Non-Manual Gesture Recognition using Transfer Learning Approach

Sameena Javaid[1*], Safdar Rizvi[1], Muhammad Talha Ubaid[2], Amara Kiran[2]
[1]Department of Computer Sciences, School of Engineering and Applied Sciences, Bahria University Karachi Campus, Karachi, Pakistan.
[2]Faculty of Information Technology and Computer Science, University of Central Punjab, Lahore, Pakistan
* **Correspondence**: Sameena Javaid Email: sameenajaved.bukc@bahria.edu.pk
Citation | Javaid. S, Rizvi. S, Ubaid. M. T, Kiran. A, "Non-Manual Gesture Recognition using Transfer Learning Approach", IJIST, Vol. 5 Issue. 4 pp 576-591, Nov 2023
Received | Oct 24, 2023 **Revised |** Nov 07, 2023 **Accepted |** Nov 17, 2023 **Published |** Nov 19, 2023.

Individuals with limited hearing and speech rely on sign language as a fundamental nonverbal mode of communication. It communicates using hand signs, yet the complexity of this mode of expression extends beyond hand movements. Body language and facial expressions are also important in delivering the entire information. While manual (hand movements) and non-manual (facial expressions and body movements) gestures in sign language are important for communication, this field of research has not been substantially investigated, owing to a lack of comprehensive datasets. The current study presents a novel dataset that includes both manual and non-manual gestures in the context of Pakistan Sign Language (PSL). This newly produced dataset consists of. MP4 format films containing seven unique motions involving emotive facial expressions and accompanying hand signs. The dataset was recorded by 100 people. Aside from sign language identification, the dataset opens up possibilities for other applications such as facial expressions, facial feature detection, gender and age classification. In this current study, we evaluated our newly developed dataset for facial expression assessment (non-manual gestures) by YOLO-Face detection methodology successfully extracts faces as Regions of Interest (RoI), with an astounding 90.89% accuracy and an average loss of 0.34. Furthermore, we have used Transfer Learning (TL) using VGG16 architecture to classify seven basic facial expressions and succeeded with 100% accuracy. In summary, our study produced two different datasets, one with manual and non-manual sign language gestures, the second with Asian faces to find seven basic facial expressions. With both the dataset, our validation techniques found promising results.

Keywords: Face Expressions, Non-Manual Gestures, Pakistan Sign Language, Region of Interest, YOLO-Face, Transfer Learning, VGG16

Introduction:

Sign Language (SL) is the primary way of communication among deaf and mute people, but most hearing people don't know the SL. Reliance on translators is inconvenient as well as costly [1]. Many researchers have experimented with automatic Sign Language Recognition (SLR) using Machine Learning (ML) and Deep Learning (DL). Generally, these efforts have been constrained due to insufficient and restrictive datasets. On the other hand, sign language gestures are not restricted to conveying meanings only through hand movements (manual gestures); instead, they require simultaneous body movement, facial expressions, gaze directions, and head orientation (non-manual gestures).

Further, the grammatical structure of sign language is supported by facial expressions in terms of grammatical and affective meanings, differentiating the lexical items, syntactic building, and intensification progression [2]. Recognition of non-manual sign language gestures can improve the SLR accuracy rate. There are cases where facial expressions can change the meaning of manual signs or intensify effective communication [3].

This work introduces the Pakistan Sign Language Manual Non-Manual (PkSLMNM) gestures dataset, which integrates manual and non-manual features in a single movie. This multimodal dataset is made up of movies that vividly illustrate seven emotive emotion-based gestures, which were meticulously captured by a diverse group of 100 people. The PkSLMNM dataset distinguishes itself by containing 700 films in .MP4 format exhibiting seven dynamic signals with a variety of expressive features such as hand motions, facial expressions, head movements, torso gestures, body movements, eye gazing, eyeball movements, and directions. The establishment of the PkSLMNM dataset is the first step in developing a complete resource for Pakistan Sign Language (PSL), to enable the development of automated solutions to serve the deaf and mute community. While static gestures have been successfully used in PSL-based communication solutions, a lack of studies and publicly available datasets have created a significant gap in the recognition of multi-modal dynamic hand gesture recognition, as well as the availability of video data for training and testing.

The PkSLMNM dataset provides numerous chances for research-level tasks in robotics, Human-Computer Interaction (HCI), Computer Vision (CV), Machine Learning (ML), and Deep Learning (DL). It opens the door to new applications and solutions that can improve communication and connection with the deaf community while also furthering the fields of artificial intelligence and HCI. This information not only meets a fundamental need, but it also catalyzes the creation of sophisticated systems that promote inclusion and accessibility for people with hearing and speech impairments. The current dataset is designed explicitly for PSL having manual and non-manual affective dynamic gesture recognition. However, it may also facilitate the training and testing of data for static or dynamic facial recognition, hand gesture recognition, and head and shoulder detection and movement. Using the latest research practices may also expedite several preprocessing techniques, action recognition, pose estimation and tracking, video streaming, frame selection, compression techniques, data augmentation, etc. Additionally, this dataset is collected in a controlled environment, with a consistent white background and specific lighting, brightness, and temperature levels. On the other hand, every person's orientation has its direction of the left or right-hand mix, with the challenging resolution, scale, and rotations of signers. This dataset contains original metadata from the camera used for collection.

Background:

SL is a native language for the deaf and mute community, similar to natural languages in its adherence to social standards and cognitive functions. Its mode of communication, however, differs greatly from vocal, auditory, and spoken languages. Sign language is a Visio temporal construct that transmits meanings through a diverse range of body motions and

gestures, with a focus on hand movements, head gestures, upper body motion, and detailed facial expressions. These components work together to build a nuanced and sophisticated language system that enables people with hearing and speech impairments to communicate and grasp complex ideas and emotions [4].

The field of robust automatic SLR is a complex and ever-changing research topic. It has made great improvement in recent years, reflecting the increased attention it has received. However, it is critical to recognize that SLR varies significantly across countries and areas because to the distinctive character of diverse sign languages. These languages have distinct syntax and semantics that distinguish them from traditional spoken and written languages [5]. Nonetheless, the techniques and approaches used in automating Sign Language detection, particularly those based on Computer Vision, Machine Learning, and Deep Learning, tend to share fundamental ideas and methodology across the sign language spectrum. While the syntax and semantics may differ, the underlying technology and approaches are frequently quite similar. In this scenario, creating a specialized dataset suited to that specific regional or local Sign Language becomes a core prerequisite when commencing the development and testing of a recognition system or model for a novel sign language. Such datasets form the foundation for training and evaluating the performance of SLR systems, assuring their adaptability and efficacy in a wide range of linguistic and cultural situations.

