ENGINEER - Vol. LIX, No. 02, pp. [99- 110], 2026

DOI: https://doi.org/10.4038/engineer.v59i2.7752

Explainability Evaluation of Transfer Learning Models Utilized in Multimodal Translation of Sign Language

H.M.L.S. Kumari and C.K. Walgampaya

Abstract: Sign language is a mode of communication for people with hearing loss and speaking disabilities; however, gesture meanings vary from region to region. The proposed study evaluates the efficacy of transfer learning for converting sign gestures into natural language text and audio. American Sign Language and Sinhala Sign Language datasets were used to demonstrate the impact of data augmentation and image preprocessing on model performance across RGB, grayscale, and binary formats. Transfer learning models, including MobileNet, InceptionV3, DenseNet121, and DenseNet201 were utilized, and results indicate that RGB data significantly outperforms other formats. MobileNet achieved the highest accuracies of 98.62% for the ASL dataset and 99.66% for the Sinhala Sign dataset. Real-time validation was conducted via webcam inference to ensure practical reliability. Furthermore, SHAP (SHapley Additive exPlanations) was used to evaluate the proposed model's interpretability. The study concludes that combining lightweight transfer learning architectures with robust data augmentation can lead to significant improvements in communication of sign language users.

Keywords: Data Augmentation, Real-Time Prediction, SHAP Interpretability, Sign Language Translation, Transfer Learning

1. Introduction

Communication is essential in social interactions, as it enhances understanding among the community, including those who are deaf and non-verbal [1]. Globally, about 360 million people suffer from hearing loss, including 328 million adults and 32 million children [2]. Disabling hearing loss occurs when there is a significant impairment of more than 40 decibels in the better hearing ear [3]. Automated recognition of hand gestures is vital for overcoming daily challenges and enhancing convenience for these individuals [4]. Sign language consists of manual and body movements including fingers, hands, arms, and facial expressions to transmit information visually. It serves as a practical communication tool for the speech-impaired in their everyday lives [5]. Even though this medium plays a crucial role, it is not widely known by the public, creating a significant and persistent barrier between signers and the hearing community [6]. Since not all words have corresponding signs, specific terms are often communicated letter-by-letter using finger spelling [7]. While sign language is a global concept, it is not a singular, universal system. Instead, it exists as a diverse collection of distinct regional languages. Among these, American Sign Language (ASL) is a natural language with its own distinct grammar, similar in origin to spoken languages, while it expresses meaning through facial expressions, body movements and actions. Currently, a universal sign language does not exist. Different sign languages are used in different regions of the world. For example, British Sign Language (BSL) is completely different from ASL. People who are familiar with ASL would not easily understand BSL. Some countries incorporate elements of ASL into their own sign language. Sign language is a method of communication for people with speech and hearing impairments.

These individuals cannot overcome being mute. To assist them and facilitate their communication with others, thereby bringing some normalcy to their lives, it was recognized that the need was to bridge this communication gap. This highlights the necessity for a sign language recognition system to help alleviate their daily communication challenges. Real-time hand gestures can allow communication between deaf people and hearing people without a translator.

Miss. H.M.L.S. Kumari, *B. Sc. (Hons) Computer Science, Instructor, Faculty of Engineering, University of Peradeniya.*
Email: lihinisangeetha99@gmail.com
https://orcid.org/0009-0001-1410-7840
Eng. (Prof.) C.K. Walgampaya, *AMIE(SL), B.Sc. Eng. (Hons)(Peradeniya), MSc. (Louisville, USA), PhD (Louisville, USA), Faculty of Engineering, University of Peradeniya.*
Email: ckw@pdn.ac.lk
https://orcid.org/0000 -03-2192-474X

As the number of individuals with deafness continues to rise, the demand for interpreters also increases. Bridging the communication gap between those who are hearing impaired and the general population becomes crucial for effective interaction. Sign language translation is progressing rapidly, with a significant role to fill in facilitating communication for people with hearing impairments.

The proposed study introduces a real-time system that takes hand gesture images as input through a webcam and converts them into respective text and audio clips. This work contributes to the field by demonstrating cross-linguistic robustness, evaluating the system on two structurally different datasets: ASL and Sinhala Sign Language (SSL). The proposed method employs transfer learning using models such as MobileNet, InceptionV3, and DenseNet, ensuring the system is effective for both large and small datasets. A primary contribution of this research is the systematic comparative analysis of three image formats—RGB, grayscale, and binary within the dataset to determine how varying levels of visual information impact classification accuracy. Data augmentation is further applied to evaluate performance stability. Finally, the model's outputs are evaluated using SHAP. By integrating Explainable AI (XAI), this study moves beyond "black-box" models to prove that the system accurately captures significant anatomical features for classification, thereby increasing the interpretability and reliability of the proposed approach.

The structure of the paper is as follows: Section 2 reviews current studies related to sign language recognition. Section 3 details the proposed methodology. The results are presented in Section 4, followed by the conclusion in Section 5.

2. Related Works

Sign language enhances communication. It is a better option for the deaf and hard-of-hearing communities, improving human-computer interaction and reducing the communication gap, as explained above. An examination of recent literature reveals the development of numerous systems designed to function as sign language translators using deep learning models.

Agrawal et al. [8] proposed a study titled "A Model for Hand Gesture Recognition Using Deep Learning." They developed a hand gesture recognition system with four steps. The first step involves capturing hand gestures via a webcam live stream, followed by extracting images from the video frames, preprocessing the images, and recognizing sign language/hand gestures to convert them into text or audio output. Image processing and Convolutional Neural Networks (CNN) were used to develop their model, which was evaluated using a Kaggle dataset, a dataset created by them, and a combined dataset. The models trained with Inception V3 demonstrated higher validation accuracy of 88.67%, attributed to its numerous layers (48) compared to the second model, which had only 4-5 layers. The highest validation accuracy of the second model was 69.6%.

