Indonesian Journal of Electrical Engineering and Informatics (IJEEI)
Vol. 11, No. 1, March 2023, pp. 286~299
ISSN: 2089-3272, DOI: 10.52549/ijeei.v11i1.4449

Enhanced Emotion Recognition in Videos: A Convolutional Neural Network Strategy for Human Facial Expression Detection and Classification

Arselan Ashraf[1], Teddy Surya Gunawan[2], Fatchul Arifin[3], Mira Kartiwi[4], Ali Sophian[5], Mohamed Hadi Habaebi[6]
[1,2,6] Department of Electrical and Computer Engineering, International Islamic University Malaysia, Malaysia
[2] School of Electrical Engineering, Telkom University, Bandung, Indonesia
[3] Department of Electronic and Informatics Engineering, Universitas Negeri Yogyakarta, Yogyakarta, Indonesia
[4] Department of Information Systems, International Islamic University Malaysia, Malaysia
[5] Department of Mechatronics Engineering, International Islamic University Malaysia, Malaysia

Article Info

Article history:

Received Jan 12, 2023
Revised Mar 8, 2023
Accepted Mar 20, 2023

Keywords:

Artificial Intelligence
Convolutional Neural Networks
Emotion Recognition
Image Processing
Machine Vision

ABSTRACT

The human face is essential in conveying emotions, as facial expressions serve as effective, natural, and universal indicators of emotional states. Automated emotion recognition has garnered increasing interest due to its potential applications in various fields, such as human-computer interaction, machine learning, robotic control, and driver emotional state monitoring. With artificial intelligence and computational power advancements, visual emotion recognition has become a prominent research area. Despite extensive research employing machine learning algorithms like convolutional neural networks (CNN), challenges remain concerning input data processing, emotion classification scope, data size, optimal CNN configurations, and performance evaluation. To address these issues, we propose a comprehensive CNN-based model for real-time detection and classification of five primary emotions: anger, happiness, neutrality, sadness, and surprise. We employ the Amsterdam Dynamic Facial Expression Set – Bath Intensity Variations (ADFES-BIV) video dataset, extracting image frames from the video samples. Image processing techniques such as histogram equalization, color conversion, cropping, and resizing are applied to the frames before labeling. The Viola-Jones algorithm is then used for face detection on the processed grayscale images. We develop and train a CNN on the processed image data, implementing dropout, batch normalization, and L2 regularization to reduce overfitting. The ideal hyperparameters are determined through trial and error, and the model's performance is evaluated. The proposed model achieves a recognition accuracy of 99.38%, with the confusion matrix, recall, precision, F1 score, and processing time further quantifying its performance characteristics. The model's generalization performance is assessed using images from the Warsaw Set of Emotional Facial Expression Pictures (WSEFEP) and Extended Cohn-Kanade Database (CK+) datasets. The results demonstrate the efficiency and usability of our proposed approach, contributing valuable insights into real-time visual emotion recognition.

Corresponding Author:

Teddy Surya Gunawan
Department of Electrical and Computer Engineering,
International Islamic University Malaysia, Malaysia.
Email: tsgunawan@iium.edu.my

1. INTRODUCTION

The facial profile is critical because it contains vital information about a person's personality. A person's face reveals a great deal about their personality, gender, age, and other distinguishing characteristics,

as well as clues to their social background and true appeal [1]. Additionally, emotional expression plays a significant role in facial communication. Human faces exhibit a variety of expressions in response to various events, which is quite helpful in conveying their emotions. Recognizing these emotions can be highly beneficial in conveying people's intentions and personalities [2]. Emotional research and prediction have been a challenge for some time because emotional characteristics vary between individuals. Emotion recognition is a technique that focuses on analyzing and recognizing human expressions. The field of emotion recognition has exploded in popularity over the last decade. Numerous studies have been conducted to ascertain emotions using various inputs such as speech, visual, and textual data [3][4]. The field of human-computer interaction, robotics, medical diagnosis, gaming, and surveillance systems benefit from video-based emotion recognition [5].

A multimodal approach to developing an emotion recognition model was developed by [6]. Multiple images and video databases were consulted, including the ADFES-BIV database. A histogram of oriented gradients and local binary patterns extracted facial features. Support Vector Machine (SVM) and K-Nearest Neighbors (KNN) were used for classification. For ADFES, the KNN+LBP and LBP+SVM methods were applied across a range of cell sizes, namely 8, 16, 32, and 64. The best accuracy for LBP+KNN was 87.44 % when cell size 32 was used, while the best accuracy for LBP+SVM was 90.23% when cell size 32 was used. The characterization experiment was conducted using the ADFES-BIV dataset [7]. They proposed a programmed framework that used a sparse-representation-based classifier and achieved an accuracy of up to 80% by considering the fleeting data inherent in video recordings.

A model for emotion recognition using deep convolutional neural networks and the TensorFlow machine learning library was developed by [8]. ADFES-BIV was used to train the model. The image frames were extracted from the video samples, and then the Viola-Jones algorithm was used to detect faces. Testing and validation were conducted using the WSEFEP database. The model's recognition rate accuracy was determined to be 95%. In [9], the authors developed and validated a series of video enhancements with three distinct levels of emotional intensity, ranging from low to high. The video recordings were obtained from the ADFES-BIV dataset after participants completed a facial emotion recognition task that included six core emotions plus embarrassment, disgust, and pride—resulting in an accuracy of 69%. Three distinct power levels were certified, ranging from extreme-intensity expressions with the highest degree of precision and reaction time to medium-intensity expressions with a higher degree of precision and reaction time than low-intensity expressions.

Numerous conventional methods have been used to achieve emotion recognition, but the results have been mixed and continue to fall short in accuracy, precision, recall, and F1 scores. When more emotions are included in recognition models, the performance parameters of the recognition models tend to decrease. It is because a more significant number of classes results in a more negligible probability difference between them, making it more difficult for a model to detect the specific class convincingly. Classifying two emotions with comparable characteristics, such as despair and disgust, might yield false positives. With the addition of machine learning techniques to emotion recognition, a noticeable increase in performance is evident. Deep learning has raised the bar even higher in this field and prompted several researchers to conduct experiments. CNN, a deep learning technique, has demonstrated exceptional performance in video-based emotion recognition [10]. However, several open gaps still need to be addressed to improve the accuracy and generalizability of the models, as follows:

1. The internal covariate shift problem is one of the primary challenges in training deep neural networks for emotion detection in faces. The changing distribution of input data causes this issue during training, which causes the internal representations of the network to shift, making it difficult to converge on the optimal solution.
2. Another issue that requires attention is overfitting. Deep neural networks have a high capacity to memorize training data, which can result in overfitting and poor performance on new data.
3. While significant progress has been made in developing face emotion detection and classification models, their precision and generalizability remain limited. One reason is the dearth of significant, diverse training and testing datasets.

