Computers, Materials & Continua

Tech Science Press

DOI: 10.32604/cmc.2023.042886

ARTICLE

Deep Learning Approach for Hand Gesture Recognition: Applications in Deaf Communication and Healthcare

Khursheed Aurangzeb[1], Khalid Javeed[2], Musaed Alhussein[1], Imad Rida[3], Syed Irtaza Haider[1] and Anubha Parashar[4,*]

[1]Department of Computer Engineering, College of Computer and Information Sciences, King Saud University, P.O.Box 51178, Riyadh, 11543, Kingdom of Saudi Arabia

[2]Department of Computer Engineering, College of Computing and Informatics, University of Sharjah, Sharjah, 27272, United Arab Emirates

[3]Laboratory Biomechanics and Bioengineering, University of Technology of Compiegne, Compiegne, 60200, France

[4]Department of Computer Science and Engineering, Manipal University Jaipur, Jaipur, 303007, India

*Corresponding Author: Anubha Parashar. Email: anubhaparashar@gmail.com

Received: 15 June 2023 Accepted: 11 October 2023 Published: 30 January 2024

ABSTRACT

Hand gestures have been used as a significant mode of communication since the advent of human civilization. By facilitating human-computer interaction (HCI), hand gesture recognition (HGRoc) technology is crucial for seamless and error-free HCI. HGRoc technology is pivotal in healthcare and communication for the deaf community. Despite significant advancements in computer vision-based gesture recognition for language understanding, two considerable challenges persist in this field: **(a)** limited and common gestures are considered, **(b)** processing multiple channels of information across a network takes huge computational time during discriminative feature extraction. Therefore, a novel hand vision-based convolutional neural network (CNN) model named (HVCNNM) offers several benefits, notably enhanced accuracy, robustness to variations, real-time performance, reduced channels, and scalability. Additionally, these models can be optimized for real-time performance, learn from large amounts of data, and are scalable to handle complex recognition tasks for efficient human-computer interaction. The proposed model was evaluated on two challenging datasets, namely the Massey University Dataset (MUD) and the American Sign Language (ASL) Alphabet Dataset (ASLAD). On the MUD and ASLAD datasets, HVCNNM achieved a score of 99.23% and 99.00%, respectively. These results demonstrate the effectiveness of CNN as a promising HGRoc approach. The findings suggest that the proposed model have potential roles in applications such as sign language recognition, human-computer interaction, and robotics.

KEYWORDS

Computer vision; deep learning; gait recognition; sign language recognition; machine learning

1 Introduction

Human-computer interaction (HCI) has emerged in recent years as a vital component of our daily lives as it has become a crucial component of HCI in the recognition of actions and activities

in our daily lives [1–3]. In recent years, hand gestures have become one of the most popular forms of communication with machines [3]. Individuals can directly interact with machines using only the movement of their palms and fingers in HCI [4]. For effective interaction between humans and robots, machines must realize human gestures [5,6]. As a result, HGRoc has become an essential and popular research topic in today's world. HGRoc has a variety of applications, such as simulated reality, robot self-control, virtual games, and simple client interfaces.

The field of HGRoc is rapidly advancing, and it has many applications in medicine. Utilizing technologies such as computer vision, machine learning, and sensor technology, it translates hand movements into digital commands that are understandable to digital systems. A non-contact, interactive interface facilitated by this technology in healthcare reduces the spread of infectious diseases. To manipulate digital images, such as CT scans or patient records, a surgeon can use hand gestures rather than physically touching a screen or keyboard during a surgical procedure near a patient. Sterility is crucial during surgical procedures, so keeping the operating room clean is important.

Additionally, HGRoc is useful as a rehabilitation tool. As a result, patients recovering from stroke and musculoskeletal injuries can be monitored and evaluated using technology so therapy plans can be adjusted accordingly. By tracking and analyzing their progress, they can keep track of their progress [7]. Additionally, the technology can be used to develop interactive therapy games that engage patients in recovery. Further, it can assist in communication for individuals with hearing impairments. By recognizing and interpreting sign language, the technology can provide real-time translation services, facilitating communication and promoting inclusivity. Lastly, for individuals with severe motor impairments or paralysis, HGRoc, combined with assistive robotic technology, can help perform daily tasks. Simple gestures could command robotic arms to carry out activities, promoting independence and improving quality of life.

Sign expression is a visible form of communication that utilizes organized, communicative finger motions to convey thoughts, and it is the only means of communication for individuals with speech and hearing impairments [8]. As per statistics from the healthcare center, approximately 5 percent of the global residents (or around 360 million people) suffer from medium to severe hearing loss and communicate only through sign language (WHO, 2015) [9]. However, sign expression is not simply understood by the listener to the community, resulting in an interaction difference between the speech and hearing-reduced and hearing populations [10]. Recently, deep learning (DL) has achieved encouraging outcomes in different domains, such as activity recognition [11,12], disease recognition [13], and energy forecasting [14]. Therefore, HGRoc, using computer knowledge, can serve as a translator for sign motion translation, creating a bridge between these two communities. This would be advantageous in many ways, such as creating a more inclusive environment and allowing citizens with speech and hearing impairments to cooperate more effectively with the hearing community [15]. It could also provide opportunities for individuals with hearing impairments to access education and work opportunities they may not have had before. Furthermore, this technology has the potential to revolutionize the way we interact with machines, making them more accessible and user-friendly for everyone [16].

The classification of finger movements in symbol expression can be segregated into dual distinct categories: stationary and dynamic. In static gestures, the hands and fingers are positioned in a specific space without any temporal movement, while dynamic gestures involve continuous hand movements over time. A contact-based recognition technique as well as a vision-based recognition technique, can be used for translating sign languages [17]. An electronic circuitry device used in contact-based recognition identifies changes in movement and transmits the data to a computer so that further

processing can be carried out. Despite the favorable recognition results that have been demonstrated in the literature, this approach is relatively expensive and impractical for regular interaction between humans and computers. On the other hand, a vision-based method is easier to use since it takes a picture of the signer and analyzes it using an image-processing algorithm rather than sensory devices. Human-computer interactions are more practical and cost-effective with this method [18].

Using a vision-based model, we propose a vision-based approach to recognize static hand gestures in sign language translation systems in this study. By taking RGB video streams from cameras and processing them to reduce the dimensions from three color channels to one, we have proposed a method for addressing the vision challenges associated with hand gesture identification. Using the proposed system, the real-time recognition of hand gestures is achieved by extracting salient features and accounting for shape, size, lighting conditions, and occlusion variations. HVCNNM is introduced in the research based on the CNN architecture of the VGG16 model. Additionally, the paper analyzes the sign language dataset, which is more complex because it includes both alphabetic and numerical categories. As a result of this dataset, the proposed approach demonstrates better accuracy and less training time than current methods. According to this study, a vision-based approach can be used to recognize dynamic hand gestures in sign language translation systems, which can help people with hearing loss and speech and hearing impairment communicate better. Our framework for hand gesture recognition consists of three key components: the preprocessing, a lightweight CNN model, and a text-to-speech engine. The refining step reduces image noise and dimensions for efficient feature extraction, followed by classification via an optimized CNN model suited for real-time applications. The identified gestures are then converted to audio feedback through a text-to-speech engine, enhancing usability and accessibility. The crucial supportive contributions of the proposed HVCNNM are the following:

