Artificial Intelligence Model Interpreting Tools: SHAP, LIME, and Anchor Implementation in CNN Model for Hand Gestures Recognition

Chung-Chian Hsu[1, 2], S. M. Salahuddin Morsalin[1, *], MD Faysal Reyad[2], Nazmus Shakib[2]

[1] Department of Information Management,
National Yunlin University of Science and Technology, Taiwan
[2] Graduate School of Artificial Intelligence,
National Yunlin University of Science and Technology, Taiwan
{hsucc, D10813004, m11023054, m11023053} @yuntech.edu.tw

Abstract. Explainable AI (XAI) are the tools and frameworks of artificial intelligence applications that make it easier to trust the results and outcomes produced by machine learning algorithms. Additionally, XAI helps with debugging, enhancing model performance, and describing the behavior of models to others. This paper presents an innovative approach for hand-gesture detection using an Explainable AI Convolutional Neural Network (XAI-CNN) and SHAP (Shapley Additive Explanations) values, LIME (Local Interpretable Model-agnostic Explanations), and Anchor as Explainable AI tools. The XAI-CNN model is specifically designed for ten different classes of hand-gesture accurate recognition, including palm moved, C, ok, I, fist, index, palm, thumb, down, and fist moved symbols. The proposed XAI-CNN architecture, built upon the previous CNN model, demonstrates an impressive accuracy of 99.98%. Furthermore, the SHAP (XAI tools) values, LIME, and Anchor integration enable the interpretation and visualization of the model's decision-making process separately, enhancing the transparency and trustworthiness in the hand-gesture recognition process. This research contributes to the robustness and interpretable AI systems for hand-gesture recognition, empowering users with accurate and understandable AI technology.

Keywords: Explainable AI, Shapley Additive Explanations, Local Interpretable Model-agnostic Explanation, Anchor, Hand-gesture Recognition, Image Classification, Convolutional Neural Network.

1 Introduction

The XAI tool sophisticates the interpretability and transparency of complex machine learning models, such as Convolutional Neural Networks (CNNs) widely used for hand gesture recognition. The human body key-point detection and recognition [1] are challenging tasks in human-computer interaction and computer vision. XAI tool is an enabler in many sectors and helps to make them auto-controlled from remote places.

However, the complexity of recognizing hand gestures due to the diversity in shape, size, capturing angle, lighting conditions, and background distortion developed the Baseline CNN model [2]. XAI techniques provide insights into the decision-making process of these models, enabling users to understand the underlying factors contributing to predictions. We can create more potentiality, trustworthiness, and accountability by integrating XAI into the CNN model, empowering the users' capability to understand the performance and verify the model's results.

XAI tools integration techniques in the CNN model with SHAP values, LIME, and Anchor not only enhance the interpretability but also contribute to increasing the trustworthiness in hand gesture detection [3], [4]. By utilizing the CNN model, which excels at extracting complex spatial patterns from input data. We capture intricate hand-gesture features that might be challenging for traditional algorithms. When combined with XAI tools, which provide feature-level explanations, XAI-CNN models become more robust and precise in their predictions. The SHAP values, LIME, and Anchor implementation enable fine-tuning the XAI-CNN model's parameters for increased accuracy by finding the essential aspects that affect the model's decision-making process. This paper explores the synergistic relationship between XAI-CNN and XAI tools, demonstrating their effectiveness in enhancing hand-gesture detection accuracy while maintaining transparency and interpretability by following Trustworthiness in Artificial Intelligence (TAI) guidance.

2 Related Works

The capabilities of deep learning and its effectiveness in extracting and classifying the detail characteristics from input data have received more attention in recent work. An explainable artificial intelligence approach [5] for unsupervised fault detection models in industrial applications and SHAP has been implemented in rotating machinery. XAI has provided scientific insights about the variability of the signals, which enhances the acceptance and trustworthiness of users, and professionals in AI-power devices [6]. The AI model architecture becomes understandable by XAI tools' additive explanation and predictive results based on the local interpretable model-agnostic explanations to illustrate the deep-learning model [7]. The deep learning models can efficiently classify retinoblastoma and fundus images with high accuracy, recall, precision, and F1 scores on the test set. The SHAP, and LIME, visualizations provide local and global explanations for the model's predictions [8], highlighting important regions for classification. A CNN base model [9] detected the hand-based near-infrared pictures database, which offered accurate and efficient recognition of human hand motions. The paper [10] presents a method that uses XAI techniques to improve the Explainable details of a classifier for air-handling unit faults. The increasing expansion of deep learning algorithms in computer vision applications to combat noise and interference in the noise picture is growing using Sub-Pixel Layer and convolution (1×1) [11] to amplify input features and fusion feature maps.