By addressing these gaps, we can enhance the performance of these models and enable their applications in numerous domains, such as healthcare, education, and entertainment. The multidisciplinary field of emotion recognition based on video input includes disciplines such as human-computer interaction, affective computing, and psychology. To generate a well-fitted model, a well-organized and presented database is required. The facial features extracted from the emotion samples in the database must be extracted with care [11]. Table 1 lists additional models for visual emotion recognition and discusses their strengths and weaknesses.

Table 1. A Review of Visual Emotion Detection Models

Method used	Strength	Limitations	Reference
Convolutional Neural Networks and Local Directional Strength Pattern (LDSP)	Local directional pattern (LDP) typically examines only the absolute values of a pixel's edge strengths to extract features. This generalization of LDP generates similar patterns for two distinct types of edge pixels. LDSP resolves this issue. It considers the binary values of the position, with the directions denoting the strongest and weakest initial strengths. The maximum strength denotes the strongest direction on the bright side of a pixel, whereas the lowest strength shows the strongest direction on the dark side. Consequently, merging the binary positions of these two directions produced more robust patterns than LDP. This model has the greatest average recognition rate, 95.42 %.	The model was trained on CNN, which can also extract image features. Nevertheless, in this method feature extraction capability of CNN was not employed.	**[12]**
Multi-Task Convolutional Neural Network	The critical issue of balancing between tasks in MTL was resolved in this model by developing a dynamic-weighting technique to dynamically assign the loss weights to each side task.	The main downside of the multiple-task strategy for emotion classification is that many training sets might degrade processing performance and consume a large amount of memory. Also, LFW and IJB-A have numerous variables other than position, such as expression and blurring, etc., that the suggested method cannot handle well.	**[13]**
Deep Comprehensive Multipatches Aggregation Convolutional Neural Networks	The proposed strategy consists primarily of two CNN segments. The first branch collects local characteristics from image patches, while the second extracts global characteristics from the entire expressional image. In the approach, local features represent expressional details, whereas holistic characteristics define the expression's high-level semantic content. Both local and global characteristics are combined prior to classification.	This model produces some false positives, and generalizability is also not promising.	**[14]**
Two-Part Convolutional Neural Network	This model employs a two-stage convolutional neural network (CNN), the first stage responsible for background subtraction and the second for facial feature vector extraction. Weights and exponents in the final perceptron layer of the two-layer CNN are refined with each iteration. The proposed model is novel and, as a result, more accurate than traditional approaches that utilize a single level of CNN.	There was a correlation between the number of layers and an increase in execution time, but this correlation did not contribute significantly to the study and is therefore not discussed.	**[15]**
Convolutional Neural Networks	The model has a total of two convolution layers, with dropouts following each. The dropout layer is utilized in order to lessen the amount of overfitting.	The accuracy of the proposed model was reported to be 78.04% which is not very promising.	**[16]**
Lightweight convolutional neural network (CNN)	The paper proposes a lightweight convolutional neural network (CNN) capable of detecting facial emotions in real-time and large quantities, with improved classification accuracy. In order to accomplish this, the system employs a multi-task cascaded convolutional network (MTCNN) to detect faces and extract their coordinates, which are then sent to the emotion classification model. The cascade detection feature allows for more efficient use of memory resources, as only one network must be used at a time. Overall, the system provides an effective, context-adaptable solution for detecting facial emotions in real-time.	On the FER-2013 dataset, the accuracy of the proposed model was only about 67 percent accurate, which was not encouraging.	**[17]**
Deep Convolutional Neural Networks	The paper presents an innovative end-to-end pipeline method for real-time facial expression recognition. The method employs two optimized convolutional neural networks (CNNs) for face recognition and facial expression recognition. The face recognition network uses two different models, one with high accuracy but low inference speed (faster region-based CNN) and another with lower accuracy but higher inference speed (single shot detector CNN). The paper uses transfer learning and fine-tuning of three CNN models (VGG, Inception V3, and ResNet) for facial expression recognition. The approach enables real-time inference	The model's generalizability is not as promising as the authors assert, as there is a significant disparity in model accuracy.	**[18]**

Method used	Strength	Limitations	Reference
	speed for the entire process, making it useful in various applications. The paper presents a promising solution for real-time facial expression recognition using CNNs.		
Deep Convolutional Neural Network	The proposed system is intended to interpret the emotional evolution over time based on facial images. The framework employs a novel technique for extracting deep features from the Fully Connected Layer 6 of the AlexNet, with a standard Linear Discriminant Analysis Classifier serving as the ultimate classifier. The system was evaluated using five benchmarking databases, including JAFFE, KDEF, and CK+, and databases containing images captured in the wild, such as FER2013 and AffectNet.	While research indicates that automatic facial expression recognition may be a promising avenue for mental health care, current state-of-the-art methods may not be precise enough and require substantial computational resources.	**[19]**
Transfer Learning in the Deep CNN	This study proposes a novel method for developing a facial expression recognition (FER) system using a very Deep Convolutional Neural Network (DCNN) model and Transfer Learning (TL) technique. The proposed method replaces the dense upper layer(s) of a pre-trained DCNN model with layers compatible with FER, followed by fine-tuning the model using facial emotion data. The research introduces a pipeline strategy for the training procedure, in which the training of the dense layer(s) is followed by successive tuning of the pre-trained DCNN blocks. This strategy has resulted in the gradual enhancement of FER's accuracy to a higher level.	The evaluation dataset may not accurately represent real-world scenarios and lacks generalizability.	**[20]**
Fine-Tuned VGG-16	This study proposes a modified convolutional neural network (CNN) based on the VGG-16 classification model, which was pre-trained on the ImageNet dataset and tuned for emotion classification. The research focuses on classifying the FER-2013 dataset, which contains more than 35,000 face images captured in natural settings for seven distinct emotions. The dataset is divided into training data distributions of 80%, validation data distributions of 10%, and testing data distributions of 10%. This dataset trains the modified CNN to recognize the various emotions displayed in the images.	The accuracy of the model was not promising at about 69.40%.	**[21]**
Convolutional Neural Networks	This study aims to develop a mobile application that recognizes emotions based on facial expressions in real-time using the Deep Learning technique Convolutional Neural Network (CNN). The MobileNet training algorithm for recognition has been implemented. The study identifies four types of emotions: happiness, sadness, surprise, and disgust.	The study aims to identify four types of emotions: happiness, sadness, surprise, and disgust. However, six fundamental emotions, including anger and fear, have not been considered in this research. In addition, the generalization of the research has not been addressed, which limits the model's applicability to real-world situations.	**[22]**