- Efforts in the hand gesture recognition domain have traditionally been concentrated on discerning static symbols, a focal point that unfortunately resulted in inadequate performance when confronted with challenges such as changing orientations, wide-ranging variations, and cluttered backgrounds. A new robust hand gesture recognition model, dubbed HVCNNM, has been developed to address these inherent limitations. Uniquely, this model demonstrates remarkable flexibility, distinguishing both numerical and alphabetical gestures, thus broadening its application potential. This adaptability positions HVCNNM as an all-encompassing solution with diverse applicability in fields like HCI, security systems, and interactive gaming environments.
- The need to process Red Green Blue (RGB) images might be redundant in specific non-contact sensor applications. The complexity associated with applying neural network-based filtering across three color channels could entail considerable computational cost. A pragmatic alternative could be to utilize grayscale visual data, achieving the desired goal with less computational demand. In our study, we adopted this strategy by feeding a single data channel through the network. This approach reduced computational complexity and delivered compelling performance results when juxtaposed with those derived from color frames.
- We undertook rigorous experimentation to substantiate the contributions above, leveraging two widely recognized and challenging datasets. In this research, we developed an innovative hand gesture recognition model. To verify its efficacy, we undertook a comparative analysis of the performance of several pre-existing networks. The classification scores thus obtained underscored the robustness of our model, conclusively demonstrating its superior performance to current state-of-the-art methods.

This paper is structured as follows: Section 2 briefly reviews the relevant literature on gesture recognition. Section 3 describes the proposed methodology and offers a comprehensive dataset overview. In Section 4, we present and analyze the experimental data. The final Section 5 of this paper presents the findings of this study in concluding remarks.

2 Related Work

Healthcare, gaming, and robotics applications depend on gesture recognition for intuitive user interfaces [19]. The following subsections will provide a brief description of each method used to detect gestures, namely machine learning and DL.

2.1 Machine Learning-Based HGRoc

A sensor-based device and machine learning-based features are used in the first category, providing higher accuracy and robustness [20]. For sensor-based hand gesture recognition, specialized hardware is required, and gloves that contain sensors or external sensors can be cumbersome. Furthermore, these systems are more expensive and less accessible than vision-based methods using standard cameras, since they require extra equipment. Last but not least, sensor-based solutions can be vulnerable to electromagnetic interference or occlusions, which reduces their robustness and reliability. In contrast, vision-based approaches, which use cameras to acquire gesture images, are more versatile and require specialized devices, but their applications are limited. As a result of vision-based sensors for hand gesture recognition, natural interaction is possible without the need for specialized hardware, which makes them more accessible, cost-effective, and user-friendly. As part of this approach, images are segmented, matched, recognized, and classified using a variety of image processing techniques. This category has been the subject of significant research because it requires relatively little specialized equipment [21]. For instance, Nagarajan et al. [22] used an edge-oriented histogram approach to identify static gestures. Their multiclass support vector machine (SVM) classifier was able to accurately classify the gestures by extracting edge histogram counts from images of sign language alphabets. According to Wang et al., edge-based features in DL and computer vision are valuable for gesture recognition. A new method using superpixels and distance metric for recognizing marker-less hand gestures was developed using the Kinect depth camera [23].

Similarly, Gupta et al. [24] proposed a system that combined the histogram of oriented gradients (HOG) algorithms with the scale-invariant feature transform algorithm (SIFT). A single array of features extracted by both algorithms was used to improve classification performance. The extracted features by both algorithms were then merged into the one-dimensional array to improve the classification. This study demonstrated the effectiveness of combining features in computer vision applications by accurately classifying hand gestures using a standard k-nearest neighbor (KNN) classifier. Traditional feature extraction approaches, however, do not capture all salient classification features. KNN, for example, requires extensive storage and computational resources, so it is not well suited to large datasets. KNN also does not perform well in computer vision and DL applications because it simply memorizes training sets without learning properly.

2.2 Deep Learning-Based HGRoc

As a result of their superior feature extraction capabilities and recent technological advancements, DL-based methods are currently a popular research direction. HGRoc is a robust gesture recognition system that uses CNN to extract skin, hand, and orientation. Li et al. [25] developed a robust system using CNN. Zi Li et al. [26] developed an autoencoder to extract features from RGB images and

principal component analysis (PCA) for human actions. Oyedotun et al. [27] proposed a stacked denoising autoencoder to identify 24 hand gestures to improve gesture recognition accuracy and reduce errors. Li et al. [28] introduced an end-to-end CNN approach that incorporates a soft attention mechanism for robust HGRoc. Ranga et al. [29] utilized a hybrid feature extraction method combining discrete wavelet transform and Gabor filter and evaluated multiple classifiers, including random forest, KNN, SVM, and CNN for recognition. Chevtchenko et al. [30] proposed a fusion approach that utilizes both CNN and traditional features for gesture recognition. Ozcan et al. [31] proposed a heuristic optimization algorithm to optimize the hyperparameters of their proposed model for HGRoc. Neethu et al. [32] introduced a method for HGRoc based on finger detection using CNN. However, it may have limitations when applied to the ASL dataset. Wadhawan et al. [33] evaluated different CNN models with different optimizers to improve sign language recognition accuracy. The authors adjusted parameters such as the number of layers and filters to achieve higher validation accuracy. Liu et al. [34] suggested a 19-layer CNN followed by a single-shot multi-box detector for hand gesture identification and classification. Dadashzadeh et al. [35] introduced a two-stage CNN that fuses the color properties and segmented information for classification. Rathi et al. [36] proposed a 2-level architecture for HGRoc. However, using multimodal data like RGBD is limited due to special depth sensors requirement. Holdout validation is not always reliable, and very deep CNN architectures are computationally expensive and need a lot of labeled data for training.

After examining the existing literature, it is observed that conventional methods of extracting features limited to capture salient information for distinguishing similar gestures. Additionally, some classifiers can be memory intensive, requiring extra storage space. In addition, the main problem statement in HGRoc, as identified in the existing literature, is to develop robust and accurate systems that can recognize and classify hand gestures in real-world environments. This involves addressing several key challenges as follows:

- **Variability in hand gestures:** The inherent diversity in hand gestures, with wide-ranging variations in shape, size, orientation, and movement patterns, presents a significant challenge in devising a system capable of accurately identifying and categorizing all conceivable permutations.
- **Variability in lighting conditions:** The influence of lighting conditions on the visual representation of hand gestures in images or video sequences cannot be overstated. This factor significantly complicates the development of systems that maintain robustness amidst varying lighting conditions, introducing a formidable challenge in the field.
- **Occlusion:** The occurrence of partial or complete occlusion of hands and fingers in images or video frames poses a formidable challenge for systems in accurately recognizing and classifying hand gestures. The potential for occlusion adds another layer of complexity, further compounding the difficulty inherent in the task.
- **Real-time processing:** A multitude of HGRoc applications, including human-robot interaction and gaming, necessitate real-time processing of hand gestures. Achieving this level of immediacy, given the intricate nature of the recognition task, presents a substantial challenge. Real-time gesture recognition calls for efficient and rapid processing systems that can handle complex computations without significant latency.

