International Journal for Scientific Research (IJSR)
المجلة الدولية للبحوث العلمية
IJSR
Vol. (5), No. (2) February 2026 الإصدار (5)، العدد (2)

Hand Gesture Recognition for human-robot interaction Based on BiLSTM-Boosted EfficientNet Network

Hawraa Abdul-elah Kadhum
Mechatronic Engineering Department, Al-Muthanna University, Al-Muthanna, Iraq
hawraakadhum@mu.edu.iq

Asmaa Mutar Jaber
MSc in Construction Management Engineering, Civil Engineering Department, Al-Muthanna University, Samawah, Iraq
asmaamutar@mu.edu.iq

Abstract

This study introduces BBE-Net, a novel BiLSTM-Boosted EfficientNet architecture designed for highly accurate and robust static hand-gesture recognition. The proposed framework integrates EfficientNet as a deep feature extractor and Bidirectional LSTM (BiLSTM) as a spatiotemporal dependency modeler, enabling the network to capture both fine-grained spatial structures and contextual relationships within gesture images. A preprocessing pipeline—consisting of standardized image resizing and histogram equalization—enhances contrast and illumination invariance, producing clearer input representations for feature extraction. EfficientNet generates multi-scale, semantically rich feature maps, which are subsequently refined by BiLSTM layers to model long-range and bidirectional correlations. The resulting discriminative features are classified through an ensemble-learning module that employs Bagging with Decision Trees and majority voting to improve stability and reduce variance. Experiments conducted on the Sebastian Marcel Static Hand Posture Database demonstrate the effectiveness of the proposed method. With extensive augmentation, 10-fold cross-validation, and repeated trials, BBE-Net achieves an accuracy of 99.70%, outperforming several recent state-of-the-art approaches. Analyses using confusion matrices, ROC curves, and class-wise metrics confirm the method's near-perfect discriminative capability.

Keywords: BBE-Net, BiLSTM-Boosted EfficientNet, Hand Gesture Recognition, Human-robot Interaction.

1. Introduction

The growing demand for natural and natural communication interface has made hand-gesture recognition a crucial research field in computer vision and human computer interaction (HCI). As the use of smart devices, wearable technologies, augmented and virtual reality products, and intelligent robots is becoming widespread, gesture-based interaction is an intuitive alternative to a physical controller, voice control, or other conventional input systems [1]. Its non-invasive and touch free design is in tandem with the current usability needs especially in the setting where hygienic, remote or contactless communication is required. This has led to the emergence of interest in the creation of precise and strong gesture-recognition systems both in academic literature and in practice. Although it is highly applicable, the hand-gesture recognition is not an easy task [2]. The real images in the world tend to be highly varying due to the lack of uniformity in lighting, or changes in skin color, awkward or untidy backgrounds, occlusions, dissimilarity of hands between users, and natural ambiguities between similar gestures. Besides, the cerebral character of gesture classification necessitates models that have the ability to identify minute distinctions in the positions of fingers, contours, and hand orientation. These constraints require recognition systems that are not just precise but also robust to noise, distortions and environmental variation [3,4].

Traditional methods of recognizing hand gestures were mainly based on manual features, like Histogram of Oriented Gradients (HOG), Scale Invariant Feature Transform (SIFT) or Local Binary Patterns (LBP), with standard classifiers. Although such methods were interpretable and efficient, they could not be used to generalize to large datasets or to different conditions as they depended on hand-designed features [5]. They did not tend to represent semantic information deeply, so they became inappropriate for handling more complex visual variations or large-scale gesture lexicons. With the introduction of deep learning, the sphere has been shifted radically, with discriminative spatial features being extracted automatically from raw data [6]. The hierarchical learning of representation and translation invariance have particularly found the use of CNNs especially

influential. VGG, ResNet, and EfficientNet are also examples of such architectures, and they have also shown impressive performance on various vision tasks. Nevertheless, CNN based models are naturally limited because they center attention on spatial data. Although good at detecting edges, texture, and patterns, CNNs have difficulty in capturing inter-regional interactions or long-range contextual interactions that could be important in differentiating near-similar gestures [7].

To counteract these drawbacks, scholars have been considering hybrid constructions that involve CNNs with sequential models more and more. Recurrent Neural Networks (RNNs) are especially Long Short-Term Memory (LSTM) and Bidirectional LSTM (BiLSTM), can be used to model sequential or temporal data. They can generate contextual relationships between feature sequences that CNNs cannot generate when applied on image derived feature sequences [8]. Previous studies that use CNNs together with LSTM variants have demonstrated an increase in recognition accuracy particularly for complex gestures. Nevertheless, most of these frameworks are not efficient, contain numerous parameters, take a long time to train, or do not combine spatial and contextual information effectively. The other issue with gesture recognition using deep learning is achieving consistent and credible classification [9]. The performance of a classifier can be affected by overfitting and class imbalance or noisy data even in the presence of powerful feature extractors. Conventional fully connected layers can frequently fail to be robust in such a state of affairs, and one motivation is the incorporation of more resilient classification methods. Aggregation of decision boundaries, using ensemble learning, especially Bagging and Boosting and randomly selected trees, has been shown to be a successful way to increase model stability [10]. Ensemble methods when used together with deep feature representations can dramatically shrink the variance as well as alleviate outlier sensitivity to make more reliable predictions. Moreover, the increased focus on real-time and resource-efficient apps implies the need to have lightweight and high-performing models. Current gesture-recognition systems have to be able to run on consumer-level hardware or embedded hardware without affecting accuracy [11]. With its compound-scaling methodology, which balances depth,

width and resolution concurrently, EfficientNet has shown a very high balance of accuracy and cost of computation. This is why it is a good candidate to be incorporated into working gesture-recognition pipelines. However, EfficientNet cannot be relied upon in tasks that demand higher contextual awareness and as such, supplementary systems must be put in place to boost feature interpretation [12]. To address these issues, this paper presents BBE-Net, a BiLSTM-Boosted EfficientNet model that is aimed at delivering an improved trade-off between accuracy, robustness, and computational efficiency in terms of hand-gesture recognition at rest. The key contributions of this study include:

1. Development of the BiLSTM-Boosted EfficientNet (BBE-Net) Architecture:

In this study, the BBE-Net architecture is introduced, in which EfficientNet for deep feature extraction is integrated with BiLSTM for modeling spatiotemporal dependencies, such that this fusion enhances the understanding of spatial and dynamic information and improves the accuracy of recognizing complex gestures.

2. Enhancement of Spatiotemporal Features via BiLSTM:

By employing the BiLSTM architecture, the features extracted by EfficientNet are used to recover bidirectional and long-term dependencies among image regions, creating a powerful mechanism for strengthening the discriminative capability of spatiotemporal features, which improves the recognition of similar and subtle gestures.

3. Utilization of Ensemble Learning for More Stable Classification:

The use of an ensemble learning approach to combine the outputs of multiple base models reduces variance, increases stability, and enhances the model's robustness against noise and gesture variations, thereby improving the final accuracy and generalization capability of BBE-Net.

The rest of this paper is structured in the following way. Section 2 is a review of related literature of hand-gesture recognition as well as hybrid deep-learning architecture. Section 3 outlines the methodology containing the preprocessing, architecture of BBE-Net, and ensemble-classification plan. Experimental results and comparative research are presented in section 4. Lastly, the paper ends with Section 5, which addresses the research directions in the future.

