OPEN ACCESS

Kashf Journal of Multidisciplinary Research
Vol: 02 - Issue 4 (2025)
P-ISSN: 3007-1992 E-ISSN: 3007-200X
https://kjmr.com.pk

REAL-TIME GENDER AND EMOTION RECOGNITION USING OPTIMIZED DUAL-PATH CNN ARCHITECTURE FOR ENHANCED HUMAN-COMPUTER INTERACTION

Muhammad Qasim Khan
Department of Computer Science, Iqra National University Peshawar, Khyber Pakhtunkhwa, Pakistan.
Fazal Malik*
Department of Computer Science, Iqra National University Peshawar, Khyber Pakhtunkhwa, Pakistan.
Afsheen Khalid
Center for Excellence in IT, Institute of Management Sciences, Peshawar, Khyber Pakhtunkhwa, Pakistan.
Dilawar Khan
Computer Science & IT Department, University of Engineering and Technology, Peshawar, Khyber Pakhtunkhwa, Pakistan.
Ashraf Ullah
Institute of Computer Science and IT, University of Science and Technology Bannu, Khyber Pakhtunkhwa, Pakistan.
Rahmat Hussain
Institute of Computer Science and IT, University of Science and Technology Bannu, Khyber Pakhtunkhwa, Pakistan.
Muhammad Javed
Institute of Computer Science and IT, University of Science and Technology Bannu, Khyber Pakhtunkhwa, Pakistan.

****Corresponding author: Fazal Malik*** (fazal.malik@inu.edu.pk)

Article Info

Abstract

The analysis of facial expressions plays a fundamental role throughout human-computer interaction together with affective computing along with behavioral sciences thereby enabling security surveillance devices as well as healthcare diagnostic systems and smart interactive programs. A research project uses optimized deep learning methods to solve ongoing real-time gender and emotion recognition difficulties stemming from facial expression diversity together with illumination fluctuations and dataset preference. Our proposed framework includes two essential CNN components: (1) a simplified two-layer CNN system for male/female classification and (2) a hierarchical four-layer CNN model for identifying happy and sad expressions together with angry face and other emotions depicted in fear, disgust, surprise and neutral. The proposed framework implements three main technical advancements that include Viola-Jones face detection and optimized convolutional layers with ReLU activation and batch normalization for efficient feature extraction and strategic max-pooling for dimensionality reduction. Performance assessments across benchmark collections including FER-2013, CK+ and KDEF and IMDB indicated 94% success in gender detection and 93% accomplishment in emotional identification together with real-time implementation capabilities. In controlled environments the system maintains strong performance yet displays confusion between emotionally related categories of disgust and fear. The research incorporates three main elements to advance the work: (1) a CNN architecture design optimized for precision and speed balance, (2) combined batch normalization and dropout regularization methods to improve feature extraction and (3) thorough multiple dataset testing to verify generalized performance. The latest networks create fundamental frameworks for future applications in intelligent surveillance and affective robotics as well as behavioral analytics yet more research needs attention mechanisms that enhance emotion recognition precision during fine-grained emotion differentiation.

Keywords:
Gender Recognition, Face Detection, Feature Extraction, Facial Expressions, Deep Learning, CNNs, Viola-Jones, HCI.

1. Introduction

The majority of human communication relies on nonverbal cues, accounting for approximately 55% to 93% of total interaction. Emotional analysis performed on the face stands as the primary key element that drives 55% of recognition capability [1]. The evaluation process of emotions uses multiple elements that include behavior along with mental state, personality and physical intention [2]. Security footage and expression recognition together with home automation systems, benefit from facial expression analysis. The technology also supports PC gaming and clinical settings for diagnosing depressive disorders and care needs. Furthermore, it helps detect anxiety levels, lies, and emotional states during psychoanalytic sessions, and paralinguistic cues. It monitors operator fatigue and robotics also utilizes this technology. Studying facial expressions helps people become better at nonverbal communication and it increases their efficiency at oral communication according to [3].

Machines find the process of recognizing emotions to be difficult. People of all genders together with those belonging to every nationality and cultural background and racial group can successfully understand emotions. Facial emotion identification performs multiple emotion categories classification consisting of neutral, happy, angry, sad, surprise, disgust and fear emotions [4]. The correct interpretation of emotions faces barriers from the inconsistencies in gender identity along with age and racial as well as ethnic backgrounds and variations in image or video clarity. Machines require an automated system that performs similar to human emotion recognition capabilities. Facial expression recognition represents a central research topic that scientists study in present times. The identification of facial representations relies on computer vision paired with image processing and machine learning algorithms according to research studies [5].

Different industries use machines on a growing scale which demands the development of better natural human-machine communication techniques. A system of effective communication needs machines particularly robots and computers to interpret human decisions. The autonomous features of robots lead to improvements in their interaction abilities. Machine perception uses algorithms to imitate human senses thus allowing interaction with environments [6, 7]. The combination of sensors with cameras allows environment data gathering that machines process using effective algorithms to achieve enhanced perception. The field benefits notably from deep learning methods which are described in studies [8, 9].