This method used an XGBoost algorithm for fault detection and classification. In addition, the XAI-based SHAP technique provides explanations for the fault's

diagnosis. We have followed a deep learning-based adaptive CNN model [12] for automatic hand gesture detection and movement recognition by injecting some modules. The proposed XAI-CNN model enhances trust-worthy in the hand-gesture recognition process model through Interpretable Model-agnostic Explanations. AI has revolutionized the way we live and work, but it can sometimes be difficult to understand how AI algorithms reach their decisions. This is where explainable AI comes in XAI provides a framework for understanding how an AI model makes decisions, increasing trust and accountability.

3 Proposed XAI-CNN Method

This section presents an overview of the previous Convolutional Neural Network (CNN) and our proposed Explainable Artificial Intelligence Convolutional Neural Network (XAI-CNN) architecture, working procedures, and a depiction comparison. Explainable AI is a set of tools and frameworks to help users understand and interpret predictions made by machine learning models, natively integrated with a number of products and services. This helps us to debug and improve the model performance, and assist others in understanding your models' behavior. Fortunately, this work offers a plethora of powerful tools and libraries that empower AI models and machine learning practitioners to address these challenges in head-on practices.

3.1 Traditional CNN for Hand Gesture Detection

The CNN model is a kind of artificial neural network designed to process data with a grid-like structure, such as images. It consists of multiple layers of interconnected nodes, including convolutional layers that extract relevant features by applying filters to local regions of the input. This hierarchical architecture enables CNNs to automatically learn and capture complex patterns and spatial relationships, making them highly effective in tasks like image recognition, including hand gesture detection. Pooling layers that minimize the spatial dimensions come after the convolutional layers and fully connected layers that carry out the final classification based on the retrieved features. By leveraging the inherently hierarchical and shared-weight structure, CNNs have revolutionized computer vision tasks and achieved good performance in various domains, for instance, object detection, recognition, classification, segmentation, and so on.

The structure of the previous CNN model consists of two convolutional layers, one convolutional 2D layer, two pooling layers, and two fully connected layers with Rectified Linear Unit (ReLU) as an activation function. The model has two stages: image preprocessing in the first stage and gesture recognition in the second stage using the CNN model with two distinct architectures. The previous model contains multiple layers with various parameters.

3.2 Proposed XAI-CNN Architecture

We have applied the tuning and adaptation approaches to the previous CNN model and modified it to robust the performance. After several optimizations of the model, we have reached our proposed XAI-CNN model structure that provided better

generalization capability than the previous CNN model. Figure 1 depicts the performing process of our proposed XAI-CNN model.

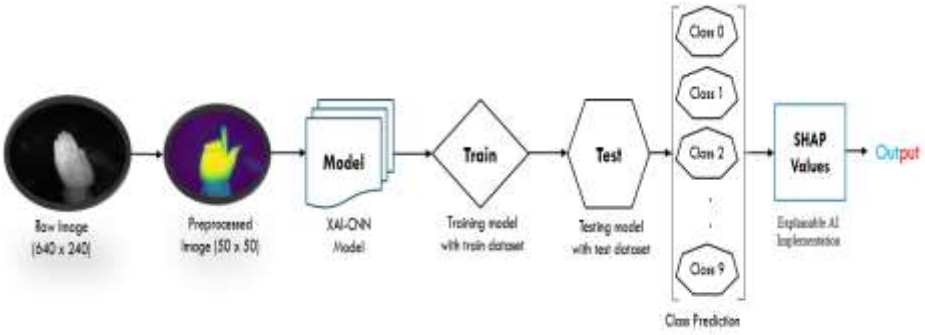


Figure 1: Proposed XAI-CNN model work process.

This transparency is crucial for regulatory compliance, ethical considerations, and gaining user acceptance. We have processed the image data, as seen in Figure 1, before being sent to the model for training. The input data goes through the image conversion process of RGB to grayscale. Each image undergoes a Grayscale to RGB conversion before being produced. Moreover, its measurements were lowered from 640*240 to 50*50 for rapid processing. The network is then given the pre-processed picture dataset and instructed to learn features and classify the photos into the assigned ten classes. We have used 20% of the test data for the proposed XAI-CNN model testing, and it has measured up to 99.98% accuracy. Figure 2 demonstrates the proposed XAI-CNN model architecture. Interpretable models play a pivotal role in machine learning, promoting trust by shedding light on their decision-making mechanisms.