2.Literature Review

In [13], the authors treat the increasing need of proper hand-gesture recognition, which is one of the elements of natural human-computer interaction. They observe that the traditional skeleton based procedures are dependent on hand crafted features or traversal rules which limit representational force and make generalization difficult. The authors suggest a solution to these problems, which is a dynamic skeleton called DyHand, which incorporates BiLSTM networks and soft-attention mechanism to minimize intra- and inter-class variability. Analyzed on DHG-14/28 and SHREC'17 setups with multiple augmentation settings, DyHand scores better, with 97.14 percent and competitive performance on all gesture categories.

Articles In [14], the authors explore the dynamics of real-time sign language recognition in a human-computer interaction context regarding the Indonesian sign language. Their rule-oriented framework interprets words that are constituted by coordinated hand movements, body-front gestures and accompaniment by mimics or poses which may consist of one- or two-hand motions. They combine the feature of holistic MediaPipe with time-based models and use LSTM and GRU to train and infer. The experimental findings also reveal that there is a strong detection of dynamic gestures and correct recognition of target vocabulary, with LSTM and GRU performance of 94 and 96 respectively.

In [15], the authors focus on the problem of dynamic gesture recognition to be applied in the virtual-reality application, and high motion variability as well as user variance are also problematic. They present a better recognition algorithm based on the ResNeXt architecture to overcome the weak performance of the available models. The approach uses 3D convolutions to obtain spatiotemporal data and uses a minimal convolutional attention structure to refocus and expedite learning. Temporal feature learning is further reinforced by deep attention submodule. The studies regarding the EgoGesture and NvGesture show a high performance, reaching a maximum of 95.03% and 86.21% accuracy, and a competitive result in RGB-only conditions.

The constraints of hand-gesture recognition in hearing- and speech-impaired communication are examined in [16], which highlights such shortcomings as the complexity of the background and the paucity of gestures covered in the previous literature. They propose a two-step deep learning architecture, which to the first step upgrades Inception-V3 to a low-energy mIV3Net to decrease the computational expenses, and to the second step fine-tunes it to accentuate the discriminative features. The design enhances the abstract representation and separability between classes. System results have been tested on five public datasets, including MUGD, ISL, ArSL, NUS-I, and NUS-II, where the system has achieved a high accuracy of up to 99.8, making it significantly better than the current HGR methods.

Within the framework of [17], researchers discuss the deployment of hand-gesture recognition to add value to sharing information in areas like e-learning and health care where gestures may help in teaching, diagnosing and communicating. They introduce a multi-stage, dynamic structure which includes video decomposition, quality improvement, skeleton tracking by SSMD, hand detection with the help of CNN, hybrid feature extraction, and optimization by population-based incremental learning, and the final stage which is gesture identification by a 1D-CNN classifier. Tested on the Indian Sign Language and WLASK datasets, the method can deliver tracking accuracies of 83.71% and 85.71%, which proves its capability to enhance the level of efficiency in communication in a variety of workplace environments.

In [18], the authors are interested in the dynamic 3D hand-skeleton data and remark that currently, skeleton-based gesture recognition algorithms tend to be inefficient in terms of performance and generalization because of the low efficiency of their spatiotemporal features. They present a multi-branch attention-based graph model that is intended to provide a complete skeleton profile by means of two graph-network channels, namely, producing spatial-temporal features and the other generating temporal-spatial features, and one more deep-learning branch that produces generic representations. Concatenated features are categorized through a fully connected layer with the assistance of position

embedding and masking. The model has been tested on MSRA, DHG, and SHREC'17, with the highest possible accuracy of 97.01.

Authors in [19] define the problem of hand gesture recognition (HGR) with electromyography (EMG) signals, overcoming the problem of inter-subject variability and noise. They assess the effect of post-processing algorithm to remove spurious predictions on spectrograms and CNN architecture models when comparing typical CNN and CNN-LSTM to investigate the influence of memory cells. The results of experiments on the EMG-EPN-612 dataset indicate that post-processing greatly increases accuracy, which is 41.86 and 24.77 percent on CNN and CNN-LSTM respectively, and 90.55 percent on CNN-LSTM with post-processing. This research points out new avenues of HGR research not necessarily restricted to traditional feature extraction and classification.

In [20], the authors present a dynamic hand gesture recognition approach called Depth-Aware Spatiotemporal Fusion (DASF) framework to improve human computer interaction. To better represent features, the method uses a U-Net network to isolate hand regions in RGB frames and uses depth data to learn them. A 3D-CNN is used to obtain spatiotemporal features which are then combined with depth cues through a multi-level LSTM network. When tested on the 20BN-Jester and Sebastian Marcel Dynamic Hand Posture datasets, DASF achieves validation accuracy of 97.8% and 98.5% with better results than the previous methods and with high potential to be applied in accurate and effective gesture recognition in the real world.

In [21], the authors discuss issues in human-robot interaction that arise due to the inconsistency of hand gesture spatial and temporal aspects. Their framework is a hybrid which consists of Spectral-Political CNN with Inception module and LSTM layers to identify multi-scale spatial information and temporal relationships. Another Context-Augmented Scaled Dot-Product Attention mechanism still helps increase attention to significant areas of input. Moreover, the collaborative interactions are enabled with the help of a customizable robot architecture. The algorithm that is tested with Jochen-Triesch ASL, Sebastian Marcel Static Hand Posture and Custom Indian Sign Language data demonstrates

a maximum accuracy of 99.77% thus setting the state-of-the-art performance in recognizing gestures within human-robot systems.

In [22], the authors explore resilient hand gesture recognition in real-world applications, to counteract the constraint of models that have been trained in a controlled laboratory environment. They point out the issues with changing backgrounds, lighting, and position of gestures, which are in many cases underrepresented in current datasets. To enhance flexibility and precision, the research uses image augmentation based on deep learning, such as replacing the background, geometric distortions, and lighting changes, and a green-screen method. Findings support that these techniques improve model performance significantly in variety of settings and it is necessary to optimize augmentation strategies to the properties of datasets in order to achieve their successful implementation in a real world.

3.Methodology

In this study, we propose an innovative and efficient architecture, termed BiLSTM-Boosted EfficientNet (BBE-Net), for hand-gesture recognition. The primary objective of this architecture is to jointly leverage the strong capability of EfficientNet in extracting deep, multi-scale, and semantically rich visual features, and the strength of BiLSTM in modeling spatiotemporal dependencies within feature sequences. The synergy between these two components enables the model to effectively capture localized spatial information alongside the dynamic inter-relations among different regions of the input image.

In the first stage, input images undergo a set of preprocessing operations aimed at enhancing image quality and preparing them for more precise analysis. The most critical step in this stage is histogram equalization, which redistributes the intensity levels to improve image contrast and highlight gesture-related details. Beyond improving visual clarity, this process creates a more suitable basis for EfficientNet to extract fine-grained and discriminative features, resulting in markedly improved performance under diverse illumination conditions. Subsequently, hierarchical and multi-layer features are extracted using the

EfficientNet backbone. Owing to its compound scaling strategy—simultaneously scaling depth, width, and resolution—EfficientNet is capable of generating highly discriminative feature representations with a relatively small number of parameters. This characteristic significantly enhances the model's ability to identify subtle patterns associated with various hand gestures. The extracted features are then passed to the BiLSTM network. By processing the feature vectors bidirectionally, BiLSTM enables the model to capture reciprocal interactions, long-range dependencies, and spatial correlations across different image regions. This step enriches the feature representation and improves the model's capacity to distinguish between visually similar and complex hand gestures. Finally, the enhanced features produced by the BBE-Net architecture are fed into an ensemble-learning-based classifier. Ensemble learning combines the outputs of multiple base learners, thereby reducing variance, stabilizing decision-making, and increasing robustness against noise and gesture variability. This strategy leads to a substantial improvement in the overall performance and recognition accuracy. Figure (1) illustrates the diagram of the proposed methodology.