The identification of human emotions stands as a fundamental necessity for affective computing to provide robots with abilities to better support their users [10]. Robots used in hospital wards and aged care settings need environmental awareness because facial expressions help robots understand the internal condition of patients [11]. By processing sequences of facial images with deep learning techniques computers acquire the ability to identify moods thus improving both human-machine dialogue along with human-machine relationship. Robotics systems achieve better natural intelligence-based interactions through these technological approaches that help them learn autonomously [12, 13]. The transformation of video material enables better indexation and retrieval operations and produces summaries and detects actions and performs facial examinations. Each video consists of consecutive frames that contain movement features in addition to both structural elements and color patterns. Facial features work as identifiers for people and emotional states to help develop smart surveillance along with virtual reality technology and medical diagnostics as well as robotics and elderly care applications. The expressions from our faces indicate sadness and happiness together with anger facilitating behavioral science operations and human-machine interaction. The human face displays six fundamental expressions of anger, disgust, fear, happiness, sadness together with surprise [14].

The Viola-Jones object/face detection method [15] uses algorithms to locate faces together with important sub-parts including the nose and lips. Real-time accuracy for facial recognition systems gets

improved through the implementation of this system. The process of pre-processing video frames enhances input quality by removing disturbances and blurring artifacts [16]. The cascade object detection model finds face occurrences which the multi-object tracker KLT follows in sequences before the deep CNN model analyzes them to identify gender along with facial expressions. The success of human-robot interaction depends on automated systems that identify gender in addition to recognizing human emotions. The high degree of sample variation makes ML interpretation difficult which generates models with millions of parameters. The human ability to identify seven mood categories in facial expressions reaches a maximum of 65% accuracy. Part of this task complexity becomes apparent through FER (2013) dataset classifications which use Figure 1 to group emotions into "Angry, Disgust, Fear, Happy, Sad, Surprise, and Neutral" categories and Figure 2 separates expressions into "Man" and "Woman" classifications.

Figure 1. Example Images From The FER 2013 Dataset For Face Emotion Recognition

Figure 2. Example Images From The IDMB Dataset For Gender Classification

The analysis of human facial expressions serves as an essential component for developing human-computer devices as well as affecting computing solutions and healthcare systems and surveillance applications. The challenge of identifying gender and emotion in real-time stems from multiple sources which include variations in facial expressions as well as changes in lighting and the existence of obstacles as well as imbalanced datasets. Real-time deployment challenges deep learning models because manual feature extraction restricts adaptability yet conventional methods demand manual feature extraction to achieve accurate results. Current systems experience similar performance problems when dealing with emotions that are comparable (fear vs disgust) and variations between different user demographics. A select CNN architecture with weight reduction features must be developed for achieving high accuracy within acceptable energy requirements for deployment applications.

The proposed study develops an optimized deep learning approach for real-time systems which recognize facial expressions together with gender identification. A dual-path CNN performs gender detection alongside seven facial expression recognition through batch normalization and ReLU activation and max-pooling techniques. The model achieves real-time capabilities by enhancing its ability to handle various datasets (FER-2013, CK+, KDEF, IMDB) while preventing errors in the classification of similar expressions.

The proposed workflow contains four distinct steps starting with Data Preprocessing where Viola-Jones detects faces while the images undergo resizing normalization and augmentation procedures. 2) Dual-Path CNN: A two-layer CNN for gender classification (male/female) with batch normalization and max-pooling; a four-layer CNN for fine-grained emotion recognition (7 expressions). The optimization process utilizes three stages for training and testing with dropout and batch normalization on FER-2013, CK+, KDEF and IMDB datasets (80% training, 20% testing). The system performs real-time processing by selecting video frames one by one for direct implementation. The proposed system generates two outputs which include Gender Classification (Male/Female) together with Facial Expression Recognition (7 expressions).

A research proposal creates an optimized CNN framework which combines computational effectiveness with real-time ability for affective computing as well as healthcare diagnostics, intelligent surveillance and industrial safety use cases.

The subsequent sections detailed existing research methods in Literature Review (Section 2) before describing methodology framework in Section 3 followed by results and discussion in Section 4 and recommendations for future work in Conclusion (Section 5).

2. Literature Review

Scientists within the computer vision community have united their research efforts toward facial analysis through collaboration between neurology and psychology with computer science and cognitive science [17]. The extensive research into Facial Emotion Recognition (FER) takes place because of its broad set of applications. The Facial Action Coding System (FACS) developed in 1978 depends on Action Units (AUs) to detect six universal emotions including fear and sadness together with surprise and happiness along with contempt and anger [18]. The performance of FER technology needs to be highly stable for real-time systems operating in low-resolution environments including smart meetings and surveillance applications. A variety of demographic groups is represented in the benchmark dataset created for this purpose. Multiple studies prove emotions enhance decision-making therefore underlining the need for real-time FER applications [19, 20].

