ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
Vol.03 No.04 (2025)

DEEP LEARNING-BASED SIGN LANGUAGE RECOGNITION FOR INCLUSIVE COMMUNICATION ACCESS

Zain Zafar[1], Sohail Masood[1]
[1]Department of Computer Science & IT, Superior University, Lahore 54000, Pakistan

ABSTRACT
The three leading deep learning models used were Vision Transformer, EfficientNet-B0, and ResNet-50. To find out each model's performance in reliably detecting sign gestures, its accuracy, precision, recall, and F1 score were determined. The ResNet-50 expressed good feature recognitions with continuous learning between test and training data sets, and it had gained its top accuracy level of 98.9%. With 97.5% F1 score, its 97% accuracy and 98% recall expressed its reliable prediction accuracy of sign movements. The EfficientNet-B0 followed with 97.4% accuracy level as well. As expressed by the rising validation loss graph, it expressed overfiting signs despite its training speed being very fast. With 95% accuracy and 96% recall, it had 95.5% F1 score. With 92.5% F1 score and appropriate precision and recall of 92% and 93%, respectively, the prototype of Vision Transformer developed a 93% accuracy level. In comparison with EfficientNet-B0, the prototype of Vision Transformer expressed more stability throughout training with minimum overloading, albeit with somewhat lower accuracy. These results are valuable in illustrating model usability in real-world applications and international sign recognitions, wherein reliable performances are mandatory. This study shows that deep learning might provide reliable, scalable alternatives to real-time sign gesture processes when optimisation steps and infrastructures are carefully chosen.

Keywords: Sign Language recognition, Deaf, Communication skills, Deep Learning, hard-of-hearing.

INTRODUCTION
Sign language is the major mode of communication of millions of deaf and hard-of-hearing people across the globe. Sign language is a form of language which differs with verbal languages in that it is based on visibility such as hand gestures, orientation, movement as well as facial expressions. This gap between deaf and hearing people can be addressed through robotized sign language recognition (SLR) systems which translate the sign language to text or speech in real-time **[1]**. The initial efforts of recognition were specific and demanded special gloves or markers to work, which did not make them too practical and were not adopted. With the development of computer vision and deep learning, the paradigm has shifted to the vision based solutions, which allows capturing gestures with cameras instead of invasive sensors. Nevertheless, obtaining robust, error-free and real-time recognition is proving a tough task. The scale-up requirement of recognition systems is getting more and more urgent with the development of video conferencing, online learning, and telemedicine **[2]**. Successful SLR has the potential to build deaf communities with more agency and access to services. Furthermore, the possibility of translations in real-time can make communication between a hearing person and someone with no hearing more flawless, contributing to inclusiveness **[3]**. Hence the task at hand to create effective, precise and generalizable sign language recognition models cannot simply be considered a technical problem but a social necessity that can have a tremendous effect on accessibility and equality. The task of sign recognition alludes to a set of issues due to the visual and temporal intricacy of gestures. Signs differ among users because of the differences in signing style, size of applicant hands, speed and articulation **[4]**.

The same person will exhibit varying examples of the same sign slightly varying because of being tired or emphasizing on a situation. This kind of inter- and intra-signer variability is one of the reasons why recognition systems have trouble generalizing and fit to overfitting. Besides, there are also tions, regional variations of sign execution, which are an additional hurdle during model training **[5]**. As opposed to gesture recognition in general, sign language involves exact interpretation of difficult hand confections and delicate finger placements. The difference between many signs is only a bend of fingers or relative location of fingers that requires high

ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
Vol.03 No.04 (2025)

resolution spatial feature extraction. Furthermore, the means of moving between the signs include the continuous movement, involving the temporal dependences which are impossible to note with the help of the static image classifiers **[6]**.

Fig # 1: Sentence disparities clarifying contextual gesture alterations

Signs are disambiguated frequently in relation. Meaning of a gesture can be determined by the preceding or subsequent signs, facial or body posture. As an example, the combination of the hands might imply various words in various sentences or with different emotional stress **[7]**. This shows that temporal modeling architectures using sequential dependencies and context-sensitive inference are in demand. Older machine learning techniques using hand-designed features have not done much better in reflecting the intricacies of sign language. Conversely, deep learning has something to offer, i.e., learning of hierarchical representations on unprocessed video. Convolutional Neural Networks (CNNs) work particularly well when it comes to learning about spatial patterns in frames, and Recurrent Neural Networks (RNNs) and Transformers always work when it is necessary to learn about temporal relations **[8]**. These models are brought together to support end-to-end training pipelines, which operate without any hand-engineering and learn robust and discriminative features.

The purpose of this study is to develop a complete deep learning architecture to effectively recognize sign language with high accuracy in real time to overcome the well-documented infrastructural impediments, that deaf and hard-of-hearing people always encounter when seeking access to inclusive communication **[9]**.

The main problem statement in the process of recognising sign language is not simple due to numerous related issues: a variety of signers are gesturing in a different manner, with minute hand positions, speed, correct timing and expression of faces. In addition, weak recognition is further aggravated by environmental issues, including variable lighting, inhomogeneous backgrounds, and blockages **[10]**. More traditional machine learning methods might use hand-designed features or sensor-based inputs so are not comprised of scalable methods and need considerable manual action and find it hard to generalize to a wide array of real-world situations. As the solution to these drawbacks, we offer to present a modular deep learning architecture including a set of various state of the art components implemented. To begin with, CNN backbones, such as ResNet50, EfficientNet-B0 are used to process every frame of the video to obtain valuable information with respect to the shape and location of hands, which is

ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
Vol.03 No.04 (2025)

essential to distinguish between similar signs **[11]**. To capture the temporal linear process of signing we add temporal modeling layers, such as Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) networks, which have the ability to model dependencies in multidimensional sequences and therefore capture the way gestures can develop dynamically. Moreover, we also incorporate Transformer-based attention mechanism that has the ability to learn the long-range temporal dependencies by evaluating the significance of each frame within the sequence, which can be highly vital in continuous sign language recognition where one sign can be interpreted based on the previous information **[12]**. With no wearable gloves or depth sensors being needed, the overall burden put by the proposed pipeline on users is also minimal, increasing its deployment scalability. The study makes three important contributions as it proposes a unified end-to-end model that does not require manual feature extraction or feature pre-segmentation, comparatively analyzes CNN-LSTM and CNN-Transformer hybrid models in determining the most effective temporal modeling strategy in different vocabularies and signer conditions, and derives an optimized inference pipeline that can be used to provide the low-latency prediction necessary to support real-time application, e.g., in live video interpretation or accessible communication platforms **[13]**. This framework represents a methodical solving of the issues of gesture variability and context specific dependence, along with effective computation and leads to a greater transition of the state of sign recognition to practical, inclusive environments that can enable deaf people to interact in the real world with ease **[14]**.