Data preprocessing: Our dataset contains almost 20,000 images of different hand gestures and the original image size was 640x240. As a result, we have decreased the size of the image to 50x50 and converted them into grayscale to perform efficiently by the machine. Model modification: We have modified the layers of the previous ADCNN model as follows:

- We have changed the feature maps of the first layer of 128 to increase the shallow features of the input.
- We have added 3 layers (Convolutional 2D, ReLU, MaxPooling (2x2), and Dropout layer) before the flattening layer. As a result, MaxPooling decreases the spatial feature size, reducing the number of parameters and computations in the network. The next layer is the dropout configuration to randomly exclude 20 percent of neurons to reduce overfitting.
- Next, we have added another dense layer and ReLU activation function to receive the previous feature map 128-dimensional output as the first fully connected layer.
- The final part of our proposed XAI-CNN structure is the output layer which comprises a Softmax activation function and contains 10 neurons, one for each hand gesture recognition and classification.

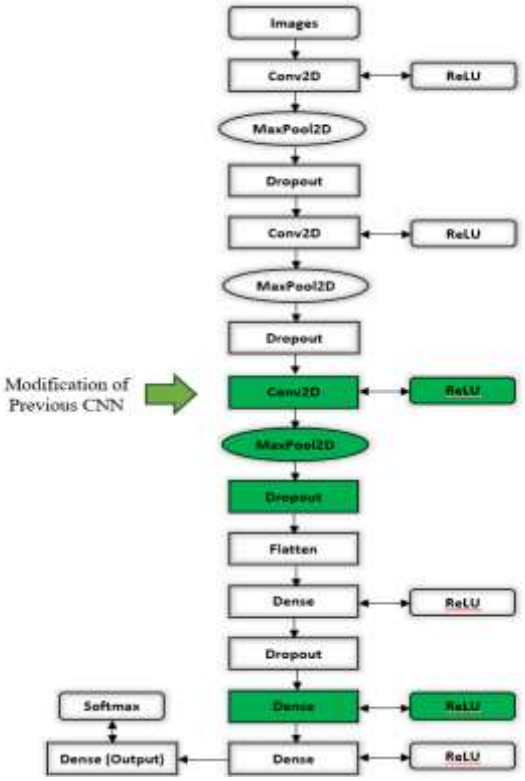


Figure 2: Proposed XAI-CNN model structure.

3.3 Data Augmentation

Data augmentation addresses a variety of challenges when training the model, such as limited or imbalanced data, overfitting, and variation and complexity. This approach is essential to train the XAI-CNN model and improve its capacity to generalize variations in hand-gesture appearances. It also generates additional training data by applying various transformations to the original dataset, such as rotations, translations, scaling, and flipping. By augmenting the dataset, we increase diversity and variability, which helps the model generalize better and become more robust to variations in hand-gesture presences. Proper kernel initialization contributes to efficient training and prevents vanishing. Kernel regularization techniques, such as regularization, mitigate overfitting, allowing the model to generalize better and capture essential hand-gesture features. It also helps prevent overfitting, as it introduces noise and variations that discourage the model from relying too heavily on specific information or patterns in the original data. Moreover, data augmentation expands the dataset size, which is particularly beneficial when working with limited labeled data, improving CNN's ability to learn representative and discriminative hand-gesture features. We have modified hyper-parameters in our proposed module to customize the XAI-CNN model. All the final selected parameters as listed in Table 1.

Table 1. Proposed XAI-CNN Model parameters

Layers	Configurations
Optimizer	Adam
Loss	categorical_crossentropy
Metrics	['accuracy']
Conv2D	128 filters, 5x5 kernel, and ReLU
Max-Pooling	2x2 kernel
Dropout	20%
Conv2D	32 filters, 3x3 kernel, and ReLU
Max-Pooling	2x2 kernel
Dropout	20%
Conv2D	32 filters, 3x3 kernel, and ReLU
Max-Pooling	2x2 kernel
Dropout	20%
Flatten	--
Dense	128 Neurons and ReLU
Dropout	20%
Dense	128 Neurons and ReLU
Dense	64 Neurons and ReLU
Dense (Output)	Softmax 10 Classes

4 Experimental Design

4.1 Dataset arrangement

The interpretable model is trained to mimic the behavior of the black box model around the data point. AI has revolutionized work, but it can sometimes be difficult to understand how AI algorithms make decisions. This is where explainable AI comes in XAI provides a framework for understanding how an AI model makes decisions, increasing trust and accountability. The local model is trained to generate an explanation of the prediction, highlighting the most important features that contributed to the prediction. The idea behind XAI is to explain the prediction model by training a local, interpretable model around the data point. We trained and tested our proposed XAI-CNN model using data obtained by the leap motion controller on hand sign recognition, known as "leapGestRecog" [13]. Twenty thousand photographs of ten distinct hand gestures made by ten different subjects—five men and five women—are included in the collection. Each type of hand gesture has a corresponding symbol. The experiment comprised all 10 of the dataset's unique gesture photos. The hand gesture identification procedure is often challenging if the image of datasets does not belong to normal illumination. However, it might be difficult to distinguish specific hand actions in these photographs because of low lighting. We have divided 80 % and 20 percent of the dataset into training datasets and testing sets. It led to a split of 16,000 photographs as training sets and 4,000 photos for testing all datasets. The picture format is PNG with 640 × 240 pixels resolution. Ten static hand gestures have been employed in the testing and recorded with various scale, rotation, and translation parameters, shown in Figure 3.