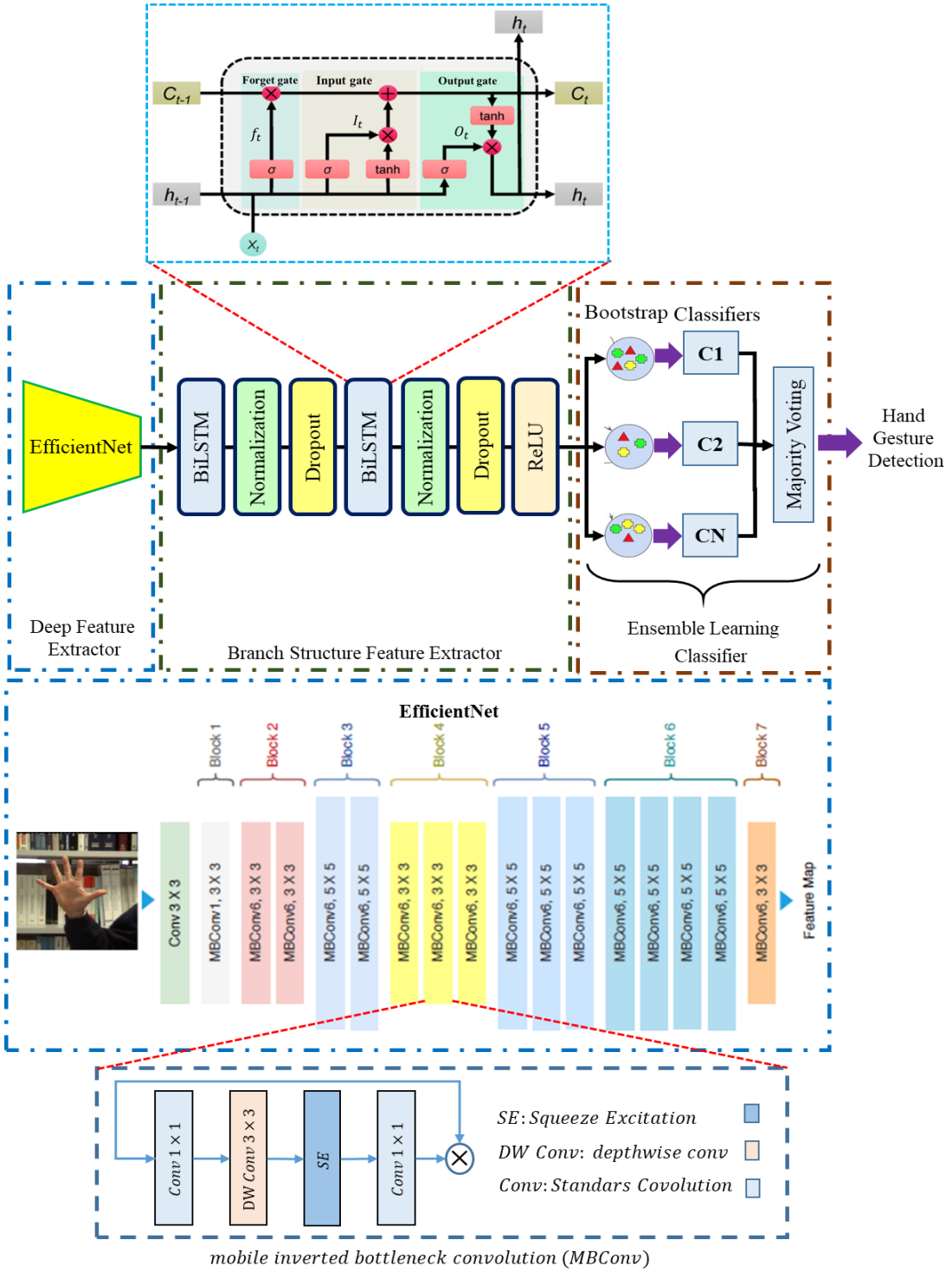

Figure (1). Diagram of the proposed method

3.1. Image Preprocessing:

The main goal of the preprocessing stage is to convert the raw hand-gesture images into a precise and standardized form and then run them through the proposed BBE-Net architecture, which greatly contributes to improving the quality of the input data and the overall model output. To start, the gesture images are resampled (resized) to a standardized and fixed resolution that matches EfficientNet input requirements. This action not only captures the structural consistency within the entire dataset, but also avoids possible problems associated with variation in scale and distortion in the aspect ratio which are likely to deteriorate the feature extraction quality. Once the dimensions are aligned, histogram equalization is used to enhance the quality of the visual display and augment the discriminability of gesture-related patterns. This operation effectively boosts the contrast and makes prominent areas clearer in the images by reallocating the values of intensity around the dynamic range of the image. Consequently, structural and textural elements, including the general form of hands, the shape of fingers, and the shapes of curves, are better highlighted, and the negative effects of poor or uneven lighting situations are minimized. These two preprocessing operations combined guarantee that the input images attain acceptable degrees of consistency, sharpness, and informational value, thus allowing the proliferation of more effective and dependable feature extraction of the BBE-Net architecture.

3.2. BBE-Net Architecture:

A new architecture, BiLSTM-Boosted EfficientNet (BBE-Net) is developed to obtain the desired comprehensive, deep, and structurally coherent feature extractions of hand-gesture images. The processed images are then input into the EfficientNet deep network in this architecture. By taking advantage of its optimized compound-scaling approach, which scales both network depth and width, as well as the resolution of network inputs, EfficientNet can learn multi-scale, hierarchical, and highly discriminative spatial features across its multiple layers. These characteristics encompass the major patterns pertaining to a general shape of hands, configurations of fingers, structural limits, and textures of

gestures. The feature maps, which EfficientNet produces, are then sent to a Bidirectional Long Short-Term Memory (BiLSTM) network in the next stage. BiLSTM allows modeling the complex relationships, the long-range relationships as well as the non-local relations within the feature vector in both directions (past to future and vice versa). The design is bi-directional such that it not only takes advantage of the local spatial information that is acquired by the convolutional layers, but also the sequential aspect of the obtained features, hence detecting latent patterns that result due to the interaction of various parts of the image.

Spatial feature extraction (EfficientNet) and contextual-sequential analysis (BiLSTM) yield a very rich, powerful, and discriminatory feature representation. This synergy at once exploits the qualities of convolutional models in mining illumination-invariant, noise-tolerant spatial features, the features of recurrent models and learning complex inter-feature interactions and contextual dependencies. The net effect is the ability of the model to differentiate between visually similar gestures, and to correctly identify the hard-to-identify and difficult gesture patterns is greatly enhanced.

3.2.1. EfficientNet Network:

EfficientNet is among the most developed architectures which were proposed in the area of deep convolutional neural networks and are aimed primarily at reaching a desirable point between high accuracy and computation efficiency. The central invention of this architecture is its compound scaling approach, where rather than just multiplying the network depth by an arbitrary degree, or the network width by arbitrary degree, all three key dimensions of the architecture network depth, network width, and input resolution are uniformly and systematically scaled. This hierarchical scaling has allowed EfficientNet to produce more detailed and accurate image representations than a number of classical models, with a much smaller number of parameters.

