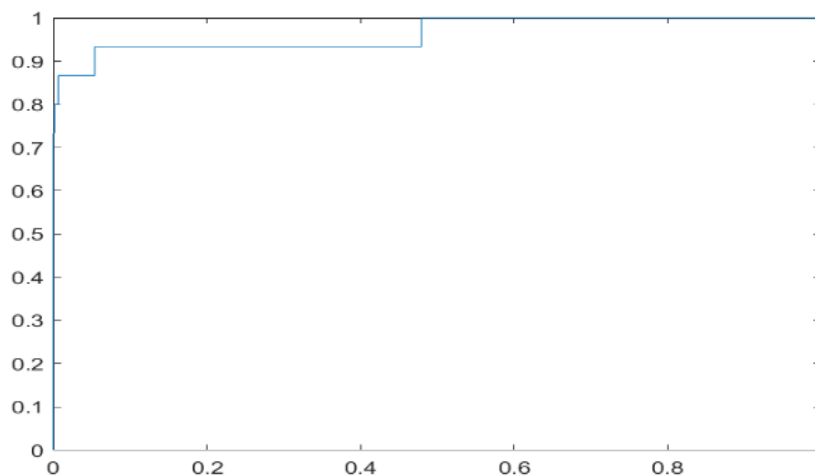


Figure 5: ROC Showing Classification Performance of ResNet101

The evaluation results, as presented in confusion matrices (Figures 6-8), provide insights into the performance of the models across different classes. A total of 632 sample images, comprising 30% of the entire dataset of alphabet

and diagraph images, were used for testing. This sample included 26 single-word alphabet classes, 15 two-word diagraph classes, and one three-word diagraph class, each contributing 30% of the test images.

Among the models evaluated, ResNet101 demonstrated the highest True Positive Rate (TPR) at 66.5% across all classes, indicating its capability to correctly identify positive instances. Conversely, it recorded a False Negative Rate (FNR) of 33.5%, indicating instances where positive instances were incorrectly classified as negative. In contrast, ResNet18 exhibited the lowest TPR of 61.7% and a corresponding FNR of 38.3% across all classes, suggesting its comparatively lower performance in correctly identifying positive instances.

Lastly, ResNet50 achieved a TPR of 64.5% and an FNR of 35.5%. These results collectively highlight ResNet101's superior performance in terms of correctly identifying positive instances, with a lower false negative rate compared to ResNet50.

