Journal of Artificial Intelligence and Technology, (Ahead of Print)
https://doi.org/10.37965/jait.2026.0984

ISTP
RESEARCH ARTICLE

Pseudo-Temporal 3D CNN Fusion of Gradient and Deep Spatial Features for Hand Gesture Recognition

M. Keerthi Kumar and B. D. Parameshachari
Department of Electronics and Communication Engineering, Nitte Meenakshi Institute of Technology, Nitte (deemed to be a university), Yelahanka, Bengaluru, India

(Received 27 October 2025; Revised 02 May 2026; Accepted 18 May 2026; Published online 20 June 2026)

Abstract: Communication between people with disabilities and those who do not understand sign language is a growing social need and a challenging task. The usage of deep learning (DL) techniques acts as a gateway for people with communication impairments to bridge the communication gap. This research develops an integrated approach using DL architectures to recognize hand images and facilitates effective communication. Features from the raw data are extracted using the histogram of oriented gradients (HOG). HOG evaluates the magnitude and orientation of the gradient of the input image based on its outline, which is used as the edge direction. The extracted features are classified using the proposed integrated model, which comprises MobileNet V2 and a three-dimensional convolutional neural network (3D CNN). MobileNet V2 is utilized for extracting spatial features, while the 3D CNN detects spatial data in three dimensions to facilitate better classification accuracy. The proposed model fuses HOG-based gradient descriptors with deep spatial features from MobileNetV2 using a pseudo-temporal 3D CNN, enabling superior static sign language recognition. Experimental analysis shows that the proposed method achieves an accuracy of 99.55%, which is higher than that of existing techniques.

Keywords: Deep learning; hand gesture; histogram of oriented gradients; integrated model; sign language recognition

I. INTRODUCTION

Hand gestures play an important role in sign language recognition. Hand gestures are a significant component of human–computer interaction (HCI) in areas such as virtual reality, monitoring, and detection systems [1–3]. Human gesture recognition is specifically used in the communication process between deaf individuals and other people. Techniques used in hand gesture recognition are classified as either sensor- or vision-based. In the sensor-based systems, gloves with sensors are used to describe hand and finger gestures [4,5]. However, these systems are expensive and cannot be used in varying environments. In vision-based systems, hand gesture recognition is performed using cameras to recognize motions [6]. Sign language has a unique linguistic structure that varies from one region to another. Hand gestures are a fundamental tool for sign language communication. Unlike other natural languages, sign language utilizes body motions to communicate messages through gestures [7,8]. Each symbol denotes a letter, word, or emotion, demonstrating that sign language is a fully formed natural language with grammar and sentence structure [9]. Most of the technologies introduced in sign language recognition systems are being utilized as an assistive tool for hearing-impaired people. Many algorithms based on machine learning and deep learning (DL) techniques provide a gateway to achieving superior recognition accuracy in sign language [10,11]. Existing DL architectures, such as convolutional neural network (CNN), help to learn relationship characteristics among images during training; however, this remains a challenging task in gesture recognition. CNN-based DL architectures exhibit commendable performance in scenarios involving complex backgrounds, illumination, and varying sizes of images [12]. Egocentric hand gestures involve fingertips, where hand gestures are a significant factor in identifying the fingerprint's position [13]. Moreover, CNN architectures are known for their capability to analyze complex patterns, eliminating the need to develop handcrafted features [14,15]. Unlike traditional sign language recognition algorithms that rely on handcrafted features or DL models independently, this work introduces a pseudo-temporal feature fusion strategy using a three-dimensional CNN (3D CNN) for static image classification. The proposed model integrates histogram of gradients (HOG)-based gradient descriptors with deep spatial features captured from MobileNetV2, where concatenated features are considered as a pseudo-temporal sequence. This enables the 3D CNN to learn inter-feature dependencies rather than temporal dynamics. This adaptation enhances feature representation and improves recognition performance while maintaining computational efficacy.

II. LITERATURE REVIEW

This section presents recent research based on sign language recognition from hand gestures using different techniques, along with their advantages and drawbacks.

Ma *et al.* [16] developed the two-stream mixed (TSM) method with feature extraction and fusion to enhance the correlation among the images for dynamic gestures. The TSM-CNN was comprised of a preprocessing block for image resizing and data augmentation. Moreover, the fusion feature map acquired by concatenation in the

Corresponding author: M. Keerthi Kumar (e-mail: keerthim.k123@outlook.com/ keerthim.k123@gmail.com).

TSM block was utilized as input for classifiers. Finally, the classification took place with the help of a CNN. The suggested TSM was employed as an effective recognition system for hearing-impaired individuals. However, the computational complexity of the model was relatively higher due to the combination of architectures. Fregoso *et al.* [17] suggested a hybrid technique where particle swarm optimization (PSO) was utilized in optimizing the parameters of CNN. The suggested technique experienced two types of approaches based on the convolutional layer and batch size. First, the reliability of parameters among each layer was observed, where the objective function denoted the recognition rate. In the second one, the variation occurred in the convolutional layer, while the objective function offered a better recognition rate. The utilization of the CNN architecture alone did not provide better results due to the poor detection of optimal CNN parameters. Jain *et al.* [18] developed an effective recognition framework with support vector machines (SVMs) and CNNs. This research developed an optimistic filter size for single- and double-layer CNNs. Initially, the preprocessing was performed to clean the data, and the features from a dataset were captured by SVM with several kernels, and on the other hand, the CNN with single and double layers was employed for training the model. However, some errors occurred while recognizing the gestures for each channel in the architecture. Xiao *et al.* [19] introduced a modified Capsule Network (CapsNet) known as CapsNet for sign language recognition. The CapsNet was an alternative to CNN that considered the spatial relationship and feature orientations of entities. CapsNet offered high generalization ability and was robust in detecting the routing iteration for varying numbers of hand gestures. Further, the suggested framework was effective in handling the sign language digits and alphabet in a similar period. Nonetheless, the suggested architectural framework was incapable of addressing dynamic sign language recognition tasks. Bousbai *et al.* [20] suggested an effective hand gesture recognition framework using an ensemble approach of CNN and CapsNet. The ensembling architecture of CNN and CapsNet was utilized in the process of improving the efficiency of the recognition systems. The combination of outcomes achieved from the suggested framework exhibited that the ensembling of the model with minimal augmentation of data provided optimal outcomes. The suggested model minimized dimensionality by using principal component analysis (PCA), while the classification took place with the help of SVM. Yet, the ensembling of DL approaches was only effective for large amounts of data due to the huge training memory.