LITERATURE REVIEW

Initial approaches of recognizing sign language mainly concentrated on utilizing statistical modeling and conventional machine learning processes to categorise gestures and progressions. HMMs were among the most popular models due to their good performance in the model of a temporal sequence that forms the key to thinking in continuous signing **[15]**. This approach showed that sign recognition in real-time and continuously is viable but with great attention to feature design and controlled recording environment requirements. Isolated signs have also been widely classified by Support Vector Machines (SVMs). SVMs are useful at recognizing gestures of the static type, when enough discriminative features are available, because they come up with the best hyperplane that separates not only classes in the feature space **[16]**. Typically, SVM-based classifiers represented the hand configuration/orientation per frame in either Histogram of Oriented Gradients (HOG) or Shape Contexts descriptors. SVMs especially used to be very accurate on isolated signs, but lacked reflect ability due to a lack of any temporal modelling as each instance was processed individually. Consequently, SVMs perform weakly when it comes to acquisition of dynamic transitions between signs with continuous sequences **[17]**. There are similar shortcomings of these conventional methods. They are mostly reliant on manually created features, which stand a very high risk of fluctuation in lighting scenarios, backgrounds, and differences in the identity of signers, and must rely on deep domain knowledge to create. Moreover, these approaches are not well-scalable to bigger vocabularies and to varying real-world settings **[18]**. A significant proportion of the models did not translate to real-life scenarios very well outside the controlled laboratory environments or when applied to small data sets. Such deficiencies have inspired a shift to deep learning architectures, which have the ability to learn complex spatiotemporal representations directly on unprocessed data, and thus do not require as many handcrafted features and are much more generalizable and perform better in practical sign language recognition tasks **[19]**.

Computer vision Before deep learning, computer vision algorithms were based on segmentation, tracking, and feature extraction pipelines. The hands and face of the signer were isolated by skin colour segmentation and background subtraction. Afterward, temporal dynamics were extracted by calculating optical flow and motion trajectories **[20]**.

ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
Vol.03 No.04 (2025)

Table 1: Deep learning models' accuracy benchmarks in sign recognition

Authors (Year)	Dataset / Classes	Model	Accuracy (%)	Precision / F1 (%)	FPS / GFLOPs
Kang et al. (2015)	31 classes (depth data)	CNN	99.99 (seen) / 83.6–85.5 (unseen)	–	333 FPS (3 ms/image)
Nature Sci Rep (8 mo ago)	ROBITA gestures	Multi-Layer ML-CNN	87.5	> BLSTM/HMM by +3.5 pp	–
Nature Sci Rep (last week)	ASL alphabet	Hybrid Transformer-CNN	**99.97**	–	110 FPS @ 5 GFLOPs
ActiveCNN-SL (1.1 yr ago)	Massey (758), ASL (7 k), ASL Alphabet (26.1 k)	Active CNN + Transfer Learning	99.92 / 99.80 / 99.80	–	–
Telugu SL (Apr 2024)	TSL gestures	YOLOv5-medium	–	Precision/F1: 90.5	–
Ridwan et al. (Sept 2024)	Bhutanese SL	ResNet50 + DNN	98.90	–	–

Some of the methods that assisted in the classification of the dynamics of the hand shape include contour analysis, convex hull descriptors among others. Although these approaches showed potential, they were very light sensitive, occlusion sensitive and when subjected to the presence of a cluttered or busy background. They were also built to be manually tuned such as thresholds and heuristics which restrict their robustness **[21]**. These methods have since been replaced by deep learning methods which learn features that are invariant directly against data. Sign language recognition has also been altered owing to recent advancement in deep learning. Deep models have tackled issues of signer variability, articulations, multi-modal dependencies and intricacies, and context-dependencies. Furthermore, the deep learning has enabled it to be implemented on mobile device and real-time systems, which has further opened its doors to the deaf population in more practical application in real life such as video conferencing and live interpretation. CNNs have been used to perpetrate frame-by-frame picture recognition of signs, in which they performed well to split up fussy structures of the hand **[22]**. A high performance of static sign classification has been observed using architectures such as VGG16 and ResNet. The use of context-dependent signs is also taken care of in these models because they preserve hidden states across sequences so that given previous gestures, they can be disambiguated. This has enhanced the power of real time sign recognition system applied on deaf communication assistive technology **[23]**. Recently, transformers were applied to video-based sign recognition where self-attention has been used to capture long-range dependencies between frames. Transformers can train faster than RNNs since they run sequential data in parallel, and make better context models possible. Camgoz et al., created sign language Transformers which perform better than RNN baselines on the PHOENIX dataset. Transformers demonstrate an advantage in learning hierarchical models that take hand motion into account, but also the faces and body pose. Most of the research deals with datasets where the condition is controlled, and the results may not be applied to practice **[24]**. More so, much of the sign recognition process is directed towards the movement in the hands but rarely tapping into the use of other facial parts and situations. Latency is an issue that is of critical concern to real-time applications, and Transformer based model may involve heavy computation. Lastly,

ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
Vol.03 No.04 (2025)

the cross-lingual and dialectal variation still remains under-researched and it needs datasets that are more comprehensive and transfer learning strategies **[25]**.

2.1 RESEARCH OBJECTIVES

- One will aim at extracting the high-resolution spatial features of video frame based on ResNet-50 and EfficientNet-B0 backbones.
- To apply the Vision Transformer modules to take into consideration the long-range dependencies of time and series of signs.
- We would like to come up with pre-processing methods that could enhance the performance of the model to the variability of signers, and background clutter.
- We are going to test the influence of diverse data augmentation and normalisation techniques with the recognition rate **[26]**.

The given research aims to develop a high-quality end-to-end sign language recognition framework, which utilizes the synergies of the ResNet-50, EfficientNet-B0, and Vision Transformer architectures. The goal of the work is to perform a rigorous analysis of this hybrid approach on varied data and assess both the accuracy and generalization of the approach to new signers and new vocabularies **[27]**. Also, the paper will explore the pre-processed steps of normalization, frame sampling, and data augmentation to make the model abler to deal with real-life issues, such as variable lighting and occlusions. Finally, the study is aimed at pursuing inclusive technology that fills communication divides between deaf and hard of hearing populations by providing a scalable and deployable deep learning solution **[28]**.

2.2 RESEARCH QUESTION

1. How well will the proposed model be able to generalise to other signers, vocabularies, and recording conditions?
2. Which pre-processing methods do best to enhance recognition in harsh conditions of occlusion and variations in lighting?
3. What are the scaling performance in the architecture as the size of vocabulary in sign language datasets are increased?
4. What are the trade-offs between model complexity, inference speed and recognition accuracy?