Convolutional Neural Networks (CNN) are crucial and advantageous for feature extraction and classification. In pursuit of promising results in video-based emotion recognition, we have developed a deep learning model that focuses on recognizing five emotions: anger, happiness, neutrality, sadness, and surprise. We will assess the performance of our model using metrics such as accuracy, precision, recall, F1-score, and confusion matrix. This paper is structured as follows: Section 2 delves into the methodology behind our video-based emotion recognition model and outlines the various pre-processing steps. Section 3 presents the results and discussions, while Section 4 offers a concise conclusion to the paper.

2. RESEARCH METHOD

Figure 1 illustrates the proposed architecture for developing a video-based emotion recognition model using deep learning. In the case of video-based emotion recognition, the input visual samples are processed through a series of pre-processing steps before the face features are extracted. Emotion recognition is accomplished by extracting facial features from image frames. CNN is used for training and classification. The proposed CNN model comprises the following layers: an input image layer, multiple convolutional layers, pooling layers, fully connected layers, a softmax unit, dropout layers, and an output classification layer. Five emotions are recognized: anger, happiness, neutrality, sadness, and surprise.

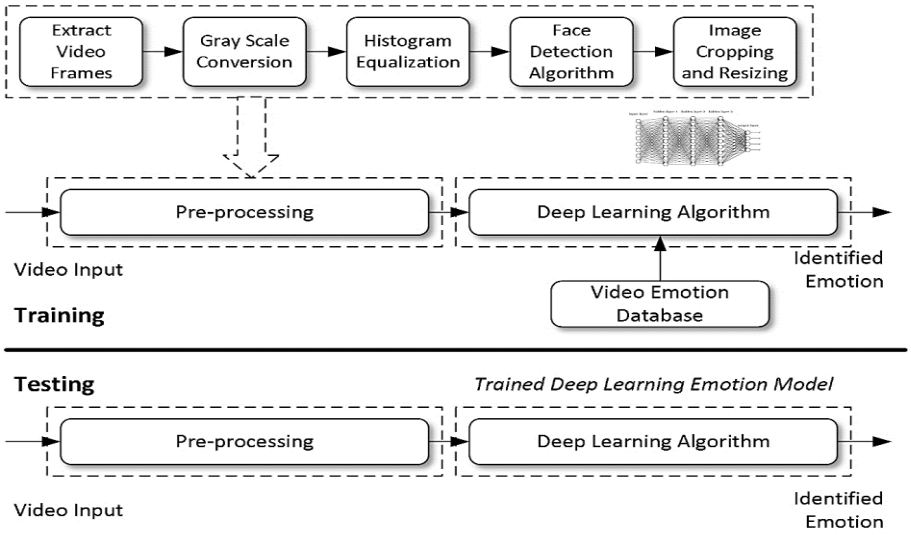

Figure 1. Proposed Model

2.1. Dataset Preparation

The Amsterdam Dynamic Facial Expression Set – Bath Intensity Variations dataset was used in this study. The ADFES-BIV variant of the Northern European arrangement of ADFES boosts includes recordings demonstrating the nine feelings in three distinct normalized powers (low, moderate, and high). The set contains 370 brief recordings (1040ms each), ten specifically designed for training and acclimation during examinations (one for each enthusiastic articulation classification). Sadness, surprise, happiness, pride, contempt, anger, disgust, fear, embarrassment, and neutral are all enthusiastic articulations demonstrated. The first step is to extract image frames from each video sample and organize them into distinct folders, as illustrated in Figure 2. Creating distinct folders for each emotion under consideration simplifies data acquisition. For this research, we selected five primary emotions: anger, happiness, neutrality, sadness, and surprise. Although the dataset contains more emotions, the class difference for extracting features is insignificant. For example, emotion classes like sadness, contempt, disgust, and embarrassment result in a smaller probability difference between them, making it more difficult for a model to detect the specific class convincingly, resulting in more false positives hence degrading the performance. Also, most researchers have considered these five emotions, and we can benchmark our proposed model more effectively. The image samples from the Warsaw Set of Emotional Facial Expression Pictures (WSEFEP) and Extended Cohn-Kanade Database (CK+) were used for Testing to generalize the proposed model performance.

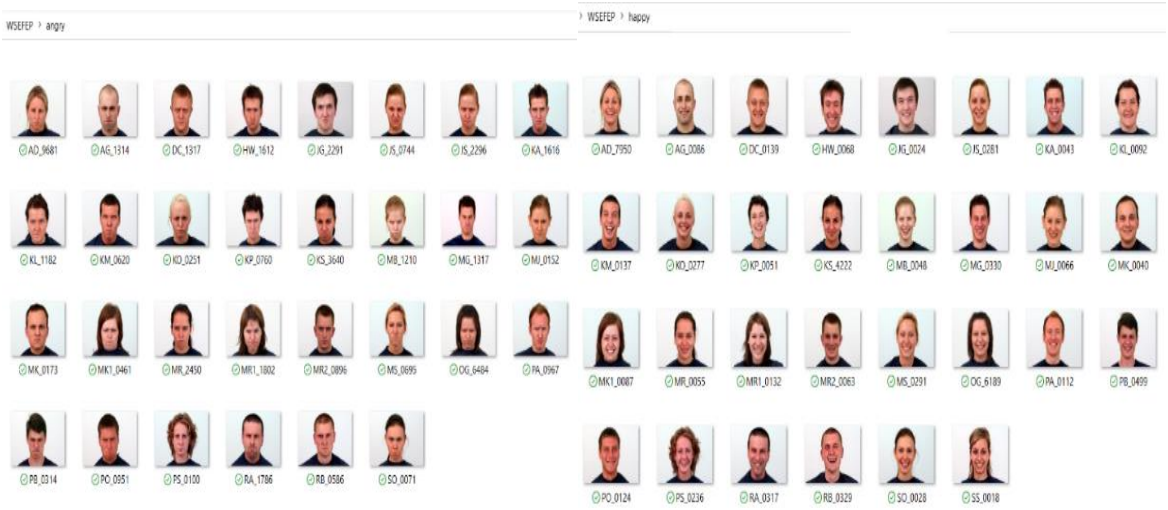

Figure 2. Extracted Frames for Anger and Happiness

2.2. RGB to Gray Scale Conversions

Figure 3 depicts a grayscale image that has been converted. These images are shades ranging from dark to light and black to white [23]. A grayscale image is created by determining the light intensity in each pixel over a specified electromagnetic range (for example, infrared or visible light). Converting RGB images to Grayscale is an efficient way to reduce processing time.