In "Optimization of Transfer Learning for Sign Language Recognition Targeting Mobile Platforms," Rathi [9] proposed a real-time ASL recognizer based on a mobile platform to enhance accessibility and ease of use. This technique employs transfer learning on new hand gesture data for ASL alphabets, leveraging various pre-trained high-end models. The goal is to optimize the best model for mobile platform performance while addressing its limitations. The dataset contains 27,455 images, each representing one of 24 ASL letters. The optimized model, designed for memory efficiency on mobile devices produces a 95.03% recognition accuracy.

Buttar et al. [10] proposed a work titled "Deep Learning in Sign Language Recognition: A Hybrid Approach for the Recognition of Static and Dynamic Signs." They developed a real-time American Sign Language detector using a skeleton model, which reliably categorizes continuous signs in most cases through a deep learning approach. Additionally, they created a sign language detector for static signs using YOLOv6. This application is particularly beneficial for sign language users and learners to practice in real-time. After independent training on static and continuous signs, both algorithms were combined into a hybrid model. The proposed model, consisting of an LSTM with MediaPipe holistic landmarks, achieves approximately 92% accuracy for continuous signs, while the YOLOv6 model achieves 96% accuracy for static signs.

Daroya et al. [11] proposed a work titled "Alphabet Sign Language Image Classification Using Deep Learning," which introduces a method to classify RGB images of static hand poses representing letters in sign language using a CNN inspired by the DenseNet architecture. The proposed network achieved an accuracy of 90.3%, which is comparable to other studies, including those utilizing both

depth and RGB images. They used a dataset based on the American fingerspelling format. To enhance the variety within the dataset, data augmentation techniques were applied. These techniques included random rotations up to 5 degrees, random horizontal and vertical shifts, random shearing, and random zooming.

Saleh and Issa [12] utilized transfer learning with pre-trained networks, specifically VGG-16 and ResNet152, to enhance the accuracy of classifying 32 hand gestures from the ArSL dataset (ArSL2018). To address class imbalance, they applied random under-sampling, reducing the number of images from 54,049 to 25,600. This technique effectively minimized the disparity in class sizes. Their approach achieved impressive accuracy rates of 99.4% with VGG-16 and 99.6% with ResNet152 during testing.

Dardina et al. [13] conducted real-time hand gesture recognition in depth images using a CNN. Their proposed CNN model architecture aims to address the communication barrier faced by deaf individuals, achieving an accuracy of 94.61% in recognizing 11 different gestures using depth images. The pre-processed hand gesture images were converted into binary images and then trained using the CNN model.

Romala et al. [14] proposed a work titled "Sign Language Recognition System Using Convolutional Neural Network and Computer Vision." In this study, users can capture hand gesture images from a web camera and predict the particular hand gesture. They utilized a CNN for training and classifying images, allowing for the recognition of 10 ASL alphabet gestures with high accuracy. The model attained over 90% accuracy, demonstrating its high performance.

Cui et al. [15] proposed a work titled "Fast American Sign Language Image Recognition Using CNNs with Fine-tuning." Given CNNs' proven excellence in image classification tasks, the authors suggested the application of a fine-tuned CNN model to classify ASL images. First, the CNN architecture was optimized to better suit the ASL classification task. Subsequently, the models were trained on this optimized structure using test images. The average accuracy of the proposed model over 5-fold cross-validation was 79% for trained CNNs from the CIFAR-10 model and 89% for the fine-tuned CaffeNet model.

Recent studies are focusing on using XAI to go beyond just using standard CNN-based classification. This leads to understand how deep learning models work with sign language. A study “Explainable AI for sign language recognition models: Integrating Grad-Cam LIME and Integrated Gradients" by Fatima and team [16] looked into using Grad-CAM, LIME and Integrated Gradients for ASL recognition. This study found that Grad-CAM gives heatmaps that show general hand areas. LIME provides explanations but can be inconsistent across different samples. Integrated Gradients was the most accurate giving information that is helpful for analyzing small gesture features in ASL recognition. ‘Deep Neural Network-Based Sign Language Recognition: A Comprehensive Approach Using Transfer Learning with Explainability’ proposed by Ridwan et al. [17] explains Deep Neural Network-Based Sign Language Recognition used SHAP (Deep Explainer) to see which features were important. This method helped the authors know exactly which input areas affected predictions of the model. This made the decision-making process much clearer for ASL recognition. These advancements are a step forward, in making sign language systems trustworthy. However, there is still a gap. Most current research only uses one dataset; it does not compare explainability methods across different data types. This suggests that we need standardized evaluation frameworks in the field of Explainable AI for Sign Language Recognition.

Overall, recent studies show a clear move toward using transfer learning and real-time webcam systems to help bridge communication gaps. While these models have reached impressive accuracy especially with ASL, there are still important areas that remain unexplored.

3. Methodology

3.1 Data Preparation

The static hand gesture dataset ASL dataset [18] from the IEEE Data Port was used in this study. The dataset was created by capturing static gestures of the ASL alphabet from 8 participants, excluding the letters J and Z, which involve dynamic movements. Figure 1 shows a sample from the static hand gesture ASL dataset. Images were captured using a Logitech Brio webcam with a resolution of 1920×1080 pixels in a university laboratory under artificial lighting. The final dataset images were cropped to a 400×400-pixel area, focusing solely on the hand region. This dataset is available in three versions: RGB (colour), binary, and grayscale images. All three versions were used in this study with the proposed

model. Sample images from each version are shown in Figures 3, 4, and 5. The dataset consists of 24 classes, covering all the letters of the Alphabet, except for J and Z. It includes a total of 960 images, with each class balanced to contain 40 images. This consistency is maintained across all three datasets: the RGB (colour) dataset, the grayscale dataset, and the binary dataset. Both sequential and affine data augmentation techniques were applied in this proposed study.