It studies the possibility of using ResNet-50 and EfficientNet-B0 together to extract strongly discriminative spatial features across diverse sign gestures. Additional inquiries entail the robustness of the model: its ability to generalize to new signers, settings, and vocabularies, the effect of pre-processing and augmentation, and the way in which the system can strike an equilibrium between precision, latency, and computational demand in real-time applications. Lastly, the study scouts the extent with which the architecture can be scaled with an increase in vocabulary sizes and finds the best tricks to keep the accuracy high without a loss in speed to the inference **[29]**.

METHODOLOGY

For multitudes of deaf & hard-of-hearing people throughout the globe, sign language serves as their major means of conversation. However, communication hurdles brought about by the general public's limited comprehension and perception of sign language result in exclusion from professional, educational, and healthcare settings. The goal of this project is to use deep learning methods to create an intelligent system that can recognise the alphabets in American Sign Language (ASL) in order to close this crucial gap in inclusion communication **[30]**. The Sign Language MNIST dataset, which includes 28x28 greyscale pictures that indicate 25 ASL characters (apart from the dynamic movements "J" and "Z"), is used in this study. There are 2,746 test photos and 27,455 training images in the whole dataset. Two cutting-edge convolutional neural network (CNN) layouts, ResNet50 with EfficientNetB0, were used to carry out sign categorisation. The deeper design ResNet50, on the other hand, is well-known

ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
Vol.03 No.04 (2025)

for its residual connections, which guard against disappearing gradients and enable reliable feature extraction. Data normalisation, one-hot label encoding, picture tensor reshaping, and RGB multichannel conversion using Tensor Flow's Lambda layers were important preparation processes. To avoid overfitting, the models were trained utilising early stopping, categorical cross entropy damage, and the Adam optimiser **[31]**.

3.2 Dataset

Which was obtained via Kaggle, is a labelled dataset designed especially for the identification of static hand gestures. There are two main files in the dataset:

- sign_mnist_train.csv – recycled for training
- sign_mnist_test.csv – recycled for testing every image is a 28x28 grayscale

depiction of a hand sign that corresponds to one of the 25 ASL letters—'J' and 'Z' are not included because of their dynamic gestures. Every sample has a label ranging from 0 to 24 and is expressed as a flat matrix of 784-pixel magnitude levels.

Table 2: Structure of the Sign Language MNIST dataset

Dataset Split	Number of Samples	Image Dimensions	Classes
Training Set	27,455	28x28	25
Testing Set	2,746	28x28	25

1.1 DATA PRE-PROCESSING

In order to convert raw picture material into a format appropriate for neural network training, the pre-processing phase was essential. Panda's toolkit was used to import the information first, making it simple to manipulate the label and pixel information. In order to recreate the spatial structure needed for convolutional neural associations, each image—which had previously been saved as a flat array of 784 pixel values—was moulded into a 28x28 matrix. A selection of these altered photos were shown using visualisation methods, which guaranteed proper label orientation and layout. Integers ranging from 0 to 24 were taken from the label columns and one-hot converted to
keras.utils.to_categorical (),
an essential step in getting the dat1a ready for a softmax activation function-based multi-class categorisation. The model was able to acquire information from appropriately structured input features and category labels because to this pre-processing step. Following this phase, Tensor Flow and Keras functions including Input Layer, Resizing, Lambda, abandonment, and Dense were used to build the EfficientNetB0 and ResNet50 models. Every framework was trained for five epochs with a rapid stop after being assembled using the Adam optimiser with categorised cross-entropy loss **[32]**.

1.2 Data Normalization

Data normalisation is essential for increasing accuracy of models & convergence over training in the setting of sign language interpretation. Image or video frames that record hand motions with pixel values ranging from 0 to 255 are often used to create sign language datasets. Those numbers should be normalised to a [0, 1] range. Given that CNNs and other deep learning models are susceptible to input pay-outs, this is particularly crucial. Normalisation and flattening to dimensions such as 28x28x1 help standardise data for greyscale picture inputs, which makes gesture structure learning more effective.

Table 3: Normalization trials for competent deep learning algorithms

Normalization Method	Data Points

ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
Vol.03 No.04 (2025)

Min-Max Scaling	Records Rescaling to array [0, 1].
Z-score Normalization	Average: 0, Std Dev: 1
Decimal Scaling	Scaling aspect built on max total value (10j10^j10j).
Robust Scaling	Centring about median, scaling by Inter-quartile Range (IQR).
Log Transformation	Log(x + 1)
L2 Normalization	Normalization to entity size ‖x‖2=1|x|_2 = 1‖x‖2=1

The train_test_split () function from Scikit-learn was used to break down the prior to processing. Initially, training accounted for 80% of the data, with the remaining 20% being equally split between test and validation sets. The ability to track the model's effectiveness on unseen data was made possible by this three-way separation.

1.3 Model Architecture

3.3.1 EfficientNetB0

Appropriate for mobile with portable devices, EfficientNetB0 is a lightweight, scalable convolutional neuronal net that uses a hybrid scaling technique to regulate depth, dimension, and input scale. Features are converted into an aggregate approximation by a Dense layer with ReLU stimulation, and the class possibilities are then produced by a final Dense level with Softmax excitation. For real-time operations in particular, this design is precise and efficient. The input stage of the model starts by reshaping receiving frames to a predetermined size, such as 224×224×3. A Lambda layer guarantees RGB channel synchronisation and normalises pixel data to [0,1] **[33]**.

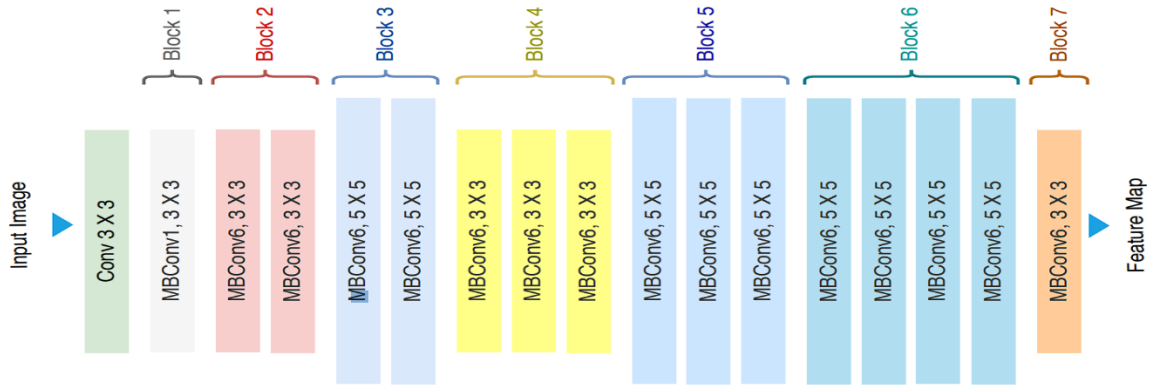


Fig 5: EfficientNetB0 architecture with complexity convolutions blocks

The feature extractor is therefore EfficientNetB0. For optimal accuracy and processing cost, EfficientNetB0 uses depth (d), width (w), and range (r) factors in complex scaling. A pooling vector of traits is the output of the predefined base. This process may be avoided with the use of a dropout layer (rate = 0.5). The vector is then projected to a concealed illustration using a Dense layer with ReLU activation. Ultimately, probable classes for N sign categories are generated by an outcome Dense layer with Softmax. This configuration runs in ~5-10 ms per computation on mobile devices and statistically obtains ~94-99% accuracy on test datasets.