The main objective of this paper is to develop a lightweight DL model that can be used to recognize and classify hand gestures accurately from images and video frames to address these challenges. In recent years, it has become increasingly popular to develop systems capable of recognizing dynamic hand gestures, such as sign language, which involve continuous movements of hands and fingers.

Dynamic Hand Gesture Recognition has developed new techniques and datasets still being used in this field.

3 The Proposed Method

Over the past decade, it has been primarily conventional techniques used to extract features to detect hand signs. However, the problem with these techniques is that they require a lengthy feature engineering process and have limited performance for detecting hand gestures. In addition, they tend to produce a high loss, particularly in surveillance settings with shadows, varying lighting conditions, and colored objects. To copy these challenges, we conducted a thorough study of DL architectures for precise hand gesture recognition. Drawing inspiration from current advancements and the promise of deep features, we explored a wide range of CNNs to enhance the accuracy of recognition and minimize the loss rate. Our method, the first step involves converting the input images into the Luminance Chrominance Blue Chrominance Red (YCRCB) color space. Subsequently, a Gaussian blur is applied, followed by skin color selection. To ensure continuity, any holes in the image are filled using dilation techniques. Once these preprocessing steps are completed, the images are converted into grayscale. The processed images are then passed through the proposed method for feature learning. A HGRoc framework is presented in Fig. 1, which shows the main steps.

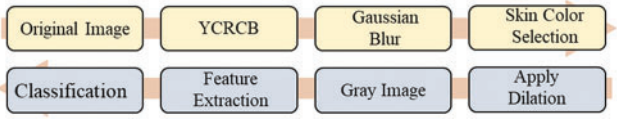


Figure 1: The main steps in the proposed model

A detailed description of the core components of our framework can be found in this section, and a visual illustration is in Fig. 2. With the proposed framework, humans and machines can interact naturally and intuitively by using gesture recognition. We have three main components: preprocessing, lightweight CNN models, and text-to-speech. The input image is first preprocessed. During this step, noise and the image's dimensions are reduced, making it easier to extract features effectively. By doing this, the classification step is more accurate in the future. The preprocessed images are then fed into our proposed lightweight CNN model, which maps the input images to their respective classes. A text-to-speech engine (TTSE) module is used to output the predicted gesture class in real-time, ensuring high accuracy at low computational cost. TTSE allows humans and machines to interact naturally and intuitively. This can be especially useful for people with disabilities since it provides audio feedback. As a result, users can communicate more easily with machines with a human-like interaction.

3.1 Motivation behind the Proposed Architecture

Developed from the intricate process of visual perception observed in living organisms, CNN is an advanced neural network architecture. It is designed to mimic how living creatures recognize and interpret visual information. CNN can recognize and learn complex patterns and features within images by processing and odelling data in multiple layers. Computer vision tasks such as image classification, object detection, and others can be performed with it effectively. In recent years, CNNs have proven highly effective in solving a variety of problems since the invention of LeNet, the first widely recognized DL architecture for classifying hand-written numbers. These computational challenges involve various computer vision tasks, including object reidentification [37], power estimation [38]. As a result of its superior performance over traditional feature-based approaches when dealing with

large-scale datasets, CNNs have gained significant traction in image classification. As well as using classifier learning paradigms, their success can be attributed to the ability to learn intricate and multi-scale features directly from raw data. In computer vision as well as other areas, CNNs have built-in capabilities that enable them to identify complex patterns and features within images.

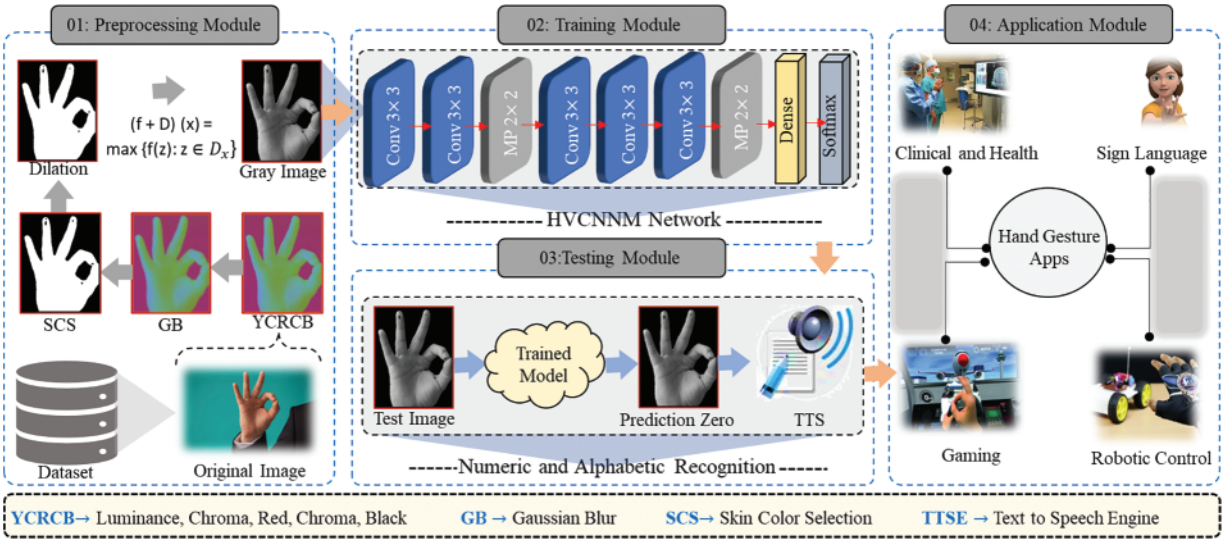


Figure 2: There are four main modules in the proposed HVCNNM for HGRoc: First, a grayscale image is created by processing different RGB images using a preprocessing technique. For representative, discriminative, and robust features, a lightweight model is proposed in the second module. A model based on unseen data is then tested for outcomes and converted to speech on unseen data. Finally, the proposed model has applications in the real world, particularly in healthcare centers

In the 2D convolution operation, multiple kernels (3×3) of the same size are used to generate feature maps based on the input data size (64×64). These feature maps are then passed on to the next operation. To reduce the dimensionality of feature vectors while also achieving translation invariance, the maximum activation values in a small odellingood are selected during this operation. In addition to convolution and subsampling, the fully connected layer is another critical component of the CNN pipeline. This layer is responsible for odelling high-level abstractions from the input data. During the training process, the neurons within both the convolution and fully connected layers adjust their weights to improve the representation of the input data. This process allows CNN to learn and recognize increasingly complex patterns and features within images, making it a highly effective tool for a wide range of computer vision tasks.

3.2 Preprocessing Module

Preprocessing is an essential step in gesture recognition systems because it improves the accuracy and efficiency of subsequent processing steps. First, the input image is converted into YCRCB color channels to select the skin color for extracting only the foreground of the hand. Then a morphological operation is applied to improve the contrast and detect edges, making it easier to identify important regions in the image. The bitwise operations are used to extract regions of interest and filter out background noise. Finally, the process image is converted from three RGB channels into one dimension, such as grayscale, as shown in Fig. 3. Their detailed mathematical representation of the proposed preprocessing is given in the Eqs. (1)–(4).