Each region or country has historically evolved its own distinct Sign Language, such as "American Sign Language" (ASL), "British Sign Language" (BSL), "Turkish Sign Language" (TSL), "Chinese Sign Language" (CSL), "Indian Sign Language" (ISL), and "Arabic Sign Language" (ArSL), among many others. "Pakistan Sign Language" (PSL) is the predominant language of the deaf and mute minority in Pakistan. A publicly accessible dataset for Pakistan Sign Language, on the other hand, is uncommon. This lack of datasets poses a substantial difficulty to scientific research in the field of sign language recognition and communication [6]. Continuous sign language recognition is critical in today's fast-evolving world of real-time applications. The intricacy of sign language, combined with the requirements of video processing, highlights the importance of large datasets and studies in this area. The creation of datasets that reflect regional and native sign languages is critical for stimulating innovation in sign language technology and facilitating successful communication with deaf and mute people

Numerous researchers have been working intensively on the computational automation of PSL, but their achievements have not been made publicly available. The Deaf Reach Program, in partnership with the PSL organization, embarked on a revolutionary project in the early 20s by building a repository comprising both static and dynamic signs, totaling a stunning 5000 signals [7]-[8]. This program was essential for the deaf and mute community since it provided access to a large collection of signs. It should be noted, however, that this repository was compiled by a single signer and concentrated on individual sign activities in a controlled and somewhat less complex context.

The demand for datasets that capture the intricacies of real-world settings is high in the fields of computer vision and deep learning. From 2019 to 2022, significant development has occurred to address this demand. Several academics have taken significant steps during this period to make datasets available to computer vision, machine learning, and deep learning communities [6]. These datasets have the potential to considerably advance research and development activities in Pakistan Sign Language automation, paving the path for improved communication and interaction for those who have hearing and speech disabilities.

In 2019, Saad Butt and his fellow researchers published a dataset named PSL-(OpenPose) for automatic sign recognition on Kaggle. This dataset includes over 2000 images of 37 Urdu alphabets and 700 photos of different commonly used Urdu words collected by a

webcam. Nine subjects recorded these signs. They provided the dataset in the form of a JSON file consisting of skeletal critical points of each data point of human pose detection estimation using the OpenPose library. We found this dataset less challenging in terms of data robustness and domain-specificity [9]. Similarly, another dataset was published on Mendeley in 2021 by Ali Imran and his fellows; they collected 40 images of a single-hand configuration through a webcam having multiple orientations. They targeted Static Urdu alphabets and covered 37 classes with single-subject hand images [10]. In the current year 2022, another research study contributed in a new dimension of emotion representations of sign language using hand signs with seven basic emotions, they evaluated good results with transfer learning [11].

Previous datasets have mostly concentrated on static sign language motions displayed as still photos, with none containing dynamic sign language gestures given in multi-modal video forms. Furthermore, these datasets frequently have scope and variety restrictions, which might result in trade-offs between the efficiency and flexibility of the recognition systems they enable. The primary purpose of the present dataset, PkSLMNM, is to establish a new standard in Pakistan Sign Language by providing a unique blend of dynamic gestures, and manual and non-manual modalities. The language of this dataset is centered on dynamic hand movements, which sets it apart from its predecessors. There is, to the best of our knowledge, no publicly available dataset that covers dynamic gestures.

The new dataset is also notable for its large lexicon, resilience, high-quality recordings, and a unique syntactic trait that allows for the recognition of dynamic hand motions. This dataset, in addition to acting as a baseline for dynamic sign language recognition, is critical in improving the field's research and development activities. During the validation of our dataset, we discovered an additional small dataset focusing on facial expressions. This was accomplished by extracting Regions of Interest (ROIs) from movies, thereby expanding the resources available for study in the fields of sign language recognition and affective computing.

Objectives and Novelty:

The primary objective of this research is to produce the PkSLMNM dataset, which is a novel and comprehensive dataset for PSL recognition. This dataset attempts to meet the demand for dynamic sign language data by addressing both manual and non-manual modalities and providing a diverse lexicon of signs. This dataset aims to improve the area of sign language recognition by focusing on dynamic gestures, high-quality recordings, and grammatical features, giving researchers and developers a significant resource for the progress of PSL recognition technology.

The PkSLMNM dataset adds several novel aspects to the field of sign language recognition. It is notable for having a vast lexicon of dynamic signs, which provides a broader choice of vocabulary and phrases for research and development. Because the dataset concentrates on dynamic hand motions and grammatical features, sophisticated sign language structures can be recognized. Furthermore, the high-quality recordings in the dataset improve its usefulness for training and assessing recognition models. During the validation process, another tiny dataset focusing on facial expressions was discovered, broadening the dataset's possible applications. This expansion into affective computing and the understanding of emotional cues in sign language gives the PkSLMNM dataset a new dimension, making it a versatile resource for sign language recognition and affective computing researchers.

In summary current research study has the following objectives:

- To develop a dataset having manual and non-manual combined modalities named PkSLMNM.
- To evaluate the dataset by extracting face ROI using Yolo-Face architecture.
- To develop or extract another dataset of only face expressions from our newly developed dataset of PkSLMNM

- To evaluate the facial expression dataset for facial expression recognition using transfer learning.

Materials & Methods:

Equipment and Setup:

Two cameras are used to record the complete dataset. Camera 1 is a versatile feature-set EOS M50 digital camera from CANON, and Camera 2 is EOS 700D digital camera from CANON. Three Newer LED 500 Ultra High-Power dimmable lights are used; one of them is used to illuminate the subject and two others for background radiance and shadow elimination. Throughout the dataset creation, the lighting remained cold and with stable color temperature, and as a backdrop white project or a white background wall was consistent.

High-definition 4K videos are recorded using Canon M50 and Canon 700D, with a 1920 x 1086 at 25 frames per second for Canon M50 and 59.94 frames per second using Canon 700D, saved in .MP4 file format. Tab. 1 expands the details of the equipment used. The first thirty people are recorded with Canon M50, records while the remaining seventy people are recorded with Canon 700D. Overall, the set of videos includes 34,650 frames recorded for seven classes of dynamic gestures. The size of videos for every gesture varied from 1 second to 3 seconds due to the speed of every individual performing any motion or action and the gesture's versatility.

Table 1: Details of equipment and videos for dynamic signs

Features	Camera 1	Camera 2
Name of camera	CANON EOS-M50	CANON EOS-700D
Frame Width	1920	1920
Frame Height	1086	1086
Data Rate/ Bit Rate	67589kbps	58473kbps
Frame Rate	25 fps	59.94 fps

Objective and Procedure:

The PkSLMNM dataset was collected following an approved protocol by the "Ethical Review Committee (ERC) of the School of Engineering and Applied Sciences at Bahria University Karachi Campus, under the application ERC/ES/005". This thorough ethical oversight guaranteed that all ethical principles and guidelines were followed during the data collection method. All 100 people who helped to create the dataset, gave their informed agreement to be a part of this study, underscoring the importance of ethical data collection and use.

In the interest of ethical research procedures, it is critical to state that the PkSLMNM dataset was obtained solely for research purposes and that the data-collecting technique was not damaging to the participants in any manner. Each participant was filmed for seven dynamic videos, with each recording session lasting no more than 15 minutes. The dataset includes a broad sample of male and female individuals ranging in age from 20 to 50 years old, reflecting the adult population.