3.2 Data Augmentation

To enhance the diversity and robustness of the training data, this study utilizes a combination of sequential and affine data augmentation. Data augmentation is essential for the proposed study to prevent model overfitting and to ensure high generalization, particularly given the subtle anatomical differences in hand gestures and the limited size of the regional Sinhala Sign Language dataset. By artificially expanding the training set, the model is forced

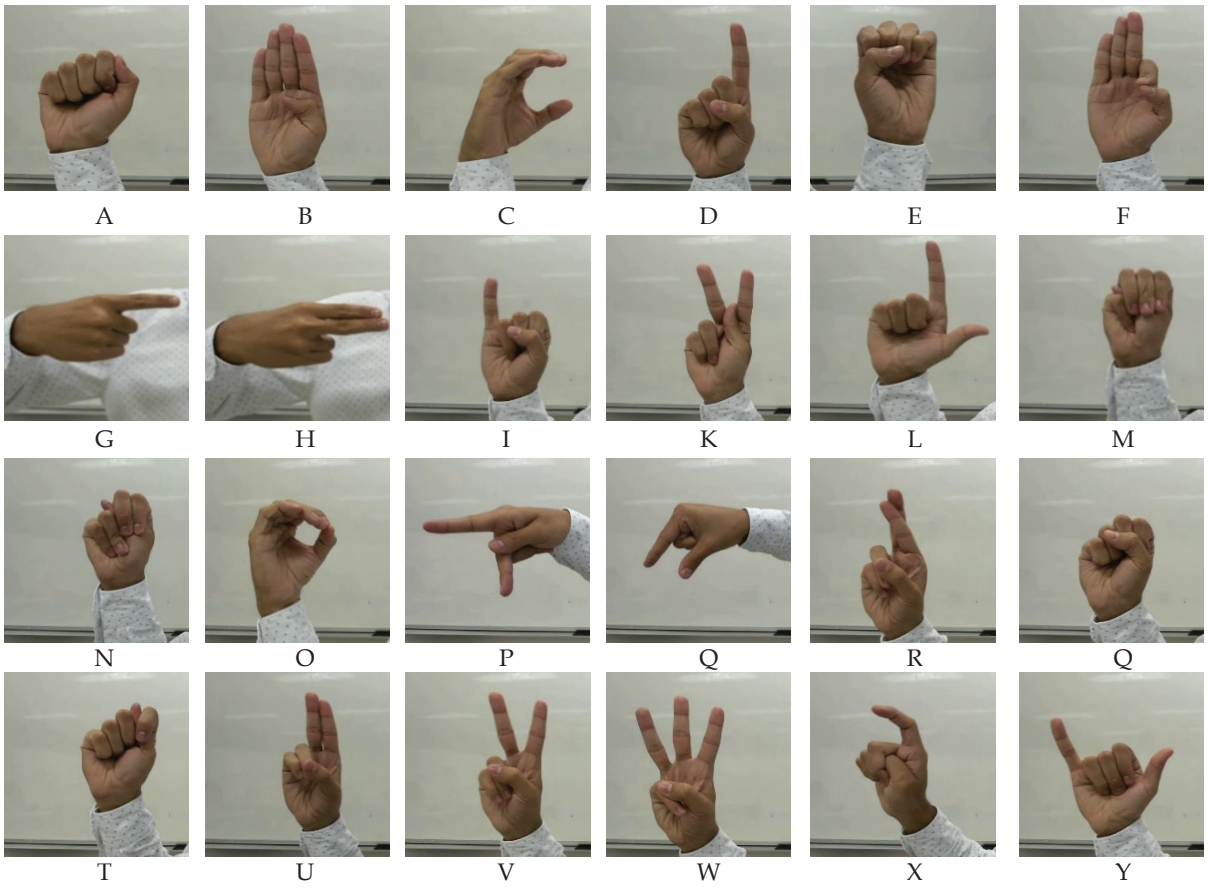


Figure 1 – Samples of each letter in the static hand gesture ASL dataset [16]

The second dataset used was a Sinhala Sign language dataset [19] from Kaggle. The dataset mainly consists of classes for different numbers as well as WH-questions (e.g. What, When, How, etc.) and consisted of 33 classes in total. some of the signs are dynamic movements as well, and they were made into classes with suffixes for the rank of the frame they have in each case. Some images had signs representing multiple characters, in which case they are separated using an '&' character. Some of the classes have different lighting compared to some other classes, while at the same time having consistent lighting throughout each class. The person appears to be wearing a ring, and this is kept consistent as well.

background noise or specific lighting conditions.

The augmentation pipeline was implemented using a Sequential meta-augmenter with random_order=True. This approach is crucial because it ensures that the model never sees the same transformation sequence twice, effectively simulating the infinite variety of real-world environments. The specific techniques were selected based on the following rationales:

Sequential data augmentation: Horizontal flips (Fliplr) and random crops account for users signing with the natural handedness of different users or at varying framing positions. Gaussian blur, contrast normalization, and additive noise are used to simulate low-quality

webcam sensors and poor lighting—common challenges in real-time home applications.

Affine data augmentation: Affine transformations, including scaling (zoom), translation (movement), rotation (up to 25°), and shearing, were selected to replicate natural human variance. These simulate different distances from the camera, diverse hand sizes, and the natural tilt or slant of a user's hand while signing.

Each training image was augmented into 10 distinct variations. This factor was chosen to provide a sufficient data volume for deep transfer learning—which typically requires thousands of samples to converge—without introducing excessive redundancy that could lead to biased gradients.

Regarding the dataset organization, the data was partitioned into a 70% training, 20% testing, and 10% validation split. This 70:20:10 ratio was adopted to provide a substantial foundation for the transfer learning process while reserving a statistically significant portion (20%) for a final, objective evaluation of accuracy on unseen data. The 10% validation set allowed for unbiased hyperparameter tuning and early stopping to prevent over-training. Importantly, data augmentation was applied exclusively to the training set to ensure the integrity of the evaluation process, keeping the testing and validation sets strictly representative of original, real-world data.