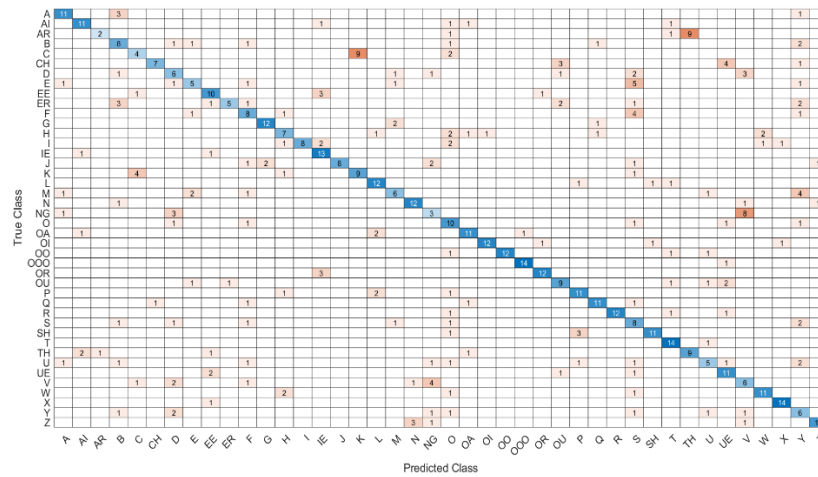


Figure 6: Confusion Matrix Showing Model Evaluation for ResNet18

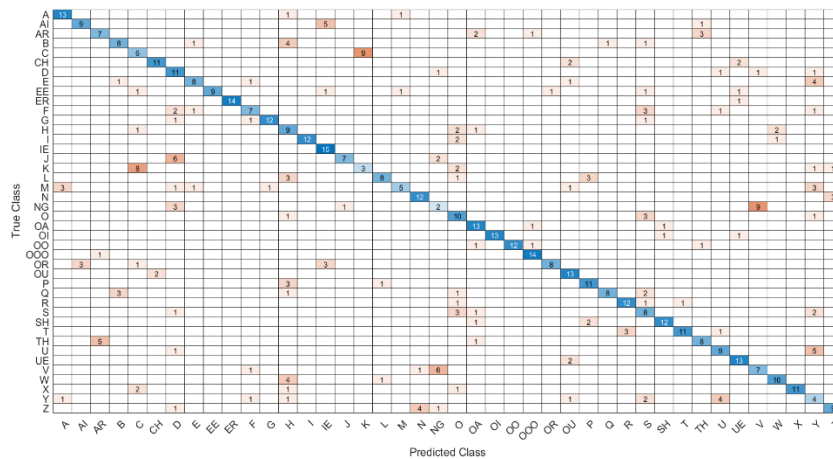


Figure 7: Confusion Matrix Showing Model Evaluation for ResNet50

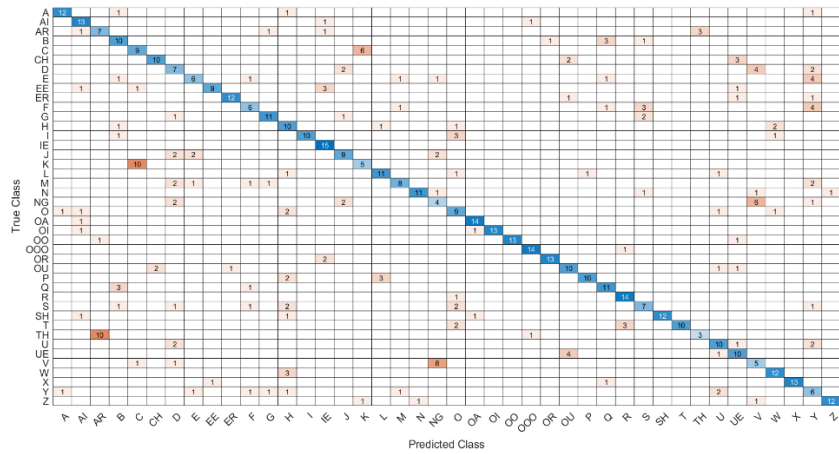


Figure 8: Confusion Matrix Showing Model Evaluation for ResNet101

These experiments demonstrated that the model achieved strong performance due to the quality of the input data and the effectiveness of the training process. The results further validate the proposed model's accuracy in interpreting alphabet and diagraph sign language, as evidenced by the findings presented in Figures 9-11. Results shows that few test images were wrongly predicted. The reason for this is the limited number of training data to enable the classifier learn to classifier into 42 groups.