Selda Bayrak *et al.* [21] presented a complex-valued deep neural network (CVDNN) model able to process a feature vector composed of complex numbers across layers. CVDNN was a strong method able to address complex optimization problems of conventional deep networks more effectively. CVDNNs that utilize complex numbers as input data and complex activation functions in every layer have obtained superior performance in robotic systems, biometric technologies, disease diagnosis, and telecommunications.

Vasileios Kouvakis *et al.* [22] introduced a novel method for image-based semantic communications that used a variant of the quadrature amplitude modulation (QAM) scheme. This modulation scheme was developed from actual 32-QAM by eliminating eight peripheral symbols and was able to obtain superior error performance in American Sign Language (ASL) applications. Moreover, the semantic encoder depended on a CNN that efficiently used the ASL alphabet.

Ismail Taha Ahmed *et al.* [23] developed a hand gesture recognition model, which integrated the Residual Network-50 (ResNet50) model feature extraction with the Tamura texture descriptor and utilized the adaptability to represent intricate interactions among features. Experiment results were done by a publicly available dataset that includes images of ASL gestures.

Abdullah Baihan *et al.* [24] introduced SLR by modifying DL and a hybrid optimization algorithm. The spatial and geometric-based features were captured by the Visual Geometry Group 16 (VGG16), and motion features were captured by the optical flow algorithm. The CNNSa-LSTM was an integration of CNN, self-attention (SA), and long short-term memory (LSTM) for identifying sign language. The introduced method was developed to feature extraction through integrating CNNs for spatial analysis to SA mechanisms to focus on relevant features, while LSTMs efficiently capture temporal dependencies. The CNNSa-LSTM method improved performance in tasks including complex, sequential data.

Overall, the existing methods face issues related to poor classification accuracy, with the inability to retain detailed information and recognition errors. The aforementioned issues are overcome with the proposed integrated model, which is the combination of DL architectures. By considering the advantageous properties of CNN, this research puts forward an effort to build an enhanced CNN architecture for sign recognition from hand gestures. In this research, the integration of HOG-based handcrafted gradient descriptors with deep spatial features from MobileNetV2 is fused by a 3D CNN in a pseudo-temporal model for static sign language recognition. Unlike conventional models that employ 3D CNNs to sequential data, this manuscript uses 3D convolutions to process stacked feature maps from complementary sources, enabling the capture of complex inter-feature relationships in static images. This fusion strategy uses HOG's precision in gradient and edge-based representation in MobileNetV2 for high-level spatial abstraction, obtaining superior recognition accuracy with minimized computational complexity.

A. PROBLEM STATEMENT

Existing sign language recognition methods struggle to capture the spatial and structural data in static hand gesture images, resulting in reduced accuracy and generalization ability. When 3D CNNs are generally utilized for temporal sequence extraction, their application for fusion features in static image classification remains unexplored.

B. OBJECTIVE

This manuscript aims to develop an integrated DL model combining HOG-based gradient descriptors, MobileNetV2 for spatial feature extraction, and 3D CNN-based feature fusion to improve recognition accuracy for static sign language gestures.

C. CONTRIBUTION

The major findings of this research are listed as follows:

- This research develops an integrated model using two DL techniques, MobileNet V2 and 3D CNN.
- The HOG-based feature extraction approach is utilized to evaluate the magnitude and orientation of gradients in the input image. The HOG features imply which object and its outline are used in assessing edge directions.

- MobileNet V2 extracts spatial features, while the 3D CNN concatenates these features, considering their 3D relationships and helping in recognizing complex image patterns.
- This research incorporates pseudo-temporal 3D CNN fusion of HOG and MobileNetV2 features, extracting complex inter-feature relationships in static images for highly precise sign language recognition.

The remainder of the manuscript is structured as follows: The proposed methodology of this research is detailed in Section II. Section III presents the experimental outcomes achieved by assessing the proposed method. While the conclusion of this research is described in Section IV.

D. MODEL SELECTION

The developed model is designed to leverage the advantages of HOG, MobileNetV2, and 3D CNN. HOG is chosen for its ability to extract fine-grained gradient and edge data, which is essential for differentiating subtle variations in hand shapes. MobileNetV2 is adopted for its effective extraction of high-level spatial features with low computational cost, making it suitable for lightweight models. The 3D CNN is incorporated to fuse feature maps along a pseudo-temporal depth dimension, enabling the network to capture complex inter-feature relationships that conventional 2D convolutions overlap. This integration is chosen to increase performance while maintaining computational efficacy for static sign language recognition.

III. PROPOSED METHODOLOGY

This research develops an effective sign language recognition scheme using an integrated DL approach comprised of MobileNetV2 and 3D CNN. The overall process takes place in three stages: data collection, feature extraction, and classification. This section deliberates on the detailed process involved in the recognition of sign language using the proposed model. The workflow of the process involved in sign language recognition in the integrated model is presented in Fig. 1.