The combination of deep learning techniques and Ekman's Facial Action Coding System (EFACS) allows real-time gender and emotion recognition to map facial emotions onto facial muscle movements [21]. Performance enhancement results from applying preprocessing techniques which include normalization for noise reduction along with histogram equalization and face extraction for noise reduction along with image alignment. Dimensionality reduction enables better performance by selecting the Region of Interest (ROI). The deep learning system identifies six basic emotions through its operation as deep learning models perform the classification. Graphical face generation gains strength by implementing Ekman's facial units [22]. Reducing the number of required features is made possible through feature extraction methods. The real-time classification system based on geometry shapes depends on mathematical analysis of eyes along with relationship patterns across forehead and nose components with eyebrows and lips and chin regions. Research that employs Active Appearance Models (AAM) constitutes many studies in geometric-based facial feature extraction methods for tracking facial points. Version two of the AAM reduction model increases data convergence rates together with multi-layer perception methods tackling high dimensionality problems [23].

The elastic graph matching procedure starts the detection process for face points before the Kanade-Lucas-Tomasi tracker takes over to manage tracking operations. Recognition performances increase through the combination of triangle forms alongside point and line shapes that employ multi-class AdaBoost [24]. Face normalization through an ASM-based system happens by using ASM landmarks

followed by appearance-based classification. The algorithm employs KL weights estimation as a discriminative weight estimation method together with appearance-based features extraction through 2D-DCT. Evaluation of performance takes place through the JAFFE and CK databases testing regime. The implementation of deep learning algorithms for real-time gender and emotion detection improves facial analysis precision in real-time use. This technique depends on statistical descriptors that include LBP, SIFT and Gabor functions and HOG. The high-dimensional structure emerges from the integration between face alignment features and additional features. A Deep Sparse Autoencoder (DSAE) uses unsupervised learning for meaningful information retrieval and produces output through Softmax classification. The DSAE achieves its best results through HOG when detecting 7-class and 8-class expressions [25, 26].

An integrated system uses grayscale face pictures and depth measurements and Local Binary Patterns (LBP) analysis to identify six basic emotions. The recognition performance is improved by depth images while Local Binary Patterns work to extract appearance-based features. The improved random forest classifier analyzes emotions through geometric and textural feature unification [27]. Support Vector Machines (SVM) enable promising results for the identification of spontaneous and posed smiles through local features and geometry analysis [28]. The Part-based Hierarchical Bidirectional Recurrent Neural Networks (PHRNN) model divides facial landmarks into four components to extract sequential features using subnets and temporal features and geometric features through its processing mechanisms [29]. The methodology combines angle-based triangular regions generated from facial landmarks where the classification steps involve Conditional Random Field (CRF) together with K-Nearest Neighbors (KNN) for referencing purposes. A set of trained feature vectors serves dynamic emotion detection functions [30]. The Euclidean distance method in geometric feature extraction enables identification of anger and happiness and sadness and surprise. The evaluation of 3D video frames using this method achieved a strong real-time performance through a discrimination ratio of 85%. The appearance-based methods in Facial Emotion Recognition overcome landmark localization limitations that affect geometric approaches because they extract emotional features from the visual face characteristics directly. Low-resolution images benefit from LBP that delivers superior emotional detection compared to Gabor wavelets. Linear Discriminant Analysis (LDA) enhances LBP while using Gabor filter Tsallis entropy to analyze JAFFE expressions with improved performance [31, 32].

The analysis of image resolution impacts FER performance through integrated features based on geometry and appearance extraction methods. Research performed on the Facial Expression and Body Gesture (FABO) and Cohn-Kanade (CK) datasets at multiple resolutions verifies resolution plays a critical role in correct emotion identification work [33]. Local Binary Pattern (LBP) faces limitations due to detail loss and thus Compound Local Binary Pattern (CLBP) integrates sign and magnitude information while using Support Vector Machine (SVM) classification for testing on CK and Japanese Female Facial Expression (JAFFE) datasets. Research findings validate that the Compound Local Binary Pattern (CLBP) delivers better results than conventional methods [34]. A research paper demonstrates how important facial characteristics can be found within salient sections for better emotion detection capabilities. A fast facial landmark detection method achieves effective one-to-one classification on both the CK+ and JAFFE datasets in relation to the facial patches [35].

The applications of advanced machine learning techniques include COVID-19 pneumonia diagnostic systems as well as sentiment analysis solutions and accident prediction systems and cybersecurity systems. The research combines data augmentation techniques with GitHub X-ray datasets for COVID-19 diagnosis through optimized Random Forest and AdaBoost and XGBoost together with Convolutional Neural Networks (CNNs) [36-39]. Three methods including XGBoost and AdaBoost and Artificial Neural Networks (ANNs) enhance the Google Play Store review classification process [40].

Random Forest produces superior results than AdaBoost in dark data-driven crash identification tasks [41]. The combination of XGBoost and AdaBoost within cybersecurity practices improves URL detection by reducing instances of incorrect positive and negative results. An enhanced threat detection system involves multiple phases which increases cybersecurity threat detection capabilities [42-44].

The proposed work in an approach uses texture features to extract information for both posed and spontaneous emotion detection. First the Sobel filter generates gradient components which the Elongated Quinary Pattern descriptor uses for quantizing local gradients. The Multi-Classifier System boosts recognition accuracy by surpassing other established methods [45]. A research study utilizes texture descriptors for FER while first detecting facial components then scales the detected features into blocks for subsequent extraction. The research demonstrated that using a multiclass SVM classifier produces better results than established techniques while testing CK, KDEF and FEED datasets [46]. Investigating human facial expressions serves as a fundamental requirement for identifying emotions while measuring behaviors because these capabilities drive the development of human-machine interfaces within applications such as security authentication and surveillance as well as the improvement of customer satisfaction. The research adopts Ensemble CNN for performing real-time emotion and gender recognition [47]. Socket when it comes to enhancing security since these systems are commonly employed in airports and banks they encounter various computer vision obstacles. Research has analyzed the capabilities of CNNs in comparison to QCNNs because quantum computing solves problems that CNNs have with extensive datasets [48].