Figure 3: Sample image of "leapGestRecog" dataset.

4.2 Image Processing

The photos are changed to grayscale for real-time classification, as seen in Figure 4. As a result, the picture pre-processing lowers the number of parameters in the first convolutional layer and less the computational load. Additionally, we have used color space conversion and reshaping for one-color channel processing to compare three RGB channels.

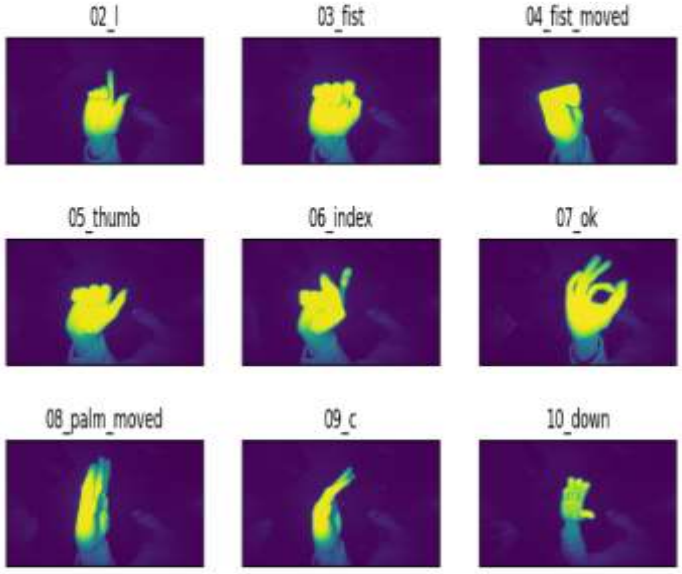


Figure 4: Randomly selected images from training data.

4.3 Evaluation Metrics

We employed a variety of criteria to evaluate the performance of the previous CNN and XAI-CNN models. Accuracy, Precision, Recall, and F1-score were these. TP: True Positive, FP: False Positive, FN: False Negative, TN: True Negative.

Precision: Another term for it is the Positive Predictive Value. The proportion of accurate forecasts to all positively anticipated class values is known as precision. We calculate the accuracy by using the equation below.

$$Precision = \frac{TP}{TP+FP} \tag{1}$$

Recall: It can be referred to as sensitivity. Calculating recall involves dividing the percentage of accurate predictions by the total number of correct class predictions. We calculate the Recall by the following equation below.

$$Recall = \frac{TP}{TP+FN} \tag{2}$$

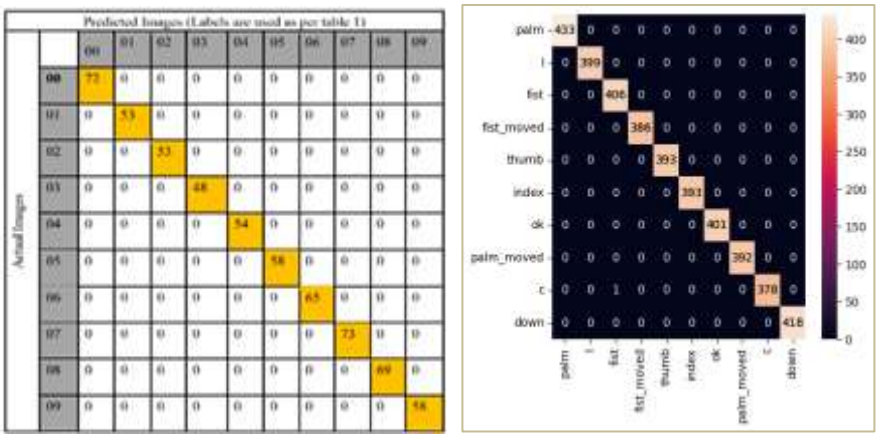
$$F1 = \frac{2\times Precision \times Recall}{Precision + Recall} \tag{3}$$

$$Accuracy = \frac{TP+TN}{TP+\mathrm{TN}+FP+FN} \times 100 \tag{4}$$

F1-score: The F1-score is also known as the F-measure or the F-score. The F1 score illustrates how Precision and Recall are balanced. Only when the values of Precision and Recall are both high values of the F1-score increase. Accuracy is the percentage of correct classifications.