Participants were taught to keep the desired facial expression for a brief period while being recorded while facing forward to their body position. During data collection, strict quality control methods were implemented, including real-time image review to ensure participants did not move, blink, or display unintentional expressions. These reviews also confirmed the availability and quality of all relevant photographs. Notably, the dataset retains realism by not regulating the subjects' apparel (shirts), allowing for slight variations in the individual's position and facial states. All other areas of data collecting, on the other hand, were thoroughly managed to assure data quality and consistency.

Dynamic gestures offered distinct obstacles during data collection, with individuals moving at variable speeds, resulting in video lengths spanning from 1 to 3 seconds. Certain elements also

contributed to data robustness. Occlusion induced by persons wearing glasses, male subjects with beards, and certain female participants wearing a headscarf are among these issues. Furthermore, the backdrops in the sample vary significantly due to variances in apparel color and style. Furthermore, the dataset includes a wide range of skin tones and facial features, which adds to the complexity of algorithm evaluation in the context of sign language detection.

a. Disgust gesture

b. Neutral gesture

c. Happy gesture

d. Sad gesture

e. Scared gesture

f. Angry Gesture

g. Surprise Gesture

Figure 1. Glimpses of the developed dataset

Lexicons of Dataset:

The current PkSLMNM dataset is a comprehensive collection of both manual (hand-based) and non-manual (facial and body-based) gestures. As depicted in Figure 1, the dataset encompasses a set of seven adjectives—bad, best, sad, glad, scared, stiff, and surprise—each of which portrays one of seven primary expressions: disgust, neutral, happiness, sadness, fear, anger, and surprise, respectively. All the recorded video clips have an average duration of 2 seconds, and the variations in input clip sizes, ranging from 3MB to 15MB, can be attributed to the utilization of two different types of cameras, variations in individual speed, and the versatility of the gestures [12].

Facial expressions are complex, involving the activation of minute muscle movements in the face, known as micromotor muscles. These muscular movements are critical in determining an individual's mood or feeling. Facial Action Coding System (FACS), also known as Facial Action Units (FAUs), was a way to categorize an individual's emotional state in 1978. This technique takes into account several face units such as the forehead, brows, eyelids, nose, cheeks, and lips. Table 2 contains a full explanation of the facial action units and associated hand movements, as well as minor body movements, that correspond to all of the gestures or signs employed in Pakistan Sign Language (PSL), which form the basis of our research.

The current dataset is visually reviewed and collected in the presence of domain experts, data is also cleaned up, and all erroneous signs are deleted. Balanced distribution among classes with a proper labeling scheme is followed.

Methodology:

During the first phase of our research, we carefully collected an independent dataset to address the lack of understanding of manual and non-manual gestures in Pakistan Sign Language (PSL). For the validation of this newly produced dataset, we used the YOLO-Face architecture, which incorporates face identification as well as exact Region of Interest (ROI) selection and cropping. Figure 2 shows the flow of methodology.

This procedure has verified that our dataset was accurate and reliable. Following that, we classified seven fundamental expressions using Transfer Learning, leveraging prior knowledge to obtain a high level of accuracy in expression classification. Figure 2 depicts our methodology's sequential procedure, demonstrating the systematic advancement from dataset collection to face detection, ROI selection, and the final stage of expression categorization.

YOLO-Face Architecture (Face ROI Selection):

The method for Face ROI recognition, as depicted in Figure 3 comprises four distinct stages: face detection, data preprocessing, feature extraction, and face ROI cropping.

- The first stage, face detection, is in charge of recognizing and finding faces within the input image or video frame.
- The second stage involves data preprocessing, in which image processing techniques are used to prepare the discovered images for subsequent processing by machine learning models.
- The third stage focuses on feature extraction, in which significant features from the prepared photos are retrieved, allowing for a thorough depiction of the facial traits.
- ROIs are detected and cropped in the final phase, face recognition, by comparing the features retrieved from the input data based on their facial qualities.

Face Detection:

The section delves into the initial step of our methodology, which is devoted to detecting faces in input data even in spontaneous and dynamic circumstances. Face detection at this level requires the ability to deal with a slew of nonlinear elements in the input images, such as fluctuations in luminance, facial posture, and facial magnitude. The face detection

method was chosen after a thorough study of numerous approaches appropriate for real-time applications, such as YOLO-Face.

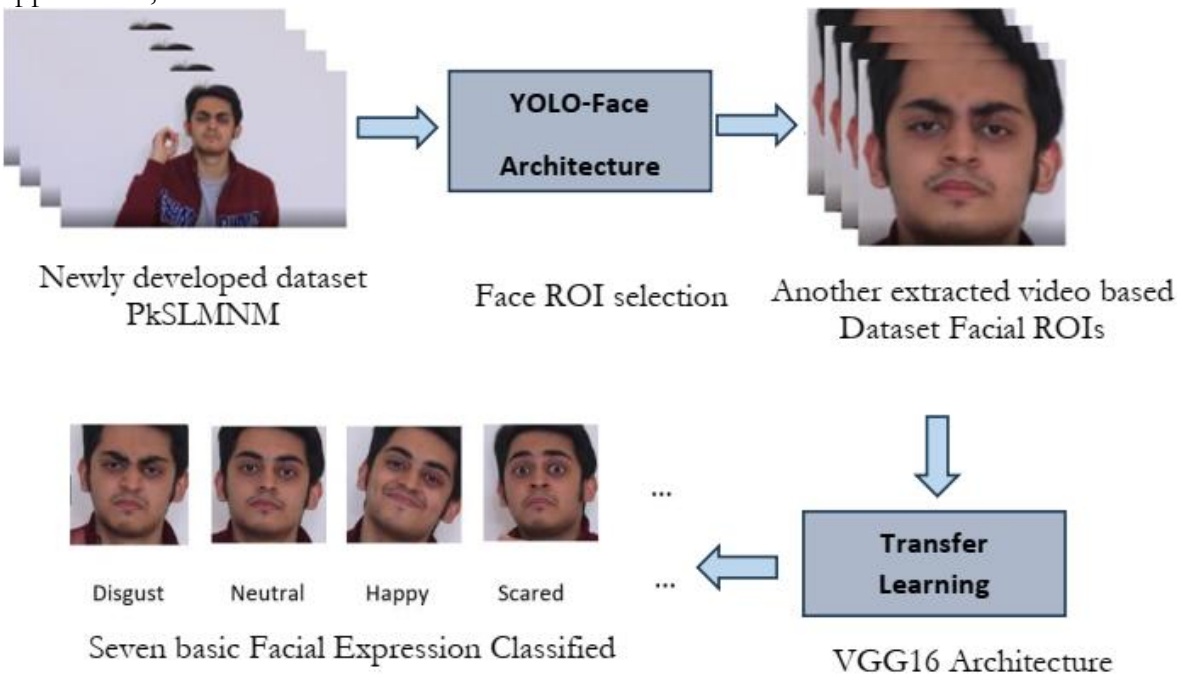

Figure 2: Workflow of the Research

The proposed method uses a one-phase deep learning module, as opposed to a two-phase procedure such as Faster-RCNN [13]. YOLO, or "You Only Look Once," has the unusual ability to forecast classes connected with respective bounding boxes at the same time. One of the key benefits of the one-phase strategy is its incredible speed, which outperforms two-phase systems. This speed and efficiency are especially beneficial in real-time circumstances, making it an excellent choice for our face identification phase.