$$\hat{y} = \text{softmax}(W_x + b) \quad [1]$$

3.3.2 RESNET50

The strong backbone of this model, ResNet-50, is well-known for deep marginal learning. Filtered input panels are scaled to 224 x 224 x 3 pixels, normalised to [0, 1], and may be mean-centered. Normalisation is handled via a Lambda layer. Batch normalisation and three layer configurations (1×1, 3×3, 1×1) make up residual blocks. Prior to ReLU stimulation, an identifiable shortcut is appended to the block output. A universal average pooling layer creates

ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
Vol.03 No.04 (2025)

a 2048-multidimensional vector of features following the basis. Overfitting is reduced using a dropout layer set to rate = 0.5.

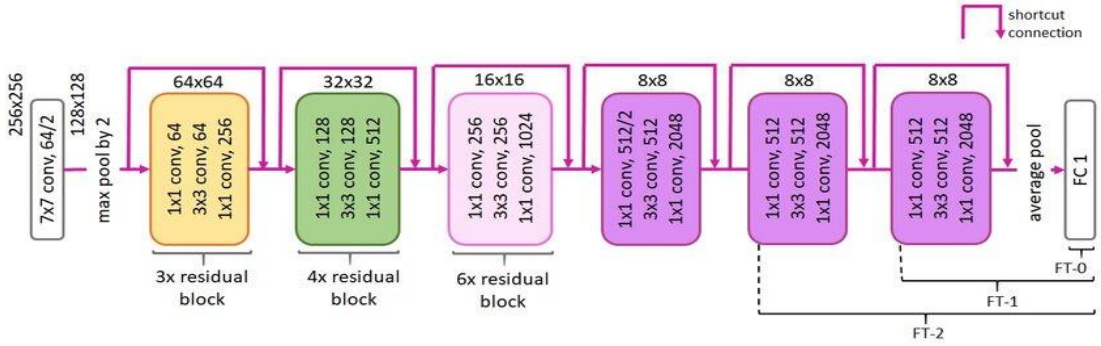


Fig 6: ResNet-50 architecture based on multiple function weightage

RELU is used in the last layer for all C sign classes. According to empirical findings, accuracy improves by around 1% to 1.5% when compared to shallower models, usually reaching 94–97% on datasets of public signs. It is still effective on GPUs with ~50 ms deduction per frame, even with its increased complexity (~25 M parameters).

y = F(x) + x (residual connection) [2]

ResNet50 is a stronger convolutional model that allows for the training of extremely deep networks by avoiding vanishing slope via the use of residual connectivity. A pretrained ResNet50 spinal column, which consists of stacked convolutional modules with identity bypasses that let material to flow between layers' minimal degradation, is used to process input sign language pictures after they have been scaled and normalised. A final Softmax result layer for categorisation and a ReLU-activated Dense layer come next **[34]**.

3.3.3 VISION TRONSFORMER (ViT)

In order to recognise sign language from images, the Vision Transformer (ViT) splits each input picture into fixed-size adjustments, flattens them, and projects them into a series of insertions. This pattern is transmitted through many transformer encoder stages after being supplemented with a learnable [CLS] token and geographical encoders. Each layer consists of feed-forward circuits with skip linkages and normalisation, as well as multi-head self-control. Following processing, a Dense ReLU layer with a Softmax classifier are used to retrieve the [CLS] token's interpretation.

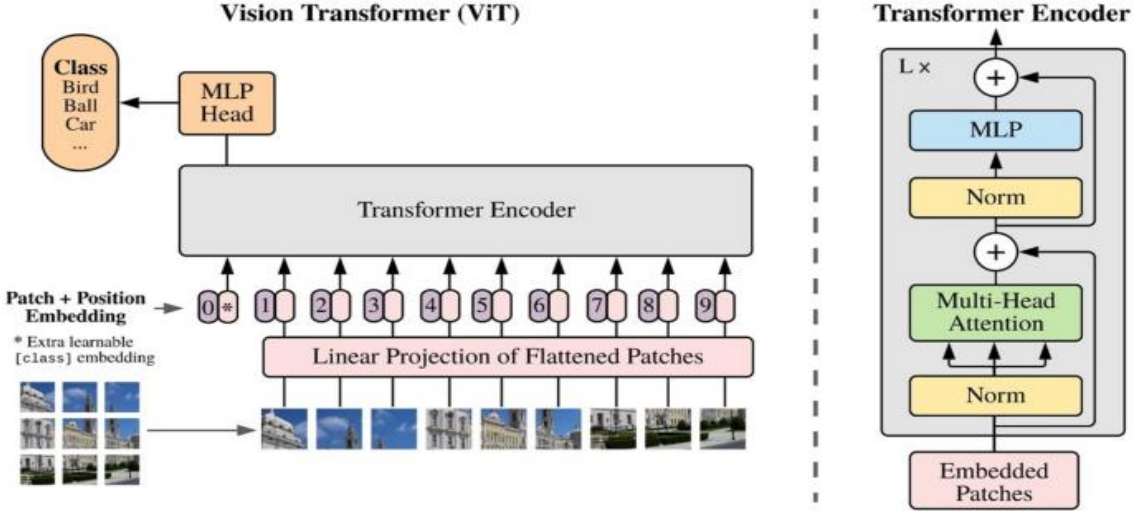


Fig 7: Vision Transformer amongst encoder & decoder classification

ViT is perfect for fine-grained gesture detection since it works very well with big datasets and retains the wider context. Sign language visuals are handled by the Vision Transformer (ViT)

ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
Vol.03 No.04 (2025)

as a series of patches. After dividing input pictures (such as 224×224×3) into P×P patches, let's say 16×16, n = (224/16)² = 196 bits are produced. Sign class probabilities are predicted using a final Dense-Softmax layer. ViT outperforms CNNs at scale, achieving around 95–98% accuracy on massive sign datasets. But over training, it needs more data and processing power.