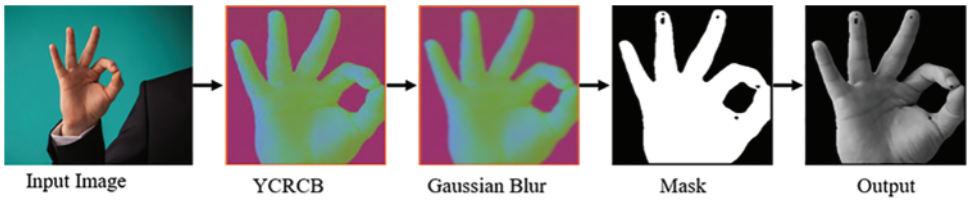


Figure 3: Details of the preprocessing techniques applied on the input image for further pre-processing

The conversion from an RGB color space to YCbCr can be performed using the following formulae, assuming R, G, and B are in the range [0, 255]:

$$Y = 0.299 \times R + 0.587 \times G + 0.114 \times B \tag{1}$$

$$Cb = -0.169 \times R - 0.331 \times G + 0.500 \times B + 128 \tag{2}$$

$$Cr = 0.5 * r - 0.419 * g - 0.081 * b + 128 \tag{3}$$

A Gaussian blur is usually applied using a convolution operation between the image and a Gaussian kernel (a square matrix). The Gaussian kernel can be represented as:

$$\mathrm{G}(\mathrm{x}, \mathrm{y}) = \frac{1}{2\pi\sigma^2} \mathrm{e}^{-\frac{\mathrm{x}^2+\mathrm{y}^2}{2\sigma^2}} \tag{4}$$

3.3 Proposed Lightweight CNN Model

In computer vision, HGRoc has always been a critical area, with numerous challenges emanating from the need for robust and precise models to extract detailed feature descriptors from input images. Given the complex and varied nature of hand gestures, the task necessitates models capable of deciphering high-level features from images. CNNs have demonstrated efficacy in this regard due to their sophisticated capability to learn and extract such high-level features. The present study introduces a lightweight CNN model specifically designed for HGRoc. The intent behind this model is to strike a balance between high accuracy and robustness in the classification of hand gestures. The architecture of this model is intricately composed of twelve layers, which include convolutional layers, pooling layers, dropout, and fully connected layers, each of which plays a distinctive role in the model's overall functionality.

The journey of an input image through the model begins with the first Conv2D layer, equipped with four filters. These filters carry out a 2D convolution operation on the image, bringing the lower-level features, such as edges and textures, to the fore. Following this, the image encounters another Conv2D layer, a MaxPooling2D layer, and a Dropout layer. The objective of these layers is to extract more complex, high-level features from the processed image, reduce the dimensionality of the image representation from $224 \times 224 \times 3$ to $50 \times 50 \times 1$ (thereby making the model computationally efficient), and prevent overfitting by randomly dropping out neurons during training. This pattern of layers is reiterated in two more sets of convolutional, pooling, and dropout layers. The repetition ensures a deep and comprehensive understanding of the image, allowing the model to learn a hierarchy of features that escalate in complexity, ultimately equipping it with the capability to distinguish between different hand gestures.

Following the extraction and learning of features, the model employs a Flatten layer, which transforms the 2D feature map into a 1D vector. This vector is then passed through two dense layers,

which further enhances the model's learning. The final dense layer, equipped with 36 neurons, is a critical component as it is responsible for outputting the probabilities for the corresponding 36 hand gesture classes. Training the proposed CNN model involves a challenging dataset that incorporates both numeric and alphabetic gestures, featuring a range of variations in hand postures, lighting conditions, and backgrounds. The optimizer used is the Adam optimizer combined with stochastic gradient descent. This combination efficiently navigates the model through the solution space to find the optimal weights. Furthermore, to gauge the difference between the predicted and actual probability distributions, the model uses the categorical cross-entropy loss function. This quantification of prediction error forms the crux of the learning process, guiding the model's parameter adjustments to reduce this error over time. Evaluation of the model's performance involves using a test set comprising images not exposed to the model during training. This step ensures an unbiased assessment of the model's generalization capability. The evaluation metrics employed include accuracy, precision, recall, and F1-score, each offering a unique perspective on the model's performance. The cumulative insights from these metrics provide a holistic view of the model's effectiveness in hand gesture classification.

3.4 Application Module

Gesture recognition technology finds valuable applications in healthcare and various domains. It enhances HCI, facilitates sign language translation, enables precise control in robotics and prosthetics, improves surgical training and assistance, and promotes hygiene and infection control. In virtual reality and augmented reality, gesture recognition enhances immersive experiences. It allows users to manipulate virtual objects and execute commands intuitively. In robotics and prosthetics, hand gesture recognition enables precise control over robotic systems and prosthetic limbs, enhancing mobility and independence. In surgical settings, it assists surgeons in controlling virtual simulations and surgical instruments with intuitive hand gestures, improving skills and precision. Additionally, in healthcare environments, touchless interfaces based on hand gestures promote hygiene and infection control by reducing the risk of cross-contamination. These applications demonstrate the broad potential of gesture recognition technology in enhancing various aspects of healthcare and other domains. Their visual representation is shown in Fig. 4.

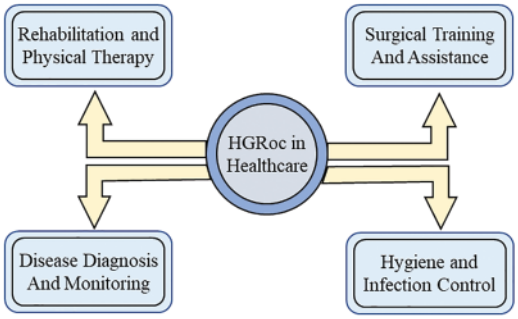


Figure 4: Application of HGRoc in healthcare in different domains

4 Experimental Results

This section briefly elaborates on the hand gesture recognition datasets with visual samples. Furthermore, comprehensive results are generated to prove the strength of the proposed model. Eventually, the obtained classification score is fairly compared with the state-of-the-art techniques. The details of these subsections are given below.

4.1 Dataset

In this part of the research, a brief explanation of two challenging hand gesture recognition datasets is provided below with visual samples for the reader's understanding.

4.1.1 Massey University Dataset (MUD)

For HGRoc there are different datasets available, but we have chosen a challenging dataset that includes both numbers and alphabets, and researchers actively used and benchmarked their proposed model in recent articles. A MUD dataset is used, which contains 36 classes of sign characters, including 26 and 0–9 alphabets and numerals. This dataset consists of 70 images for each class, resulting in 2520 images with scale, illumination, and rotation variations. Therefore, in this article, we have selected this dataset to compare the performance of our method with existing state-of-the-art methods. Samples of the dataset are shown in Fig. 5.