EfficientNet is built upon basic building blocks, which are MBConv modules and are combinations of depthwise separable convolutions, expansion and squeeze layers, and Swish activation. In these blocks, dimensions of each channel are later increased in a controlled fashion and then depthwise convolution is applied, which

is an operation that significantly lowers the amount of computation required when compared to conventional convolutions. 1x1 convolutions are then used to reduce the channels in order to create a structure that is able to extract both low- and high-level semantic patterns and features with high efficiency. Another important part of EfficientNet is the use of the Squeeze-and-Excitation (SE) mechanism of attention in most of the blocks. This process recalibrates feature channels by channel weights according to their relative significance, boosting the most informative responses and removing noise or irrelevant patterns. This selective emphasis comes in handy especially in hand-gesture recognition, in which SE permits the model to concentrate on discriminative aspects like finger shapes, fine motions, and joint edges.

EfficientNet down-samples with MBConv blocks, which down-sample the spatial resolution. Moreover, in line with the contemporary CNN designs, Batch Normalization is used throughout to stabilize the training process and speed up the convergence process. Due to its systematic and organized structure, EfficientNet can derive a hierarchical sequence of features starting with the initial edges, corners, and textures to the more sophisticated ones that describe the hand shape, finger ratios, and regions of the gesture associated with the specific gesture. The derived feature vectors are very expressive and give a sound base on which further processing is to take place. Within the suggested BBE-Net model, the EfficientNet output is an effective multi-level feature of the input image and becomes the accurate input of the sequence-based analysis processed by the BiLSTM module. Attributes that are generated by EfficientNet are robust to lighting and structural, as well as, illumination variations and changes in the appearance of a hand, as well as sufficiently information rich to enable effective spatiotemporal dependency modeling in subsequent network layers. This eventually increases the ultimate recognition accuracy of the model and increases its strength in differentiating between various and visually comparable hand gestures. Figure 2 shows the design of the EfficientNet network.

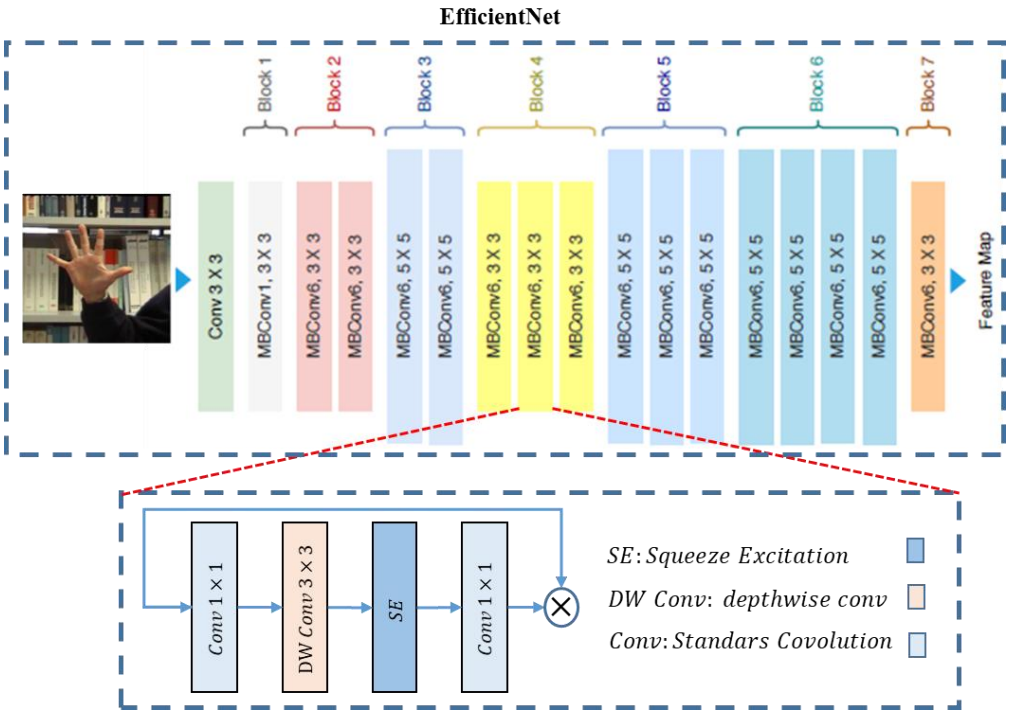

Figure (2). EfficientNet architecture

3.2.2. BiLSTM Network:

The layers of the suggested BiLSTM are arranged in the following way: a BiLSTM layer is followed by a normalization layer, a Dropout layer, another BiLSTM layer, with normalization and Dropout, and lastly a ReLU activation function. This architecture aims to derive short-term and long-term temporal dependencies, improving training stability, avoiding overfitting, and augmenting sequence-based features. In order to explain the inner workings of BiLSTM, one needs to specify the main workings of an LSTM unit. The working of each LSTM cell is based on three gates and a cell memory. The forget gate will decide what parts of the old memory need to be remembered or forgotten:

$$f_t = \sigma(W_f[x_t.h_{t-1}] + b_f) \quad (1)$$

The incorporation of new information in the cell memory is then controlled by the input gate:

$$i_t = \sigma(W_i[x_t.h_{t-1}] + b_i). \quad \widetilde{C_t} = \tanh(W_c[x_t, h_{t-1}] + b_c). \quad (2)$$

The new and old information is fused to become the new cell state:

$$C_t = f_t \odot C_{t-1} + i_t \odot \widetilde{C}_t. \quad (3)$$

Lastly, the output gate decides which components of the information in the processed information to proceed to the next time step:

$$o_t = \sigma(W_o[x_t, h_{t-1}] + b_o). \quad h_t = o_t \odot \tanh(C_t). \quad (4)$$

This is the mechanism that allows the long-term dependencies to be preserved and extract the complex temporal patterns, which can be significantly useful in hand-gesture recognition where the sequential motions and subtle changes in time are important. The input sequence in BiLSTM layers is handled both forward and backward which allows the model to access both past and future time steps. This bidirectional processing helps in the enrichment of the feature representation and increases the capacity of the network in recognition of the complicated temporal configurations. The general architecture of the BiLSTM network is shown in Figure (3).

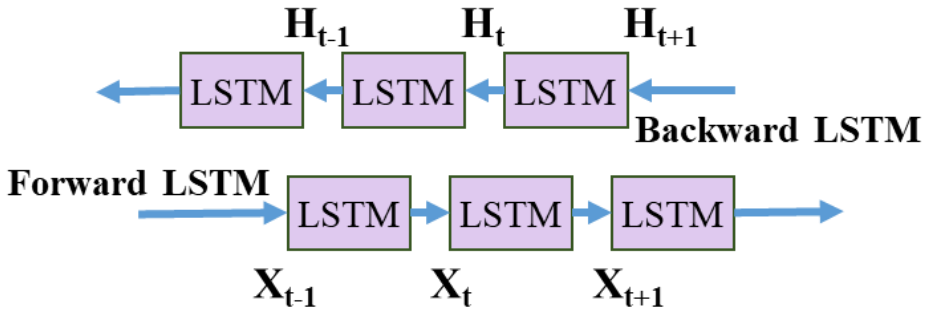

Figure (3). General structure of the BiLSTM network

The hidden state of the time (t) is calculated in the forward direction as follows:

$$\overrightarrow{h_t} = \text{LSTM}(x_t . \overrightarrow{h_{t-1}}) \quad (5)$$

Similarly, in the backward direction:

$$\overleftarrow{h_t} = \text{LSTM}(x_t . \overleftarrow{h_{t+1}}) \quad (6)$$

Final output of time (t) = forward and backward hidden states concatenation:

$$h_t = [\overrightarrow{h_t}; \overleftarrow{h_t}]. \quad (7)$$

After every BiLSTM layer, the normalization layer is used to stabilize the data flow and avoid sudden changes in the gradient. To normalize an input (z), the normalization is done by the following:

$$\hat{z} = \frac{z-\mu}{\sqrt{\sigma^2+\epsilon}} \quad (8)$$

Representational flexibility is then recovered by a set of learnable parameters (y) and (β):

$$y = \gamma\hat{z} + \beta. \quad (9)$$

The mechanism is used to increase the rate of convergence and stability of training by imposing a statistical alignment of features.