A. DATA ACQUISITION

In this research, the data are collected from the sign language MNIST [25] dataset, similar to the classic MNIST dataset. Each individual training and testing case denotes a label in a one-to-one map, in alphabetical order. Moreover, it has 25 classes, with 1350 instances per class, and the images in the dataset have dimensions of $28 \times 28 \times 28$ pixels, with grayscale values between 0 and 255. Each row in a dataset comprises 25 labels, from 0 to 24, that denote ASL data, corresponding to 784 pixel values. Although the sign language MNIST dataset is mainly used as a benchmark for hand gesture recognition, it has low-resolution grayscale images with minimal background complexity. However, it provides a controlled environment for evaluating the effectiveness of feature extraction and classification mechanisms. In this manuscript, it is used to validate the ability of the proposed pseudo-temporal feature fusion algorithm under standardized conditions. In addition to the sign language MNIST dataset, a more complex ASL dataset with variations in lighting, background, and hand shapes is considered. While MNIST provides a controlled benchmark for validating the model, the ASL dataset is utilized to evaluate the robustness and generalization ability of the proposed algorithm. A sample of the

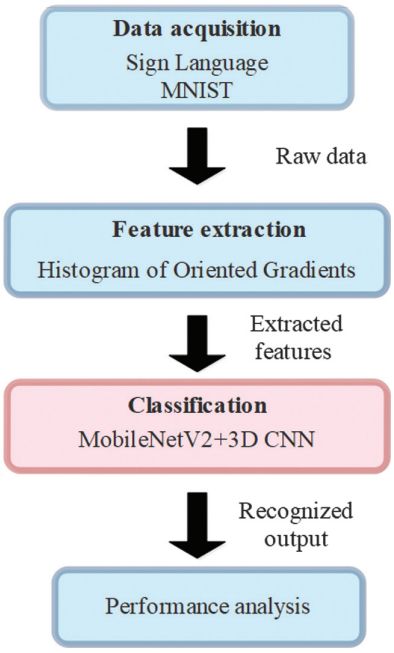

Fig. 1. Process involved in the recognition of hand gestures using MobileNetV2 + 3D CNN.

Table I. Sample of sign language MNIST dataset

Label	Pixel 1	Pixel 2	Pixel 3
3	107	118	127
6	155	157	156

MNIST dataset is listed in Table I, and the image samples are presented in Fig. 2.

1). ASL DATASET. This dataset is a collection of images from the ASL [26], which contains 26 letters and 3 gesture signals, such as space, delete, and nothing. This dataset contains 87,000 images with 3,000 uniformly distributed images for each class under different lighting, backdrop, and hand-shape conditions.

Fig. 2. Sample image from sign language MNIST.

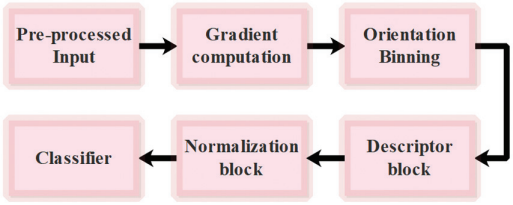

Fig. 3. The process of HOG feature extraction.

B. FEATURE EXTRACTION

The next stage of data collection involves extracting features from the raw data using the histogram of oriented gradients (HOG) [27,28]. In this research, preprocessing is not performed to preserve the originality of the data, as only minor distortions are present in the images. Therefore, this research directly uses HOG feature extraction, which helps evaluate the level and direction of the gradient of the input image. The HOG features capture the object and its outline, which are used to assess edge directions. HOG feature extraction is performed by segmenting an image window into 64 blocks, where each block is composed of 2×2 cells. The process involved in HOG-based feature extraction is diagrammatically presented in Fig. 3. The manipulation is performed for each pixel based on the orientation of edges and the direction of the gradient. The HOG feature extraction process is described below.

1). COMPUTATION OF GRADIENT. The gradient computation is performed to obtain the image gradient by shifting it with horizontal and vertical dimensions of distinctive masks based on Equation (1). The horizontal and vertical masks are denoted as D_x and D_y, respectively. Moreover, the derivatives x and y are evaluated based on the convolution operation presented in Equation (2). The gradient of the image is represented as I, and the gradient of x and y derivatives are represented as I_x, I_y, respectively. The magnitude of the gradient is represented as $|G|$, which is evaluated using Equation (3). Next, the orientation of the gradient is evaluated based on Equation (4), which helps in figuring out the accurate angle of the images, which is effective in the precise recognition of hand gestures:

$$D_x = [-1 \quad 0 \quad 1] \; and \; D_y = [1 \quad 0 \quad -1] \tag{1}$$

$$I_x = I \times D_x \; and \; I_y = I \times D_y \tag{2}$$

$$|G| = \sqrt{I_x^2 + I_y^2} \tag{3}$$

$$\theta = arctan\left(\frac{I_x}{I_y}\right) \tag{4}$$

2). ORIENTATION BINNING. After the computation of the gradient, cells for the histograms are generated. Here, each pixel value is evaluated based on the histogram channel. The cells are rectangular, and the histogram channels are spread over 0° to 180° or 0° to 360° based on a signed or unsigned gradient.

3). DESCRIPTOR BLOCKS. The descriptor blocks are used in the regional normalization process, which helps in varying the illumination, color, and the strength of the gradient. Regional normalization is performed to group the cells into a spatially connected block. When the aforementioned steps are performed in normalization, the histograms are integrated into a single feature vector [29,30]. Here, v is considered the non-normalized vector comprised of histograms in the block $[v_k]$, where $k = 1,2$ and e is the smaller constant. The factor f is known as a normalization factor, which is computed based on the numerical Equations (5) and (6). Here, v is the non-normalized vector and e is a small constant, which is known as the smaller constant value. After the HOG-based feature extraction stage, different classes of hand gestures are recognized using the proposed classification model:

$$\text{L1 } norm: f = \frac{v}{[v] + e} \tag{5}$$

$$\text{L2 } norm: f = \frac{v}{\sqrt{[v]^2 + e^2}} \tag{6}$$

The features of the proposed model are extracted from HOG and MobileNetV2, which are not utilized individually but concatenated to develop a unified representation, which is next processed through a 3D CNN for fusion and classification. MobileNetV2 captures high-level spatial features from its global average pooling layer, while HOG extracts gradient and edge-based descriptors. The concatenated feature maps are considered as a 3D tensor, where the third dimension is considered as pseudo-temporal depth, enabling 3D CNNs to extract inter-feature dependencies for temporal extraction in video data. This proposed model allows the network to exploit complementary spatial and gradient-based data, as the inputs are static images.