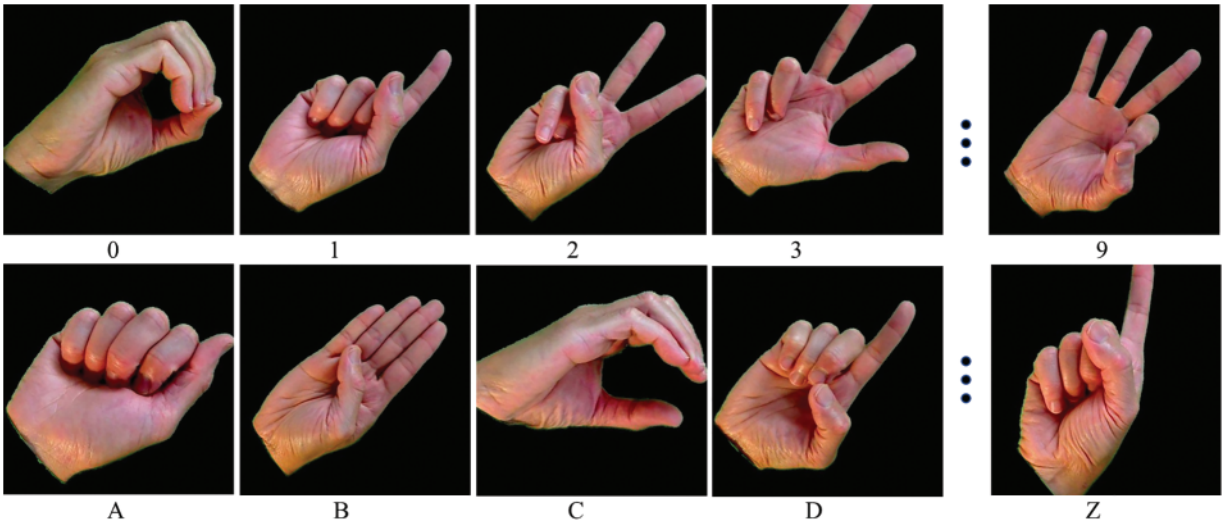


Figure 5: The dataset evaluates the proposed model and comparison with existing state-of-art models. Here, the first and second rows represent a numerical and alphabetic of MUD, while the last row samples are taken from ASLAD

4.1.2 ASL Alphabet Dataset (ASLAD)

The ASL Alphabet database contains 87,000 RGB hand gesture images, which are categorized into 29 different classes. Among these classes, 26 represent individual letters of the alphabet, while one represents a space character and another represents a deleted character. Additionally, a specific default image is included in the dataset that represents no symbols. Overall, the dataset comprises 3000 images per class, each representing a unique hand gesture. The last row in Fig. 5 visually represents some of the images included within the ASL Alphabet database, showcasing the variety of hand gestures present within the dataset.

4.2 Evaluation Metrics

The proposed framework is assessed by standard evaluation metrics, namely Precision, Recall, F1-score, and Accuracy. The mathematical representation of these evaluation metrics can be found in

Eqs. (5) through (8). By utilizing these metrics and equations, we were able to quantitatively assess the framework's performance and determine how well it performed in meeting our objectives.

$$A = \left(\frac{\text{Tpositive} + \text{Tnegative}}{\text{Tpositive} + \text{Tnegative} + \text{Fpositive} + \text{Fnegative}}\right) \tag{5}$$

$$P = \left(\frac{\text{Tpositive}}{\text{Tpos} + \text{Fpos}}\right) \tag{6}$$

$$R = \left(\frac{\text{Positive}}{\text{Tpos} + \text{Fneg}}\right) \tag{7}$$

$$F1 = 2* \left(\frac{P * R}{P + R}\right) \tag{8}$$

4.3 Implementation Detail and System Configuration

We implemented the proposed framework using the Python programming language and utilized TensorFlow DL on a computing system equipped with an NVIDIA GeForce RTX 3090 GPU to accelerate computation time. We divided each dataset into three subsets to carry out the training process: 70% of the data was used for training. In contrast, the remaining 10% and 20% were reserved for model validation and testing. We conducted the training process for a total of 25 epochs, with a batch size of 64 and a learning rate of 0.001 with Adam optimizer. These specific values were chosen based on their effectiveness in achieving optimal training results. By utilizing these parameters, we were able to effectively train the framework and evaluate its performance on the given dataset.

4.4 Ablation Study over MUD

HGRoc is effectively used in different applications, therefore, to overcome the scalar sensor-based HGRoc needs efficient computer vision-assisted DL methods for accurate recognition. In this direction, different techniques are developed [39,40], however, the mainstream methods used computational complex models for HGRoc which is unable to deploy for resource-constrained devices such as raspberry pi to assist the IoT for real-time recognition. Therefore, to overcome these challenges a lightweight model is required to efficiently balance the trade-off between computational complexity and performance for real-time sign language recognition.

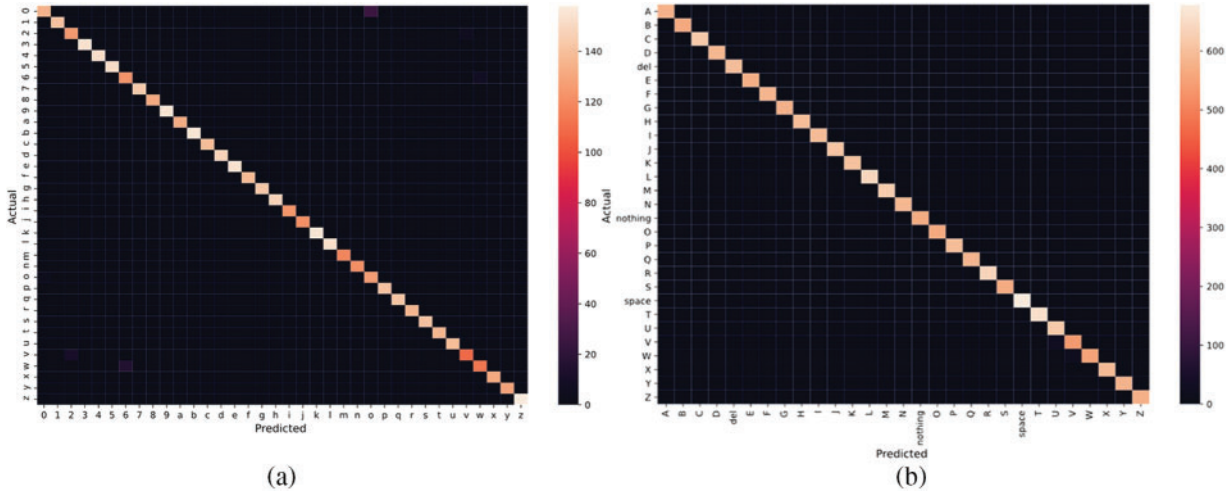
In the literature, different CNN models are used to benchmark the performance of their proposed model for HGRoc. The results from the evaluation of the different pre-trained CNN models for hand gesture recognition are presented in Table 1. Overall, the results show that the proposed model outperformed all other models in terms of various evaluation metrics, achieving a Precision score of 0.99, a Recall score of 0.99, and an F1-score of 0.99. The main reason for the superior performance of the proposed model is to decrease the spatial dimension of the input image from 224 × 224 to 50 × 50. As a result, it requires limited computation and considers the most suitable choice for edge devices such as IoT. Furthermore, we study from the experimentation the existing models are recognized HGRoc from the three channels, such as RGB (Red, Green, Blue).