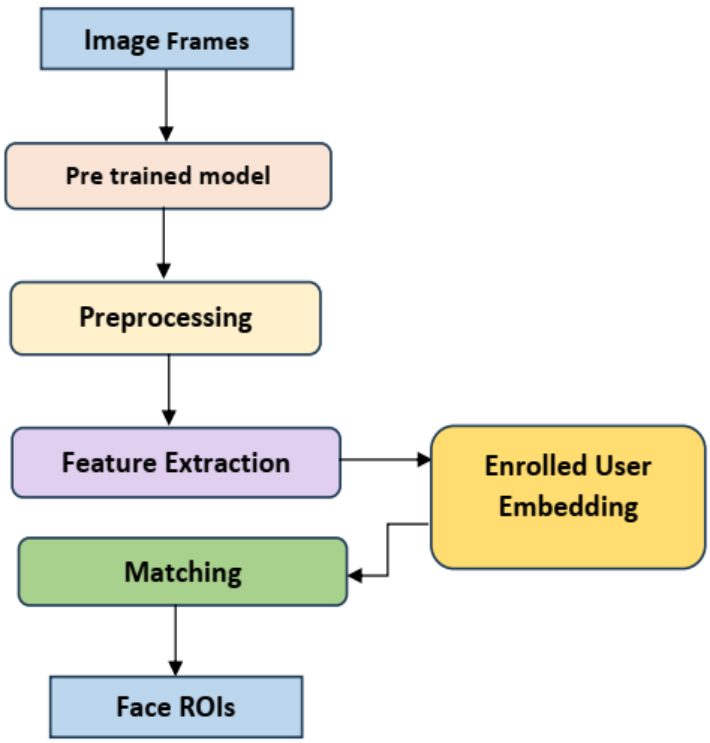

Figure 3. Real time Face ROI extraction workflow

YOLO Face:

The importance of machine learning and deep learning cannot be neglected in this era specifically for image processing [17]-[33]. The section explains the primary characteristics of the proposed model that utilizes YOLOv3 [13] as the detection model, which depends on object recognition methods using deep learning. Following this mechanism, face detection modules perform significantly better.

Prediction of Classes. YOLO-Face works on binary classification which means it only recognizes two classes; face or no face, instead of multiclass. The proposed system is based on an object detection model, which reduces the error at the output of the model using a multipart loss function. There are four parts of this loss

function which include loss of classification, regression, and confidence in anchors with and without objects. These four parts of loss are divided with a 2:1:0.5:0.5 ratio.

Network for extracting features. Features are extracted in YOLO-Face using the darknet-53 network [14], it is a concatenation of two other feature-extracting networks; darknet-19 of YOLO-V2 and residual network of newfangled with 53 layers. When the feature extractor of the proposed network is compared with the darknet-19 of YOLO-v2. The first network outperforms the latter. In terms of efficiency, it significantly worked better when compared to ResNet-101 and ResNet-152. The performance of YOLO-v3 degrades while detecting small-scale objects in YOLO-v3, which is the only main problem. Hence, the proposed network incremented the number of layers of the actual darknet-53 network in the two initial residual blocks. These amendments in the structure of the network are made before the characteristics of small objects reduce more on the feature map resulting in an increase in the quality of the features of the face. The number of layers in the initial two blocks has been incremented to 4x and 8x from 1x and 2x. The feature map of the network is scaled down with stride 2 of the convolutional layer which is considerably bigger compared to YOLO-v3. The proposed network uses deeper Dark Net which is a set of 71 convolutional layers. Figure 4 depicts the entire network structure. YOLO-Face has an equivalent recognition speed as Face Attention Network (FAN) according to [15] results, however, FAN is more inefficient.

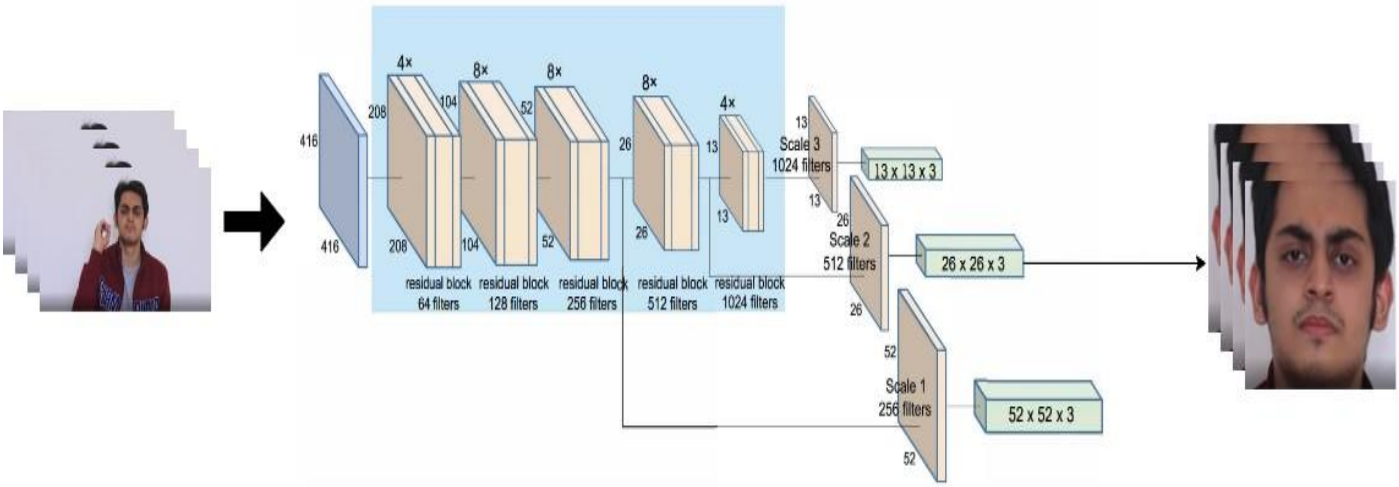

Figure 4. YOLO-Face architecture is adapted to find face ROIs in our dynamic PkSLMNM dataset

Prediction of Various Scales:

YOLO-Face uses a notion equivalent to FPN [16] as shown in Fig. 3, it detects different faces and predicts their respective bounding boxes around three various scales. Using multiple sizing feature maps, the approach outperforms multi-scale detection. As the feature map generator, the outcome of the last residual layer is chosen that is embellished to create a pyramid structure. The detection mechanism is achieved when combined with the two preceding blocks which have not been sampled by 2x yet. The anchor boxes of the object recognition systems are modified to make them competent for face detection, which implies that anchor boxes are created narrow and tall rather than with heights less than width.