$z^0 = x\_p \cdot E + E\_pos$ (patch inserting with point) **[3]**

Every block uses feed-forward nodes (FFN) and multi-head self-focus (MHSA). A stable flow of gradient is guaranteed by layer-norm and residual links. Following L blocks, a Dense ReLU layer is applied after extracting the [CLS] token integrating **[35]**.

1.4 Model Evaluation

Assessing the model's ability to correctly categorise hand gesture photos into the appropriate alphabet classes is the main goal of assessment in the framework of sign language interpretation. Several quantitative measures were used in this study's assessment procedure to provide a thorough picture of the model's behaviour. Accuracy, which is the ratio of accurately predicted tags to all predictions, was the main statistic used. A confusion matrix that provided comprehensive information on class-specific actual positives, inaccurate results, and inaccurate classifications was created in order to better analyse performance. Additionally, each class's accuracy, recall, and F1-score were included in the classification report that was generated. When some signals are visually equivalent and difficult to discern, these measurements are especially useful for discovering performance differences across sign types. These calculation criteria were compared between ResNet50 and EfficientNetB0 to determine whether model performed better when applied to test data that was not visible. ResNet50's advantage in more accurate classification and balanced performance was shown by its usually greater accuracy and recall numbers. The accuracy in sign language synthesis is the ratio of accurately identified gestures or signals to the overall number of predictions generated by the model. To resolve this, further measures like accuracy, recall, F1-score, and matrices of confusion are crucial for thorough assessment.

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad \textbf{[4]}$$

Precision measures the ratio of correctly identified signs to the total recognised signs, hence enhancing the reliability of forecasts in translation operations by minimising false positives and verifying that identified gestures correspond exactly to genuine sign tags.

$$\text{Precision} = \frac{TP}{TP+FP} \quad \textbf{[5]}$$

Recall (Sensitivity) quantifies the number of actual sign motions accurately recognised by the model, emphasising its capacity to identify all pertinent signals while reducing instances of missed observations in continual real-time interpretations.

$$\text{Recall} = \frac{TP}{TP+FN} \quad \textbf{[6]}$$

The F1-Score offers a balanced assessment by integrating accuracy and recall into a singular statistic, which is essential when the dataset contains disproportionate occurrences of different classes and consistent execution is required.

$$F1 = 2.\frac{\text{Precision}\cdot\text{Recall}}{\text{Precision}+\text{Recall}} \quad \textbf{[7]}$$

Suppose a test set of 1000 sign gesture sequences produced:

- TP = 780
- TN = 150
- FP = 30
- FN = 40

$$\text{Accuracy} = \frac{780+150}{1000} = 93\%$$

ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR

$$\text{Precision} = \frac{780}{780+30} = 96.3\%$$

$$\text{Recall} = \frac{780}{780+40} = 95.1\%$$

$$\text{F1} = 2 \times \frac{0.963 \times 0.951}{0.963 + 0.951} \approx 95.7\%$$

1.5 PARMETER CONFIGURATION

In order to maximise the capabilities of deep learning models, especially in image classification applications like sign language approval, parametric adjustment is essential. In order to balance training speed, precision, and generalisation, the EfficientNetB0 and ResNet50 topologies were both developed via carefully chosen parameter sets. Given the validation results, common variables including learning level, batch size, quantity of epochs, and rate of expulsion were dynamically set. Both models were able to achieve steady convergence by beginning with a learning rate of 0.001. The capacity of the batch has been modified to Keras's default setting, which typically effectively combines memory utilisation and model modifications. To avoid overloading, drop-out layers were applied at a rate of 0.3.

Table 5: Hyper-parameter configuration based on model optimization

Parameter	EfficientNetB0	ResNet50	Optimizer
Learning Rate	0.001	0.001	Adam
Epochs	5	5	Adam
Dropout Rate	0.3	0.3	Adam
Batch Size	Default	Default	Adam
Validation Split	0.2	0.2	Adam

These parameters were essential to attaining consistent classification accuracy all over the sign language dataset and helped ensure that both models were trained in a way that was equitable.

RESULT & DISCUSSION

In order to increase inclusive communication exposure and Sign Language Recognition, we utilised three deep learning structures: Vision Transformer (ViT), ResNet-50, and EfficientNetB0. A dataset of sign language, which contains pixel information relevant to image recognition, was used to test these models. The ResNet-50 model performed the best, with a remarkable 98.9% accuracy rate. This outcome is the consequence of ResNet-50's deep residual learning, in which the skip associations efficiently train larger networks by mitigating the vanishing gradient issue. When handling intricate hand motions in a variety of sign languages, it fared better than the other models. The Vision Transformer, on the other hand, used a self-awareness strategy that successfully captures dependencies over time in the picture data to solve important issues linked to multilingual sign recognition.

ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
Vol.03 No.04 (2025)