C. CLASSIFICATION USING THE INTEGRATED MODEL

Following the stage of HOG-based feature extraction, the classification takes place with the help of the proposed integrated model of MobileNetV2 and a 3D CNN. The MobileNetV2 is utilized to deal with the spatiotemporal data, and the 3D CNN is utilized in concatenating the features from HOG and 3D CNN, thereby representing the three dimensions of width, height, and time. The HOG features imply the object and its outline, which are used in assessing the edge directions. Similarly, MobileNetV2 is utilized in extracting the spatial features. Thus, the features extracted from HOG and MobileNetV2 are provided for the stage of classification, which is performed using a 3D CNN. The HOG extracts features based on the grayscale level and gradient of the image. The features from the MobileNetV2 are extracted using the average pooling layer. The MobileNetV2 is utilized in the process of spatial extraction of features with the dimensions of $1 \times 8 \times 8$. Moreover, the 3D CNN is utilized in recognizing the precise spatial data and helps in better classification by considering the three-dimensionality. The convolutional layer comprises 3×3 with two strides. The separable operation is performed in MobileNet using depthwise and pointwise convolution layers. Moreover, batch normalization and rectified linear unit (ReLU) activation operations are employed after each convolution layer. While extracting the spatial features, these depthwise separable convolution layers have a significant role in the convolutional operation. Based on this, the length and width of a graph and kernel remain the same. The evaluation of standard convolution (St_c) and the separable convolution (Se_c) is numerically represented in Equations (7) and (8). By evaluating the standard and separable convolution, the overall assessment is mathematically represented in Equations (9) and (10):

$$St_c = Df \times Df \times Dk \times Dk \times M \times Df \times Df \times M \times N \quad (7)$$

$$Se_c = Df \times Df \times Dk \times Dk \times M \times N \quad (8)$$

$$\frac{St_c}{Se_c} = \frac{Df \times Df \times Dk \times Dk \times M \times Df \times Df \times M \times N}{Df \times Df \times Dk \times Dk \times M \times N} \quad (9)$$

$$\frac{St_c}{Se_c} = \frac{1}{N} + \frac{1}{Dk^2} \quad (10)$$

Where the total count of input and output channels is presented as M and N, the side length of the graph and the convolutional kernel are represented as *Df* and *Dk*, respectively. The optimization of basic units in the network is performed using the linear bottleneck and inverted residual structures. Bottleneck layers present in the architecture are comprised of residual bottleneck separable convolution layers. The final layer of MobileNetV2 is comprised of a global average pooling layer to aggregate the spatial data, along with a fully connected layer and the softmax function.

The architectural presentation of the proposed 3D CNN is presented in Fig. 4. After the extraction of spatial features from the fully connected layers of MobileNet V2, the 3D CNN model is utilized in recognizing spatial data, which is utilized in classifying hand gestures. Hence, the extracted features are based on three dimensions, whereas the 3D CNN serves as an effective approach in recognizing the complex patterns with better accuracy values. The 3D CNN is comprised of partial connections and weight sharing. The neuron at the output layer is connected to a local input region, which helps to diminish the number of parameters and minimize the issues related to overfitting. Moreover, the convolutional layer helps to retain the spatial patterns, which is a major factor in hand gesture recognition. The weight-sharing property of the integrated model shares the weight of the convolutional kernel across the spatial region and helps in improving the generalization ability of the model. The pooling layer is introduced among the successive convolutional layers to minimize the spatial size of feature maps. The 3D convolutional layer is effective in learning local patterns in the network. The usage of more kernels in the network enhances the ability to learn deeper features, further facilitating the recognition of the hand gestures. The 3D CNN effectively analyzes the changes in hand position and movement, resulting in understanding the temporal dynamics of sign language gestures. Moreover, it is well suited to capture the spatial relationship and inherit complex patterns in sign language gestures. Thus, the concatenation of features using a 3D CNN helps in recognizing the three-dimensionality so as to achieve preferable results in recognizing hand gestures with commendable accuracy. The sign language MNIST is static, grayscale 28×28, and without temporal sequences; the proposed model uses a 3D CNN in a non-traditional manner by introducing pseudo-temporal depth from stacked feature channels. The spatial features from MobileNetV2 and gradient descriptors from HOG are concatenated to develop a 3D tensor. The third dimension does not represent original time steps but a structured arrangement of complementary feature maps. 3D convolutional kernels process across this depth, enabling the model to learn joint spatial-inter-feature relationships, which a 2D CNN is unable to capture efficiently. This model treats various feature scores to temporal slices, allowing richer feature fusion and enhanced recognition performance even in static images.