5 Experimental Results

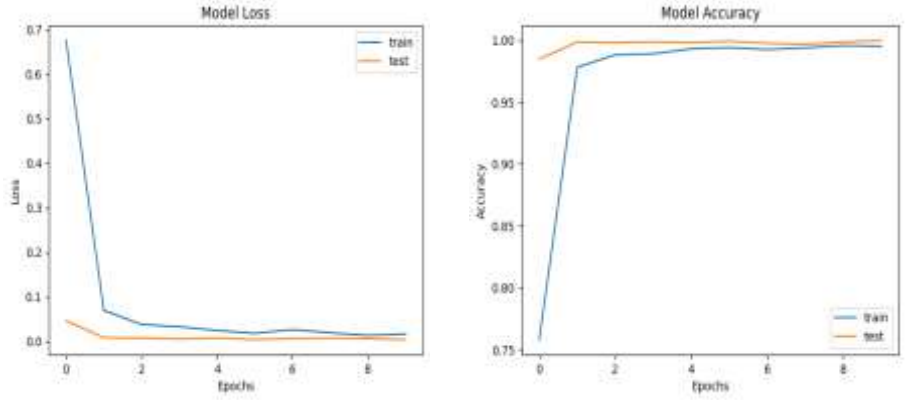
This section demonstrates the experimental outcomes of the proposed XAI-CNN architectures and the result comparison. We have used the Python platform to conduct the experiments utilizing the Numpy, Keras, and Scikit-learn modules. As explained in Section III, the training set from the "leapGestRecog" dataset was used to learn features from the training data and to test the models using the testing data. Figure 5 shows the comparison of traditional CNN and our proposed XAI-CNN confusion matrix result difference.



(a) (b)

Figure 5: Confusion matrix of (a) previous model, (b) proposed XAI-CNN.

Figure 6 (a) shows the proposed XAI-CNN model training loss. Figure 6 (b) demonstrates the proposed XAI-CNN model accuracy.



(a) (b)

Figure 6: Performance of the proposed XAI-CNN (a) Model loss, (b) Model accuracy.

Table 2 lists all the performance comparisons of the previous CNN models and the proposed XAI-CNN based on the values of various performance parameters. The baseline CNN model got 95.73% accuracy, the adaptive CNN model had 99.73 percent accuracy, and the deep CNN model acquired 99.73% accuracy. Whereas the proposed XAI-CNN model has achieved the best accuracy of 99.98 percent for the same epoch numbers.

Table 2. Performance comparison of the CNN model.

Parameters	Baseline CNN	Adapted CNN	Deep CNN	Proposed XAI-CNN
Epoch	10	10	10	10
Precision	0.96	0.98	0.98	0.99
Recall	0.96	0.99	0.99	0.99
F1 Score	0.96	0.99	0.99	0.99
Accuracy	95.73	99.73	99.75	99.98

5.1 Shapley Additive Explanations

The SHAP is a valuable XAI tool for training our proposed XAI-CNN model for hand-gesture detection and computer vision tasks. SHAP values illustrate the contribution or the importance of each feature on the prediction of the model. These values provide insights into the contribution of each feature to the model's output prediction. While dealing with image data, the SHAP (Shapley Additive Explanations) values provide insights into the contribution of each pixel or region of the image to the model's output prediction. These values quantify the importance and relevance of different image features for the decision-making process of the model. The method can be applied to visualize the Shapley values, identify the most important features, and quantify the model's bias towards certain groups or classes.

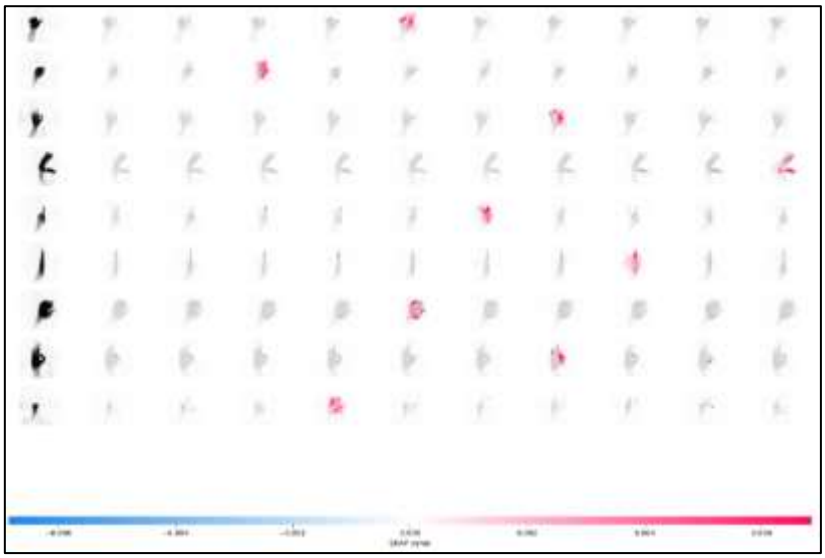


Figure 7: SHAP Values implementation on XAI-CNN.