The dropout layer is then used to minimize overfitting by randomly shutting down a number of neurons:

$$mask \sim Bernoulli(p) \quad (10)$$

The dropout mask is applied to the inputs:

$$h' = h \odot mask \quad (11)$$

This regularization makes the network learn distributed representations as opposed to relying excessively on a restricted number of neurons, which enhances generalization.

Lastly, the ReLU activation function is utilized to project extracted features into an appropriately nonlinear space to allow modeling of complex relationships to be more effectively represented. ReLU also alleviates the problem of vanishing gradient and improves training stability.

$$\text{ReLU}(x) = \max(0, x) \quad (12)$$

These layers combined make it possible for the BiLSTM network to capture the rich, stable, and robust temporal features. The BiLSTM component is important in the overall BBE-Net architecture because it prevents overfitting and provides support for an efficient gradient flow.

3.4. Classification Using Ensemble Learning:

The last phase in the proposed BBE-Net architecture is to use an ensemble-learning mechanism to obtain a more stable, accurate, and noise-tolerant recognition system. The purpose of applying such a method is to achieve the advantages of several base classifiers and leverage the variation of their decision boundaries to provide improved system performance. The last features that the BBE-Net framework extracts are rich and discriminative spatiotemporal structures; the representations are input into the ensemble module to ensure the classification is of very high reliability. Decision Tree is used at this stage as the foundation classifier. The decision tree is also suitable for the complex characteristics of the hand-gesture features due to its inherent capacity to subdivide data according to hierarchical principles and nonlinear decision-boundaries. Moreover, it has a low feature scaling sensitivity and great ability to capture irregular patterns hence it could be used in this framework.

The Bagging (Bootstrap Aggregating) method is used in order to add more diversity to the models and not to over-rely on certain data samples. There are several training subsets in this method that are randomly sampled with replacement of the original dataset. The subsets learn independent copies of a decision tree, and the relative independence of classifiers lowers the sensitiveness of the final model to noise and outliers. The variety in the training data of classifiers is necessary since each is trained on a different subset of the data, and it is plausible that some will work well with challenging samples; this diversity becomes one of the factors that promotes system robustness. Once several independent classifiers have been trained, their results are pooled together with the help of Majority Voting. In this approach every classifier gives its prediction and the end result is that of the class with the highest number of votes. This combination process lowers the variance of the decisions, rationalizes the predictions and decreases the error caused by unpredictable choices of any single classifier. Combining Bagging with Majority Voting helps to enhance the generalization of the models and allows the proposed framework to be capable of discriminating between hand gestures that are close to each other and fine-grained

and delicate patterns. Finally, this ensemble-classification module is crucial for finalizing the BBE-Net architecture and increasing the performance of the latter to the level of stability, accuracy, and reliability.

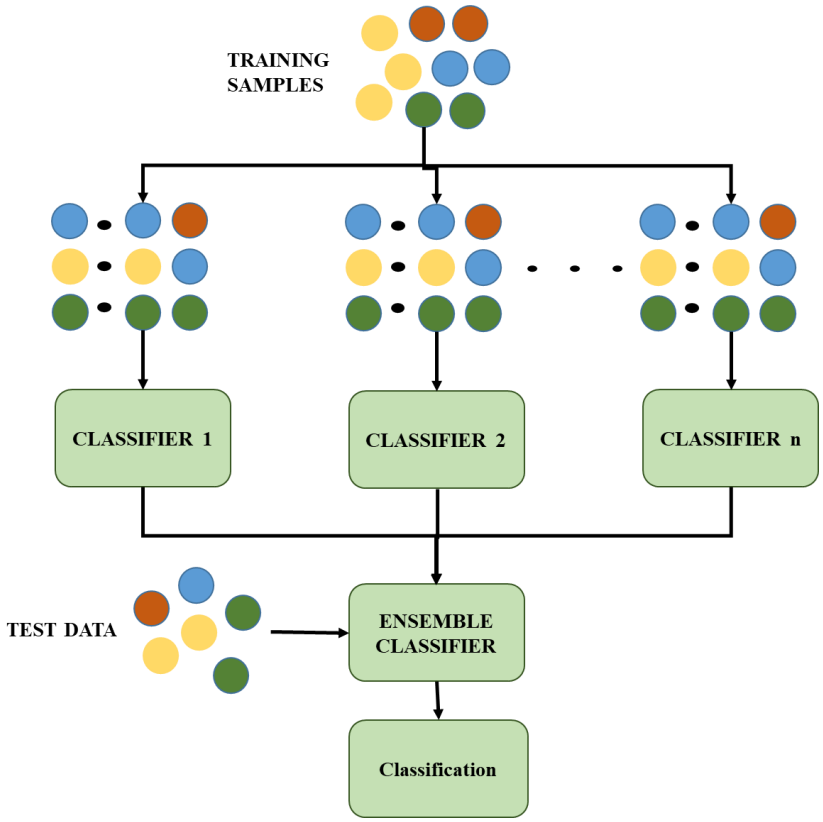

Figure (4). Ensemble learning diagram

4. Experimental Results

In this part, the assessment procedure of the proposed approach is introduced, as well as comparisons to other methods, model simulation, and analysis of experimental results. The accuracy, precision, recall, and F-score performance measures are employed to determine the capability of the model to detect hand gestures. The findings have been presented in tables and charts. To conduct the experiments, data augmentation techniques were applied to the set of images and the size of the dataset was increased to 22,124 images, thus enabling the network to learn a wider range of gesture variations and enhancing model's generalization. This augmented dataset was then divided into training and testing datasets, with 70 percent (15,478 images) of them being used to train the model and 30 percent

(6,637 images) to assess its capabilities in unknown situations. In training, network weights were updated using training subset, and the learning process was controlled by it.

In a bid to establish the reliability and strength of the findings further, 10-fold cross-validation was used. In this approach, the model is tested using a variety of data partitions and this would give a better estimate in varying conditions of the dataset. The evaluation metrics of the proposed model in the first experiment were obtained, and it is possible to conduct a statistical analysis of the accuracy, precision, recall, and F-score. The proposed model simulations were made in MATLAB 2023a, and the experiments were conducted on a computer with Nvidia graphics, 16GB of RAM, Intel Core i5 processor, and Windows 11. The experiment was done 40 times to increase the credibility of the results and the reported outcomes are the statistical average of all 40 times. This method guarantees the stability of the final performance and reduces the impact of the accidental deviations of the evaluation process.