IV. RESULTS AND DISCUSSION

This section presents the results achieved by the integrated DL model by evaluating the outcomes based on its efficiency in recognizing hand gestures. The Python 3.10.12 software is utilized in an implementation of the proposed model, and the system is specified with the configuration of Windows 10 OS, Intel Core i5, and 8 GB of random access memory. The suggested framework is assessed based on accuracy, precision, recall, and F1-score. Accuracy: It is defined as the total count of correctly categorized data divided by the total number of data, which is evaluated using Equations (11–14). Precision: It is defined as the ratio of correct predictions to the total number of correct classes. Recall: It is defined as a ratio of the total number of positives that are properly categorized as positive to the total number of positive and negative values. F1-score: It is assessed by utilizing both recall and precision; when the values of precision and recall are 1, the F1-score is one:

$$Accuracy = \frac{TN + TP}{TN + FP + TP + FN} \quad (11)$$

$$Precision = \frac{TP}{TP + FP} \quad (12)$$

$$Recall = \frac{TP}{TP + FN} \quad (13)$$

$$F1score = 2 \times \frac{Precision * Recal}{Precision + Recall} \quad (14)$$

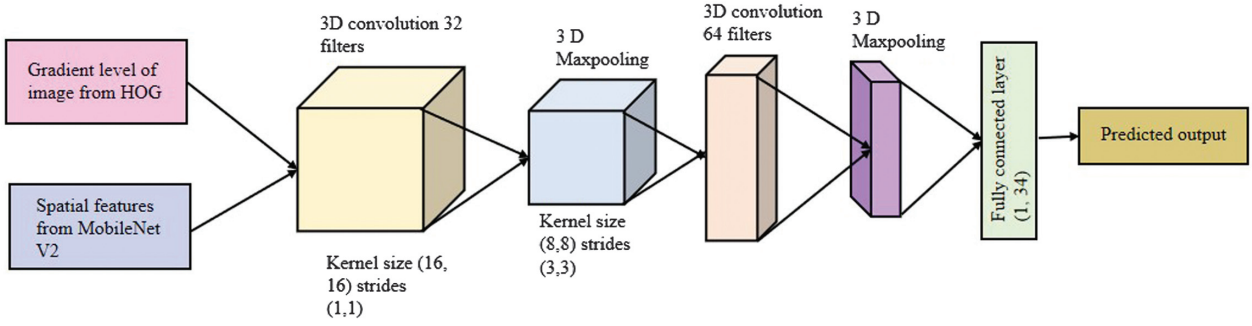

Fig. 4. Architecture of the proposed 3D CNN.

A. PERFORMANCE ANALYSIS

This subsection presents the outcomes achieved by the suggested integrated model and the feature extraction method utilized in mining out the relevant features. The evaluation is performed based on different feature extraction techniques and different classifiers.

The existing techniques, namely Gray-Level Co-occurrence Matrix (GLCM), Speeded Up Robust Features (SURF), Local Phase Quantization (LPQ), and Local Binary Pattern (LBP), are used for comparison. The experimental outcomes displayed in Table II exhibit the efficiency of the model based on different feature extraction techniques utilized in mining data from the sign language MNIST dataset. From the obtained outcome, it is seen that the HOG method achieves a better accuracy of 99.55%, as opposed to GLCM, SURF, LPQ, and LBP with accuracies of 91.33%, 94.50%, 89.55%, and 85.67%, respectively. The HOG features aid in capturing the orientation of local gradients and help in effectively detecting hand gestures. Moreover, the HOG descriptors help in encoding the significant shape and texture, which enables them to recognize gestures. The graphical depiction of results based on different feature extraction techniques for the proposed classifier is presented in Fig. 5.

1). EVALUATION BASED ON DIFFERENT CLASSIFIERS. In this subsection, the outcomes are evaluated based on different classifiers used in the classification of hand gestures. The state-of-the-art classifiers, namely Gated Recurrent Unit (GRU), ResNet50, MobileNetV2, and VGG16, are utilized in assessing the efficiency of the suggested classifier. Table III presents the experimental outcomes based on different classifiers for recognizing hand gestures. The experimental results in Table III exhibit the efficiency of classifiers while recognizing hand gestures.

Table II. Evaluation of outcome based on different feature extraction techniques

Method	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
GLCM	91.33	93.00	91.00	92.00
SURF	94.50	96.00	95.00	96.50
LPQ	89.55	91.00	88.00	90.00
LBP	85.67	86.00	84.00	85.00
HOG	99.55	99.99	99.80	99.90

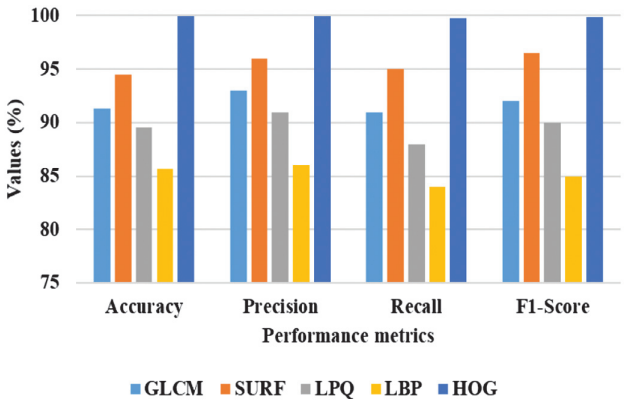

Fig. 5. Evaluation of results based on different feature extraction for the proposed classifier.

Table III. Evaluation of outcome based on different classifiers

Classifiers	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
GRU	95.00	95.50	94.50	95.00
ResNet50	98.00	98.00	98.00	98.00
MobileNetV2	96.00	97.00	95.00	96.00
VGG16	86.50	87.00	86.00	86.50
Integrated model	99.55	99.99	99.80	99.90

The integrated model achieves superior results for overall metrics considered for evaluation. Specifically, the classification accuracy of the proposed integrated model is 99.55%, whereas the state-of-the-art classifiers, such as GRU, ResNet50, MobileNetV2, and VGG16, attain commendable classification accuracies of 95%, 98%, 96%, and 86.50%, respectively. These outcomes of the integrated model are higher as a result of the combination of MobileNet V2 for extracting spatial features, whereas 3D CNN is employed for capturing temporal features helpful for the effective classification of hand gestures. The graphical depiction of results based on the proposed classifier is presented in Fig. 6.