Face classification and identification models have been improving in the past few years by modifying structure and optimizing the loss function, such networks include VGG-Face, Face Net, Cosface as well as Ring Loss. For modeling the multiclass classification tasks, Softmax with cross-entropy loss function has drawn a lot of attention and can be measured by equation 1:

$$L(y,\hat{y}) = -\frac{1}{N}\sum_{i=1}^{N} yi^{T}\log(\hat{y}i) = -\frac{1}{N}\sum_{i=1}^{N}\log(\hat{y}ig) \quad (1)$$

Here total number of identified samples is represented by N, one-hot encoded expression of the ith sample is labeled as yi∈{0,1}j, with Softmax and ground truth class the predicted probabilities are represented by ŷig.

The study aims to improve the precision of the classification results of the Face Net through L2 normalization and softmax with cross entropy as the loss function. The modification in the loss function has made CNN learn features between various classes more distinguishably, subsequently getting smaller and larger intra-class distances. When compared with the traditional methods, the proposed network outperformed work effectively better as shown in the results that in the LFW database, the final precision is enhanced.

Transfer Learning (VGG 16 Architecture)

Transfer Learning with the VGG16 architecture is a powerful deep-learning approach, especially when dealing with image recognition or classification applications. VGG16, a CNN architecture, is well-known for its depth and ability to capture intricate hierarchical structures inside images. Transfer Learning aims to capitalize on VGG16's knowledge gained by pre-training on a large and diverse dataset like ImageNet [11].

Model: "VGG16 Transfer Learning Model"

Layer (type)	Output Shape	Param #
input_1 (Input Layer)	[(None, 224, 224, 3)]	0
block1_conv1 (Conv2D)	(None, 224, 224, 64)	1792
block1_conv2 (Conv2D)	(None, 224, 224, 64)	36928
block1_pool (MaxPooling2D)	(None, 112, 112, 64)	0
block2_conv1 (Conv2D)	(None, 112, 112, 128)	73856
block2_conv2 (Conv2D)	(None, 112, 112, 128)	147584
block2_pool (MaxPooling2D)	(None, 56, 56, 128)	0
block3_conv1 (Conv2D)	(None, 56, 56, 256)	295168
block3_conv2 (Conv2D)	(None, 56, 56, 256)	590080
block3_conv3 (Conv2D)	(None, 56, 56, 256)	590080
block3_pool (MaxPooling2D)	(None, 28, 28, 256)	0
block4_conv1 (Conv2D)	(None, 28, 28, 512)	1180160
block4_conv2 (Conv2D)	(None, 28, 28, 512)	2359808
block4_conv3 (Conv2D)	(None, 28, 28, 512)	2359808
block4_pool (MaxPooling2D)	(None, 14, 14, 512)	0
block5_conv1 (Conv2D)	(None, 14, 14, 512)	2359808
block5_conv2 (Conv2D)	(None, 14, 14, 512)	2359808
block5_conv3 (Conv2D)	(None, 14, 14, 512)	2359808
block5_pool (MaxPooling2D)	(None, 7, 7, 512)	0
flatten (Flatten)	(None, 25088)	0
fc1 (Dense)	(None, 4096)	102764544
fc2 (Dense)	(None, 4096)	16781312
predictions (Dense)	(None, 1000)	4097000
Total params: 138,357,544 Trainable params: 138,357,544 Non-trainable params: 0		

The above given architecture summary of the VGG16 model, including its layers and the number of parameters in each layer. The model consists of convolutional layers (Conv2D), max-pooling layers (MaxPooling2D), and fully connected layers (Dense). The "blockX_convY" layers represent convolutional layers within different blocks, and "blockX_pool" represents max-pooling layers. The final layers include a flattening layer

followed by three fully connected layers ("fc1," "fc2," and "predictions"), with the last one having 1000 output nodes for ImageNet's 1000 classes. The model has a total of 138,357,544 parameters, all of which are trainable.

Results & Discussion:

Face ROI Selection (YOLO-Face Results):

Our newly developed dataset PkSLMNM is tested using the YOLO-Face methodology to detect face ROIs. Different metrics are used to evaluate the performance like precision, recall, F1-Score, accuracy, and current average loss. The values are 0.89, 0.96, 0.84, and 90.89% for precision, recall, F1-Score, and accuracy respectively.

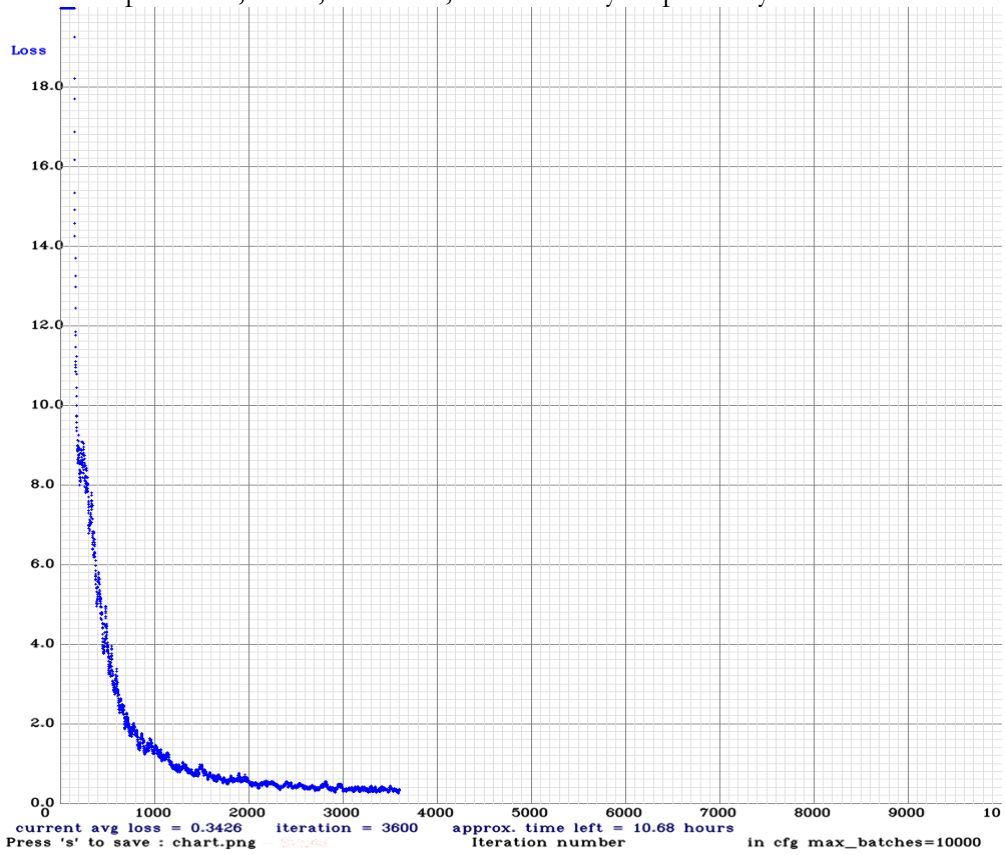

Figure 5. Average Loss Graph of the proposed model

Figure 5 represents a loss graph with the current average loss of 0.34. This object localization and classification loss classify the face region pixels into a binary class. Further, we extracted or cropped the regions to make a new dataset of face expressions only derived from PkSLMNM. We got another useful dataset for facial expression or emotion detection in the spatio-temporal domain through this semantic segmentation in videos. Further, figure 6 depicts the accuracy graph of training and validation of the proposed model, both the training and validation accuracy lines converge and stabilize at a high degree of accuracy, showing that our model is learning effectively without overfitting.