4.1. Dataset:

The Sebastian Marcel Static Hand Posture Database [23] was used in this research to train and test the proposed methodology. This data set contains hand-pose images of ten individuals, taken under both simple and highly cluttered backgrounds. The changes in light and size of the image also make the dataset more complex, providing a realistic and challenging scenario for testing the effectiveness of models in the real world. To conduct the research, a dataset that provides clear and semantically significant hand postures in heterogeneous backgrounds was required; thus, the Sebastian Marcel dataset was used as a suitable reference point on which to experiment and verify the given architecture.

This data consists of six main hand postures denoted as A, B, C, Point, Five and V, some of which are shown in Fig. (4). The total number of images present in the dataset is 5,531, out of which 4,872 images were placed in the training folder to train the model and create its parameters. As a test sample, 277 Complex and 382 Uniform images are given as performance evaluation, to allow the independent evaluation of the model's behavior in cluttered and uniform backgrounds. This

separation helps to closely examine the accuracy and stability of the system in various situations and examine the model's ability to identify the hand postures under different environmental conditions.

Figure (5). Sample images of hand gesture classes from the Sébastien Marcel dataset

4.2. Evaluation Metrics:

In order to critically evaluate the work of the proposed approach, a combination of conventional quantitative indicators were applied, such as Accuracy, Precision, Recall, and F1-score. Accuracy is the count of samples correctly classified to the total number of samples indicating a general picture of the extent to which the model can identify hand postures correctly. Precision measures the number of correctly identified samples of the total number of samples classified as in a given class by the model and, therefore, represents how well the model minimizes false positives.

Recall tests how the model can correctly recognize all the examples of a given class, thus, defining its sensitivity and the possibility to eliminate false negatives. The F1-score, which is the harmonic mean of Precision and Recall, is another useful metric of evaluation where it is important and of equal importance to evaluate both the criteria. Collectively, these four measures provide a detailed and more sophisticated evaluation of the performance of the proposed method and allow making a reasonable comparison with other approaches presented in the literature. The following are the definitions of these metrics:

$$\text{Accuracy} = \frac{TP+TN}{TP+FP+TN+FN} \quad (13)$$

$$\text{Precision} = \frac{TP}{TP+FP} \quad (14)$$

$$\text{Recall} = \frac{TP}{TP+FN} \quad (15)$$

$$\text{F1-score} = \frac{2PR}{P+R} \quad (16)$$

TP, TN, FP and FN are the number of correctly predicted positive, number of correctly predicted negative, number of falsely predicted positive, and number of falsely predicted negative, respectively.

4.3. Results Evaluation:

In this section, the proposed system performance is considered in the context of the training and the testing process. In order to perform a thorough analysis, a number of commonly used machine learning performance indicators are used. Firstly, the learning curve is introduced to demonstrate how the model converges and how the training process becomes stable with the passing of the iterations. A confusion matrix is then reported to determine the sample-level classification performance of the model and to provide the opportunity to comprehensively study the patterns of misclassifications by classes. Also, ROC curve and the associated AUC values are used as measures of the discriminative ability of the model. To further estimate the accuracy, class-wise accuracy is evaluated with Precision, Recall, and F-score that give information on the strengths and weaknesses of the model in identifying certain patterns of hand-poses and that supplement the generalization behavior of the model. Lastly, the efficiency of the proposed approach on the test data is contrasted with some of the recent and notable approaches that have been reported in the literature.

4.3.1. Evaluation of the Training Process:

The learning curve of the proposed BBE-Net throughout the training is shown in Figure 6. The horizontal and vertical axes represent the training iterations and the value of the loss function respectively. As can be seen, the loss decreases at the very start of training, which means that the learning process is very fast, and the

weights are adjusted successfully within the first few training steps. By the early iterations of the first few hundreds, the rate of decline is slower and the curve is more or less flat, indicating that the model has attained a more or less steady convergence state. This action proves that the network has managed to optimize its weights and parameters in order to minimize the least prediction error on the training set. Moreover, the fact that the values of the losses are lower in subsequent iterations indicates the consistency of the learning process and the lack of strong instability and overfitting. Altogether, the learning curve confirms that the proposed BBE-Net architecture experiences a stable and efficient training procedure and reaches the optimal parameter set.

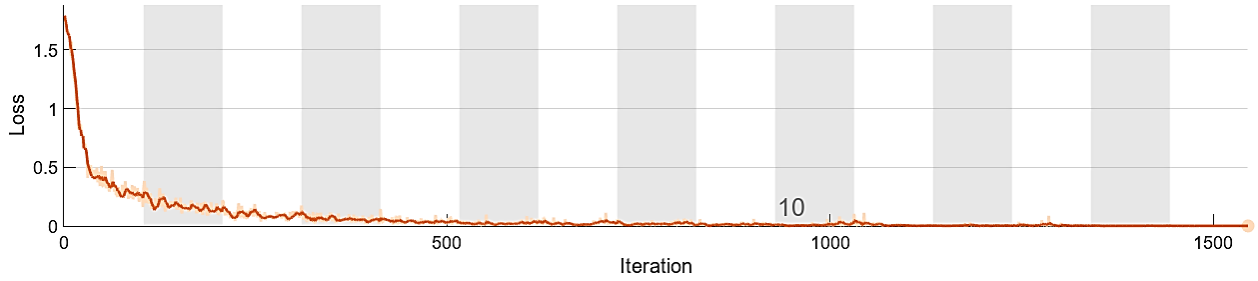

Figure (6). Convergence curve of the BBE-Net model

4.3.2. Evaluation of the Test Process:

The confusion matrix of the proposed hand-pose classifier of 6 categories (A, B, C, Point, Five, and V) is shown in Figure 7. The general performance of the model is quite impressive: the diagonal elements, which are true positive predictions, are very high across all the classes, with 1,689, 727, 832, 944, 1,815, and 610 correctly identified samples under class A, B, C, Point, Five and V, respectively. These values are high and it is through these that 100% Recall of all classes but five is to be met. There are no missed samples and classes A, B, C, Point, and V are identified in a completely accurate manner. In class five, the total misclassification rate is 1.1, and this is translated to 12 samples that were wrongly classified as A, 4 samples as C, and 4 as Point. Although this small confusion of the distinction of the class Five was made, the model retains a very high Precision of all the classes. The accuracy is 100 percent in classes B, Five, and V and in class A, C, and Point is 99.3, 99.5, and 99.6 respectively, and this means that they

would be resistant to false-positive errors. To conclude, the confusion matrix points out an extremely trustworthy and close to perfect classifier. The proposed model works perfectly in five out of six categories of poses and also has a low misclassification rate of class Five and it can be seen as a strong model with a great discriminating ability.