A HRI advancement resulted in an animatronic robot face with eighteen degrees of movement which recognizes human emotions through a special custom CNN trained with data from FER-2013, CK+ and KDEF datasets [49]. Researchers have proposed several advanced feature selection methods along with dimensionality reduction strategies for intra-class distances and discriminative feature extraction and Double LBP for robust feature extraction and weighted LBP projection to reduce misclassification errors that demonstrate effectiveness in real-time applications [50-54].

The three vital elements for successful FER include pre-processing along with visual feature extraction followed by classification. Standardized datasets such as JAFFE consisting of 213 images from seven emotions should be used instead of handcrafted features because they provide more efficient results [55]. The JAFFE dataset includes illustrations which can be seen in Figure 3.

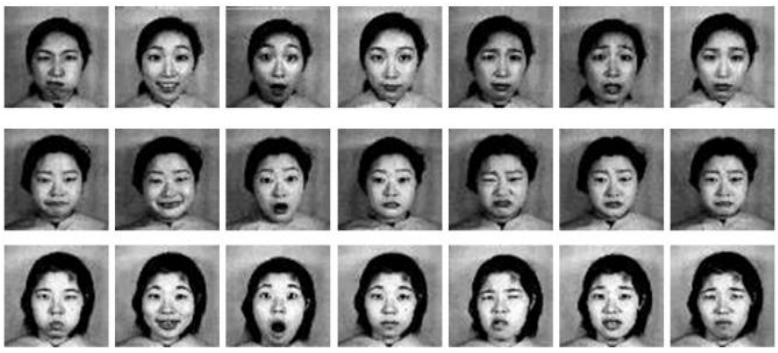


Figure 3. Example Images From The JAFFE Dataset

The MMI Face Emotion Database includes 273 video frames of high resolution images for emotion analysis and recognition and classification purposes. This database serves as a common research instrument because it captures subjects in their natural expressions while making spontaneous facial movements[56]. The presented example images can be found in Figure 4.

Figure 4. Example Images From The MMI Face Emotion Database

3. Research Methodology

Figure 5 illustrates the proposed integrated deep learning approaches used in image classification which include Gender Classification and Facial Expression Recognition as per the proposed methodology. Two

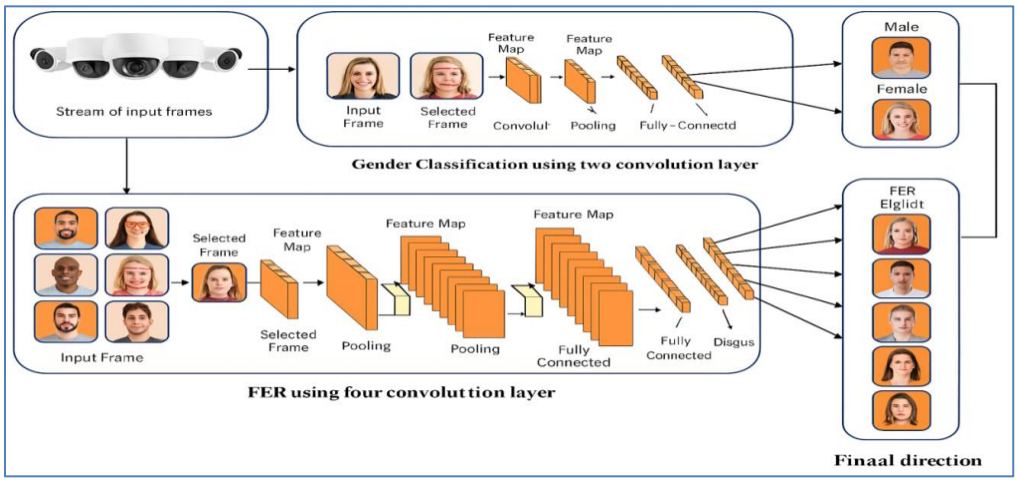


Figure 5. Propose Model Of CNN

specialized CNN models are applied because they optimize performance accuracy while maintaining low computational complexity for real-time implementation.

3.1. Deep Learning-Based Gender Classification And Facial Expression Recognition

This research investigates deep learning methodology specifically aimed at gender detection combined with facial emotion recognition through CNNs and evaluates system design choices for implementation in real-time applications.

3.1.1. Gender Classification Using Two Convolutional Layers

The model performs gender classification by identifying images either as male or female. Preprocessing of the input image starts with resizing then includes normalization steps. A series of low-level feature detection occurs in the first convolutional layer before activation functions (e.g., ReLU) activate the output and spatial reduction through pooling occurs. After the first convolutional layer the neural network extracts more abstract facial patterns while enhancing the extracted features between the layers. After normalizing the output through a fully connected layer it continues to one or more dense layers which integrate features. The softmax output unit determines the image classification as male or female.