3.3 Transfer Learning Models

Transfer learning is an advanced deep learning technique that leverages a pre-trained model for one task as a starting point to learn another similar task. For instance, training a CNN on the ImageNet dataset can take approximately 2-3 weeks using multiple GPUs. For this study, transfer learning models, including MobileNet, InceptionV3, DenseNet121, and DenseNet201, were applied to the datasets described above. MobileNet is tailored for mobile and embedded vision applications, offering both efficiency and a lightweight design. By employing depth-wise separable convolutions, it significantly reduces the number of parameters and computational demands [20]. InceptionV3 is part of the Inception family of architectures and is known for its inception modules which factorize convolutions into smaller operations. This model achieves high accuracy with fewer parameters compared to other large networks. DenseNet121 and DenseNet201 are part of the DenseNet family, which connects each layer to every other layer in a feed-forward fashion. This design improves the flow of information and gradients through the network, making them highly efficient and capable of achieving state-of-the-art performance [21].

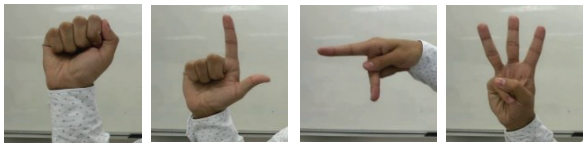
3.4 SHAP (Shapley Additive exPlanations)

In this study, SHapley Additive exPlanations (SHAP) were integrated to transition the model from a "black-box" system to a transparent, interpretable framework. The primary objective of employing SHAP was to validate the model's internal decision-making process by identifying

Figure 2 – Samples of each class in the Sinhala Sign Language Dataset

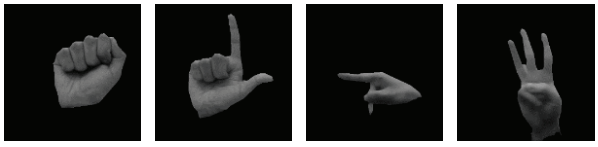
which anatomical features of the hand such as finger orientation or palm position contributed most significantly to the classification. By calculating the contribution of each pixel through game theory, SHAP allows for a direct visual audit of whether the model is focusing on relevant hand gestures or irrelevant background noise [22].

In the resulting visualizations, red regions signify positive SHAP values that strongly support the model's predicted class, while blue regions indicate negative influence [23]. However, through the evaluation of these plots, a critical finding of the proposed study is that SHAP may not be the optimal XAI tool for image-based sign language models. The analysis revealed that SHAP can occasionally produce misleading or "noisy" results that do not perfectly align with the complex spatial features of hand gestures. Consequently, SHAP was utilized here not only as a validation tool but also as a benchmark to identify the limitations of current XAI methods in this domain. This conclusion provides a clear research path for future work, where tools like Grad-CAM (Gradient-weighted Class Activation Mapping) could be explored to achieve higher interpretability and better spatial localization for deep learning models applied to image data.



A L P W

Figure 3 – The data sample of RGB dataset of Static hand gesture dataset ASL dataset



A L P W

Figure 4 – The data sample of grayscale dataset of Static hand gesture dataset ASL dataset

A L P W

Figure 5 – The data sample of binary dataset of Static hand gesture dataset ASL dataset

3.5 Proposed Methodology

The methodology section outlines the procedures and techniques used to conduct the research. Figure 5 illustrates the key steps, including data collection, preprocessing, model training, evaluation, and prediction using a webcam to generate text or audio output. First, the dataset was divided into three parts: training, testing, and validation sets, with ratios of 70%, 20%, and 10%, respectively. Next, a separate training dataset was created using the data augmentation techniques described in the data preparation section. The pre-trained models mentioned above are then employed with additional layers.

The transfer learning model is used without its top classification layer, utilizing weights pre-trained on the ImageNet dataset. This approach leverages established feature extractors to identify complex patterns while freezing the base layers to prevent the loss of pre-trained knowledge during the initial training phase. The input images have a shape of (224, 224, 3). Global average pooling is applied to the output of the transfer learning model (pooling='avg'), reducing each feature map to a single value. This pooling method was selected over a Flatten layer to significantly reduce the number of trainable parameters and minimize the risk of overfitting. A dense layer with 1024 units and a ReLU activation function is then added to the model. To prevent overfitting, a dropout layer with a dropout rate of 0.5 is incorporated. Finally, a dense layer with 24 units and a softmax activation function is added for classification into 24 classes (Dense (24, activation='softmax')). The model was compiled using the Adam optimizer with a learning rate of 0.0001, and categorical cross-entropy was utilized as the loss function to handle the multi-class classification task.

This configuration results in a new Keras model that combines the pre-trained transfer learning base with custom layers. All layers of the transfer learning model are set to be non-trainable, meaning their weights remain unchanged during training. This approach enables the model to leverage pre-trained features while training only the custom dense and dropout layers. Training was conducted with a batch size of 32 for 50 epochs, incorporating an early stopping mechanism to halt training if the validation loss failed to improve for five consecutive iterations. The new model is then evaluated using 5-fold stratified cross-validation. The stratified approach was chosen to ensure that each fold maintains a balanced representation of the 24

gesture classes, providing a more reliable assessment of model generalization. The models discussed above were trained using the original dataset without data augmentation to assess the performance of these transfer learning models in the absence of augmentation. The results of each model were then analysed based on the predictions made. In the next step, a webcam was connected to capture hand gestures as input.

The trained transfer learning models were used to identify these hand gestures, with the results displayed as both text and audio output.

SHAP was employed to interpret how predictions are generated by the trained transfer learning model.