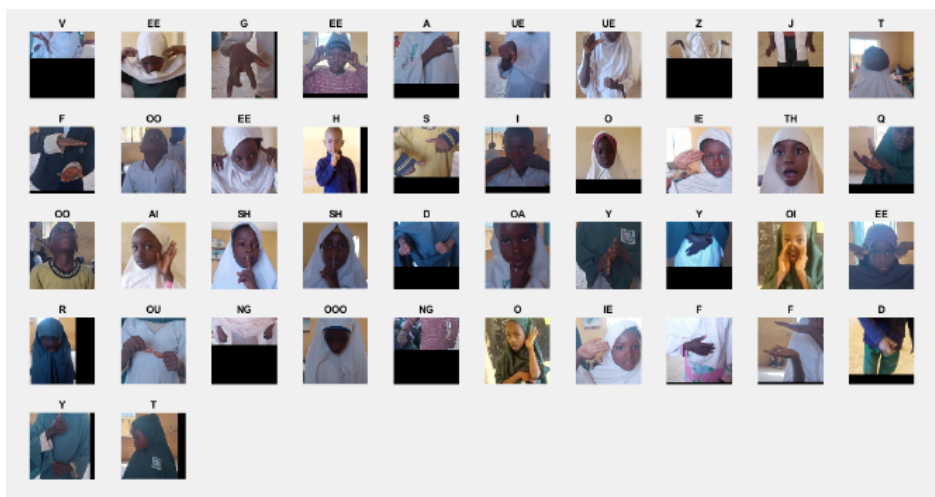


Figure 9: Validation Results for ResNet18

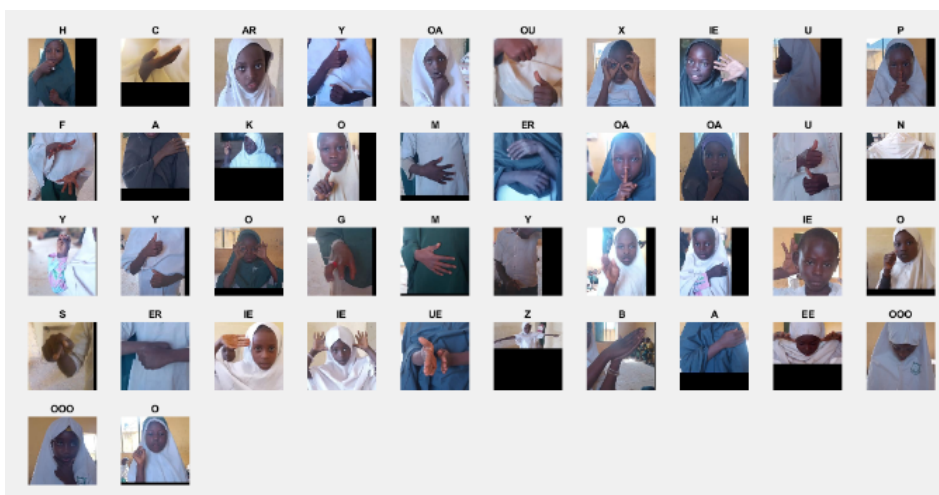


Figure 10: Validation Results for ResNet50

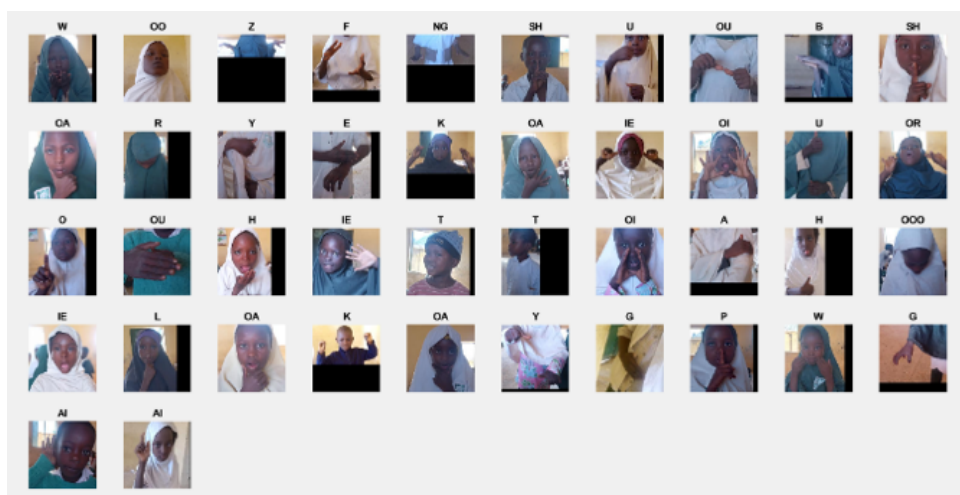


Figure 11: Validation Results for ResNet101

The results obtained from the experiments conducted on the captured dataset demonstrated the effectiveness of ResNet model. The ResNet101 model achieved the highest accuracy of 66.5% on the captured dataset compared to ResNet 18 and ResNet50 that gave 61.7% and 64.5% recognition accuracies respectively. The results indicate that the ResNet model can classify sign language images that include emotional cues.

The proposed model achieved the highest accuracy of 66.5% with ResNet101 on the self-captured dataset, highlighting the recognizing ability of ResNet101 with a higher number of layers. This result the higher layer model is particularly impressive when compared to the performance of the lower layer models. However, in terms of the performances of the classifier, ResNet50 performed best with an AUC of 99.9% when compared with the other models which gave 97.4% and 96.4% for ResNet18 and ResNet101 respectively. This difference in performance underscored the efficiency of the layer's depth.