Figure 7 explains ten outputs (each class) for nine different images. Red pixels increase the model's output probability while blue pixels decrease the output probability. The input images are shown on the left and as nearly transparent grayscale backings behind each of the explanations. SHAP values show which pixels or areas of the input image have the most effects on the model's prediction of a specific hand motion in the context of hand gesture detection using an XAI-CNN model. We can comprehend the visual signals and patterns that the model uses to generate precise predictions by examining the SHAP values. The heat-map process flows to display the SHAP values for picture data, with brighter areas denoting regions with greater significance or contribution to the model's judgment. Users may check and trust the model's outputs thanks to this visualization, which aids in understanding and comprehending the logic behind the model's predictions.

In summary, SHAP values provide an interpretable and quantitative measure of the contribution of pixels or regions in an image towards the model's decision-making process, allowing for better understanding and transparency in image classification tasks such as hand-gesture detection.

5.2 LIME Explanations

In LIME's explanation for Figure 8, the colors and their meanings typically represent the importance of different regions or pixels in the image for the model's prediction of the class. The colors provide insights into which parts of the image strongly influence the model's decision and which parts are less relevant. The green background likely indicates the regions or pixels in the image that have a positive impact on the model's prediction of the class. These regions are supportive of the model's decision and contribute to the model's confidence in predicting class.

The brighter or more saturated the green color, the stronger the positive influence of those regions on the prediction. The red background represents the regions or pixels in the image that have a negative impact on the model's prediction of the class. These regions might be misleading or contrary to the gesture, leading the model to have lower confidence in its prediction. The brighter or more saturated the red color, the stronger the negative influence of those regions on the prediction.

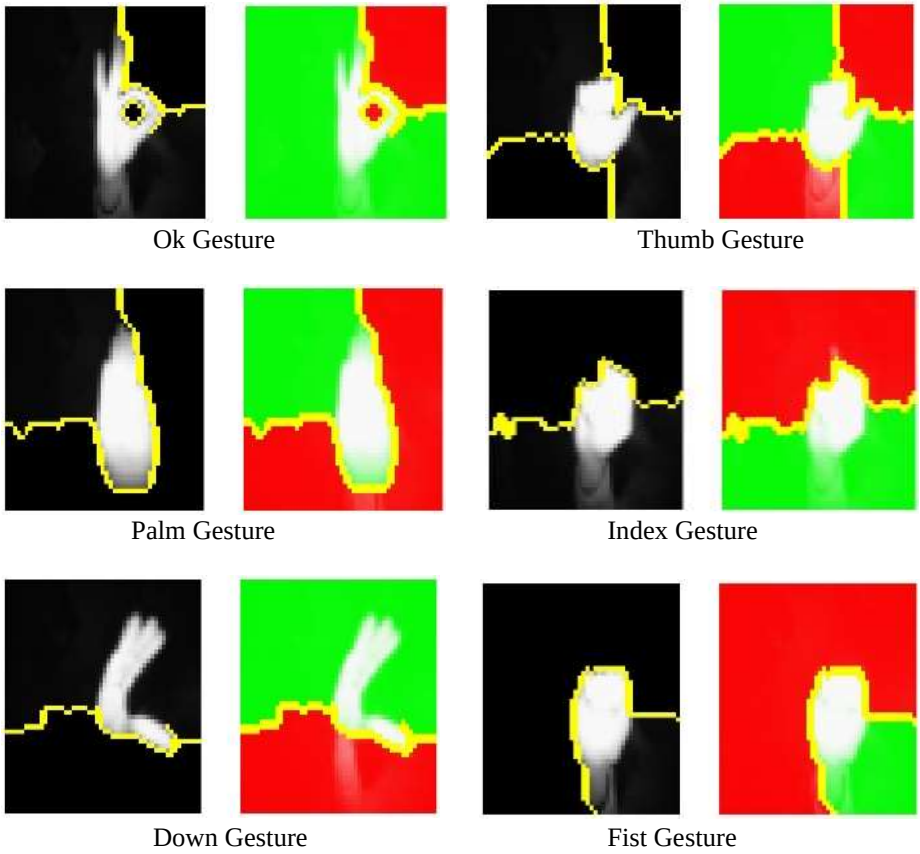


Figure 8: LINE employment in proposed XAI-CNN.