Figure 3. Gray Scale Converted Image

2.3. Histogram Equalization

A picture histogram illustrates the difference in appropriation between two computer-generated images, as illustrated in Figure 4. It plots the perceived value of pixels [24]. When viewing a histogram for a particular image, the viewer can initially assess the image's apparent spread in general.

Figure 4. Histogram Equalization, Unequalized Image Versus Equalized Image

The image's histogram prepares for a contrast adjustment via histogram equalization. This technique is frequently used to improve the overall distinction of multiple images, primarily when the image is addressed by a low-intensity source [25]. This modification allows a more accurate circled representation of the intensities on the histogram. Equalization of the histogram achieves this by effectively spreading out the intensity values.

2.4. Face Detection, Image Cropping, and Image Resizing

Faces must be recognized from video frames extracted from the video database. The Viola-Jones algorithm for face detection, as explained in [26], was used in this work. The faces identified by the face discovery calculation are trimmed to provide a more complete and precise view of the facial image. Cropping is the process of removing undesirable external locations from a photograph or exhibited image. The technique entails the ejection of a portion of an image's periphery districts to eliminate coincidental waste, improve the encompassing image, alter the viewpoint's extent or feature, or isolate the theme from its experience. The images' size fluctuates due to editing procedures on the edges [27, 28]. Hence, to achieve consistency, these edited pictures are exposed to resizing to a standard size of $M \times N$. Figure 5 shows a representative example.

Figure 5. Face Detection, Image Cropping, And Image Resizing, respectively.

2.5. CNN Architecture

CNN is a multilayer neural network with multiple layers. For instance, visual content (such as facial features) is frequently represented by a collection of feature maps obtained by convolving the input with various predetermined channels [29]. Pooling layers may be utilized after convolutional layers to extract the most vital information from convolutional. In addition, it contains layers with complete connectivity, with each neuron in the data layer connected to each neuron in the subsequent layer. Based on this representation, a Softmax layer completes the classification task. SVM was primarily used for classification and some training algorithms in previous literature. This study utilizes CNN for feature extraction, training, and classification. CNN simplifies image data processing by eliminating unnecessary features without sacrificing classification-critical information. Table 2 displays the CNN training and classification pseudocode for the proposed model.

Table 2. Pseudocode for Training and Classification

TRAINING PROCESS
INPUT:
Step1: Labelled training data say $X = [X^1, X^2, X^3, \dots, X^n]$, *% n is the total number of emotion classes.*
Step 2: CNN ← X; *% the training data is fed to the CNN for feature extraction*
Step3: (C_i, R_i, M_i) ← X; $C_i = Convolutional\ layers\ (i = 1\ to\ 3)$; $R_i = ReLU\ (i = 1\ to\ 3)$; $M_i = Max\ Pooling\ Layers\ (i = 1\ to\ 3)$; *% Every convolutional layer is connected to the rectified linear unit (ReLU) and a max-pooling layer.*
Step 4: Output from step 3 is fed to B; *%* $B = Batch\ Normalization\ Layer$.
Step 5: $(Fc1, Fc2)$ ← B; *% Batch normalization layer output is connected to fully connected layers 1 and 2, respectively.*
Step 6: Dropout = 0.5; *%. Step 5 is followed by a dropout of 50% to avoid overfitting.*
OUTPUT:
CNN'; *% Trained CNN*
CLASSIFICATION PROCESS
INPUT:
Step 1: x; *% is an image to be classified.*
Step 2: CNN' ← x; *% feed the input to the trained model*
OUTPUT:
x' ; *% classified image.*

Table 3. Novel configuration selection using trial and error method

ConvNet	Neurons	Momentum	Learning Rate	Dropout (%)	Fully Connected Layers	Recognition Rate (%)
1	16	0.3	0.1	30	1	85
1	16	0.5	0.01	40	1	83
1	16	0.8	0.001	50	2	88.45
2	16,32	0.3	0.1	30	1	87
2	16,32	0.5	0.01	40	1	91
2	**16,32**	**0.8**	**0.001**	**50**	**2**	**94**
3	**16,32,64**	**0.3**	**0.1**	**30**	**1**	**96.2**
3	**16,32,64**	**0.5**	**0.01**	**40**	**1**	**98**
3	**16,32,64**	**0.8**	**0.001**	**50**	**2**	**99.38**
4	16,32,64,128	0.3	0.1	30	1	89.35
4	16,32,64,128	0.5	0.01	40	1	91.23
4	16,32,64,128	0.8	0.001	50	2	90
5	16,32,64,128, 192	0.3	0.1	30	1	81
5	16,32,64,128, 192	0.5	0.01	40	1	83.38
5	16,32,64,128, 192	0.8	0.001	50	2	86.23

The network's input is a face image extracted from the input dataset. Three convolutional layers follow, each with polling applied. Filter sizes 16 (3×3), 32 (3×3), and 64 (3×3) make up the three layers, respectively. In order to give the model more depth and the ability to identify more complex features, second and third convolution layers have been added. A max-pooling layer of 3×3 with stride 2 connects each layer using rectified linear units (ReLU). Pooling layers consolidate the features learned by convolutional layers in CNNs. In order to limit the number of parameters and computations required in the network, it gradually reduces the spatial size of the representation. Each feature map in the pooling layer is treated independently. The parameters were chosen by trial and error throughout the CNN training stage, as shown in Table 3, within a range of recommended values derived from previous literature investigations. The best-performed configurations are highlighted in bold in Table 3. In the proposed method based on a trial-and-error approach, the best-performing CNN configuration parameters used were momentum = 0.8, learning rate = 0.001, learning rate schedule = piecewise, learning rate drop factor = 0.1, and L2 regularization = 0.0001.