Predictions were conducted using images taken from recorded predicted output of real-time input from a webcam as well as from original images from test data. This procedure helps verify whether the proposed model is functioning correctly and illustrates how specific image features contribute to the final prediction of the model.

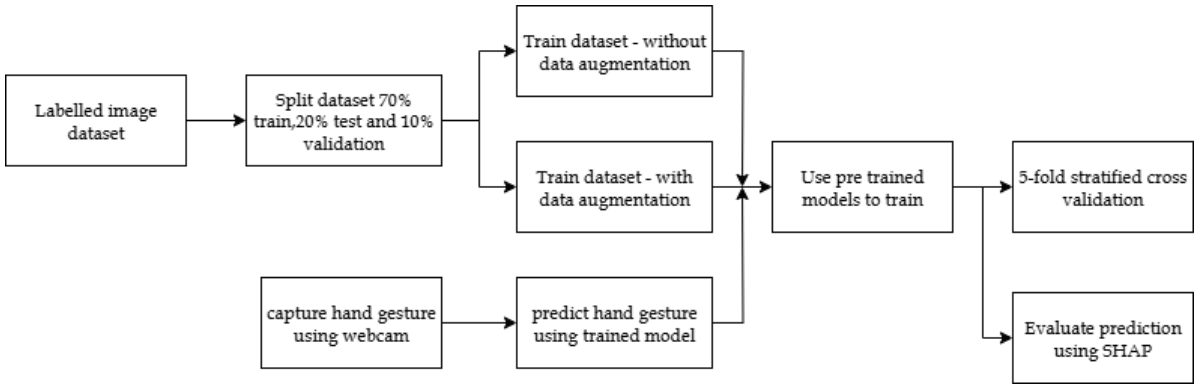


Figure 6 – Proposed methodology

This explicit implementation confirms that the model prioritizes the structural contours of the hand gestures rather than irrelevant background noise. When the model predicts correctly for a particular class, SHAP will typically highlight more red-coloured regions of the image with high positive contributions from these features. Red regions in SHAP plots represent positive SHAP values (features pulling the prediction towards the target class) whereas blue regions represent negative SHAP values (features pulling the prediction away from the target class).

4. Results & Discussion

The proposed model was evaluated using metrics such as accuracy, precision, recall and F1 score, which are crucial for assessing deep learning models in classification tasks. These metrics are computed based on true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) as shown in equations (1), (2), (3) and (4).

$$Accuracy\ (AC) = (TP+TN)/(TP+TN+FP+FN) \quad (1)$$

$$Precision\ (PR) = TP/(TP+FP) \ldots \quad (2)$$

$$Recall\ (RE) = TP/(TP+FN) \ldots \quad (3)$$

$$F1\ Score = 2*(PR*RE)/(PR+RE) \ldots \quad (4)$$

To ensure the integrity of these metrics, the dataset was partitioned into 70% for training, 10% for validation, and 20% for independent testing. While the training set optimized model weights, the validation set was used to tune hyperparameters and monitor for overfitting. To further validate the model's generalization across the entire dataset, a 5-fold stratified cross-validation was employed. This ensures that every data point is used for both training and testing across different folds while maintaining the original class distribution in each split.

Table 1 presents the average performance metrics obtained from 5-fold cross-validation on the unaugment RGB dataset using four transfer learning models: MobileNet, InceptionV3, DenseNet121, and DenseNet169.

Among these models, MobileNet achieved the highest accuracy (98.51%), precision (98.75%), and F1-score (98.51%), indicating superior performance on the RGB dataset without data augmentation

Table 1 – Average Performance Matrices from 5-fold cross validation on an unaugment RGB dataset

Transfer learning model	Mean accuracy%	Mean precision	Mean F1score
Mobilenet	0.9851	0.9875	0.9851
InceptionV3	0.8955	0.9097	0.8874
Densenet121	0.9179	0.9408	0.9163
DenseNet169	0.9478	0.9618	0.9459

Table 2 demonstrates the average performance matrices for above stated transfer learning models for unaugment grayscale. MobileNet achieved the best results indicating strong performance. DenseNet169 also performed competitively, whereas InceptionV3 recorded the lowest values across all metrics.

Table 2 – Average Performance Matrices from 5-fold cross validation on an unaugmented grayscale dataset

Transfer learning model	Mean accuracy%	Mean precision	Mean F1score
Mobilenet	0.9753	0.9759	0.9753
InceptionV3	0.8478	0.8488	0.8678
Densenet121	0.9079	0.9128	0.9063
DenseNet169	0.9365	0.9536	0.9365

Table 3 presents the performance metrics obtained from 5-fold stratified cross-validation on the unaugmented binary dataset. Similar to Tables 1 and 2, MobileNet shows the best performance, while InceptionV3 records the lowest.

Table 3 – Average Performance Matrices from 5-fold cross validation on an unaugmented binary dataset

Transfer learning model	Mean accuracy%	Mean precision	Mean F1score
Mobilenet	0.9434	0.9439	0.9432
InceptionV3	0.7674	0.8085	0.7615
Densenet121	0.8062	0.7951	0.7910
DenseNet169	0.7984	0.8107	0.7910

Table 4 presents the performance metrics obtained by the proposed model using the augmented RGB dataset with the transfer learning models described above.

Table 4 – Average Performance Matrices from 5-fold cross validation with augmented RGB dataset

Transfer learning model	Mean accuracy%	Mean precision	Mean F1score
Mobilenet	0.9862	0.9863	0.9861
InceptionV3	0.8883	0.8910	0.8876
Densenet121	0.9659	0.9668	0.9656
DenseNet169	0.9488	0.9531	0.9488

Table 5 demonstrates the performance metrics obtained by the proposed model using the augmented grayscale dataset with the transfer learning models described above.