Which are not only required high computation when the convolution operation performed but also restrict the underline models from efficiently mapping the extracted features into their corresponding classes because of processing redundancy data. Therefore, we not only converted the input image from RGB into a single channel, such as a grayscale but also reduced the spatial dimension from 224 × 224 to 50 × 50. Based on these changes, we proposed a lightweight CNN

model which achieved higher performance than existing methods but also required less computation. In comparison, the MobileNet model also performed well, achieving a Precision score of 0.96, a Recall score of 0.95, an F1-score of 0.94, and an Accuracy score of 0.95. This is likely due to the architecture of the MobileNet model, which is optimized for mobile and embedded devices and has a relatively small number of parameters compared to other models. The InceptionV3 and VGG16 models attained similar performance scores, with Precision scores of 0.94 and Recall scores of 0.94 and 0.93, respectively. These models have been widely used in computer vision tasks and are known for their high accuracy and robustness. On the other hand, the MobileNetV2 and NASNet Mobile models performed poorly, achieving low precision, recall, and F1-scores. The main reason is attributed to the fact that these models were developed for specific applications and it may not be well-suited for hand gesture recognition. Additionally, the relatively small size of the hand gesture dataset used in this study may not have been sufficient to fully leverage these models' complexity and capacity. The confusion matrix and training accuracy and loss graph for both datasets are shown in Figs. 6a and 6b.

Table 1: Comparing various backbone models with the proposed model for hand gesture recognition using MUD. The best and the second-best model performance is highlighted in bold and underlined

Models	Precision	Recall	F1-score	Accuracy
InceptionV3	0.94	0.94	0.93	0.94
VGG16	0.94	0.93	0.92	0.93
MobileNet	<u>0.96</u>	<u>0.95</u>	<u>0.94</u>	<u>0.95</u>
MobileNetV2	0.922	0.875	0.87	0.875
NASNetMobile	0.648	0.625	0.6	0.62
HVCNNM (Proposed)	**0.99**	**0.99**	**0.99**	**0.99**



(a) (b)

Figure 6: Confusion matrix generated via the proposed model for (a) MUD and (b) ASLAD

4.5 Comparative Analysis

For fair evaluation, the performance obtained via the proposed model is compared with the current attempt in HGRoc and used in the same dataset. In [30], they developed a model through the assistance of deep and traditional techniques for the gesture recognition task. They examine both the individual and integrated performance as well. The model accuracy and loss for training and testing is presented in Fig. 7, where demonstrates an accurate recognition rate.

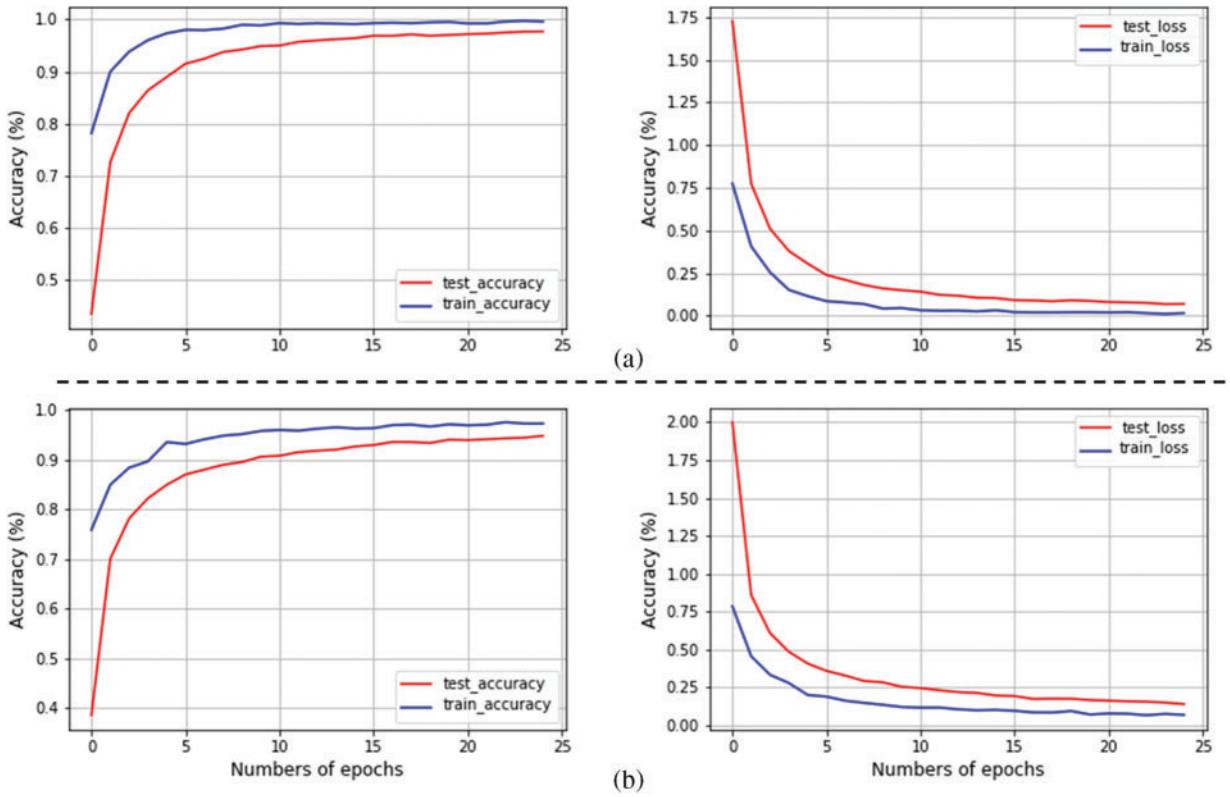


Figure 7: Graphical representation for training and validation test accuracy and loss for the proposed model with 25 epochs using (a) MUD and (b) ASLAD

The main issue of their network is to process RGB images with vast dimensions of information descriptors resulting in better, but the entire procedure is complex. Similarly, reference [39] used two different pre-trained models with a machine-learning classifier for the same task. The comparative results are reported in the Table 2.

Table 2: Performance comparison among the proposed model and state-of-the-art techniques using MUD. The best and second-best performance is represented in bold and underlined

Methods	Accuracy (%)
CNN [30]	73.86 ± 1.04
FFCN [30]	84.02 ± 0.59
AlexNet-SVM [39]	70.00

(Continued)

Table 2 (continued)

Methods	Accuracy (%)
VGG16-SVM [39]	70.00
Modified-AlexNet [21]	90.26 ± 1.35
Attention-VGG16 [41]	85.15
NMF-CS [42]	<u>98.13</u>
DSAE [43]	83.36
HVCNNM (Proposed)	**99.00**

However, the main problem in their network is utilization of a considerable number of parameters and following the complex architecture for a simple problem. Furthermore, these models practice excessive filters with diverse dimensions, making them more complicated. This work is further boosted by [21] where AlexNet is profoundly modified for domain-specific and achieved an incredible recognition score. Next, attention networks have been widely used for image-based classification to capture local features and handle complex backgrounds. To address this challenge, reference [41] proposed an attention-based framework; however, the recognition performance remains constrained. Despite leveraging pre-trained networks, several researchers have formulated their own CNN architectures for gesture recognition. For instance, in [43], the DSAE was introduced, with diverse experimental results derived from the analysis of different layers within the CNN model. The main problem in this network is that forwarding large input sizes of an image to the network and filtering dimension cannot extract more discriminative features. Motivated by a similar strategy, reference [42] introduced a novel deep learning model designed to extract more salient features, demonstrating remarkable performance on RGB data. In this study, we aim to obtain a high recognition rate over greyscale images, therefore we proposed a novel and lightweight CNN model for precise classification. Our proposed strategy achieved new state-of-the-art performance based on comprehensive experiments over various pre-trained models, which can assist the healthcare center and many more real-world applications. The overall quantative results are given in the Table 3.