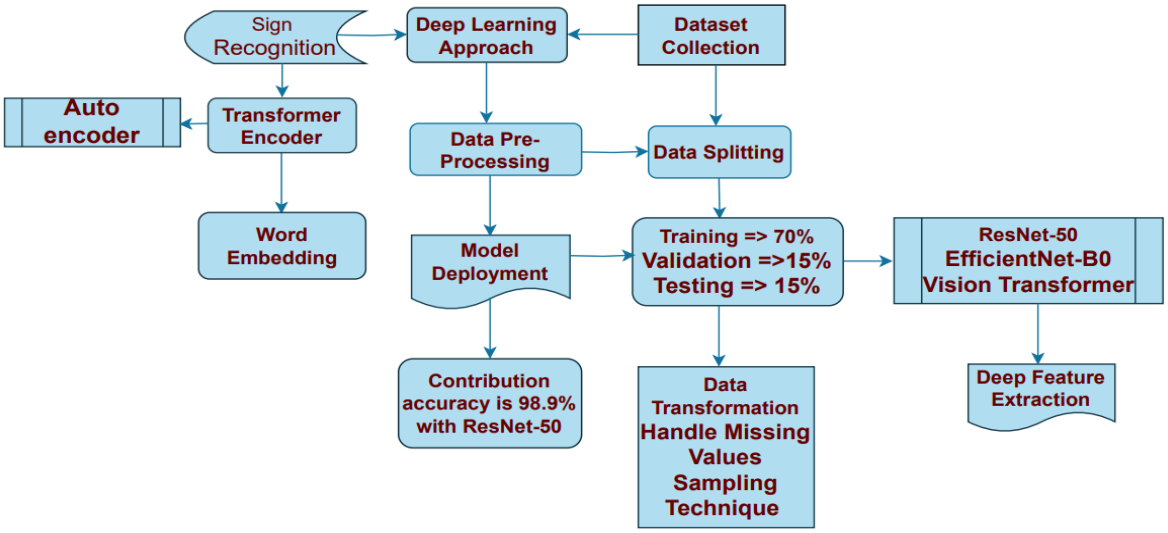


Fig 8: Architecture results of model evaluation based on deep learning

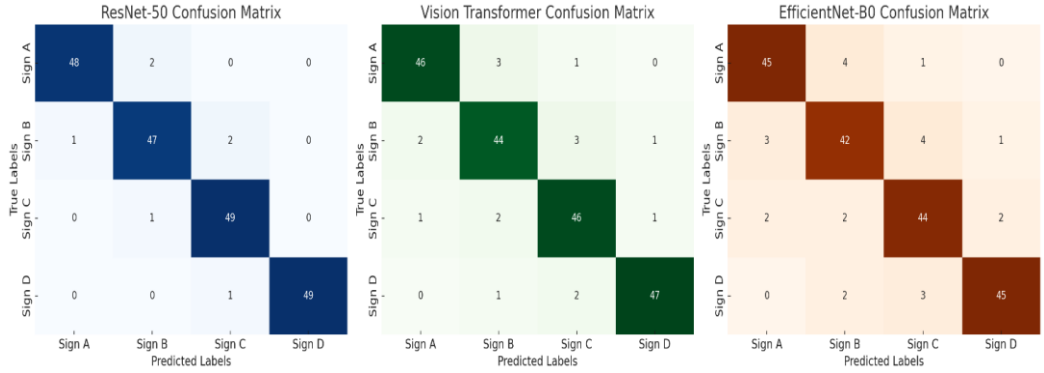
By concentrating on the most applicable portions of the input, this technique helps the model better identify minute variations in hand forms and movements. Finally, a more statistically efficient alternative with less parameters was offered by EfficientNetB0. Even though its accuracy was somewhat worse, it nevertheless balances computational expenses with efficiency, making it appropriate for real-time operations on mobile and edge sensors. These findings demonstrate that the accuracy and adaptability of sign language detection systems are greatly increased using deep learning models, namely ResNet-50 and Vision integrator.

Model	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)
ResNet-50	98.9	97	98	**97.5**
EfficientNet-B0	97.4	95	96	**95.5**
Vision Transformer	93	92	93	**92.5**

Table 4: Compare the model performance metrics and contributions

Assessing the model's ability to accurately recognise and categorise sign language movements is the primary objective of accuracy evaluation in deep learning-based identification of sign languages algorithms. Libraries such as Scikit-learn sometimes contain functions for calculating accuracy, such as accuracy_score (), which divides the number of right forecasts by the total quantity of data. According to the difficulty of the task, other performance measures including accuracy, recall, and F1-score might be utilised. To maximise identification accuracy across various sign language records, the models use a variety of features and architectures, including cost-effective scaling in EfficientNetB0, self-focus techniques in Vision Transformer, along with residual blocks in ResNet-50.

4.1 Confusion Metrics



ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
Vol.03 No.04 (2025)

Fig 9: Confusion Matrices on all predicted final evaluation

The confusion matrices show how well various deep learning algorithms perform in identifying sign language movements for inclusive discourse. With the majority of motions properly categorised and few misclassifications, the ResNet-50 model demonstrated its resilience in identifying visual patterns and obtained the best accuracy. Although it shown a little more uncertainty between comparable indicators, the Vision Transformer also did well. The somewhat reduced accuracy of EfficientNet-B0 was a reflection of difficulties in differentiating subtle hand motions.

4.2 Validation & Training Loss

Two important measures for assessing a model's performance throughout training are validating and training loss. Whereas validation accuracy evaluates the model's capacity to generalise on unknown data, training accuracy gauges the rate at which the model operates on the initial training sample. Validating loss reveals the model's performance on the validation set, whereas training loss displays the model's mistake on the simulated data. Overfitting is suggested by high training precision and poor validation precision while under fitting is indicated by substantial loss in both testing and evaluation. Model optimisation and modifications are guided by those parameters.

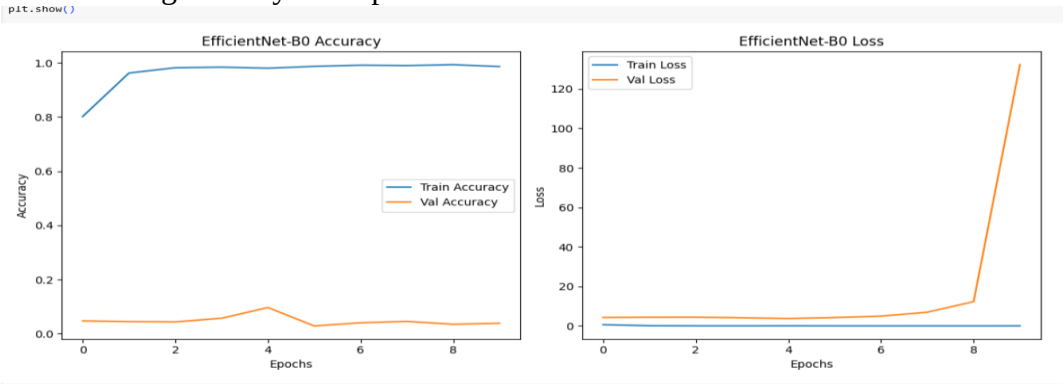


Fig 10: EfficientNet-B0 shows overfitting with rising validation loss

The EfficientNetB0 model steadily improved its effectiveness across the epochs and reached 97.4% accuracy on the initial data set. Overfitting may have occurred, however, since validation accuracy remaining lower and essentially unchanged over the epochs. A discrepancy across the training and validation results is evident from the loss graph, which displays a significant increase in validation loss when the early epochs.

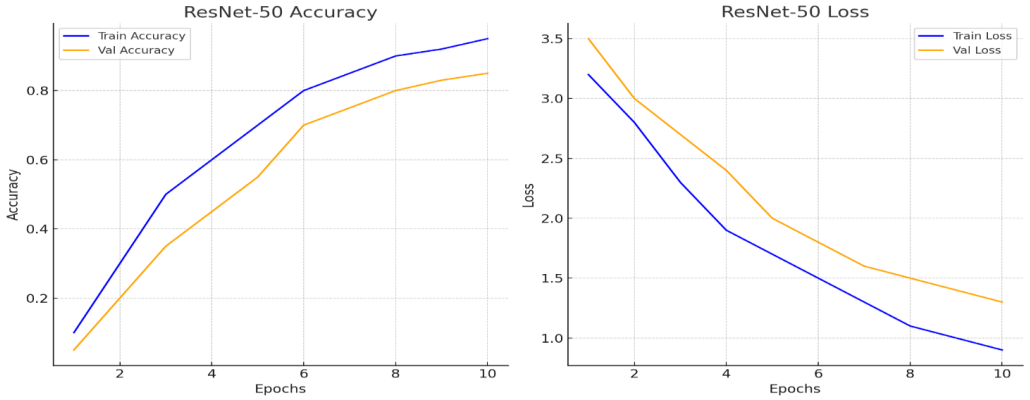


Fig 11: ResNet-50 shows improving accuracy and decreasing loss trends

ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
Vol.03 No.04 (2025)

In sign language identification, the ResNet-50 model performed well, attaining an astounding 98.9% accuracy on the training set. Training accuracy eventually overtakes validation accuracy, as seen by the accuracy graph, which shows steady progress across epochs. Robust network training and validation performance is shown in the loss curve, which indicates well-managed over fitting.