Table 5 – Average Performance Matrices from 5-fold cross validation with augmented grayscale dataset

Transfer learning model	Mean accuracy%	Mean precision	Mean F1score
Mobilenet	0.9763	0.9762	0.9767
InceptionV3	0.8722	0.8723	0.8776
Densenet121	0.9559	0.9558	0.9554
DenseNet169	0.9356	0.9355	0.9356

Table 6 presents the performance metrics obtained by the proposed model using the augmented binary dataset with the transfer learning models described above.

Table 6 – Average Performance Matrices from 5-fold cross validation with augmented binary dataset

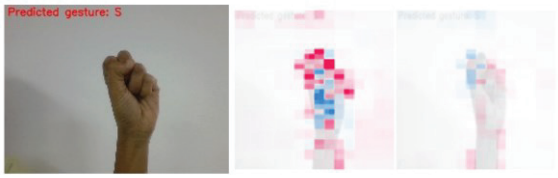
Transfer learning model	Mean accuracy%	Mean precision	Mean F1score
Mobilenet	0.9149	0.9273	0.9137
InceptionV3	0.7314	0.7594	0.7380
Densenet121	0.8531	0.8627	0.8491
DenseNet169	0.8531	0.8561	0.8502

A comparative analysis of the results across Tables 1 through 6 reveals that MobileNet consistently outperforms InceptionV3 and the DenseNet variants. This superior performance is likely due to MobileNet's (accuracy of 98.62%,) use of depth-wise separable convolutions, which allows it to capture essential spatial features of hand gestures with fewer parameters, making it less prone to overfitting on the specific geometry of sign language compared to the more complex InceptionV3 architecture. Furthermore, the results indicate that the RGB format provides the most robust feature set, as the inclusion of colour depth aids the model in distinguishing hand contours from the background more effectively than Grayscale or Binary formats. In contrast, the significant performance drop in InceptionV3 across binary datasets suggests that its wider kernels may introduce noise when processing simplified edge-based inputs.

The highest-performing proposed model was used to predict hand gestures captured via a webcam. The predicted result was provided as both an audio clip and in text. Each predicted image was saved, and the corresponding images were evaluated using SHAP plots to interpret the model's decision-making process [24].

V W

Figure 7 – W hand gesture and SHAP output



X S

Figure 8 – S hand gesture and SHAP output

H P

Figure 9 – P hand gesture and SHAP output

Figure 7 and Figure 9 show how the model correctly predicts the final output, with SHAP displaying more red coloration on the correct hand signs.

Figure 7 shows more red around the finger regions in the image representing class W than class V, demonstrating that the model uses those finger areas to predict class W rather than class V. In Figure 9, the features around the thumb, index finger, and middle finger are utilized by the proposed model to make the final prediction, as shown by the red-coloured regions in the SHAP output for the class P image.

However, some SHAP plots demonstrated incorrect or unclear outputs, such as in Figure 8. This can be due to the fact that using SHAP, it was assumed that features of an image are independent. But since in this scenario, this independence is not true, the results can be unreliable and uninformative. In certain cases, the outputs are difficult to interpret due to lack of clarity [25][26]. Table 7 shows the statistics of the model when trained on the Sinhala Sign Language dataset without using data augmentation. Table 8 shows the same for the dataset with data augmentation. Accuracy as well as loss curves are also shown on Figure 10. and Figure 11. From the outputs shown it can be seen that the MobileNet model had achieved the highest accuracy at 99.66%.

Table 7 – Average Performance Matrices from 5-fold cross validation on the Sinhala sign language dataset without data augmentation

Transfer learning model	Mean accuracy%	Mean precision	Mean F1score
Mobilenet	0.9833	0.9832	0.9832
InceptionV3	0.9523	0.9522	0.9563
Densenet121	0.9856	0.9859	0.9859
Densnet201	0.9898	0.9898	0.9892

Table 8 – Average Performance Matrices from 5-fold cross validation on the Sinhala sign language dataset with data augmentation

Transfer learning model	Mean accuracy%	Mean precision	Mean F1score
Mobilenet	0.9966	0.9966	0.9969
InceptionV3	0.9659	0.9658	0.9655
Densenet121	0.9917	0.9920	0.9917
Densnet201	0.9965	0.9965	0.9960

Similar to how SHAP was applied to the ASL dataset, SHAP was also applied to the SSL dataset.

Tables 10 and 11 represent the accuracy curve and loss curve of MobileNet model which achieved highest accuracy 99.66% which trained on augmented RGB dataset for Sinhala sign language dataset in the proposed study.

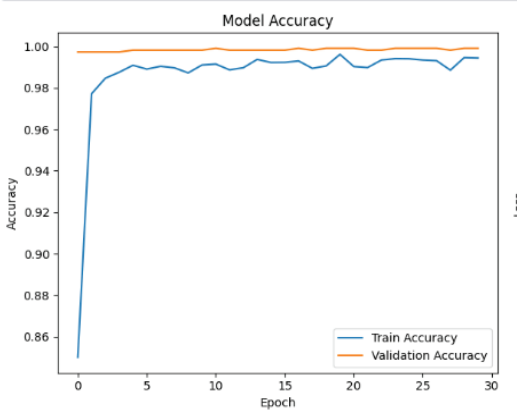


Figure 10 – Accuracy curve of MobileNet model trained on augmented data

The training and validation accuracy curves, demonstrated in Figure 10, show that the model converged extremely rapidly, achieving almost perfect accuracy in the initial phase of training and maintaining stability throughout. Similarly, the training and validation loss curves, represented in Figure 11, decrease steeply in the early phase and stabilize at very low values. The close alignment between training and validation performance suggests that the model generalizes extremely well to novel data, with hardly any signs of overfitting or underfitting.

From figures shown using Figure 12 through Figure 15, the results convey similar capabilities of the models as done previously using the ASL dataset. Fingers shown in bright red can be regarded as more crucial for the model for the prediction, while fingers shown in bright blue can be regarded as more crucial for the model against the prediction.