2). CROSS-FOLD VALIDATION. In this section, the efficiency of the suggested framework is evaluated for different K-values of 2, 3, 5, and 10. Table IV presents the experimental results achieved by the suggested approach when evaluated with state-of-the-art classifiers for different K-values. The cross-fold validation attains preferable results for the classifier when trained with a specified training and testing ratio. The proposed integrated model displays optimal results when the K-value is assigned as 10 because the training and testing take place in the ratio of 80% and 20%. Training the model in the ratio of 80:20 helps retain the continuity of features for better classification. The accuracy of the integrated model is 99.85% when the value of K is assigned as 10, whereas the evaluation based on other K-values shows poor accuracy.

Table V presents the statistical analysis performance of the proposed model across four primary metrics, such as mean values, variability, confidence intervals, and statistical significance. The proposed model obtained high mean values with lower standard deviations, showing stable and consistent performance across folds. The confidence intervals depict high reliability of results when p-values below 0.05 show that enhancements over the baseline model are statistically significant. These results exhibit

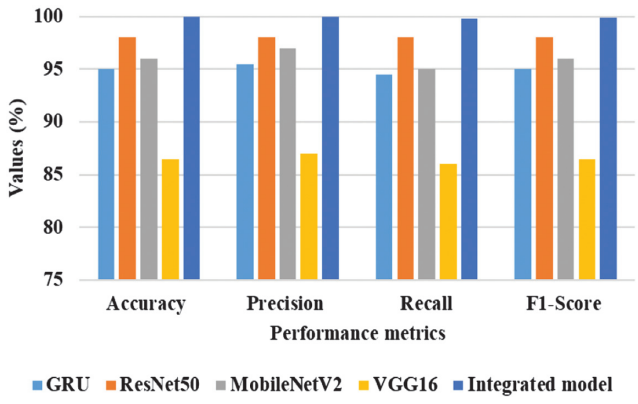

Fig. 6. Evaluation of the results based on different classifiers.

Table IV. Evaluation based on cross-fold validation

K-fold	Classifiers	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
$K = 2$	GRU	93.45	93.44	93.56	93.5
	ResNet50	96.85	96.92	96.75	96.83
	MobileNetV2	94.98	95.8	94.58	95.19
	VGG16	84.89	84.90	85.00	84.95
	Integrated model	97.78	98.94	98.70	98.82
$K = 3$	GRU	94.75	94.10	94.20	94.15
	ResNet50	97.20	97.50	97.50	97.50
	MobileNetV2	95.00	96.00	95.00	95.50
	VGG16	85.24	85.20	85.50	85.35
	Integrated model	98.90	98.10	98.10	98.10
$K = 5$	GRU	93.80	94.99	94.86	94.92
	ResNet50	98.20	97.85	98.20	98.02
	MobileNetV2	95.89	97.00	96.00	96.50
	VGG16	85.60	86.00	85.00	85.50
	Integrated model	98.68	98.85	98.45	98.65
$K = 10$	GRU	94.85	95.55	95.00	95.27
	ResNet50	98.40	98.90	98.50	98.70
	MobileNetV2	96.10	97.89	95.99	96.94
	VGG16	86.90	87.30	86.30	86.85
	Integrated model	99.85	99.70	99.40	99.55

Table V. Statistical analysis of the proposed model

Metrics	Mean (%)	Standard deviation (%)	95% confidence interval (CI)	*p*-Value
Accuracy	99.55	0.07	99.81–100.00	0.03
Precision	99.99	0.02	99.94–100.00	0.04
Recall	99.80	0.05	99.70–99.90	0.05
F1-score	99.90	0.03	99.84–99.96	0.03

the model's robustness and effectiveness in recognizing static sign language gestures.

Table VI exhibits the ablation study of the proposed model with different combinations. Replacing pseudo-temporal 3D CNN with a standard 2D CNN minimized performance, highlighting the significance of 3D convolutions in extracting richer inter-feature relationships, as shown in Table VI. Using HOG features with 3D CNN obtains less accuracy, representing a drawback of hand-crafted descriptors in high-level semantics. MobileNetV2 with 3D CNN enhances performance but lacks fine-grained edge information captured by HOG. The proposed model, integrating HOG and MobileNetV2 features through pseudo-temporal 3D CNN, obtains high accuracy, determines the strengths of gradient-based and deep spatial features, and assesses the efficacy of the pseudo-temporal 3D fusion strategy. Table VII represents the cross-dataset validation for the proposed model using the MNIST dataset for training and the ASL dataset for testing.

Table VII. Cross-dataset validation using the MNIST dataset as training and the ASL dataset for testing

Models	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
GRU	85.50	85.00	84.80	84.90
ResNet50	88.20	87.90	87.50	87.70
MobileNetV2	90.10	89.80	89.50	89.65
VGG16	89.30	89.00	88.70	88.85
Integrated model	**93.80**	**93.50**	**93.20**	**93.35**

B. COMPARATIVE ANALYSIS

This subsection describes the efficiency of the proposed approach based on the performance with the existing techniques of CNN, the combined model of SVM and CNN, CapsNet, and PSO-CNN. During the analysis of the aforementioned existing techniques, the accuracy is considered as the common metric utilized in listing out the outcomes in Table IV. The outcomes displayed in Table VIII

Table VI. Ablation study of the proposed model with different combinations

Combination of models	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
Replace 3D CNN with 2D CNN	98.50	98.70	98.30	98.50
HOG + 3D CNN	95.60	95.80	95.40	95.60
MobileNetV2 + 3D CNN	99.30	99.50	99.00	99.25
Proposed (HOG + MobileNetV2 + Pseudo-temporal 3D CNN)	99.55	99.99	99.80	99.90

Table VIII. Evaluation based on accuracy using the sign language MNIST dataset

Methods	Accuracy (%)
PSO-CNN [17]	99
SVM + CNN [18]	98.58
SLR CapsNet [19]	99.52
CNN [20]	96.32
CZM [21]	89.1
Integrated model (MobileNetV2 + CNN)	99.55

Table IX. Comparative analysis of the proposed model with existing models using the ASL dataset

Methods	Accuracy (%)
Tamura Resnet50 – Optimized GAM [23]	96.68
CNNSa – LSTM [24]	98.7
Proposed MobileNetV2 + 3D CNN	99.05

illustrate that the integrated model introduced in this research obtains better outcomes than other existing techniques. Table IX represents a comparative analysis of the proposed model with the ASL dataset. The classification accuracy achieved by an integrated model for the sign language MNIST dataset is 99.55%, whereas the existing CNN, combined model of SVM and CNN, SLR CapsNet, and PSO-CNN accomplish classification accuracies of 96.32%, 98.58%, 99.52%, and 99%, respectively. The performance on the ASL dataset validates the effectiveness of the proposed model in handling real-world variations. This represents that the model is not overfit for a simple dataset and generalizes well to complex scenarios.