We used the PkSLMNM dataset to recognize both manual and non-manual motions at the same time. This dataset contains video clips shot with an HD camera in.MP4 format, capturing a variety of facial emotions as well as hand movements. To guarantee that the data was suitable for classification, we went through a preprocessing phase in which we converted the video clips into frames and methodically removed any noise that could potentially interfere with the classification process. Each frame has a uniform resolution of 1920x1086 pixels, and all frame boundaries are properly labeled with the seven basic emotions corresponding classes.

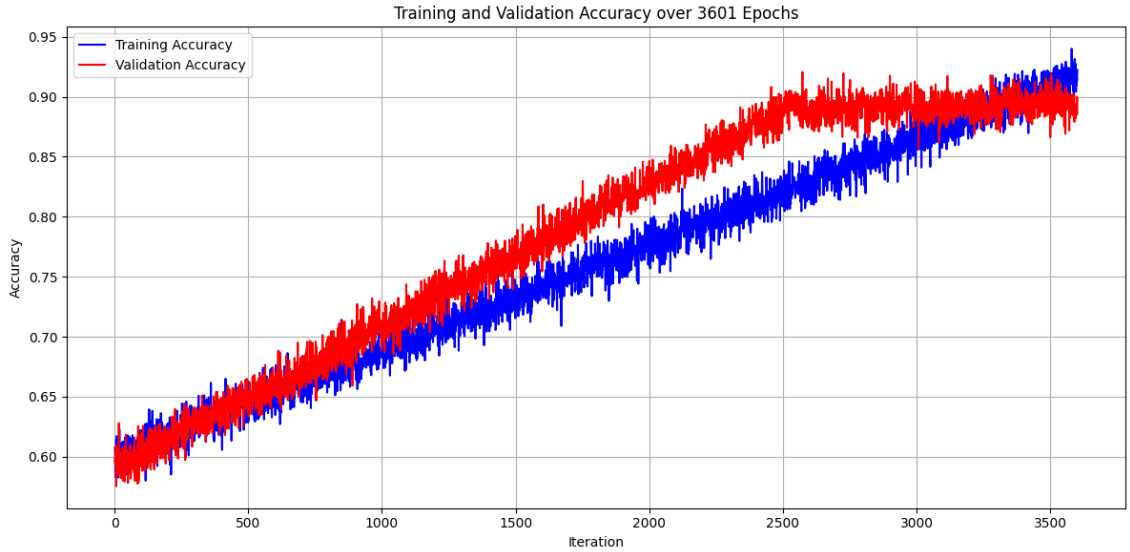

Figure 6. Training and validation accuracy graph of the proposed model

We used the You Only Look Once-Face (YOLO-Face) architecture to precisely recognize faces within the dataset and identify them as Regions of Interest (ROIs) to validate it. YOLO-Face is a hybrid technique that combines the darknet-19 architecture with a residual network composed of 53 convolutional neural network layers. To get high-quality facial characteristics, we raised the number of layers in the residual blocks by 4x and 8x, respectively, resulting in a network known as a deeper DarkNet. YOLO-Face, in particular, converts the classification challenge from a multi-class problem to a binary classification task, concentrating simply on the presence or absence of a face. Furthermore, YOLO-Face excels at detecting faces at three different scales, improving the accuracy and robustness of the system.

Face Expression Recognition (Using VGG16 Transfer Learning):

Compiling a Kera's model using categorical cross entropy as the loss function and Stochastic Gradient Descent (SGD) as the optimizer. The learning rate is set to 0.0001, and momentum is configured to be 0.9. Additionally, you've included several custom metrics for evaluation, such as accuracy, F1 score (f1_m), precision (precision_m), recall (recall_m), and AUC (Area Under the Curve).

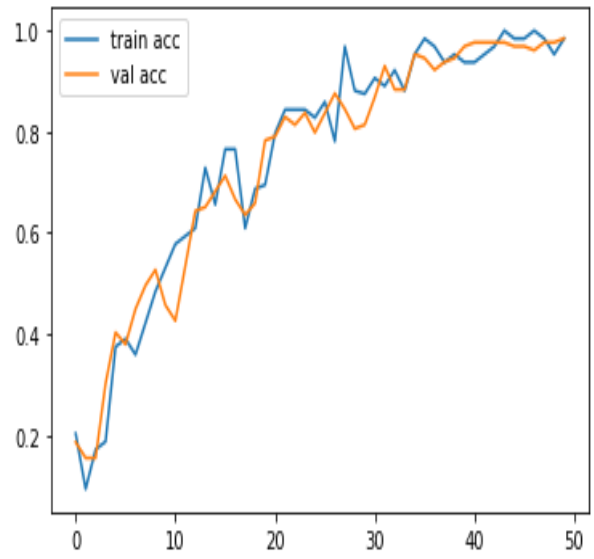

Figure 7. Training and validation accuracy graph of the facial expression classification

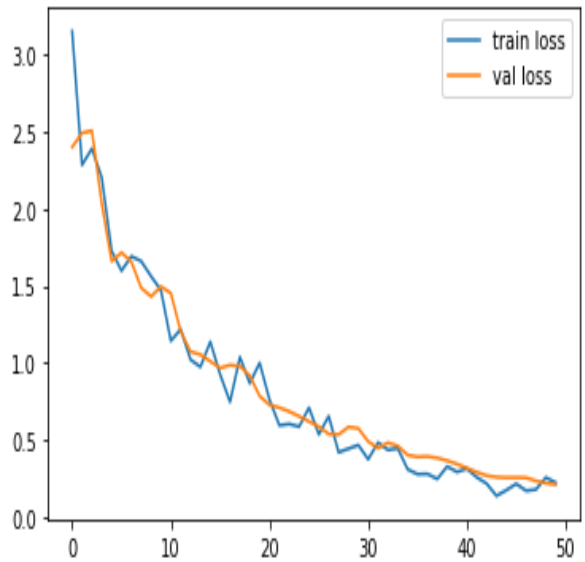

Figure 8. Training and validation loss graph of the facial expression classification

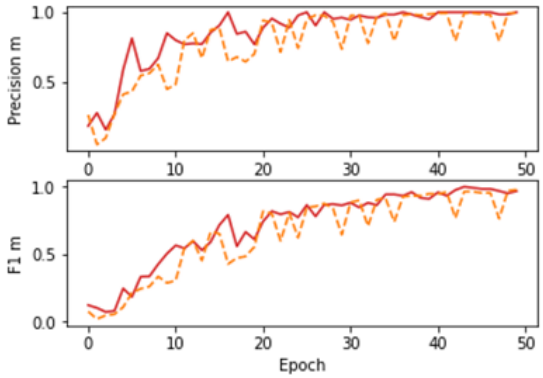

Figure 9. Training and validation Precision and F1m graph of the facial expression classification