The ResNet model's ability to achieve good accuracy on the dataset and excellent classification ability shows the robustness and adaptability of the model. This success suggests that the model can effectively learn and generalize the unique sign language for alphabet and diagraph signs, making it a more reliable choice for applications involving signs that incorporate gestures.

Olabanji & Ponnle (2021) achieved an accuracy of 95.67% on the Nigerian native sign language of the Nigeria classification using CNN. However, the study used a dataset for 15 selected words in Nigeria. The data collection procedure was not properly discussed for easy replication. The collected data only contains the hand sign avoiding the regions of the body that capture gesture information. Hence the need for this study. From the results obtained as shown from the confusion matrix of ResNet101 in Figure 8, the diagonal blue shows the true positive (TP) with diagraph IE recording highest true positive of 15 and TH recording the lowest TP of 3, while the highest false positive (FP) of 6 was recorded for alphabet C and NG diagraph. The false negative (FN) of 10 for the alphabet K and TH diagraph. Increasing layers impacted the performance of the experiments, with ResNet101 recording the highest recognition accuracy of 66.5%. This result validated what was obtained in the earlier study by Irhebhude et al. (2023).

6 Conclusion

In conclusion, this study introduced a model for recognizing alphabet and diagraph sign language, leveraging feature extraction from ResNet18, ResNet50, and ResNet101 algorithms. The system integrates hand gestures and facial movements for accurate sign language recognition. Classification into various alphabets and diagraph categories was achieved using SVM, yielding accuracies of 61.7% for ResNet18, 64.5% for ResNet50, and 66.5% for ResNet101 models, as evaluated on self-captured sign language images. This proposed model holds potential for integration into educational modules, addressing attention-related challenges and enhancing student engagement in learning processes. Given that sign language encompasses both spatial and temporal elements, with postures and gestures changing over time, integrating temporal data from video sequences may enhance recognition accuracy. It would be beneficial to capture and understand the temporal context of sign language using techniques like 3D convolutional neural networks (CNNs) or recurrent neural networks (RNNs).