The better recognition ability of the proposed integrated model is due to the advantageous factors of MobileNetV2 and 3D CNN. The MobileNetV2 aids in extracting the spatial features from the fully connected layer, while the 3D CNN model is deployed in recognizing the spatial data for classifying hand gestures.

C. RESEARCH IMPLICATIONS

The results of this manuscript have significant implications for advancing sign language recognition and related computer vision applications:

- Novel adaptation of 3D CNNs – The 3D CNNs are efficiently used for static image classification through exploring pseudo-temporal depth from multi-source feature maps by extending their applicability beyond conventional video and sequential data.
- Improved feature fusion strategy – Combining handcrafted gradient descriptors HOG with deep spatial features from MobileNetV2 by 3D convolution effectively enhances classification performance.
- Transferability to different domains – Offers a model that is adapted to different static recognition tasks, including medical diagnostics, remote sensing image analysis, and biometric verification.

D. DISCUSSION

In this section, the results obtained by evaluating the proposed integrated model are discussed along with the advantageous factors that help in achieving better outcomes. The evaluation of HOG-based feature extraction is also performed to assess the efficiency of feature selection. The HOG-based feature selection utilized in this research accomplishes a superior accuracy of 99.55%, whereas the existing feature extraction techniques, GLCM, SURF, LPQ, and LBP, accomplish accuracies of 91.33%, 94.50%, 89.55%, and 85.67%, respectively. Similarly, the proposed integrated classifier is evaluated with state-of-the-art classifiers such as GRU, ResNet 50, MobileNet V2, and VGG16, which achieved better classification accuracies of 95%, 98%, 96%, and 86.50%, respectively. Finally, the classification accuracy is used as the common metric to evaluate the efficiency of the integrated model, which obtains an accuracy of 99.55%. In comparison, the existing CNN achieves lower accuracy. The HOG feature extraction approach encodes significant shape and texture features that aid in the effective recognition of features. While recognizing the hand gestures, the integrated model extracts spatial features, while the 3D CNN model is deployed in recognizing the spatial data for classifying hand gestures. The proposed HOG + MobileNetV2 + pseudo-temporal 3D CNN is different from existing models because it integrates handcrafted gradient-based descriptors and deep spatial features in pseudo-temporal mode for static sign language recognition. Traditional DL models, including CNN, SVM with CNN, and CapsNet, rely on learned features that do not focus on fine-grained edge and shape cues. The methods used for handcrafted features, such as HOG, are unable to capture high-level semantic representations. The proposed model initially captures precise edge and gradient information through HOG and high-level spatial abstractions through MobileNetV2. It stacks complementary feature maps with pseudo-temporal dimensions and processes them with a 3D CNN. Unlike conventional 3D CNNs developed for temporal sequences, the pseudo-temporal considers static image feature channels as ordered temporal slices. This enables 3D kernels to learn inter-feature dependencies that 2D convolutions are unable to capture. This dual-phase, pseudo-temporal 3D fusion strategy in static sign language recognition causes superior accuracy and robustness compared to single-feature extraction models. The sign language MNIST dataset is simple because of its low resolution and uniform background, which contribute to high accuracy values. The consistent improvement obtained across multiple evaluation settings determines the effectiveness of the proposed feature fusion strategy. Although the majority of experiments are evaluated on the sign language MNIST dataset because of its standardized benchmark, additional validation is performed on the ASL dataset. The results acquired on the ASL dataset determine that the proposed model maintains higher accuracy under more complex conditions by showing its robustness.

V. CONCLUSION

Sign language recognition of hand gestures served as a communication bridge between deaf individuals and common people. This research developed an integrated DL approach for effective recognition of hand gestures. The integration of MobileNetV2 and 3D CNN was utilized to recognize hand gestures with higher accuracy. Data were gathered using the sign language MNIST dataset, and HOG-based feature extraction was performed to evaluate the magnitude and direction of the gradient of the input image. The HOG features implied the object and its outline for assessing the edge directions. The extracted features were fed into the stage of classifying hand gestures, which was performed using the proposed

integrated model. MobileNet V2 was used to extract spatial features, while the 3D CNN detected spatial data in three dimensions, contributing to improved classification accuracy. The suggested model achieved an accuracy of 99.55%, whereas the CNN, the combined model of SVM and CNN, SLR CapsNet, and PSO-CNN achieved classification accuracies of 96.32%, 98.58%, 99.52%, and 99%, respectively.

A. FUTURE WORK

Future work focuses on integrating attention mechanisms, such as transformer-based self-attention, to highlight discriminative regions in fused feature maps. Lightweight architectures, integrating with pruning, enable effective sign language recognition.

CONFLICT OF INTEREST STATEMENT

The author(s) declare that they have no conflicts of interest to report regarding the present study.

REFERENCES

[1] J. Shin *et al.*, "Korean sign language alphabet recognition through the integration of handcrafted and deep learning-based two-stream feature extraction approach," *IEEE Access*, vol. 12, pp. 68303–68318, 2024.

[2] A. Fateh *et al.*, "Advancing multilingual handwritten numeral recognition with attention-driven transfer learning," *IEEE Access*, vol. 12, pp. 41381–41395, 2024.