ReLU is used to prevent neural networks from overusing their computational resources. Using ReLUs prevents the vanishing gradient problem when sigmoidal functions are employed. The convolution layer is followed by a batch normalization layer and two fully connected layers with 348 and 348 neurons, respectively. Each mini-batch is normalized using batch normalization, a method for training deep neural networks. It

concludes the learning process and drastically reduces the number of training epochs required to build neural networks. This research trained the model with fewer epochs per iteration because this configuration was utilized. A dropout layer with a 50 percent dropout ratio follows two fully connected layers to prevent overfitting. If there is no dropout layer, the first batch of training data disproportionately impacts the learning process compared to subsequent batches. In turn, this would prevent learning traits that exist only in later samples or batches. Finally, the classification layer classifies emotions. Figure 6 illustrates the proposed CNN architecture.

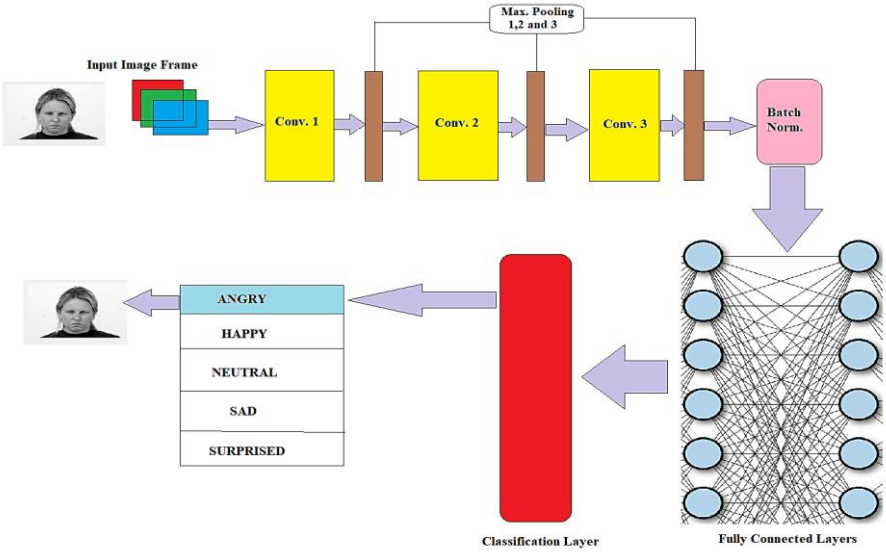

Figure 6. Proposed CNN Architectural Diagram

3. RESULTS AND DISCUSSION

In this study, we utilized a laptop equipped with the MATLAB integrated development environment (IDE) and deep learning and image processing toolboxes to perform all necessary calculations and computations. These tasks encompassed training and testing phases for our emotion recognition model. The hardware configurations employed for this research, as presented in Table 4, ensured the efficient execution of these complex processes while allowing for the optimization of the model's performance.

Table 4. Hardware Specifications

Computer	HP PAVILION 15-BC408TX
CPU	Intel Core i7-8750H (8th Gen)
RAM	8 GB DDR4 RAM
HDD	1TB
GPU	NVIDIA GeForce GTX 1050
Graphics Memory	4GB

3.1. Emotion Labelling

Accurate labeling of each emotion is crucial for precise classification and recognition in the video-based emotion recognition model. This study focuses on the following five emotions: anger, happy, neutral, sad, and surprise. We assign a unique label to each emotion, as illustrated in Table 5, to facilitate efficient classification. This systematic labeling approach aids the model in distinguishing and accurately identifying the various emotions, ultimately enhancing the overall performance of the emotion recognition system.

Table 5. Labeled Emotions for Video Database

Emotion	Files	Labels
Angry	936	1
Happy	936	2
Neutral	858	3
Sad	936	4
Surprised	936	5

3.2. Experimental Findings from Video Database

The performance of our emotion recognition system is primarily assessed based on its accuracy. We employ a feature selection algorithm to choose the most relevant features for training the network. The dataset is split into a 70-30 ratio, with 70% allocated for training and the remaining 30% designated for testing and validation purposes. The network undergoes testing over five epochs, each consisting of 104 iterations, resulting in approximately 520 iterations.

We also analyze the classification results to ensure a comprehensive evaluation using various metrics, including the confusion matrix, precision, recall, and F1 scores. Additionally, validation graphs are generated to visualize the model's performance throughout the training and Testing. This multi-faceted approach to evaluation enables us to effectively measure the system's performance, identify potential areas for improvement, and ensure its reliability in real-world applications.

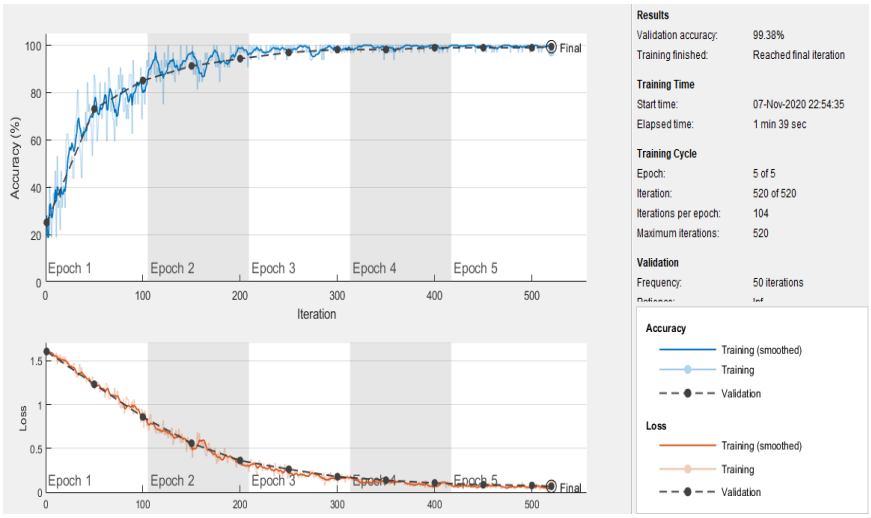

Figure 7. Training and Validation Accuracy Plot

Figure 7 plots the training and validation accuracy for the emotion recognition model. The results obtained from this dataset are highly promising, achieving a validation accuracy of 99.38%. The total processing time for the model was 1 minute and 39 seconds, indicating efficient performance. The successful outcomes can be attributed to the practical CNN configuration employed in the model and the adequate number of video frames per emotion. Furthermore, including dropout for the output from fully connected layers played a crucial role in preventing overfitting and enhancing overall performance. A critical analysis of additional performance metrics was carried out for four data subsets: training data, complete data, testing data, and validation data. The confusion matrix, recall, and precision values, as illustrated in Figures 8 and 9, provide a more comprehensive understanding of the model's performance across these data subsets. This thorough evaluation allows for a better interpretation of the results and identifies potential areas for further improvement.