References

- Akansha Tyagi, S. B. (2022). Hybrid FiST_CNN Approach for Feature Extraction for Vision-Based Indian Sign Language Recognition. *The International Arab Journal of Information Technology (IAJIT)*, 19(03), 403-411. <https://doi.org/10.34028/iajit/19/3/15>
- Aksoy, B., Salman, O. K. M., & Ekrem, Ö. (2021). Detection of Turkish Sign Language Using Deep Learning and Image Processing Methods. *Applied Artificial Intelligence*, 35(12), 952-981. <https://doi.org/10.1080/08839514.2021.1982184>
- Al-Hammadi, M., Muhammad, G., Abdul, W., Alsulaiman, M., Bencherif, M., Alrayes, T., Mathkour, H., & Mekhtiche, M. (2020). Deep Learning-Based Approach for Sign Language Gesture Recognition with Efficient Hand Gesture Representation. *IEEE Access*, 8, 192527-192542. <https://doi.org/10.1109/ACCESS.2020.3032140>
- Asonye, E., Emma-Asonye, E., & Edward, M. (2018). Deaf in Nigeria: A Preliminary Survey of Isolated Deaf Communities. *SAGE Open*, 8, 215824401878653. <https://doi.org/10.1177/2158244018786538>
- Blench, R., Warren, A., & Dendo, M. (2006). *An unreported African Sign Language for the Deaf among the Bura in Northeast Nigeria*.
- Bragg, D., Koller, O., Bellard, M., Berke, L., Boudreal, P., Braffort, A., Caselli, N., Huenerfauth, M., Kacorri, H., Verhoef, T., Vogler, C., & Morris, M. (2019). *Sign Language Recognition, Generation, and Translation: An Interdisciplinary Perspective*.
- Chao, H., Fenhua, W., & Ran, Z. (2019). Sign Language Recognition Based on CBAM-ResNet. *Proceedings of the 2019 International Conference on Artificial Intelligence and Advanced Manufacturing*, 1-6. <https://doi.org/10.1145/3358331.3358379>
- Chowdhury, A. R., Biswas, A., Hasan, S., Rahman, T. M., & Uddin, J. (2017). Bengali Sign language to text conversion using artificial neural network and support vector machine. *2017 3rd International Conference on Electrical Information and Communication Technology (EICT)*, 1-4. <https://doi.org/10.1109/EICT.2017.8275248>
- Côté Allard, U., Fall, C. L., Drouin, A., Campeau-Lecours, A., Gosselin, C., Glette, K., Laviolette, F., & Gosselin, B. (2019). Deep Learning for Electromyographic Hand Gesture Signal Classification Using Transfer Learning. *IEEE Transactions on Neural Systems and Rehabilitation Engineering: A Publication of the IEEE Engineering in Medicine and Biology Society*, PP. <https://doi.org/10.1109/TNSRE.2019.2896269>
- Daisie. (2023, June 21). *Diagraphs Explained: Comprehensive Phonics Guide*. Daisie Blog. <https://blog.daisie.com/what-is-a-diagraph-a-comprehensive-guide-to-understanding-and-using-diagraphs-in-phonics/>
- Gupta, A. K., & Singh, S. (2024). Hand Gesture Recognition System Based on Indian Sign Language Using SVM and CNN. *International Journal of Image and Graphics*, 2650008. <https://doi.org/10.1142/S0219467826500087>
- Haria, A., Subramanian, A., Asokkumar, N., Poddar, S., & Nayak, J. (2017). Hand Gesture Recognition for Human Computer Interaction. *Procedia Computer Science*, 115, 367-374. <https://doi.org/10.1016/j.procs.2017.09.092>
- Hasanah, S. A., Pravitasari, A. A., Abdullah, A. S., Yulita, I. N., & Asnawi, M. H. (2023). A Deep Learning Review of ResNet Architecture for Lung Disease Identification in CXR Image. *Applied Sciences*, 13(24), Article 24. <https://doi.org/10.3390/app132413111>
- He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Presented at the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA. doi:10.1109/cvpr.2016.90
- Irhebhude, M. E., Kolawole, A.O., & Abubakar, H. (2023). DIAGRAPH SIGN LANGUAGE RECOGNITION USING RESIDUAL NETWORK AND SUPPORT VECTOR MACHINE. *International Conference on Communication and E-Systems for Economic Stability | CeSES' 2023*. Retrieved May 9, 2024
- Irhebhude, M., Kolawole, A., & Goshit, N. (2023). *Perspective on Dark-Skinned Emotion Recognition Using Deep-Learned and Handcrafted Feature Techniques* (pp. 1-24). <https://doi.org/10.5772/intechopen.109739>
- Jain, V., Jain, A., Chauhan, A., Kotla, S. S., & Gautam, A. (2021). American Sign Language recognition using Support Vector Machine and Convolutional Neural Network. *International Journal of Information Technology*, 13, 1193-1200. <https://doi.org/10.1007/s41870-021-00617-x>