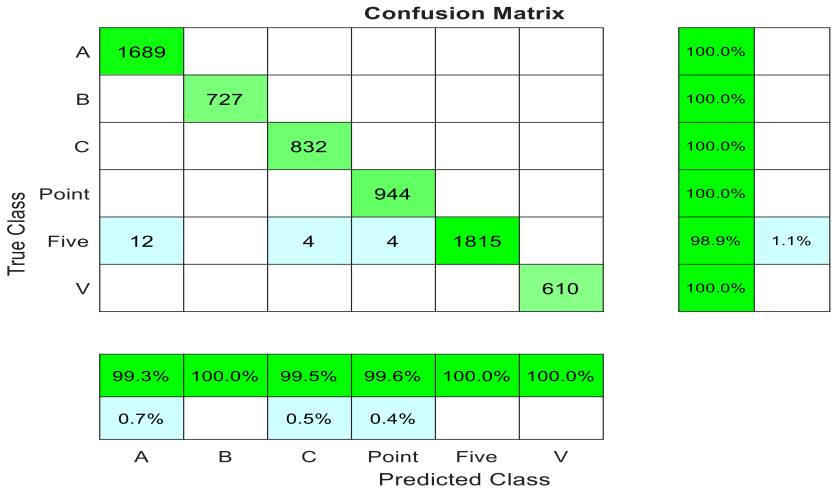

Figure (7). the confusion matrix

4.3.3. Receiver Operating Characteristic (ROC) Analysis:

As shown in Figure 7, the six classes (Class 1 to Class 6) obtained in the proposed method have Receiver Operating Characteristic (ROC) curves. The False Positive Rate (FPR) is plotted along the horizontal axis and the True Positive Rate (TPR) is plotted along the vertical axis in this plot, which is sensitivity (Recall). The examination of this figure indicates that the discriminative performance is exceptionally well and almost perfect in all classes. The gray line along the diagonal symbolizes the performance of a random classifier that has no discriminating ability. In contrast, the ROC curves of the six classes are almost identical and takes the uppermost position of the plot. Namely, the curves almost increase in the point (0, 0) to the point (0, 1) and then spread horizontally to the point (1, 1). This arrow-like shape in the upper-left corner shows that the model has a TPR of 1 (i.e., 100%) with an FPR of 0, thus showing that it is able to correctly identify all positive samples and not generate any false positives.

In this analysis, the most important indicator is the Area Under the Curve (AUC) which is indicated above the plot and has the value of 0.999743. An AUC of 0.999743 implies that a perfect classifier can be used to classify the classes with very high accuracy, given that the right decision threshold is chosen. This result supports the results of the confusion matrix, and also confirms that the method proposed has statistically close separability between the different classes of hand poses.

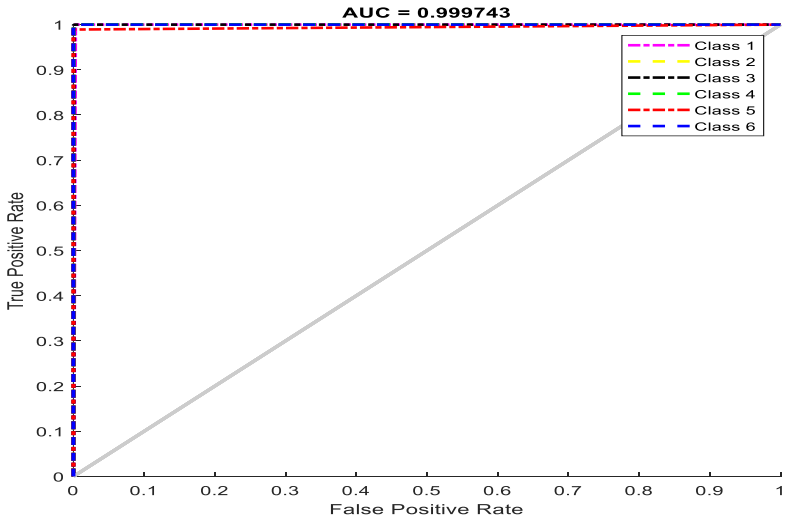

Figure (8). the ROC curve

4.3.3. Class-Based Performance Analysis:

Figure 9 shows the performance of the proposed method by class in the recognition of hand poses based on Precision, Recall, and F-score measures. As depicted, the proposed strategy has extraordinarily high and very stable performance in all six classes (A through V), and all the metric values fall in a very high range and near 100 percent. The class B and V performance is absolutely perfect; all three measures, Precision, Recall and F1-Score, are 100 percent, which means that the model correctly recognizes all positive samples without committing any Type I (False Positive) and Type II (False Negatives errors) in class B and V. Classes C and Point also have near-perfect values in all metrics, which can be seen as the presence of highly robust and balanced recognition performance.

In the case of class A, the Precision value is slightly higher than the Recall one but the two values are close to 99, and the F1-Score of 100 is an effective result as well, which indicates that the model behaves reliably even with this class. Five is the only class that has a clear distinction between its metrics. In this class, Recall is 100, meaning that all real samples of this hand pose were chosen correctly, and zero Type II errors were made. Nevertheless, the Precision of this class is only slightly less than 100, which proves that a very small number of samples of the other classes (as it has already been seen in the confusion matrix) were erroneously assigned to the Five category, which caused a few Type I errors. Even though this minor decrease in Precision has been noted, the F1-Score of the class Five is very high, indicating that overall performance is high. As a result, the class-wise assessment demonstrates that the adopted approach is an extremely discriminative and efficient classifier, which has stable and consistent performance across all target hand-pose categories.

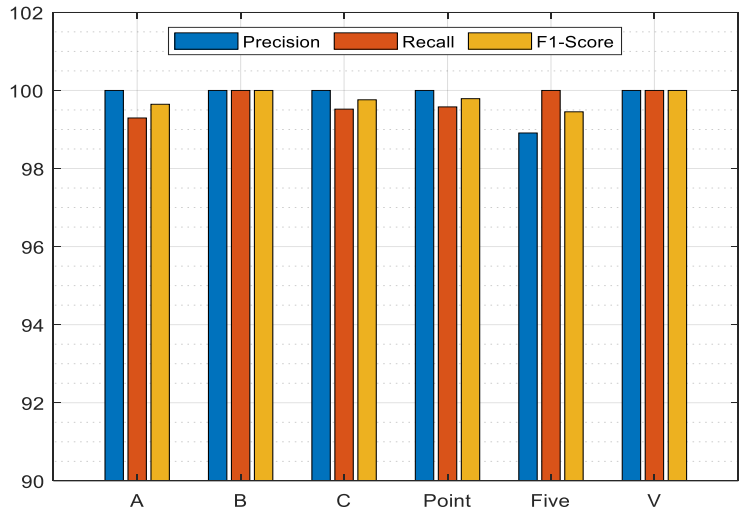

Figure (9). performance evaluation of the proposed method based on Precision, Recall, and F-score metrics

4.3.4. Comparison of Results:

The proposed method is compared with some of the state-of-the-art methods based on Accuracy, i.e., 3DCNN -LSTM [20], Attention-based Hybrid CNN [21], Hough Transform and Neural Network [21], CNN + MobileNet [21], and DNN [22] in Table 1. As has been seen, the proposed method has an accuracy of 99.70

and is doing better than any other method that has been compared. This high score is an excellent indication of the great capability of the model to identify hand poses correctly and supports the findings of the confusion matrix and ROC analysis. The Attention-based Hybrid CNN that Biswas et al. [21] suggested offered the most similar performance, as its accuracy is 99.23, meaning that attention-based structures are also effective in this sphere. Other deep-learning-based models that delivered good performance but not even close to the proposed one include 3DCNN LSTM at 98.50% and Hough Transform and Neural Network at 98.00% accuracy. Lastly, CNN + MobileNet with the highest accuracy of 94.40, and even more so the DNN approach by Awaluddin et al. [22] with 91.31, had the worst performance as compared to the rest in this comparison. The fact that the proposed one (99.70) outperforms the weak one (91.31) by more than 8 percent in hand-pose recognition speaks volumes of the overwhelming superiority of the BBE-Net architecture. Table 1 demonstrates the clear superiority of the proposed BBE-Net, in terms of accuracy, which proves that the model is superior even compared to the modern and sophisticated architectures. This shows the scientific and practical value of the suggested method.