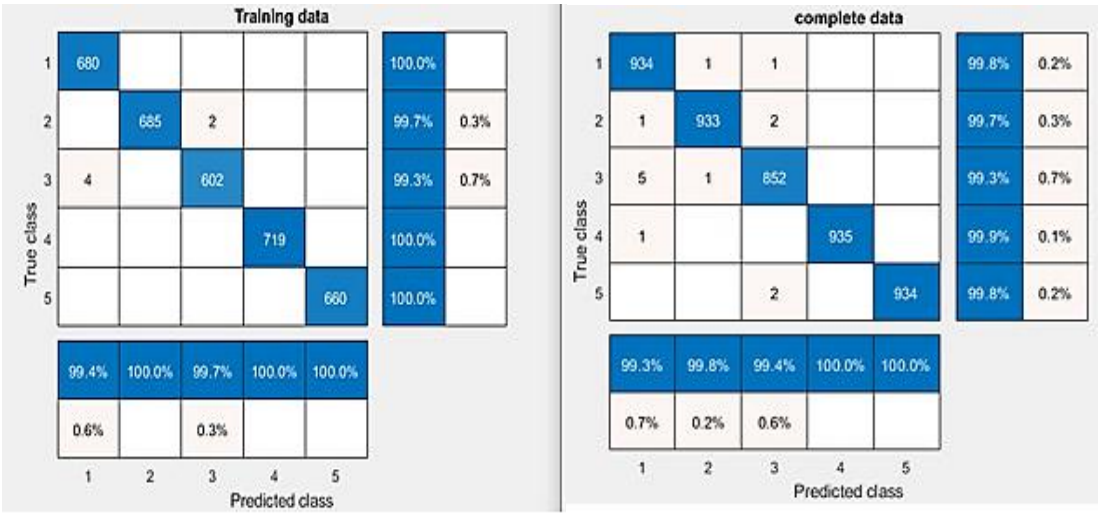

Figure 8. Performance Parameters for Training and Complete Data

Figure 9. Performance Parameters for Testing and Validation Data

Table 6 presents the recall and precision values for the five emotions across the four data subsets, offering a more precise visualization and interpretation of the model's performance. The exceptionally high recall and precision values for the testing data demonstrate the effectiveness of the proposed model in accurately predicting the results. This strong performance highlights the model's potential for successful real-world applications in emotion recognition tasks.

Table 6. Recall and Precision for Four Sets of Data

Emotion	Recall for Training Data	Precision for Training data	Recall for Complete Data	Precision for Complete data	Recall for Testing Data	Precision for Testing data	Recall for Validation Data	Precision for data
Angry	1	0.994	0.998	0.993	0.992	0.992	0.993	0.985
Happy	0.997	1	0.997	0.998	1	0.983	0.992	1
Neutral	0.993	0.997	0.993	0.994	0.983	0.992	1	0.985
Sad	1	1	0.999	1	1	1	0.991	1
Surprise	1	1	0.998	1	0.993	1	0.993	1

The F1 score provides a balance between precision and recall. The F1-Score for the above four data sets is calculated using Eq. (1) and presented in Table 7.

$$\text{F1 score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision}+\text{Recall}} \quad (1)$$

Table 7. F1-Score for Four Sets of Data

Emotion	F1-Score for Training data	F1-Score for Complete data	F1-Score for Testing data	F1-Score for Validation data
Angry	0.994	0.996	0.992	0.989
Happy	0.998	0.998	0.992	0.996
Neutral	0.995	0.997	0.988	0.992
Sad	1	1	1	0.991
Surprise	1	1	0.996	0.993

3.2.1. Testing Results from WSEFEP Database

Image samples from the Warsaw Set of Emotional Facial Expression Pictures (WSEFEP) dataset were employed to evaluate the generalization capabilities of the proposed model. Figure 10 displays the confusion matrix for the testing data, providing insights into the model's performance in predicting emotions for a different dataset. We can effectively assess the model's ability to adapt and accurately predict results beyond the original training data by examining twenty-five samples for each of the five emotions. This analysis is crucial in determining the model's practicality and reliability for real-world applications.

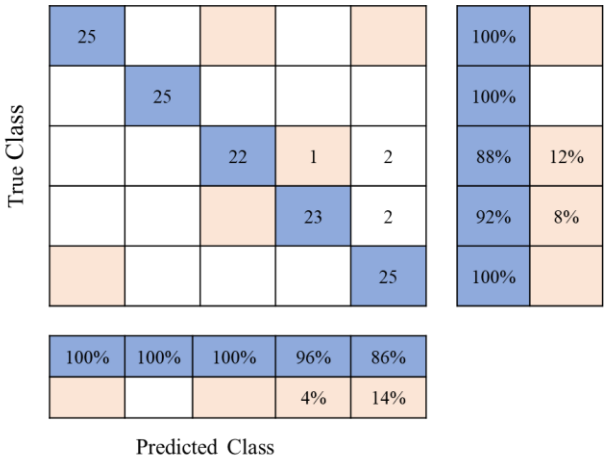

Figure 10. Confusion matrix for testing data

The model achieved an accuracy of 96% on the testing data from the WSEFEP database. Table 8 displays the recall and precision values for the five emotions derived from the test results to provide a more precise visualization and explanation. These high recall and precision values for the testing data indicate the effectiveness of the proposed model in predicting the outcomes, showcasing its robust performance when applied to a different dataset. This strong generalization ability demonstrates the model's potential for successful deployment in various real-world applications involving emotion recognition.