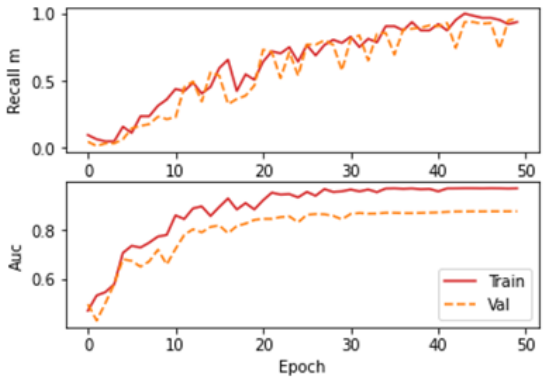

Figure 10. Training and validation Recall, and AUC graph of the facial expression classification

The model has achieved a low loss of 0.2153, indicating that it is performing well in minimizing its prediction errors. The accuracy is high at 98.45%, suggesting that the majority of predictions are correct. The F1 score is 0.9777, which is a good balance between precision and recall. Precision is high at 99.35%, indicating a low rate of false positives, while recall is 96.25%, showing a good ability to capture true positives. The AUC (Area Under the Curve) is 0.8755, which is commonly used to evaluate the performance of a binary classification model. Overall, these metrics suggest that the model is performing very well on the task it was trained for. Figures 7, 8, 9, and 10 show the graphical representations of the results. Discussion:

The PkSLMNM dataset, which focuses on Pakistan Sign Language (PSL), presents a unique resource for the study and development of Sign Language Recognition. In the discussion below, we will explore various aspects and comparisons related to this dataset.

Dynamic Gestures vs. Static Signs:

PkSLMNM stands out as it primarily focuses on dynamic hand movements, setting it apart from previous datasets that mainly concentrated on static sign language signs represented as still images. This emphasizes the importance of capturing the temporal aspect of Sign Language, making it more applicable to real-world communication scenarios.

Manual and Non-Manual Modalities:

The dataset includes both manual and non-manual modalities, which is essential for comprehensive sign language recognition. Manual modalities encompass hand and arm movements, while non-manual modalities involve facial expressions and other upper-body motions. This comprehensive approach reflects the complexity of sign language communication.

Large Lexicon:

PkSLMNM offers a significant lexicon of sign language gestures, enabling a broader range of signs to be recognized and analyzed. A large lexicon is crucial for developing practical sign language recognition systems, as it covers a wider range of vocabulary and expressions.

High-Quality Recordings:

The dataset is notable for its high-quality recordings, ensuring that the data is reliable and accurate for research and development purposes. High-quality recordings are essential for training and evaluating recognition systems, as they reduce noise and improve model performance.

Syntactic Traits:

PkSLMNM incorporates a unique syntactic trait that allows for the recognition of dynamic hand motions. This feature contributes to the dataset's versatility and adaptability, as it accommodates the complexities of sign language syntax and structure.

Real-World Sign Language:

While previous datasets focused on controlled sign activities, PkSLMNM strives to capture real-world sign language usage. This is critical for developing recognition systems that can operate effectively in everyday communication scenarios.

Robustness and Domain-Specificity:

PkSLMNM represents a valuable resource for research in sign language recognition, offering both robustness and domain-specificity. Robustness is essential for ensuring that recognition systems can handle various environmental conditions and signing styles, while domain-specificity focuses on the unique characteristics of PSL.

Emotion Representations:

The dataset includes emotional representations in sign language, allowing for the recognition of emotional cues in sign communication. This is a valuable addition as emotions play a significant role in sign language expression.

Comparisons with Other Datasets:

When compared to other sign language datasets, PkSLMNM stands out for its dynamic nature, comprehensive coverage of manual and non-manual modalities, and large lexicon. It complements existing datasets by addressing the need for real-world sign language recognition.

Potential for Advanced Recognition Systems:

PkSLMNM's features and characteristics create a foundation for the development of advanced sign language recognition systems. Researchers and developers can leverage this dataset to enhance the accuracy and applicability of sign language recognition technology.

In summary, the PkSLMNM dataset represents a significant advancement in the field of sign language recognition, offering a unique combination of dynamic gestures, comprehensive modalities, and a large lexicon. Its focus on real-world sign language usage, robustness, and emotional representations makes it a valuable resource for researchers and developers in the domain of sign language technology.

Conclusion:

In this study, our newly developed dataset underwent a comprehensive evaluation for facial expression assessment, focusing on non-manual gestures, utilizing the YOLO-Face detection methodology. This process successfully extracted faces as Regions of Interest (RoI) with an impressive 90.89% accuracy and an average loss of 0.34. Additionally, Transfer Learning (TL) was applied, employing the VGG16 architecture to classify seven fundamental facial expressions, achieving a remarkable 100% accuracy. In summary, our research yielded two distinct datasets—one featuring manual and non-manual sign language gestures, and the other comprising Asian faces for the identification of seven basic facial expressions. Our validation techniques across both datasets produced promising and noteworthy results, demonstrating the effectiveness of our approach in facial expression recognition.

Dataset Availability:

Our published dataset along with facial videos extracted as regions of interest, are present on the following links:
https://data.mendeley.com/datasets/m3m9924p3v/1
https://data.mendeley.com/datasets/h5ptg7nr54/1

Acknowledgment: We acknowledge the contribution of organizations and people who participated in PkSLMNM ingenuity as participants, organizers, or evaluators.

Funding Statement: The authors received no specific funding for this study.

Conflicts of Interest: The authors declare no competing interests. Neither financial nor personal could influence the work reported in this current paper.

References:

[1] M. Al-Qurishi, T. Khalid, and R. Souissi, "Deep Learning for Sign Language Recognition: Current Techniques, Benchmarks, and Open Issues," IEEE Access, vol. 9, pp. 126917-126951, 2021.

[2] J. Zheng, Y. Chen, C. Wu, X. Shi, and S. M. Kamal, "Enhancing Neural Sign Language Translation by Highlighting the Facial Expression Information," Neurocomputing, vol. 464, pp. 462–472, 2021.

[3] A. Singh, S. K. Singh, and A. Mittal, "A Review on Dataset Acquisition Techniques in Gesture Recognition from Indian Sign Language," in Advances in Data Computing, Communication and Security, Springer, vol. 106, pp. 305–313, 2021.

[4] S. Sharma and S. Singh, "Vision-based hand gesture recognition using deep learning for the interpretation of sign language," Expert Systems and Applications, vol. 182, p. 115657, 2021.

[5] I. Rodríguez-Moreno, J. M. Martínez-Otzeta, I. Goienetxea, and B. Sierra, "Sign Language Recognition by means of Common Spatial Patterns," in 2021 The 5th International Conference on Machine Learning and Soft Computing, Da Nang Viet Nam, 2021, pp. 96–102.

[6] H. Zahid, M. Rashid, S. Hussain, F. Azim, S. A. Syed, and A. Saad, "Recognition of Urdu Sign Language: A Systematic Review of the Machine Learning Classification," PeerJ Computer Science, vol. 8, p. e883, 2022.