The yellow color border around the sign most likely represents the regions or pixels at the image's periphery that also contribute to the model's decision. While they might have some relevance, their importance may be lower compared to the regions inside the sign. The yellow color border around the circle suggests that this region is relevant and contributes to the model's prediction of the class. However, its importance may be slightly lower compared to the central part of the sign, which is indicated by the red color. The red color fill in the center of the sign signifies that this particular region plays a significant role in the model's prediction of the class. The model heavily relies on the features or patterns in this central part to identify and classify the gesture.

5.3 Anchor Explanations

In Anchor's explanation, the colors and their meanings typically represent the important regions or conditions in the image that influence the model's prediction of the "ok" class. Each color provides insights into the specific areas or features that are significant for the model's decision. The deep green background likely indicates the regions or pixels in the image that are strongly associated with the model's prediction of the class. These regions play a crucial role in the decision-making process, and the model heavily relies on them to classify the image. The light green color covering the palm of the hand suggests that this particular area is essential for the model's prediction of the class. The features or patterns present in the palm region significantly contribute to the model's ability to recognize the sign. The light blue and deep blue color borders around the upper part of all fingers indicate that these regions are relevant and contribute to the model's prediction of the class as shown in Figure 9. The lighter blue color may represent a moderate influence, while the deeper blue color suggests a stronger impact on the model's decision. The brown color borders around the upper left and lower left parts of the image signify that these areas are important for the model's prediction.

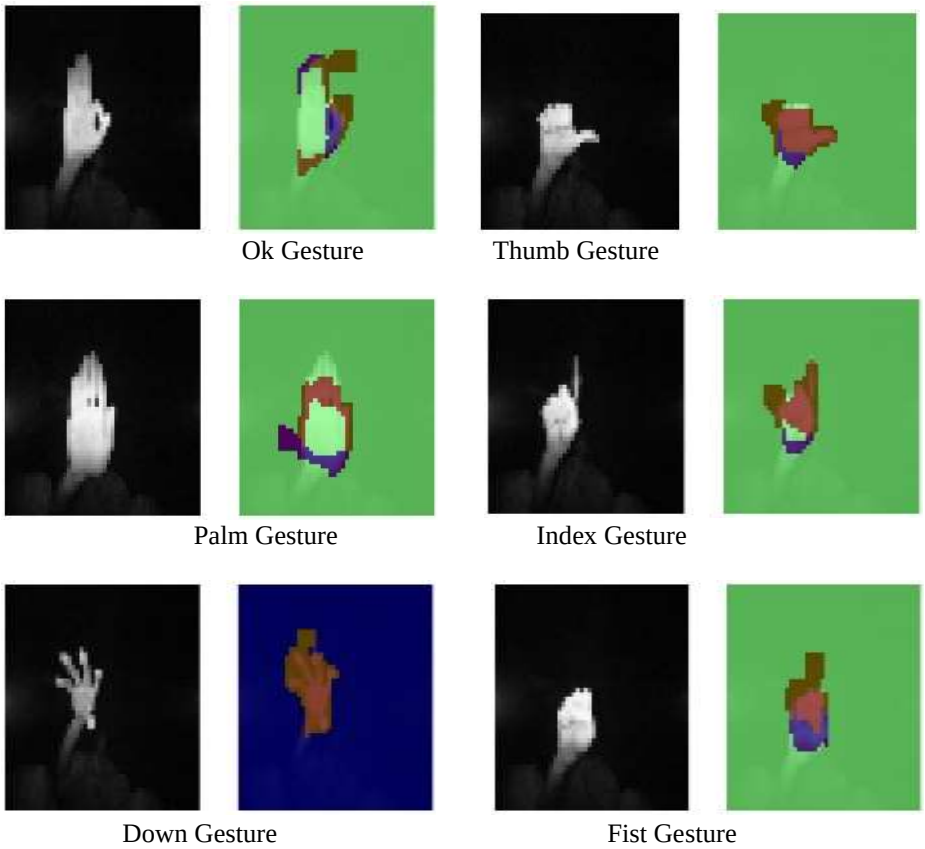


Figure 9: Anchor execution in XAI-CNN model.

The specific features or conditions present in these regions contribute significantly to the model's ability to recognize the sign. The light blue color border around the circle suggests that this region is relevant to the model's prediction. While important, its influence may be slightly lower compared to the deep green background. The deep blue color fill in the center of the sign indicates that this central part of the circle plays a critical role in the model's prediction of the class. The model heavily relies on the features or patterns in this central area to identify and classify the gesture.