It can be said that the model has learnt to identify the person's fingers separately to arrive at the predictions it has made when testing. There still existed cases that were not accurate as well. Figure 14 shows the SHAP plot for a sample image for the hand gesture for the number '5'. Yet in this case it can be seen that the model predicted it wrongly as either '5' or '14'. This is inaccurate.

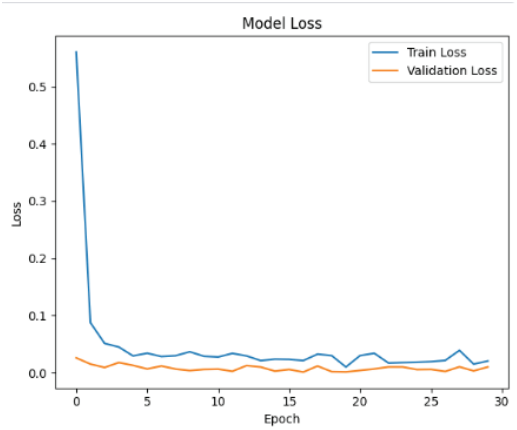
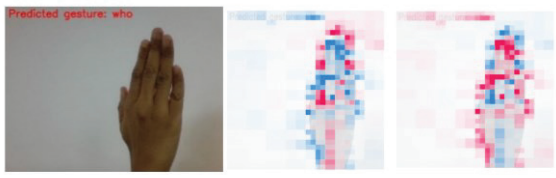


Figure 11 – Loss curve of MobileNet model trained on augmented data



Oneten_2 & Eleven_1 Who

Figure 12 – 'who' hand gesture and SHAP output

Five Four & Fourteen 2

Figure 13 – '9' hand gesture and SHAP output

Five Four & Fourteen 2

Figure 14 – '5' hand gesture and SHAP output

Oneten_2 & Eleven_1 Six

Figure 15 – '6' hand gesture and SHAP output

Table 9 – Performance comparison of the proposed model against state-of-the-art sign language recognition baselines.

Study	Methodology /Architecture	Accuracy	Key Contribution
Agrawal et al. [8]	Inception V3 (48 layers)	88.67%	Comparison of shallow vs. deep CNN architectures.
Rathi et al. [9]	Mobile-optimized Transfer Learning	95.03%	Memory efficiency for mobile-based real-time recognition.
Buttar et al. [10]	YOLOv6 (Baseline)	96.00%	Hybrid model for both static and dynamic sign detection.
Daroya et al. [11]	DenseNet (Baseline)	90.30%	Use of data augmentation on RGB fingerspelling images.
Saleh & Issa [12]	VGG-16 / ResNet152	99.4% / 99.6%	Handling class imbalance in large-scale datasets (ArSL2018).
Dardina et al. [13]	Depth-image based CNN	94.61%	Addressing communication barriers via depth-sensing.
Cui et al. [15]	Fine-tuned CaffeNet	89.00%	Optimized CNN architecture for fast ASL recognition.
Proposed Work	**Optimized Deep Learning with XAI**	**98.62% / 99.66%**	**Highest accuracy with integrated XAI (SHAP) for model transparency.**

Data augmentation consistently improves model stability across all architectures and RGB input remains the optimal data format for

maintaining high F1-scores. Rather than relying on model depth alone, the results suggest that the lightweight, efficient architecture of the proposed MobileNet-based model is better suited for the high-speed requirements of webcam-based gesture recognition.
Table 9 benchmarks the proposed model against above related studies, showing it outperforms baselines like YOLOv6 (96%) and DenseNet (90.3%) with a peak accuracy of 99.66%. Beyond performance, this study uniquely bridges the identified research gap by integrating XAI to provide transparency that traditional "black-box" models lack.

5. Conclusions

Based on the performance results of the proposed study, data augmentation significantly improves the accuracy, precision, and F1 score of transfer learning models by increasing the diversity of the training data and reducing overfitting. Results of the proposed study show that the models performed significantly better when using colour (RGB) images rather than grayscale or binary versions. For instance, MobileNet achieved an amazing accuracy of 99.66 percent on the enhanced Sinhala Sign Language data set. This goes to show that the combination of a light model with high-quality data can be incredibly effective. Therefore, the combination of coloured datasets and data augmentation methods proves to be more effective in transfer learning models.

By converting hand gestures into text or speech, the communication gap between individuals unfamiliar with sign language and those with special needs can be bridged. The SHAP plots demonstrate how evaluations were conducted in the proposed model to visualize which hand features such as finger positioning and joint angles contribute most to the prediction. This study, however, finds that SHAP may not be the best Explainable AI technique for analyzing models that use images because it can provide incorrect results since its mathematical premise, i.e., feature independence, does not hold due to the spatial relationship between the pixels within hand gestures. It should also be noted that the model's efficiency in real-time applications may be affected by other environmental factors not considered in the experimental dataset. The proposed study used only output images. However, future implementations should develop a real-time recognition model using LSTMs or 3D-CNNs.

As future work, Grad-CAM and other Explainable AI tools could be explored, as they may provide better interpretability for deep learning models applied to image data by highlighting specific localized regions of interest more clearly. These findings not only address a significant research gap regarding the application of transfer learning to regional sign languages but also have practical implications for enhancing communication accessibility worldwide.

References

1. Tolentino, L.K.S., Juan, R.S., Thio-ac, A.C., Pamahoy, M.A.B., Forteza, J.R.R. and Garcia, X.J.O., 2019. Static Sign Language Recognition Using Deep Learning. *International Journal of Machine Learning and Computing*, 9(6), pp.821-827.

2. Ojha, A., Pandey, A., Maurya, S., Thakur, A. and Dayananda, P., 2020. Sign Language to Text and Speech Translation in Real Time Using Convolutional Neural Network. *Int. J. Eng. Res. Technol. (IJERT)*, 8(15), pp.191-196.