Visiting labeled datasets allow training the model using cross-entropy loss optimized through backpropagation alongside Adam.

3.1.2. Facial Expression Recognition (FER) Using Four Convolutional Layers

The model identifies four basic expressions among happy, sad and angry among others from faces. The processing of the input image includes automatic face detection followed by resizing operations and normalization techniques. Fundamental facial elements are extracted by the first convolutional layer which proceeds to activation algorithms before performing pooling. The second layer detects advanced features from eyes and mouths before the third and fourth layers enhance these features to ensure dependable emotion identification. During the fully connected layer the system combines collected features before the softmax output layer generates classifications from various emotional categories. The model requires labeled facial expression datasets for training which utilizes a loss function together with backpropagation for optimization.

3.1.3. Comparison Of Gender Classification And FER Models

The model architecture of Gender classification operates with two convolutional layers for binary classification tasks but FER needs more complex features which are extracted through four convolutional layers. General facial structure identification forms part of the gender classification model while FER analyzes precise features including mouth curvature alongside eye shape detection. The gender classification model features two output classes (male and female) yet FER manages several distinct categories including happy, sad and angry expressions among others.

Real-time detection of gender and emotion in individual or multiple subjects becomes possible through an optimized mathematical system. The system adopts optimized CNN models which deliver performance improvement through batch normalization and ReLU activation practices. The system reduces parameters to lower its system complexity which addresses hardware constraints encountered by small CNN architectures.

The framework demonstrates an efficient method to link parameters which leads to better accuracy results as shown in Figure 5. Significant progress in facial expression analysis exists as confirmed by recent literature while every examined method brings individual value to the research field. Both gender classification and emotion detection show superior performance through CNN because this system directly processes 2D images compared to traditional deep learning architectures. Both FER-2013 and IMDB datasets undergo training and testing roles that split data into an 80:20 proportion for assessment reliability purposes.

3.2. Real-Time Gender And Emotion Recognition Workflow Using A CNN Model

The real-time system for gender and emotion identification through CNN-based AI algorithms is displayed in Figure 6 as a workflow diagram. The system starts its operation by accepting an image from either a video stream or directly as an image. The system extracts frames sequentially from video data as part of its further processing stage. Face detection occurs for each frame through implementation of Viola-Jones algorithm as well as alternative face detection methods. The system examines new frames until it identifies a face when no detection occurs on the current frame.

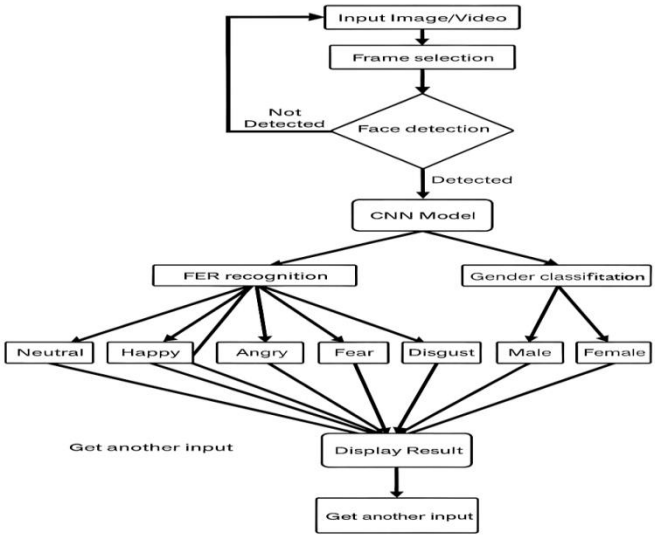


Figure 6. Workflow Diagram Of Proposed Work

After face detection the system crops regions with detected faces before submitting them for features extraction within a trained CNN model. The CNN data processing model carries out sequential operations which include FER and Gender Classification. The FER task requires the model to identify one of six facial expressions from the detected face including Neutral, Happy, Sad, Angry, Fear or Disgust. The gender classification performs an assignment of the detected face into male or female categories at the same time.

Both predicted emotion and target gender appear in the display as end results. The system retrieves the following successive frame from the continuous video stream for real-time processing after completing recognition. The structured approach enhances efficiency throughout face detection and feature extraction and classification tasks which makes it appropriate for human-computer interaction systems and security surveillance needs and behavioral analysis purposes

3.2.1. Input Image/Video

Input frames and detects faces through the Viola-Jones algorithm before the system uses CNN models to classify gender and emotional expressions. The system performs real-time evaluations which allow security functions and human-computer interaction and behavioral monitoring applications to operate effectively.

3.2.2. Frame Selection (For Video Input)

The video extraction process operates automatically to obtain frames continuously while running the same detection along with classification operations to achieve real-time updates for precise analysis.