Table (1). Accuracy comparison of the proposed method with recent state-of-the-art

Author	Method	Accuracy
Chanda et al.	3DCNN -LSTM	98.50
Biswas et al.	Attention based hybrid CNN	99.23
Biswas et al.	Hough transform and neural network	98.00
Biswas et al.	CNN + MobileNet	94.40
Awaluddin et al.	DNN	91.31
Presented	BBE-Net	99.70

5. Conclusion

This article introduced BBE-Net, which is a BiLSTM-Boosted EfficientNet model designed to promote the precision, robustness, and overall potential of the existing static hand-gesture recognition systems. The proposed framework can effectively balance the finer details of gesture images by extracting multi-scale features of EfficientNet and the contextual modeling of BiLSTM to capture both the underlying spatiotemporal relationships and the small-scale gesture details in gesture images. Additional features of tailored preprocessing pipeline further

increase the gesture visibility and illumination invariance, which allows features to be learned more effectively. Besides, the use of an ensemble-learning classifier (relying on Bagging and Decision Trees) has a major positive impact on the stability of the classification process: it lowers variances and eliminates the impact of outliers. The vast amount of experimentation conducted on the Sebastian Marcel Static Hand Posture Database, which was backed by data augmentation, 10-fold cross-validation, and repeated experiments, proved that BBE-Net is superior to various methods that had been already tested. The recognition accuracy of the model was 99.70 and modern state-of-the-art methods have lower recognition accuracy than the model itself. The model has only near-perfect separability across each gesture class. Confusion matrices, ROC and class-wise performance metrics analyses verified the high discriminative ability of the system, its ability to withstand background complexity as well as variability across the diverse types of gestures. In general, the findings confirm the usefulness of the suggested architecture and its applicability in the real-world in the context of human-computer interaction, assistive technologies, and intelligent robotic systems. This study can be advanced in the future to support dynamic gesture recognition, multimodal fusion, and running on resource-constrained platforms to make it more widely applicable.

References

[1] Sarowar, M. S., Farjana, N. E. J., Khan, M. A. I., Mutalib, M. A., Islam, S., & Islam, M. Hand Gesture Recognition Systems: A Review of Methods, Datasets, and Emerging Trends. International Journal of Computer Applications, 975, 8887.

[2] Mohamed, A. S., Hassan, N. F., & Jamil, A. S. (2024). Real-Time Hand Gesture Recognition: A Comprehensive Review of Techniques, Applications, and Challenges. Cybernetics and Information Technologies, 24(3), 163-181.

[3] Hashi, A. O., Hashim, S. Z. M., & Asamah, A. B. (2024). A systematic review of hand gesture recognition: An update from 2018 to 2024. IEEE Access, 12, 143599-143626.

[4] Murad, B. K., & Alasadi, A. H. H. (2024). Advancements and challenges in hand gesture recognition: A comprehensive review. Iraqi Journal for Electrical and Electronic Engineering, 20(2), 154-164.

[5] Hax, D. R. T., Penava, P., Krodel, S., Razova, L., & Buettner, R. (2024). A novel hybrid deep learning architecture for dynamic hand gesture recognition. IEEE Access, 12, 28761-28774.

[6] Rahman, M. M., Uzzaman, A., Khatun, F., Aktaruzzaman, M., & Siddique, N. (2025). A comparative study of advanced technologies and methods in hand gesture analysis and recognition systems. Expert Systems with Applications, 266, 125929.

[7] Kim, B., & Seo, S. (2023). EfficientNetV2-based dynamic gesture recognition using transformed scalogram from triaxial acceleration signal. Journal of Computational Design and Engineering, 10(4), 1694-1706.

[8] Hussain, A., Ul Amin, S., & Fayaz, M. (2023). An Efficient and Robust Hand Gesture Recognition System of Sign Language Employing Finetuned Inception-V3 and Efficientnet-B0 Network. Computer Systems Science & Engineering, 46(3).

[9] Rezaee, K., Khavari, S. F., Ansari, M., Zare, F., & Roknabadi, M. H. A. (2024). Hand gestures classification of sEMG signals based on BiLSTM-metaheuristic optimization and hybrid U-Net-MobileNetV2 encoder architecture. Scientific Reports, 14(1), 31257.

[10] Tchantchane, R., Zhou, H., Zhang, S., & Alici, G. (2023). A review of hand gesture recognition systems based on noninvasive wearable sensors. Advanced intelligent systems, 5(10), 2300207.

[11] Alam, M. M., Islam, M. T., & Rahman, S. M. (2022). Unified learning approach for egocentric hand gesture recognition and fingertip detection. Pattern recognition, 121, 108200.

[12] Jency Rubia, J., Babitha Lincy, R., & Sherin Shibi, C. (2024). Hybrid Convolution-Based Efficientnet-Based Hand Gesture Recognition Framework with Optimized Algorithm. International Journal of Pattern Recognition and Artificial Intelligence, 38(12), 2456008.

[13] Singh, R. P., & Singh, L. D. (2025). Dyhand: dynamic hand gesture recognition using BiLSTM and soft attention methods. The Visual Computer, 41(1), 41-51.

[14] Ilham, A. A., & Nurtanio, I. (2024). Applying LSTM and GRU methods to recognize and interpret hand gestures, poses, and face-based sign language in real time. Journal of Advanced Computational Intelligence and Intelligent Informatics, 28(2), 265-272.

[15] Xi, J., Zhang, W., Xu, Z., Zhu, S., Tang, L., & Zhao, L. (2025). Three-dimensional dynamic gesture recognition method based on convolutional neural network. High-Confidence Computing, 5(1), 100280.

[16] Karsh, B., Laskar, R. H., & Karsh, R. K. (2024). mIV3Net: modified inception V3 network for hand gesture recognition. Multimedia Tools and Applications, 83(4), 10587-10613.

[17] Al Mudawi, N., Ansar, H., Alazeb, A., Aljuaid, H., AlQahtani, Y., Algarni, A., ... & Liu, H. (2024). Innovative healthcare solutions: robust hand gesture recognition of daily life routines using 1D CNN. Frontiers in Bioengineering and Biotechnology, 12, 1401803.

[18] Miah, A. S. M., Hasan, M. A. M., & Shin, J. (2023). Dynamic hand gesture recognition using multi-branch attention based graph and general deep learning model. IEEE Access, 11, 4703-4716.

[19] López, L. I. B., Ferri, F. M., Zea, J., Caraguay, Á. L. V., & Benalcázar, M. E. (2024). CNN-LSTM and post-processing for EMG-based hand gesture recognition. Intelligent Systems with Applications, 22, 200352.

[20] Chanda, Bristy, and Hussain Nyeem. "Depth-Aware Spatiotemporal Fusion for Advancing Dynamic Hand Gesture Recognition." Available at SSRN 5011540.

[21] Biswas, Sougatamoy, et al. "Attention-enabled hybrid convolutional neural network for enhancing human–robot collaboration through hand gesture recognition." Computers and electrical engineering 123 (2025): 110020.

[22] Awaluddin, Baiti-Ahmad, Chun-Tang Chao, and Juing-Shian Chiou. "A hybrid image augmentation technique for user-and environment-independent hand gesture recognition based on deep learning." Mathematics 12.9 (2024): 1393.

[23] Idiap Research Institute. "Gesture Database." Idiap, https://www.idiap.ch/webarchives/sites/www.idiap.ch/resource/gestures. Accessed 24 Nov. 2025.