Table 8. Recall and Precision for Testing Data

Emotion	Recall for Testing Data	Precision for Testing data
Angry	1	1
Happy	1	1
Neutral	0.88	1
Sad	0.92	0.96
Surprise	1	0.86

3.2.2. Testing Results from the Extended Cohn-Kanade Database (CK+)

To evaluate the generalization capabilities of the proposed model, we used image samples extracted from video samples in the Extended Cohn-Kanade Database (CK+). This analysis investigated the model's ability to classify emotions in previously unseen data. Figure 11 presents the confusion matrix for the testing data, offering insights into the model's performance. In order to ensure a reliable evaluation, twenty-five samples were taken for each of the five emotions to determine the prediction results. This assessment is crucial for understanding the model's adaptability and effectiveness when applied to novel datasets and real-world applications.

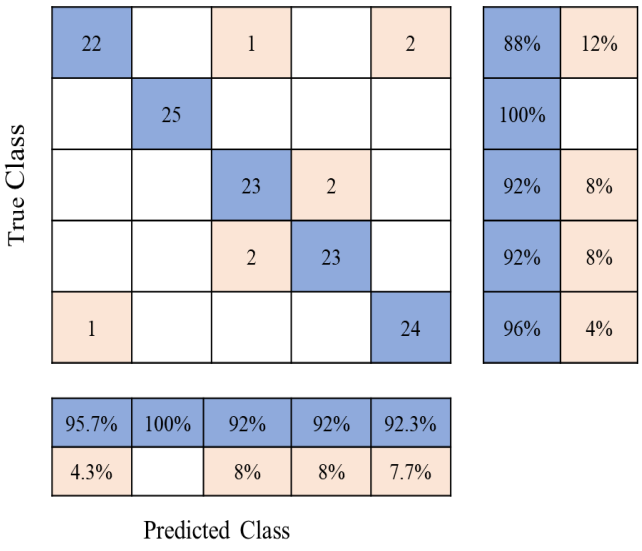

Figure 11. Confusion matrix for testing data

The Extended Cohn-Kanade Database (CK+) results revealed a testing accuracy of 93.6% for the proposed model. Table 9 presents the precision and recall values for the five emotions, based on the test results, for a more comprehensive understanding of the model's performance. The high precision and recall values achieved for the testing data suggest that the proposed model effectively predicted the testing outcomes. Consequently, we can conclude that the model demonstrated promising performance in emotion recognition when applied to the CK+ database, showcasing its potential for real-world applications.

Table 9. Recall and Precision for Testing Data

Emotion	Recall for Testing Data	Precision for Testing data
Angry	0.88	0.957
Happy	1	1
Neutral	0.92	0.92
Sad	0.92	0.92
Surprise	0.96	0.923

3.3. Discussion and Benchmarking

In this research, we have critically addressed the challenges of internal covariate shift and overfitting in our neural network architecture. We incorporated Batch Normalization after the convolutional layer to mitigate the internal covariate shift issue. Normalizing each layer's inputs ensured faster convergence during training and improved the model's accuracy. We included a Dropout layer after the fully connected layers to tackle overfitting. This regularization technique contributes to a more robust model by decreasing the interdependence of neurons. Randomly dropping out some neurons during training ensured that our model generalized effectively to new data.

Furthermore, we employed a trial-and-error approach to optimize hyperparameters, such as momentum, learning rate, and L2 regularization, within a range of recommended values. Identifying the optimal set of hyperparameters for our problem enhanced the accuracy of our model and reduced the number of training epochs needed to create neural networks. By incorporating these techniques into our architecture, we addressed the challenges of internal covariate shift and overfitting, resulting in a more robust and efficient neural network model.

Table 10. Direct Benchmarking

Method Employed	Benchmarked research accuracy (%)	Our research accuracy (%)
Convolutional Neural Networks and Local Directional Strength Pattern (LDSP) [12]	95.42	
Fine-Tuned VGG-16 [21]	69.40	
Convolutional Neural Networks [22]	85	99.38
Lightweight convolutional neural network (CNN) [17]	67	
Single shot detector CNN, ResNet, VGG, and Inception V3 [18]	97.42 (SSD), 90.14 (ResNet), 87 (VGG) and 81 (Inception V3)	

Compared to previous related works, the results obtained from the video-based emotion dataset (ADFES-BIV) exhibit greater accuracy and promise. Our research directly compared the recognition accuracy to other related studies, as shown in Table 10. Our proposed model achieved recognition accuracy of 99.38%, significantly surpassing the benchmarked studies. The superior performance of the proposed model can be attributed to the execution steps, which include an optimal CNN configuration, enhanced digital image processing methods, sufficient video frame conversions, and an optimal training environment using the trial-and-error method.

The findings of this study highlight the importance of continued exploration in the field of visual emotion recognition. Numerous approaches for enhancing emotion recognition models have emerged in recent years. Our investigation employed the CNN method, incorporating five distinct emotions: anger, happy, sad, neutral, and surprise. Generally, as more emotions are included, the model's performance tends to deteriorate. However, the proposed methodology and execution steps in our research led to superior outcomes, showcasing the effectiveness of our approach in maintaining high performance even when addressing multiple emotions.

4. CONCLUSION

This study focused on video-based emotion recognition, examining five distinct emotional states: anger, happiness, neutrality, sadness, and surprise. For this purpose, we analyzed the Amsterdam Dynamic Facial Expression Set-Bath Intensity Variations (ADFES-BIV) video database. The primary objective was identifying a person's emotional state using deep learning techniques, specifically convolutional neural

networks (CNNs). After extracting frames from video samples, we performed image pre-processing and face detection and fed the output into the CNN for feature extraction and classification. We thoroughly explained the CNN model's setup and operation. The results from the ADFES-BIV video database were highly promising, with a validation accuracy of 99.38%. Moreover, we illustrated and discussed various performance metrics, such as precision, recall, F1 scores, and the confusion matrix. To test the generalization abilities of the proposed model, we used data from the Warsaw Set of Emotional Facial Expression Pictures (WSEFEP) and the Extended Cohn-Kanade Database (CK+). Our research findings outperformed those of other recent related works when compared. The potential for future enhancements is an intriguing aspect of this study. Various multimodal deep learning algorithms can be combined with multiple architectures to overcome representational barriers. The emotional intensity scale can be expanded beyond merely perceiving emotions, potentially aiding in predicting the observed emotion's intensity. Furthermore, future work may incorporate data from various sources, such as multiple datasets and speech, visual, and textual inputs, to develop more comprehensive models. Future models can also be tested in real-world scenarios, enhancing their practical applicability.

ACKNOWLEDGMENTS

The authors would like to sincerely thank the Kulliyyah of Engineering sincerely, International Islamic University Malaysia, for providing the KOE Postgraduate Tuition Fee Waiver Scheme. They would also like to thank Universitas Negeri Yogyakarta for their generous funding and the provision of facilities for this research work.