3.2.3. Face Detection Using-Viola-Jones-Algorithm

The framework makes use of the Viola-Jones algorithm which recognizes faces through distinctive emotional features. The implementation of deep learning methods improves continuous gender and emotion recognition capabilities through automatic facial expression evaluation. Several theories and approaches started to be explored during the 1990s leading to modern systems incorporating facial appearance detection for accuracy purposes.

$$H_M(x) = \frac{\sum_m a_m \, h(x)_m}{\sum_m a_m} \quad (1)$$

3.2.4. CNN Architecture For Emotion Recognition

A four-layer CNN operates for emotion recognition with convolution and ReLU activation function and batch normalization and max pooling (2x2) functions in each layer. The first phase of the architecture applies 64 filters (3x3) to grayscale images of size 48x48 then applies ReLU and subsequent batch normalization and max pooling. The second framework increases its filters from 128 (5x5) while applying the same ReLU, batch normalization, and max pooling to the initial layer output. The third layer applies 512 filters (3x3) along with the same processing procedure. Component number four contains 512 filters (3x3) which undergo identical processing. A series of dense layers ensues including 256 features in the first and another 512 in the second with both applying ReLU and batch normalization. A SoftMax classifier categorizes emotions.

3.2.5. CNN Architecture For Gender Classification

The proposed real-time gender and emotion recognition system employs a two-layer CNN model for gender classification. The first stage combines 128 3x3 filters with the input 128x128 RGB image before performing batch normalization then ReLU activation and 2x2 max pooling. In the second layer the number of filters amounts to 256 with each filter being 5x5 while processing with ReLU and batch normalization and 2x2 max pooling. Two fully connected layers are used for gender classification where the initial layer obtains ReLU and batch-normalized 256-dense features and the final layer obtains ReLU and batch-normalized 512-dense features. Binary cross-entropy classifies genders. The designed CNN implementation follows LeNet and AlexNet models through its use of ReLU activation which prompts faster convergence rates. Our model optimizes the training efficiency by employing four CNN layers together with a single output layer instead of the eight layers used in the classical AlexNet.

3.2.6. Convolutional-Neural-Network-(CNN)

The input images within a Convolutional Neural Network (CNN) pass through 2D convolution filters known as neurons. Its organization contains convolutional layers with additional pooling layers to perform subsampling before the fully connected layers. A CNN maintains translation invariance in inputs by using three features: local connections and shared weight distributions and pooling functions that exploit 2D structures. Because CNN networks feature fewer parameters than fully connected networks they become easier to train yet achieve high performance levels [57, 58].

A. Convolutional Layer

Through the convolutional layer images receive visual feature extraction with batch processing of convolution and pooling algorithms. The weights of CNN kernels transform according to inputs of multiple image batches which hold n image elements. Convolutional layers functions on four-dimensional tensors with dimensions N, Height, Color Channel and Width. The dimension of both feature maps and kernels consists of four dimensions which include filter width or height along with input and output feature maps. This operation works by processing 4-D data formats which link to feature maps in addition to image batches. The updated image dimensions display according to Figure 7 through the following sequence:

New-image width = Image width – Kernel width + 1

New-image height = Image height – Kernel height + 1

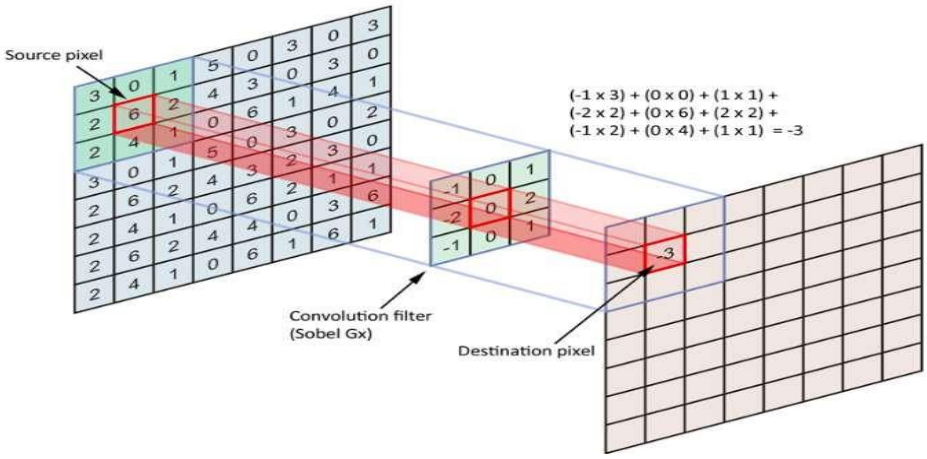


Figure 7. Convolutional Layer Operation Using Sobel Filter

B. Padding

The observation shows beneficial effects of convolutions. The corner areas tend to lose pixels in the image which becomes problematic. The small number of pixels from each convolution adds significant value in line with our previous explanation. An image with larger dimensions would simplify the process since it could be recorded easily. Such practical implementation fails to achieve this ideal outcome. Four approaches include increasing image pixels at its edges to grow image size (default edge pixels are set to zero). The original size of the 3x5, changes to a 5x7 in the following illustration. The dimension of the output matrix expands to 4x6 size while the figure demonstrates this change as depicted in Figure 8.