Table 3: Performance comparison among the proposed model and state-of-the-art techniques using ASLAD. The best and second-best performance is represented in bold and underlined

Method	Accuracy (%)
VGG16 [44]	93.71
ResNet-50 [45]	92.62
DenseNet201 [46]	91.86
SENet 2018 [47]	93.87
ORB-MLP [20]	<u>96.96</u>
ORB-KNN [21]	95.81
DeReFNet [4]	95.70
HVCNNM (Proposed)	**99.23**

The ASLAD is comparatively more straightforward than the MUD because it only contains an alphabet. Extensive work has been conducted on this dataset, where numerous traditional and DL models are widely exploited. Here also, pre-trained networks were commonly used for gesture recognition. For example, in [44], VGG16 is applied by removing the upper layer and tuning for HGRoc. Such a model comprises complex architecture with many convolutional layers and utilizes huge parameters during training. Similarly, ResNet-50 [45] is also commonly used for image classification, but here the dataset samples in each class are straightforward, so its performance is not better. In another study, a DenseNet201 [46] performance is deeply analyzed, but the recognition score is not improved compared to the other existing models. Besides this, neural network and hand-crafted based features are integrated [20] for HGRoc that obtained accuracy on a high margin. In addition, the same authors also analyzed ORB features with the KNN classifier but get a low score. Due to low performance over pre-trained models researchers are diverted to their own generated model for instance [4], which demonstrates incredible performance because of simple architecture. Motivated by such a strategy, a novel convolutional-based framework is designed to increase classification performance.

5 Conclusion and Future Research Direction

This research paper introduces a new technique for recognizing sign language hand gestures using a vision-based approach. The proposed technique involves a DL-based HVCNNM model designed to achieve a compact representation of hand gestures. Different CNN models are investigated in this study and modified for the purpose of classifying sign language hand gestures. It is easier and more accessible to recognize hand gestures with the proposed vision-based model since it does not require external hardware equipment or user dependence. Hand signs can be recognized accurately and efficiently using this approach, surpassing the performance of current approaches. An accuracy score of 99.23% was achieved for one dataset, while 99.00% for the other dataset was achieved using the proposed model. Based on these findings, the proposed vision-based model is highly effective at recognizing hand gestures in sign language and could improve communication convenience and accessibility for deaf and hard-of-hearing people. As our dataset evolves, we intend to include more signs in a variety of languages in the future. We will develop attention-based DL method for smoothly execution over the edges to ensure real-time recognition [48].

Acknowledgement: Not applicable.

Funding Statement: This research is funded by Researchers Supporting Project Number (RSPD2024 R947), King Saud University, Riyadh, Saudi Arabia.

Author Contributions: Study conception, design, and original draft writing: Khursheed Aurangzeb, Khalid Javeed; data collection: Musaed Alhussein; analysis, interpretation of results and writing review/editing: Imad Rida, Syed Irtaza Haider; draft manuscript preparation and editing: Anubha Parashar. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: Not applicable.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] A. Dayal, M. Aishwarya, S. Abhilash, C. K. Mohan, A. Kumar *et al.,* "Adversarial unsupervised domain adaptation for hand gesture recognition using thermal images," *IEEE Sensors Journal*, vol. 23, pp. 3493–3504, 2023.

[2] A. Hussain, T. Hussain, W. Ullah and S. W. Baik, "Vision transformer and deep sequence learning for human activity recognition in surveillance videos," *Computational Intelligence and Neuroscience*, vol. 2022, pp. 3454167, 2022. https://doi.org/10.1155/2022/3454167

[3] A. Hussain, S. U. Khan, N. Khan, I. Rida, M. Alharbi *et al.,* "Low-light aware framework for human activity recognition via optimized dual stream parallel network," *Alexandria Engineering Journal*, vol. 74, pp. 569–583, 2023.

[4] J. P. Sahoo, S. P. Sahoo, S. Ari and S. K. Patra, "DeReFNet: Dual-stream dense residual fusion network for static hand gesture recognition," *Displays*, vol. 77, pp. 102388, 2023.

[5] S. Liu, Y. Li and W. Fu, "Human-centered attention-aware networks for action recognition," *International Journal of Intelligent Systems*, vol. 2022, pp. 10968–10987, 2022.

[6] S. Liu, S. Huang, W. Fu and J. C. W. Lin, "A descriptive human visual cognitive strategy using graph neural network for facial expression recognition," *International Journal of Machine Learning and Cybernetics*, pp. 1–17, 2022. https://doi.org/10.1007/s13042-022-01681-w

[7] T. L. Dang, T. H. Pham, Q. M. Dang and N. Monet, "A lightweight architecture for hand gesture recognition," *Multimedia Tools and Applications*, vol. 2023, pp. 1–19, 2023.

[8] T. R. Gadekallu, G. Srivastava, M. Liyanage, M. Iyapparaja, C. L. Chowdhary *et al.,* "Hand gesture recognition based on a Harris hawks optimized convolution neural network," *Computers and Electrical Engineering*, vol. 100, pp. 107836, 2022.

[9] M. Oudah, A. Al-Naji and J. Chahl, "Hand gesture recognition based on computer vision: A review of techniques," *Journal of Imaging*, vol. 6, no. 8, pp. 73, 2020.

[10] S. Bhushan, M. Alshehri, I. Keshta, A. K. Chakraverti, J. Rajpurohit *et al.,* "An experimental analysis of various machine learning algorithms for hand gesture recognition," *Electronics*, vol. 11, no. 6, pp. 968, 2022.

[11] A. Hussain, H. Ullah, A. Ullah, A. S. Imran and M. Y Lee, "Anomaly based camera prioritization in large scale surveillance networks," *Computers, Materials & Continua*, vol. 70, no. 2, pp. 2171–2190, 2021.

[12] S. Habib, A. Hussain, W. Albattah, M. Islam, S. Khan *et al.,* "Abnormal activity recognition from surveillance videos using convolutional neural network," *Sensors*, vol. 21, no. 24, pp. 8291, 2021. https://doi.org/10.3390/s21248291

[13] A. Hussain, M. Ahmad, I. A. Mughal and H. Ali, "Automatic disease detection in wheat crop using convolution neural network," in *Int. Conf. on Next Generation Computing*, Vung Tau City, Vietnam, pp. 301–304, 2018.

[14] A. Hussain, Z. A. Khan, T. Hussain, F. U. M. Ullah, S. Rho *et al.,* "A hybrid deep learning-based network for photovoltaic power forecasting," *Complexity*, vol. 2022, pp. 7040601, 2022. https://doi.org/10.1155/2022/7040601

[15] S. Sharma and K. Guleria, "A deep learning based model for the detection of Pneumonia from chest X-ray images using VGG-16 and neural networks," *Procedia Computer Science*, vol. 218, pp. 357–366, 2023.

[16] R. Z. Khan and N. A. Ibraheem, "Hand gesture recognition: A literature review," *International Journal of Artificial Intelligence & Applications*, vol. 3, no. 4, pp. 161–174, 2012.