REFERENCES

[1] Olszanowski, M., Pochwatko, G., Kuklinski, K., Scibor-Rylski, M., Lewinski, P., & Ohme, R. K. (2015). Warsaw set of emotional facial expression pictures: a validation study of facial display photographs. Frontiers in Psychology, 5, 1516.
[2] Ashraf, A., Gunawan, T. S., Riza, B. S., Haryanto, E. V., & Janin, Z. (2020). On the review of image and video-based depression detection using machine learning. Indonesian Journal of Electrical Engineering and Computer Science (IJEECS), 19(3), 1677-1684.
[3] Gunawan, T. S., Ashraf, A., Riza, B. S., Haryanto, E. V., Rosnelly, R., Kartiwi, M., & Janin, Z. (2020). Development of video-based emotion recognition using deep learning with Google Colab. TELKOMNIKA, 18(5), 2463-2471.
[4] Nezhad, Z. B., & Deihimi, M. A. (2020). Sarcasm detection in Persian. Journal of Information and Communication Technology, 20(1), 1-20.
[5] Basavaiah, J., & Patil, C. M. (2020). Human activity detection and action recognition in videos using convolutional neural networks. Journal of Information and Communication Technology, 19(2), 157-183.
[6] Najah, G. M. S. (2017). Emotion estimation from facial images, Atilim University, 10.13140/RG.2.2.25113.62565.
[7] Sönmez, E. B. (2018). An automatic multilevel facial expression recognition system. Süleyman Demirel Üniversitesi Fen Bilimleri Enstitüsü Dergisi, 22(1), 160-165.
[8] Abdulsalam, W. H., Alhamdani, R. S., & Abdullah, M. N. (2019). Facial emotion recognition from videos using deep convolutional neural networks. International Journal of Machine Learning and Computing, 9(1), 14-19.
[9] Wingenbach, T. S., Ashwin, C., & Brosnan, M. (2016). Validation of the Amsterdam Dynamic Facial Expression Set–Bath Intensity Variations (ADFES-BIV): A set of videos expressing low, intermediate, and high intensity emotions. PloS one, 11(1), e0147112.
[10] Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems, 25.
[11] Noroozi, F., Marjanovic, M., Njegus, A., Escalera, S., & Anbarjafari, G. (2017). Audio-visual emotion recognition in video clips. IEEE Transactions on Affective Computing, 10(1), 60-75.
[12] Uddin, M. Z., Khaksar, W., & Torresen, J. (2017). Facial expression recognition using salient features and convolutional neural network. IEEE Access, 5, 26146-26161.
[13] Yin, X., & Liu, X. (2017). Multi-task convolutional neural network for pose-invariant face recognition. IEEE Transactions on Image Processing, 27(2), 964-975.
[14] Xie, S., & Hu, H. (2018). Facial expression recognition using hierarchical features with deep comprehensive multipatches aggregation convolutional neural networks. IEEE Transactions on Multimedia, 21(1), 211-220.
[15] Mehendale, N. (2020). Facial emotion recognition using convolutional neural networks (FERC). SN Applied Sciences, 2(3), 1-8.
[16] Pranav, E., Kamal, S., Chandran, C. S., & Supriya, M. H. (2020, March). Facial emotion recognition using deep convolutional neural network. In 2020 6th International conference on advanced computing and communication Systems (ICACCS) (pp. 317-320). IEEE.
[17] Zhou, N., Liang, R., & Shi, W. (2021). A Lightweight Convolutional Neural Network for Real-Time Facial Expression Detection. IEEE Access, 9, 5573-5584. https://doi.org/10.1109/ACCESS.2020.3046715.
[18] Melinte, D. O., & Vladareanu, L. (2020). Facial Expressions Recognition for Human-Robot Interaction Using Deep Convolutional Neural Networks with Rectified Adam Optimizer. Sensors, 20(8), 2393. https://doi.org/10.3390/s20082393.

[19] Fei, Z., Yang, E., Li, D. D.-U., Butler, S., Ijomah, W., Li, X., & Zhou, H. (2020). Deep convolution network based emotion analysis towards mental health care. Neurocomputing, 388, 212-227. https://doi.org/10.1016/j.neucom.2020.01.034.
[20] Akhand MAH, Roy S, Siddique N, Kamal MAS, Shimamura T. Facial Emotion Recognition Using Transfer Learning in the Deep CNN. Electronics. 2021; 10(9):1036. https://doi.org/10.3390/electronics10091036.
[21] Kusuma, G. P., & Lim, A. P. (2020). Emotion recognition on FER-2013 face images using fine-tuned VGG-16. Advances in Science, Technology and Engineering Systems Journal, 5(6), 315-322.
[22] Badrulhisham, N. A. S., & Abu Mangshor, N. N. (2021). Developing a mobile-based application for emotion recognition using facial expression in real-time. Journal of Physics: Conference Series, 1962(1), 012040. https://doi.org/10.1088/1742-6596/1962/1/012040.
[23] Cai, Y., Zheng, W., Zhang, T., Li, Q., Cui, Z., & Ye, J. (2016). Video based emotion recognition using CNN and BRNN. Chinese conference on pattern recognition,
[24] Hum, Y. C., Lai, K. W., & Mohamad Salim, M. I. (2014). Multiobjectives bihistogram equalization for image contrast enhancement. Complexity, 20(2), 22-36.
[25] Naik, S. K., & Murthy, C. (2003). Hue-preserving color image enhancement without gamut problem. IEEE Transactions on image processing, 12(12), 1591-1598.
[26] Al-Sumaidaee, S., Dlay, S., Woo, W., & Chambers, J. (2015). Facial expression recognition using local Gabor gradient code-horizontal diagonal descriptor.
[27] Boubenna, H., & Lee, D. (2018). Image-based emotion recognition using evolutionary algorithms. Biologically inspired cognitive architectures, 24, 70-76.
[28] Zhou, D., Shen, X., & Dong, W. (2012). Image zooming using directional cubic convolution interpolation. IET image processing, 6(6), 627-634.
[29] Valueva, M. V., Nagornov, N., Lyakhov, P. A., Valuev, G. V., & Chervyakov, N. I. (2020). Application of the residue number system to reduce hardware costs of the convolutional neural network implementation. Mathematics and Computers in Simulation, 177, 232-243.