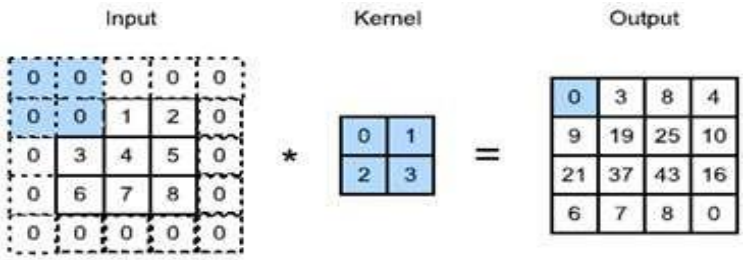


Figure 8. Effect Of Padding On Convolution Output

The dimension output will show the result of padding entire ph rows on height boundaries while pw columns on width boundaries using this calculation for total output dimensions "(nw-kw+pw+1)x(nh-kh+ph+1) "

Padding in convolutional neural networks (CNNs) accomplishes the task of maintaining input-output dimension compatibility. The padding variables ph and pw set to kh−1 and kw respectively will keep the output dimension the same as input size. When kh equals an odd number add ph=2 rows as symmetric padding on both height edges and when kh appears even provide ph=2 rows of padding to the top and bottom heights. Apply the identical logic approach to dimensions of width. Equalized kernel dimensions (kw and kh) work best for generating symmetric padding throughout. The element X[i,j] within a 2D array X crosses-correlates with the kernel centered at X[i,j] in order to generate Y[i,j] as the output. The calculation of CNN layer outputs becomes easier with this method.

C. Stride

A convolution window starts its cross-correlation calculations from the top-left section of the input array while it traverses the array from left to right and bottom to top as Figure 9 displays.

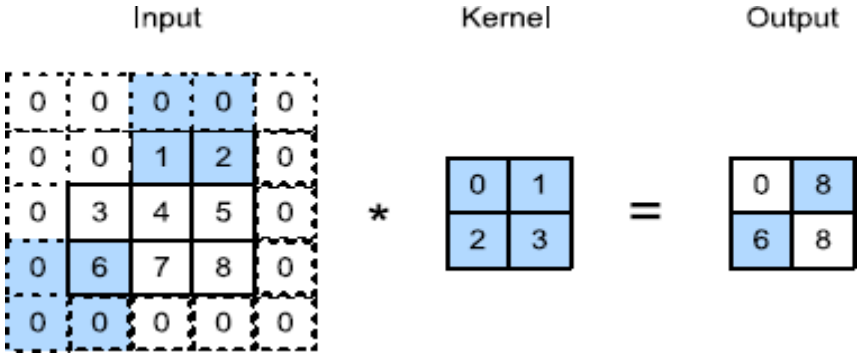


Figure 9. Illustration of Stride in Convolutional Operations

The stride uses settings of one for both height and width values. The two-dimensional cross-correlation process is altered when using a three-column-by-two-row stride configuration. If the output resides in the second position of the first column the sliding window will descend three rows. The second element in the first row undergoes a two-column right movement when performing the operation. The breadth of the window prevents any output from being produced when input elements extend beyond its limits.

The cross-correlation method is applied to height and width with defined strides of 3 and 2. The computational values from the input array and core array and output temporal show a final result of 8 while the input core total equates to 6. The formula structure for output calculation is “(nh - kh + ph + sh)/sh x (nw - kw + pw + sw)/sw” where sh and sw correspond to height and width strides. The input dimensions simplify to [(nh+sh-1)/sh] x [(nw+sw-1)/sw] when pw=kw-1 and ph=kh-1 is applied. The output structure maintains a rectangular dimension of (nw/sw) x (nh/sh) when sh and sw divide the input height and width dimensions.

D. Activation Functions

The sigmoid activation function receives substitution with ReLU which operates as an efficient computational alternative because it omits exponentiation operations. The application of ReLU brings simplified model training because it maintains a gradient value of 1 within positive zones but the sigmoid function generates near-zero gradients near output values of 0 and 1. The model becomes unreachable due to disappearing gradients when improper initialization occurs with the sigmoid function. ReLU provides reliable training stability to allow efficient back propagation and parameter updates according to Figure 10.

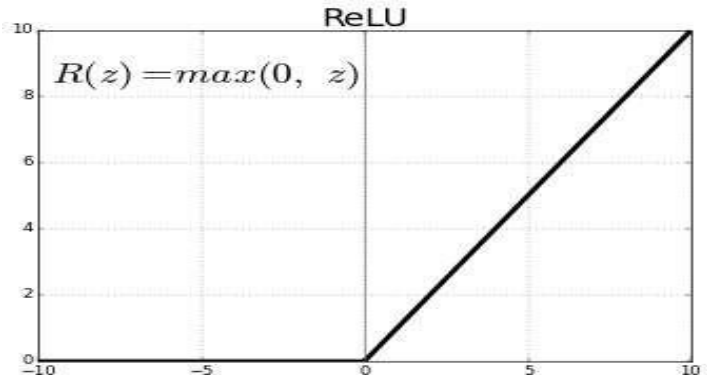
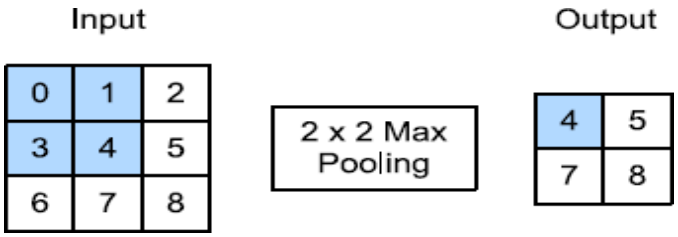