[7] "Deaf Reach Schools and Training Centers in Pakistan." https://www.deafreach.com/ [Online] (accessed May 19, 2022).

[8] "PSL." https://psl.org.pk/ [Online] (accessed May 19, 2022).

[9] A. Imran, A. Razzaq, I. A. Baig, A. Hussain, S. Shahid, and T.-U. Rehman, "Dataset of Pakistan Sign Language and Automatic Recognition of Hand Configuration of Urdu Alphabet through Machine Learning," Data in Brief, vol. 36, p. 107021, Jun. 2021, doi: 10.1016/j.dib.2021.107021.

[10] "Pakistan Sign Language Dataset - Open Data Pakistan." [Online] https://opendata.com.pk/dataset/pakistan-sign-language-dataset (accessed May 19, 2022).

[11] S. Javaid, S. Rizvi, M. T. Ubaid, A. Darboe and S. M. Mayo, "Interpretation of Expressions through Hand Signs Using Deep Learning Techniques", International Journal of Innovations in Science and Technology, vol. 4, no. 2, pp. 596-611, 2022.

[12] S. Javaid and S. Rizvi, "A novel action transformer network for hybrid multimodal sign language recognition," Computers, Materials & Continua, vol. 74, no.1, pp. 523–537, 2023.

[13] R. Gavrilescu, C. Zet, C. Foşalău, M. Skoczylas, and D. Cotovanu, "Faster R-CNN: An Approach to Real-Time Object Detection," in 2018 International Conference and Exposition on Electrical and Power Engineering (EPE), Iasi, Romania, 2018, pp. 0165–0168.

[14] J. Redmon and A. Farhadi, "YOLOv3: An Incremental Improvement," ArXiv Prepr. ArXiv180402767, 2018.

[15] W. Chen, H. Huang, S. Peng, C. Zhou, and C. Zhang, "YOLO-Face: A Real-Time Face Detector," The Visual Computer, vol. 37, no. 4, pp. 805–813, 2021.

[16] S.-W. Kim, H.-K. Kook, J.-Y. Sun, M.-C. Kang, and S.-J. Ko, "Parallel Feature Pyramid Network for Object Detection," in Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 2018, pp. 234–250.

[17] W. I. D. Mining, "Data mining: Concepts and techniques," Morgan Kaufinann, vol. 10, pp. 559–569, 2006.

[18] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A Unified Embedding for Face Recognition and Clustering," in Proceedings of the IEEE conference on computer vision and pattern recognition, Boston, MA, USA, 2015, pp. 815–823.

[19] A. Jain, K. Nandakumar, and A. Ross, "Score Normalization in Multimodal Biometric Systems," Pattern Recognition, vol. 38, no. 12, pp. 2270–2285, 2005.

[20] Javaid, Sameena and Rizvi, Safdar. 'Manual and Non-manual Sign Language Recognition Framework Using Hybrid Deep Learning Techniques'. 1 Jan. 2023: 3823 – 3833.

[21] H. Mohabeer, K. S. Soyjaudah, and N. Pavaday, "Enhancing The Performance Of Neural Network Classifiers Using Selected Biometric Features", SENSORCOMM 2011: The Fifth International Conference on Sensor Technologies and Applications, Saint Laurent du Var, France, 2011.
[22] S. Khan, H. Rahmani, S. A. A. Shah, and M. Bennamoun, "A Guide to Convolutional Neural Networks for Computer Vision," Synthesis Lectures on Computer Vision, vol. 8, no. 1, pp. 1–207, 2018.
[23] E. Kremic and A. Subasi, "Performance of Random Forest and SVM in Face Recognition." Int Arab Journal of Information Technology, vol. 13, no. 2, pp. 287–293, 2016.
[24] E. Setiawan and A. Muttaqin, "Implementation of k-Nearest Neighbors Face Recognition on Low-Power Processor," Telecommunication Computing Electronics and Control, vol. 13, no. 3, pp. 949–954, 2015.

[25] M. T. Ubaid, A. Kiran, M. T. Raja, U. A. Asim, A. Darboe and M. A. Arshed, "Automatic Helmet Detection using EfficientDet," 2021 International Conference on Innovative Computing (ICIC), Lahore, Pakistan, 2021, pp. 1-9, doi: 10.1109/ICIC53490.2021.9693093.

[26] M. A. Arshed, H. Ghassan, M. Hussain, M. Hassan, A. Kanwal, and R. Fayyaz, "A Light Weight Deep Learning Model for Real World Plant Identification," 2022 2nd Int. Conf. Distrib. Comput. High Perform. Comput. DCHPC 2022, pp. 40–45, 2022, doi: 10.1109/DCHPC55044.2022.9731841.

[27] M. T. Ubaid, M. Z. Khan, M. Rumaan, M. A. Arshed, M. U. G. Khan and A. Darboe, "COVID-19 SOP's Violations Detection in Terms of Face Mask Using Deep Learning," 2021 International Conference on Innovative Computing (ICIC), Lahore, Pakistan, 2021, pp. 1-8, doi: 10.1109/ICIC53490.2021.9692999.

[28] M. A. Arshed, H. Ghassan, M. Hussain, M. Hassan, A. Kanwal and R. Fayyaz, "A Light Weight Deep Learning Model for Real World Plant Identification," 2022 Second International Conference on Distributed Computing and High Performance Computing (DCHPC), Qom, Iran, Islamic Republic of, 2022, pp. 40-45, doi: 10.1109/DCHPC55044.2022.9731841.

[29] M. A. Arshed, A. Shahzad, K. Arshad, D. Karim, S. Mumtaz, and M. Tanveer, "Multiclass Brain Tumor Classification from MRI Images using Pre-Trained CNN Model", VFAST trans. softw. eng., vol. 10, no. 4, pp. 22–28, Nov. 2022.

[30] A. Shahzad, M. A. Arshed, F. Liaquat, M. Tanveer, M. Hussain, and R. Alamdar, "Pneumonia Classification from Chest X-ray Images Using Pre-Trained Network Architectures", VAWKUM trans. comput. sci., vol. 10, no. 2, pp. 34–44, Dec. 2022.

[31] H. Younis, Muhammad Asad Arshed, Fawad ul Hassan, Maryam Khurshid, and Hadia Ghassan, "Tomato Disease Classification using Fine-Tuned Convolutional Neural Network", IJIST, vol. 4, no. 1, pp. 123–134, Feb. 2022.

[32] M. Mubeen, M. A. Arshed, and H. A. Rehman, "DeepFireNet - A Light-Weight Neural Network for Fire-Smoke Detection," Commun. Comput. Inf. Sci., vol. 1616 CCIS, pp. 171–181, 2022, doi: 10.1007/978-3-031-10525-8_14/COVER.

[33] Q. Zhu, Z. He, T. Zhang, and W. Cui, "Improving Classification Performance of Softmax Loss Function based on Scalable Batch-Normalization," Applied Sciences, vol. 10, no. 8, p. 2950, 2020.