[17] X. Liu and K. Fujimura, "Hand gesture recognition using depth data," in *IEEE Int. Conf. on Automatic Face and Gesture Recognition*, Waikoloa Beach, USA, pp. 529–534, 2004.

[18] S. S. Rautaray and A. Agrawal, "Vision based hand gesture recognition for human computer interaction: A survey," *Artificial Intelligence Review*, vol. 43, pp. 1–54, 2015.

[19] M. Muneeb, H. Rustam and A. Jalal, "Automate appliances via gestures recognition for elderly living assistance," in *Int. Conf. on Advancements in Computational Sciences*, Lahore, Pakistan, pp. 1–6, 2023.

[20] M. M. Damaneh, F. Mohanna and P. Jafari, "Static hand gesture recognition in sign language based on convolutional neural network with feature extraction method using ORB descriptor and Gabor filter," *Expert Systems with Applications*, vol. 211, pp. 118559, 2023.

[21] J. P. Sahoo, A. J. Prakash, P. Pławiak and S. Samantray, "Real-time hand gesture recognition using fine-tuned convolutional neural network," *Sensors*, vol. 22, no. 3, pp. 706, 2022.
[22] S. Nagarajan and T. Subashini, "Static hand gesture recognition for sign language alphabets using edge oriented histogram and multi class SVM," *International Journal of Computer Applications*, vol. 82, no. 4, pp. 28–35, 2013.
[23] C. Wang, Z. Liu and S. C. Chan, "Superpixel-based hand gesture recognition with kinect depth camera," *IEEE Transactions on Multimedia*, vol. 17, no. 1, pp. 29–39, 2014.
[24] B. Gupta, P. Shukla and A. Mittal, "K-nearest correlated neighbor classification for Indian sign language gesture recognition using feature fusion," in *Int. Conf. on Computer Communication and Informatics*, Coimbatore, India, pp. 1–5, 2016.
[25] H. I. Lin, M. H. Hsu and W. K. Chen, "Human hand gesture recognition using a convolution neural network," in *IEEE Int. Conf. on Automation Science and Engineering*, New Taipei, Taiwan, pp. 1038–1043, 2014.
[26] S. Z. Li, B. Yu, W. Wu, S. Z. Su and R. R. Ji, "Feature learning based on SAE-PCA network for human gesture recognition in RGBD images," *Neurocomputing*, vol. 151, pp. 565–573, 2015.
[27] O. K. Oyedotun and A. Khashman, "Deep learning in vision-based static hand gesture recognition," *Neural Computing and Applications*, vol. 28, no. 12, pp. 3941–3951, 2017.
[28] Y. Li, X. Wang, W. Liu and B. Feng, "Deep attention network for joint hand gesture localization and recognition using static RGB-D images," *Information Sciences*, vol. 441, pp. 66–78, 2018.
[29] V. Ranga, N. Yadav and P. Garg, "American sign language fingerspelling using hybrid discrete wavelet transform-gabor filter and convolutional neural network," *Journal of Engineering Science and Technology*, vol. 13, no. 9, pp. 2655–2669, 2018.
[30] S. F. Chevtchenko, R. F. Vale, V. Macario and F. R. Cordeiro, "A convolutional neural network with feature fusion for real-time hand posture recognition," *Applied Soft Computing*, vol. 73, pp. 748–766, 2018.
[31] T. Ozcan and A. Basturk, "Transfer learning-based convolutional neural networks with heuristic optimization for hand gesture recognition," *Neural Computing and Applications*, vol. 31, pp. 8955–8970, 2019.
[32] P. Neethu, R. Suguna and D. Sathish, "An efficient method for human hand gesture detection and recognition using deep learning convolutional neural networks," *Soft Computing*, vol. 24, pp. 15239–15248, 2020.
[33] A. Wadhawan and P. Kumar, "Deep learning-based sign language recognition system for static signs," *Neural Computing and Applications*, vol. 32, pp. 7957–7968, 2020.
[34] P. Liu, X. Li, H. Cui, S. Li and Y. Yuan, "Hand gesture recognition based on single-shot multibox detector deep learning," *Mobile Information Systems*, vol. 2019, pp. 1–7, 2019.
[35] A. Dadashzadeh, A. T. Targhi, M. Tahmasbi and M. Mirmehdi, "HGR-Net: A fusion network for hand gesture segmentation and recognition," *IET Computer Vision*, vol. 13, no. 8, pp. 700–707, 2019.
[36] P. Rathi, R. Kuwar Gupta, S. Agarwal and A. Shukla, "Sign language recognition using resnet50 deep neural network architecture," in *Int. Conf. on Next Generation Computing Technologies (NGCT-2019)*, pp. 7, 2020. https://doi.org/10.2139/ssrn.3545064
[37] S. U. Khan, I. U. Haq, N. Khan, K. Muhammad, M. Hijji *et al.,* "Learning to rank: An intelligent system for person reidentification," *International Journal of Intelligent Systems*, vol. 37, no. 9, pp. 5924–5948, 2022.
[38] S. U. Khan, N. Khan, F. U. M. Ullah, M. J. Kim, M. Y. Lee, *et al.,* "Towards intelligent building energy management: AI-based framework for power consumption and generation forecasting," *Energy and Buildings*, vol. 279, pp. 112705, 2023.
[39] A. A. Barbhuiya, R. K. Karsh and R. Jain, "CNN based feature extraction and classification for sign language," *Multimedia Tools and Applications*, vol. 80, no. 2, pp. 3051–3069, 2021.
[40] S. Sharma and S. Singh, "Vision-based hand gesture recognition using deep learning for the interpretation of sign language," *Expert Systems with Applications*, vol. 182, pp. 115657, 2021.
[41] A. A. Barbhuiya, R. K. Karsh and R. Jain, "Gesture recognition from RGB images using convolutional neural network-attention based system," *Concurrency and Computation: Practice and Experience*, vol. 34, no. 24, pp. e7230, 2022.

[42] H. Zhuang, M. Yang, Z. Cui and Q. Zheng, "A method for static hand gesture recognition based on non-negative matrix factorization and compressive sensing," *IAENG International Journal of Computer Science*, vol. 44, no. 1, pp. 52–59, 2017.
[43] V. Kumar, G. C. Nandi and R. Kala, "Static hand gesture recognition using stacked denoising sparse autoencoders," in *Int. Conf. on Contemporary Computing*, Noida, India, pp. 99–104, 2014.
[44] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," arXiv preprint arXiv:1409.1556, 2014.
[45] K. He, X. Zhang, S. Ren and J. Sun, "Deep residual learning for image recognition," in *IEEE Computer Vision and Pattern Recognition*, Las Vegas, Nevada, USA, pp. 770–778, 2016.
[46] G. Huang, Z. Liu, L. Van Der Maaten and K. Q. Weinberger, "Densely connected convolutional networks," in *IEEE Computer Vision and Pattern Recognition*, Honolulu, Hawaii, USA, pp. 4700–4708, 2017.
[47] J. Hu, L. Shen and G. Sun, "Squeeze-and-excitation networks," in *IEEE Conf. on Computer Vision and Pattern Recognition*, Salt Lake City, Utah, USA, pp. 7132–7141, 2018.
[48] S. Liu, Y. Li and W. Fu, "Human centered attention aware networks for action recognition," *International Journal of Intelligent Systems*, vol. 37, no. 12, pp. 10968–10987, 2022.