3. Tarafder, K.H., Akhtar, N., Zaman, M.M., Rasel, M.A., Bhuiyan, M.R. and Datta, P.G., 2015. Disabling Hearing Impairment in the Bangladeshi Population. *Journal of Laryngology & Otology*, 129(2), pp.126-135.

4. Cheok, M. J., Omar, Z., and Jaward, M. H., "A Review of Hand Gesture and Sign Language Recognition Techniques," *Int. J. Mach. Learn. Cybern.*, Vol. 10, No. 1, pp. 131–153, 2019, https://doi.org/ 10.1007/s13042-017-0705-5.

5. Hossen, M.A., Govindaiah, A., Sultana, S. and Bhuiyan, A., 2018, June. Bengali Sign Language Recognition using Deep Convolutional Neural Network. *7th International Conference on Informatics, Electronics & Vision (iciev) and 2018 2nd International Conference on Imaging, Vision & Pattern Recognition (icIVPR) (pp. 369-373)*. IEEE.

6. Rastgoo, R., Kiani, K. and Escalera, S., 2020. Hand Sign Language Recognition Using Multi-View Hand Skeleton. *Expert Systems with Applications,* 150, p.113336.

7. Chong, T.W. and Lee, B.G., 2018. American Sign Language Recognition Using Leap Motion Controller with Machine Learning Approach. *Sensors*, 18(10), p.3554.

8. Agrawal, M., Ainapure, R., Agrawal, S., Bhosale, S. and Desai, S., 2020, October. Models for Hand Gesture Recognition Using Deep Learning. IEEE *5th International Conference on Computing Communication and Automation (ICCCA)* (pp. 589-594). IEEE.

9. Rathi, D., 2018. Optimization of Transfer Learning for Sign Language Recognition Targeting Mobile Platform. arXiv preprint arXiv:1805.06618.

10. Buttar, A.M., Ahmad, U., Gumaei, A.H., Assiri, A., Akbar, M.A. and Alkhamees, B.F., 2023. Deep Learning in Sign Language Recognition: A Hybrid Approach for The Recognition of Static and Dynamic Signs. *Mathematics*, 11(17), p.3729.

11. Daroya, R., Peralta, D. and Naval, P., 2018, October. Alphabet Sign Language Image Classification Using Deep Learning. *TENCON 2018-2018 IEEE Region 10 Conference* (pp. 0646-0650). IEEE.

12. Saleh, Y. and Issa, G., 2020. Arabic Sign Language Recognition Through Deep Neural Networks Fine-Tuning.

13. Tasmere, D., Ahmed, B. and Das, S.R., 2021. Real-Time Hand Gesture Recognition in Depth Image Using CNN. *International Journal of Computer Applications*, 174(16), pp.28-32.

14. Walizad, M. E., & Hurroo, M. (2020). Sign Language Recognition System Using Convolutional Neural Network and Computer Vision. *Int. J. Eng. Res. Technol*, 9(12), 59-64.

15. Cui, Q., Zhou, Z., Yuan, C., Sun, X. and Wu, Q.J., 2018. Fast American Sign Language Image Recognition Using CNNs with Fine-tuning. *Journal of Internet Technology*, 19(7), pp.2207-2214.

16. El-Qoraychy F-Z, Mualla Y, Zhao H, Dridi M, Créput J-C, Longo L (2025) Explainable AI for Sign Language Recognition Models: Integrating Grad-CAM, LIME and Integrated Gradients. *PLoS One* 20(12): e0336481. https://doi.org/10.1371/journal.pone.0336481

17. Ridwan, A. E., Chowdhury, M. I., Mary, M. M., & Abir, M. T. C. (2024). Deep Neural Network-Based Sign Language Recognition: A Comprehensive Approach Using Transfer Learning with Explainability. *arXiv preprint* arXiv:2409.07426.

18. Pinto Jr, R.F., Borges, C.D., Almeida, A.M. and Paula Jr, I.C., 2019. Static Hand Gesture Recognition Based on Convolutional Neural Networks. *Journal of Electrical and Computer Engineering*, 2019(1), p.4167890.

19. Dilshan, T. (2019). Sinhala Sign Language Dataset – TDJ [Dataset]. Kaggle. https://www.kaggle.com/datasets/thamindudilshan/sinhala-sign-language-dataset-tdj.

20. Howard, A.G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M. and Adam, H., 2017. MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications. *arXiv preprint* arXiv:1704.04861.

21. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J. and Wojna, Z., 2016. Rethinking the Inception Architecture for Computer Vision. *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 2818-2826).

22. Huang, G., Liu, Z., Van Der Maaten, L. and Weinberger, K.Q., 2017. Densely Connected Convolutional Networks. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 4700-4708).

23. Awan, A. A. (2023, June 28). An Introduction to Shap Values and Machine Learning Interpretability. DataCamp. https://www.datacamp.com/tutorial/introduction-to-shap-values-machine-learning-interpretability

24. Mummidivarapu, S. R. (2024, July 4). Hands-on shap: Practical Implementation for Image, Text, and Tabular Data. Medium. https://medium.com/@shreeraj260405/hands-on-shap-practical-implementation-for-image-text-and-tabular-data-f74b488f8d71

25. Jethani, Neil, et al. "Fastshap: Real-Time Shapley Value Estimation." *International Conference on Learning Representations*. 2021

26. Payrovnaziri, S. N., Chen, Z., Rengifo-Moreno, P., Miller, T., Bian, J., Chen, J. H., Liu, X., and He, Z., ''Explainable Artificial Intelligence Models using Real-World Electronic Health Record Data: A Systematic Scoping Review,'' *J. Amer. Med. Inform. Assoc.*, Vol. 27, No. 7, pp. 1173–1185, Jul. 2020