6 Conclusion

In conclusion, this paper highlights the significance of Explainable AI (XAI) tools techniques, specifically focusing on SHAP values, LIME, and Anchor integration in the proposed XAI-CNN model for hand gesture recognition. The SHAP values effectively extract and classify high-level features from data, enabling accurate recognition of hand gestures. In addition, LIME as an XAI tool enhances interpretability, transparency, and accountability in the decision-making process of the proposed XAI-CNN model. Integrating Anchor provides valuable insights into the contribution of individual hand gesture features, improving interpretability and facilitating trust in the model's outputs. Understanding the effects of feature characteristics, biases, and errors can be identified and addressed to refine the model's performance. Overall, the hand gesture recognition process becomes more precise and understandable when the XAI tools, notably SHAP values, LIME, and Anchor combined with the XAI-CNN model. This development improves our comprehension of the decision-making process and promotes AI and human cooperation. The experimental results have presented responsible and dependable AI systems by providing users with technology for accurate hand gesture recognition.

References

[1] M. -H. Sheu, S. M. S. Morsalin, C. -C. Hsu, S. -C. Lai, S. -H. Wang and C. -Y. Chang, "Improvement of Human Pose Estimation and Processing with the Intensive Feature Consistency Network," in *IEEE Access*, vol.11, pp. 28045-28059, 2023, doi: 10.1109/ACCESS.2023.3258417.

[2] C. J. L. Flores, A. E. G. Cutipa and R. L. Enciso, “Application of convolutional neural networks for static hand gestures recognition under different invariant features,” in *Proc. of 2017 IEEE XXIV International Conference on Electronics, Electrical Engineering and Computing (INTERCON)*, Cusco, Peru, 2017, pp. 1-4, doi: 10.1109/INTERCON.2017.8079727.

[3] M. T. Ribeiro, S. Singh, and C. Guestrin. “Why Should I Trust You?”: Explaining the Predictions of Any Classifier, in *Proc. of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics*, San Diego, California, pp 97–101, 2016.

[4] M. T. Ribeiro, S. Singh, and C. Guestrin. “Anchors: High-Precision Model-Agnostic Explanations.” in Proc, The Thirty-Second AAAI Conference (AAAI-18), February 2018, pp 1527–1535.

[5] L. C. Brito, G. A. Susto, J. N. Brito, and M. A. Duarte, “An explainable artificial intelligence approach for unsupervised fault detection and diagnosis in rotating machinery,” *Proc. of* 2017 *IEEE XXIV International Conference on Electronics, Electrical Engineering and Computing (INTERCON)*, Cusco, Peru, 2017, pp. 1-4, doi: 10.1109/INTERCON.2017.8079727.

[6] N. Gozzi, L. Malandri, F. Mercorio, and A. Pedrocchi, “An explainable artificial intelligence approach for unsupervised fault detection and diagnosis in rotating machinery” in *Mechanical Systems and Signal Processing*, vol. 163, p. 108105, 2022.

[7] M. Bhandari, P. Yogarajah, M. S. Kavitha, and J. Condell, “Exploring the Capabilities of a Lightweight CNN Model in Accurately Identifying Renal Abnormalities: Cysts, Stones, and Tumors, Using LIME and SHAP," in *Applied Sciences*, vol. 13, no. 5, p. 3125, 2023.

[8] B. Aldughayfiq, F. Ashfaq, N. Jhanjhi, and M. Humayun, “Explainable AI for Retinoblastoma Diagnosis: Interpreting Deep Learning Models with LIME and SHAP," in *Diagnostics*, vol. 13, no. 11, p. 1932, 2023.

[9] A. G. Mahmoud, A. M. Hasan, and N. M. Hassan, “Convolutional neural networks framework for human hand gesture recognition,” in *Bulletin of Electrical Engineering and Informatics*, vol.10, no.4, pp. 2223-2230, 2021.

[10] M. Meas et al., “Explainability and Transparency of Classifiers for Air-Handling Unit Faults Using Explainable Artificial Intelligence (XAI)," in *Sensors*, vol. 22, no. 17, p. 6338, 2022.

[11] M. -H. Sheu, S. M. S. Morsalin, S. -H. Wang, Y. -T. Shen, S. -C. Hsia and C. -Y. Chang, "FIBS-Unet: Feature Integration and Block Smoothing Network for Single Image Dehazing," in IEEE Access, vol. 10, pp. 71764-71776, 2022, doi: 10.1109/ACCESS.2022.3188860.

[12] A. A. Alani, G. Cosma, A. Taherkhani and T. M. McGinnity, “Hand gesture recognition using an adapted convolutional neural network with data augmentation," *Proc. of 2018 4th International Conference on Information Management (ICIM)*, Oxford, UK, 2018, pp. 5-12, doi: 10.1109/INFOMAN.2018.8392660.

[13] H. K. Sharma, P. Kumar, P. Ahlawat, and Y. Manchanda, "Deep Learning Based Accurate Hand Gesture Recognition Using Enhanced CNN Model," in *Proc. of Second International Conference on Computing*, 2021.