[3] D. Kwon *et al.*, "Machine learning-based smartphone grip posture image recognition and classification," *Appl. Sci.*, vol. 15, no. 9, p. 5020, 2025.

[4] O. Yaseen *et al.*, "Evaluation of benchmark datasets and deep learning models with pre-trained weights for vision-based dynamic hand gesture recognition," *Appl. Sci.*, vol. 15, no. 11, p. 6045, 2025.

[5] R. Malhotra and M. T. Addis, "End-to-end handwritten Ge'ez multiple numerals recognition using deep learning," *SICE J. Control Meas. Syst. Integr.*, vol. 17, no. 1, pp. 122–134, 2024.

[6] N. Anithadevi *et al.*, "MediaPipe-LSTM-enhanced framework for real-time dynamic sign language recognition in inclusive communication systems," *Eng. Rep.*, vol. 7, no. 7, p. e70142, 2025.

[7] N. Al Mudawi *et al.*, "Innovative healthcare solutions: Robust hand gesture recognition of daily life routines using 1D CNN," *Front Bioeng. Biotechnol.*, vol. 12, p. 1401803, 2024.

[8] A. S. M. Miah *et al.*, "Hand gesture recognition for multi-culture sign language using graph and general deep learning network," *IEEE Open J. Comput. Soc.*, vol. 5, pp. 144–155, 2024.

[9] M. A. Mosleh *et al.*, "Hybrid deep learning and fuzzy matching for real-time bidirectional Arabic sign language translation: Toward inclusive communication technologies," *IEEE Access*, vol. 13, pp. 94118–94136, 2025.

[10] Q. Ma *et al.*, "Intelligent hand-gesture recognition based on programmable topological metasurfaces," *Adv. Funct. Mater.*, vol. 35, no. 1, p. 2411667, 2025.

[11] Y. Wen, W. Ke, and H. Sheng, "Improved localization and recognition of handwritten digits on MNIST dataset with ConvGRU," *Appl. Sci.*, vol. 15, no. 1, p. 238, 2024.

[12] R. Fratti *et al.*, "A multi-scale CNN for transfer learning in sEMG-based hand gesture recognition for prosthetic devices," *Sensors*, vol. 24, no. 22, p. 7147, 2024.

[13] M. Aly and I. S. Fathi, "Recognizing American sign language gestures efficiently and accurately using a hybrid transformer model," *Sci. Rep.*, vol. 15, no. 1, p. 20253, 2025.

[14] J. Qin and M. Wang, "Sign language recognition based on dual-channel star-attention convolutional neural network," *Sci. Rep.*, vol. 15, no. 1, p. 27685, 2025.

[15] M. R. K. Kadavath, M. Nasor, and A. Imran, "Enhanced hand gesture recognition with surface electromyogram and machine learning," *Sensors*, vol. 24, no. 16, p. 5231, 2024.

[16] Y. Ma, T. Xu, and K. Kim, "Two-stream mixed convolutional neural network for American sign language recognition," *Sensors*, vol. 22, no. 16, p. 5959, 2022.

[17] J. Fregoso, C. I. Gonzalez, and G. E. Martinez, "Optimization of convolutional neural networks architectures using PSO for sign language recognition," *Axioms*, vol. 10, no. 3, p. 139, 2021.

[18] V. Jain *et al.*, "American sign language recognition using support vector machine and convolutional neural network," *Int. J. Inf. Technol.*, vol. 13, pp. 1193–1200, 2021.

[19] H. Xiao *et al.*, "Sign language digits and alphabets recognition by capsule networks," *J. Ambient Intell. Hum. Comput.*, vol. 13, no. 4, pp. 2131–2141, 2022.

[20] K. Bousbai *et al.*, "Improving hand gestures recognition capabilities by ensembling convolutional networks," *Expert Syst.*, vol. 39, no. 5, p. e12937, 2022.

[21] S. Bayrak, V. Nabiyev, and C. Atalar, "American sign language recognition model using complex Zernike moments and complex-valued deep neural networks," *IEEE Access*, vol. 12, pp. 193001–193013, 2024.

[22] V. Kouvakis, S. E. Trevlakis, and A. A. A. Boulogeorgos, "Semantic communications for image-based sign language transmission," *IEEE Open J. Commun. Soc.*, vol. 5, pp. 1088–1100, 2024.

[23] I. T. Ahmed *et al.*, "Enhancing hand gesture image recognition by integrating various feature groups," *Technologies*, vol. 13, no. 4, p. 164, 2025.

[24] A. Baihan *et al.*, "Sign language recognition using modified deep learning network and hybrid optimization: A hybrid optimizer (HO) based optimized CNNSa-LSTM approach," *Sci. Rep.*, vol. 14, no. 1, p. 26111, 2024.

[25] Link for sign language MNIST dataset Available: https://www.kaggle.com/datasets/datamunge/sign-language-mnist

[26] "ASL dataset," Kaggle Dataset, Available: https://www.kaggle.com/datasets/ayuraj/asl-dataset

[27] F. M. Alsuhimat and F. S. Mohamad, "A hybrid method of feature extraction for signatures verification using CNN and HOG: A multi-classification approach," *IEEE Access*, vol. 11, pp. 21873–21882, 2023.

[28] J. Anil, "A novel fast hybrid face recognition approach using convolutional kernel extreme learning machine with HOG feature extractor," *Meas. Sens.*, vol. 30, p. 100907, 2023.

[29] I. Assali *et al.*, "CNN-based classification of epileptic states for seizure prediction using combined temporal and spectral features," *Biomed. Signal Process. Control*, vol. 82, p. 104519, 2023.

[30] E. Asghar, A. Ratti, and T. Tolio, "An automated approach to reuse machining knowledge through 3D–CNN based classification of voxelized geometric features," *Procedia Comput. Sci.*, vol. 217, pp. 1209–1216, 2023.