- Jiang, X., & Zhu, Z. (2019). *Chinese Sign Language Identification via Wavelet Entropy and Support Vector Machine*. 726–736. https://doi.org/10.1007/978-3-030-35231-8_53
- Kashef, R. (2021). A boosted SVM classifier trained by incremental learning and decremental unlearning approach. *Expert Systems with Applications*, 167, 114154. <https://doi.org/10.1016/j.eswa.2020.114154>
- Kothadiya, D., Bhatt, C., Sapariya, K., Patel, K., Gil-González, A.-B., & Corchado, J. M. (2022). Deepsign: Sign Language Detection and Recognition Using Deep Learning. *Electronics*, 11(11), 1780. <https://doi.org/10.3390/electronics11111780>
- Liao, Y., Xiong, P., Min, W., Min, W., & Lu, J. (2019). Dynamic Sign Language Recognition Based on Video Sequence with BLSTM-3D Residual Networks. *IEEE Access*, 7, 38044–38054. <https://doi.org/10.1109/ACCESS.2019.2904749>
- Lin, K., Zhao, Y., Gao, X., Zhang, M., Zhao, C., Peng, L., Zhang, Q., & Zhou, T. (2022). Applying a deep residual network coupling with transfer learning for recyclable waste sorting. *Environmental Science and Pollution Research*, 29(60), 91081–91095. <https://doi.org/10.1007/s11356-022-22167-w>
- Ma, R., Zhang, Z., & Chen, E. (2021). Human Motion Gesture Recognition Based on Computer Vision. *Complexity*, 2021, 1–11. <https://doi.org/10.1155/2021/6679746>
- Morgan, R. Z. (2002). Maganar Hannu: Language of the Hands: A Descriptive Analysis of Hausa Sign Language (review). *Sign Language Studies*, 2(3), 335–341. <https://doi.org/10.1353/sls.2002.0011>
- Olabanji, A., & Ponnle, A. (2021). Development of A Computer Aided Real-Time Interpretation System for Indigenous Sign Language in Nigeria Using Convolutional Neural Network. *European Journal of Electrical Engineering and Computer Science*, 5, 68–74. <https://doi.org/10.24018/ejece.2021.5.3.332>
- Oyedotun, O. K., Ismaeil, K. A., & Aouada, D. (2021). Training very deep neural networks: Rethinking the role of skip connections. *Neurocomputing*, 441, 105–117. <https://doi.org/10.1016/j.neucom.2021.02.004>
- Sahoo, J., Ari, S., & Patra, S. (2021). *A user independent hand gesture recognition system using deep CNN feature fusion and machine learning technique* (pp. 189–207). <https://doi.org/10.1016/B978-0-12-822133-4.00011-6>
- Sharma, S., & Singh, S. (2020). Vision-based sign language recognition system: A Comprehensive Review. *2020 International Conference on Inventive Computation Technologies (ICICT)*, 140–144. <https://doi.org/10.1109/ICICT48043.2020.9112409>
- Shi, X., Jiao, X., Meng, C., & Bian, Z. (2022). 3D Sign language recognition based on multi-path hybrid residual neural network. *2022 14th International Conference on Machine Learning and Computing (ICMLC)*. <https://doi.org/10.1145/3529836.3529943>
- Sreemathy, R., Turuk, M., Chaudhary, S., Lavate, K., Ushire, A., & Khurana, S. (2023). Continuous word level sign language recognition using an expert system based on machine learning. *International Journal of Cognitive Computing in Engineering*, 4, 170–178. <https://doi.org/10.1016/j.ijcce.2023.04.002>
- Venugopalan, A., & Reghunadhan, R. (2023). Applying Hybrid Deep Neural Network for the Recognition of Sign Language Words Used by the Deaf COVID-19 Patients. *Arabian Journal for Science and Engineering*, 48(2), 1349–1362. <https://doi.org/10.1007/s13369-022-06843-0>
- Wang, J., Wang, X., Li, X., & Yi, J. (2023). A Hybrid Particle Swarm Optimization Algorithm with Dynamic Adjustment of Inertia Weight Based on a New Feature Selection Method to Optimize SVM Parameters. *Entropy*, 25(3), 531. <https://doi.org/10.3390/e25030531>
- Wen, F., Zhang, Z., He, T., & Lee, C. (2021). AI enabled sign language recognition and VR space bidirectional communication using triboelectric smart glove. *Nature Communications*, 12(1), 5378. <https://doi.org/10.1038/s41467-021-25637-w>
- Xu, Y., Yang, W., Wu, X., Wang, Y., & Zhang, J. (2022). ResNet Model Automatically Extracts and Identifies FT-NIR Features for Geographical Traceability of Polygonatum kingianum. *Foods*, 11(22), 3568. <https://doi.org/10.3390/foods11223568>
- Zhang, W., Quan, H., Gandhi, O., Rajagopal, R., Tan, C.-W., & Srinivasan, D. (2020). Improving Probabilistic Load Forecasting Using Quantile Regression NN with Skip Connections. *IEEE Transactions on Smart Grid*, 11(6), 5442–5450. <https://doi.org/10.1109/TSG.2020.2995777>