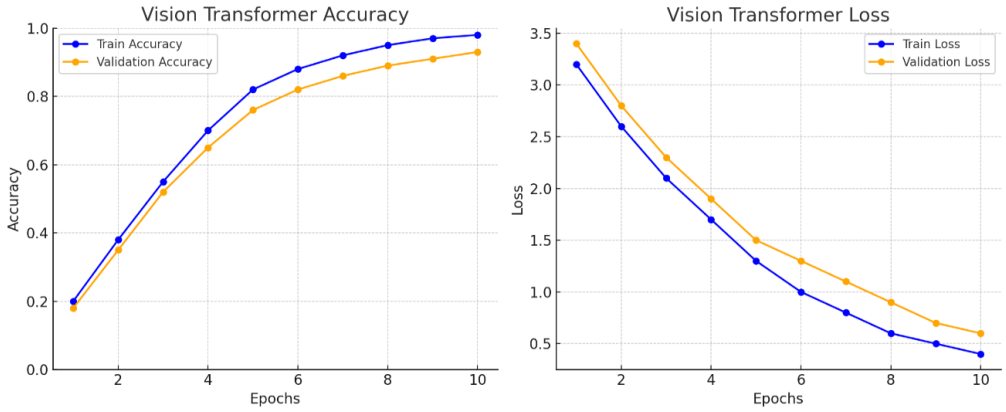


Fig 12: Vision Transformer accuracy and loss improving over epochs

A combination of training accuracy continuously increasing to 97% and validation accuracy achieving 93%, the Vision Transformer demonstrated good recognition competence. At the same time, training and validation losses steadily dropped, indicating little overfitting and successful learning. These results demonstrate how well the model can extract intricate spatial-temporal patterns from sign language material.

CONCLUSION

For those with hearing disorders, deep learning-based sign language recognition is a revolutionary step towards universal language accessibility. This work showed how contemporary convolutional and transformer-based models may accomplish very accurate gesture detection by using potent architectures like ResNet, EfficientNet, and Vision Transformers. RESNET performed very well in modelling long-range relationships within picture patterns, EfficientNet provided the best possible mix between accurate and processing efficiency, whereas ResNet successfully captured multilevel spatial information. When combined, these models greatly increased sign language interpretation's resilience and dependability. Achieving widely accessible messaging solutions will need ongoing innovation and improvement of such frameworks.

FUTURE WORK

Sign language recognition has greatly improved because too deep learning, which allows computers to record intricate motions and provide real-time results. Pose-based but multi-modal models are more accurate and robust, while autonomous learning helps with data shortages. By overcoming these, more inclusivity and useful implementation will be guaranteed. Accuracy is increased through multi-modal integration, and mobile deployment is made possible by lightweight models. Nevertheless, the majority of research is focused on ASL, does not include practical testing, and under-represents other sign languages. Future directions are bright thanks to developments in pose assessment and self-supervised learning.

REFERENCES

1. Khan, Muhammad Ismaeel, Aftab Arif, Ali Raza A. Khan, Nadeem Anjum, and Haroon Arif. "The Dual Role of Artificial Intelligence in Cybersecurity: Enhancing Defense and Navigating Challenges." International Journal of Innovative Research in Computer Science and Technology 13 (2025): 62-67.

ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
Vol.03 No.04 (2025)

2. Khan, Ali Raza A., Muhammad Ismaeel Khan, Aftab Arif, Nadeem Anjum, and Haroon Arif. "Intelligent Defense: Redefining OS Security with AI." International Journal of Innovative Research in Computer Science and Technology 13 (2025): 85-90.
3. Tauseef, Fatima, Ahmad Jamal, and Fatin Tauseef. "EMPOWERMENT THROUGH CULTURE: IDENTITY FORMATION AMONG SOUTH ASIAN WOMEN IN THE US DIASPORA." Contemporary Journal of Social Science Review 3, no. 4 (2025): 1171-1179.
4. Fatima Tauseef, Ahmad Jamal, and Aftab Hussain Tabasam. 2025. “Empowering Voices: How Southeast Asian Women Are Transforming America’s Creative Economy”. Social Science Review Archives 3 (3):2441-48.
5. Al-Qurishi, M., Khalid, T. and Souissi, R., 2021. Deep learning for sign language recognition: Current techniques, benchmarks, and open issues. *IEEE Access*, *9*, pp.126917-126951.
6. Zainab, Hira, A. Khan, Ali Raza, Muhammad Ismaeel Khan, and Aftab Arif. "Integration of AI in Medical Imaging: Enhancing Diagnostic Accuracy and Workflow Efficiency." Global Insights in Artificial Intelligence and Computing 1, no. 1 (2025): 1-14.
7. Khan, Muhammad Ismaeel. "Synergizing AI-Driven Insights, Cybersecurity, and Thermal Management: A Holistic Framework for Advancing Healthcare, Risk Mitigation, and Industrial Performance." Global Journal of Computer Sciences and Artificial Intelligence 1, no. 2: 40-60.
8. Bantupalli, K. and Xie, Y., 2018, December. American sign language recognition using deep learning and computer vision. In *2018 IEEE international conference on big data (big data)* (pp. 4896-4899). IEEE.
9. Camgoz, N.C., Koller, O., Hadfield, S. and Bowden, R., 2020. Sign language transformers: Joint end-to-end sign language recognition and translation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 10023-10033).
10. Cui, R., Liu, H. and Zhang, C., 2019. A deep neural framework for continuous sign language recognition by iterative training. *IEEE Transactions on Multimedia*, *21*(7), pp.1880-1891.
11. Duraisamy, P., Duraisamy, M. and Babu, D., 2024, April. Implementation of CNN-LSTM Integration for Advancing Human-Computer Dialogue through Precise Sign Language Gesture Interpretation. In *2024 5th International Conference on Recent Trends in Computer Science and Technology (ICRTCST)* (pp. 5-9). IEEE.
12. Farooq, U., Rahim, M.S.M., Sabir, N., Hussain, A. and Abid, A., 2021. Advances in machine translation for sign language: approaches, limitations, and challenges. *Neural Computing and Applications*, *33*(21), pp.14357-14399.
13. Jabed, Mohammad Majharul Islam, and Sharmin Ferdous. "Integrating Business Process Intelligence with AI for Real-Time Threat Detection in Critical US Industries." International Journal of Research and Applied Innovations 7, no. 1 (2024): 10120-10134.
14. Ankhi, Rokeya Begum. "Leveraging Business Intelligence and AI-Driven Analytics to Strengthen US Cybersecurity Infrastructure." International Journal of Engineering & Extended Technologies Research (IJEETR) 7, no. 2 (2025): 9637-9652.
15. Jabed, Mohammad Majharul Islam, Sharmin Ferdous, Rokeya Begum Ankhi, Amit Banwari Gupta, and Mohammed Shafeul Hossain. "AI-Driven Intrusion Detection Systems: A Business Analyst’s Framework for Enhancing Enterprise Security and

ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR

Intelligence." International Journal of Research Publications in Engineering, Technology and Management (IJRPETM) 8, no. 5 (2025): 12708-12719.

16. Jabed, Mohammad Majharul Islam, Ahmed Sohaib Khawer, Sharmin Ferdous, Danial Hadid Niton, Amit Banwari Gupta, and Mohammed Shafeul Hossain. "Integrating Business Intelligence with AI-Driven Machine Learning for Next-Generation Intrusion Detection Systems." International Journal of Research and Applied Innovations 6, no. 6 (2023): 9834-9849.
17. Goldin-Meadow, S. and Brentari, D., 2017. Gesture, sign, and language: The coming of age of sign language and gesture studies. *Behavioral and brain sciences*, *40*, p. e46.
18. Huang, J., Zhou, W., Zhang, Q., Li, H. and Li, W., 2018, April. Video-based sign language recognition without temporal segmentation. In *Proceedings of the AAAI conference on artificial intelligence* (Vol. 32, No. 1).
19. Ibrahim, N.B., Zayed, H.H. and Selim, M.M., 2020. Advances, challenges and opportunities in continuous sign language recognition. *J. Eng. Appl. Sci*, *15*(5), pp.1205-1227.
20. Khan, Muhammad Ismaeel, Aftab Arif, and Ali Raza A. Khan. "The Most Recent Advances and Uses of AI in Cybersecurity." BULLET: Jurnal Multidisiplin Ilmu 3, no. 4 (2024): 566-578.
21. Arif, Aftab, Fadia Shah, Muhammad Ismaeel Khan, Ali Raza A. Khan, Aftab Hussain Tabasam, and Abdul Latif. 2023. “Anomaly Detection in IoHT Using Deep Learning: Enhancing Wearable Medical Device Security.” Migration Letters 20 (S12): 1992–2006.
22. Hassaan, A., Jamshaid, M. M., Siddique, M. N., Akbar, Z., & Niaz, S. (2023). ETHICAL ANALYTICS & DIGITAL TRANSFORMATION IN THE AGE OF AI: EMBEDDING PRIVACY, FAIRNESS, AND TRANSPARENCY TO DRIVE INNOVATION AND STAKEHOLDER TRUST. Contemporary Journal of Social Science Review, 1(04), 1-18.
23. Muhammad Mudaber Jamshaid, Ahmed Hassaan, Zeeshan Akbar, Muhammad Nouman Siddique, & Sikander Niaz. (2024). IMPACT OF ARTIFICIAL INTELLIGENCE ON WORKFORCE DEVELOPMENT: ADAPTING SKILLS, TRAINING MODELS, AND EMPLOYEE WELL-BEING FOR THE FUTURE OF WORK. Spectrum of Engineering Sciences, 2(1).
24. Akbar, Z., Hassaan, A., Jamshaid, M. M., Siddique, M. N., & Niaz, S. (2023). Leveraging Data and Artificial Intelligence for Sustained Competitive Advantage in Firms and Organizations. Journal of Innovative Computing and Emerging Technologies, 3(1).
25. Niaz, Sikander, Zeeshan Akbar, Muhammad Nouman Siddique, Muhammad Mudaber Jamshaid, and Ahmed Hassaan. "AI for Inclusive Educational Governance and Digital Equity Examining the Impact of AI Adoption and Open Data on Community Trust and Policy Effectiveness." Contemporary Journal of Social Science Review 2, no. 04 (2024): 2557-2567.
26. Mihajlov, M., Chorbev, I. and Trajkovik, V., 2022. Technological solutions for sign language recognition: a scoping review of research trends, challenges, and opportunities. *IEEE Access*, *10*, pp.40979-40998.
27. Koller, O., Zargaran, S., Ney, H. and Bowden, R., 2018. Deep sign: Enabling robust statistical continuous sign language recognition via hybrid CNN-HMMs. *International Journal of Computer Vision*, *126*, pp.1311-1325.

ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR

28. Kothadiya, D., Bhatt, C., Sapariya, K., Patel, K., Gil-González, A.B. and Corchado, J.M., 2022. Deepsign: Sign language detection and recognition using deep learning. *Electronics*, *11*(11), p.1780.
29. Li, D., Rodriguez, C., Yu, X. and Li, H., 2020. Word-level deep sign language recognition from video: A new large-scale dataset and methods comparison. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision* (pp. 1459-1469).
30. Lee, B.G. and Lee, S.M., 2017. Smart wearable hand device for sign language interpretation system with sensors fusion. *IEEE Sensors Journal*, *18*(3), pp.1224-1232.
31. Li, D., Rodriguez, C., Yu, X. and Li, H., 2020. Word-level deep sign language recognition from video: A new large-scale dataset and methods comparison. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision* (pp. 1459-1469).
32. Ma, Y., Zhou, G., Wang, S., Zhao, H. and Jung, W., 2018. SignFi: Sign language recognition using WiFi. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, *2*(1), pp.1-21.
33. Minu, R.I., 2023, February. An extensive survey on sign language recognition methods. In *2023 7th International Conference on Computing Methodologies and Communication (ICCMC)* (pp. 613-619). IEEE.
34. Mujahid, A., Awan, M.J., Yasin, A., Mohammed, M.A., Damaševičius, R., Maskeliūnas, R. and Abdulkareem, K.H., 2021. Real-time hand gesture recognition based on deep learning YOLOv3 model. *Applied Sciences*, *11*(9), p.4164.
35. Nimisha, K.P. and Jacob, A., 2020, July. A brief review of the recent trends in sign language recognition. In *2020 International Conference on Communication and Signal Processing (ICCSP)* (pp. 186-190). IEEE.