Figure 10. ReLU Activation Function Representation

E. Pooling Layer

The workload of pooling functions similarly to the convolutional approach by producing results from separate sections within it's pooling window. The processing method of pooling layers finds either the top or mean value within the window which makes up maximum pooling and average pooling. The method differs from cross-correlation in convolutional layers. As the sliding window begins on the top-left part of the input array system it picks the most prominent value to become the output array value as illustrated in Figure 11. Using pooling reduces the data size while making computations less complex and preventing overfitting. The three possible methods for subsampling are minimum, average and maximal pooling. The study implements max pooling operations using 2x2 block sizes after every convolution stage to find maximum pixel values from 2x2 grid areas thus altering only the image size dimensions.



Figrue 11. 2×2 Max Pooling Operation

The output array displays two height units along with two width units according to the illustration. Four elements are determined by the maximum value found within the max parameter. Max (0, 1, 3, 4) = 4, Max (1, 2, 4, 5) = 5, Max (3, 4, 6, 7) = 7, Max (4, 5, 7, 8) = 8.

F. Bach Normalization (BN)

An excessive number of layers beyond 100 within deep models produce training complications that involve slow convergence and enhanced complexity. The Batch Normalization method proposed provided a solution to the encountered problems. The procedure of BN calculates mini-batch mean and standard deviation to both smooth optimization landscapes and speed the learning process toward local minima for intermediate neural network outputs [59]. Careful attention must be exercised to prevent machine learning risks according to research[60]. Evidence shows that BN operations do not actually impact the occurrence of internal covariate shift and sometimes have effects that are reverse to expectations[61] . The implementation of BN allows processing layer inputs from x into outputs of y.

$$\mathrm{BN}(\mathbf{x}) = \gamma \odot \frac{\mathbf{x} - \hat{\mu}}{\hat{\sigma}} + \beta \quad (2)$$

The normalization process through Batch Normalization (BN) applies scaling coefficient γ and offset β to regulate activations around zero mean (μ) and unit variance (δ) in order to prevent intermediary divergence. Aggressive learning rates become possible through this method which supports non-standardized data entry. The training process calculates μ and δ from the mini-batch "B" where the results are μβ and δβ thus handling statistical variations. The combination of small data sample activations in BN provides stable training with high efficiency. Better deep learning models for instant gender and emotion detection become possible through BN because it maintains normalized activations which also speed up convergence.

$$\hat{\mu}_{\mathcal{B}} \leftarrow \frac{1}{|\mathcal{B}|}\sum_{\mathbf{x}\in\mathcal{B}} \mathbf{x} \text{ and } \hat{\sigma}^2_{\mathcal{B}} \leftarrow \frac{1}{|\mathcal{B}|}\sum_{\mathbf{x}\in\mathcal{B}} (\mathbf{x} - \mu_{\mathcal{B}})^2 + \epsilon \quad (3)$$

G. Dropout

Neural networks receive Dropout as a regularization technique [62]. During training Dropout disables neurons at random intervals which makes them absent from both forward calculations and weight modification processes. The technique stops neurons from developing over dependent relationships with particular contexts because it avoids complex co-adaptations. This behavior protects against model overfitting. The network develops improved generalization capabilities through dropout because it requires other neurons to step in and fulfill the role of lost components. The method fortifies the model against overfitting while making it less dependent on specific weights of individual neurons. Dropout proves especially successful for making deep learning models operate in real-time gender and emotion recognition as illustrated in Figure 12.

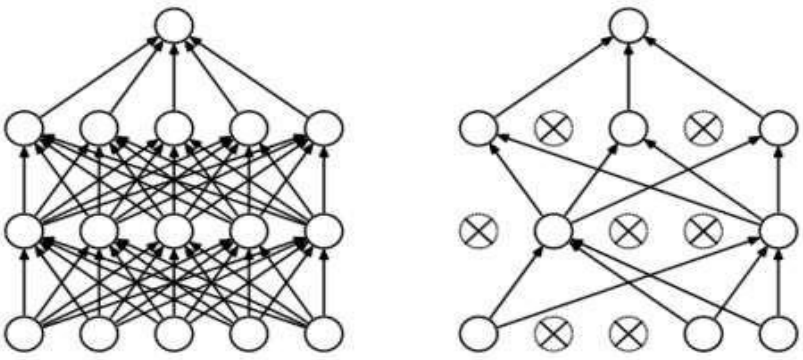


Figure 12. Standard Neural Network vs. Dropout Regularization

H. Fully Connected-Layers

The FC layer links all vegetative cells between layers to change 2D elements into a single dimensional representation. Through their combined operations Convolutional and FC layers extract hierarchical features which discover primitive pattern elements starting with edges and moving to textures and shapes at their early stages. The bottom layers of depth structures capture worldwide visual abstractions at higher levels. The convolutional layer accepts heterogeneous kernels which generate image features that become more precise after incorporating additional filters. Adding more kernels to the filters both improves their performance and raises the complexity of computations as depicted in Figure 13.

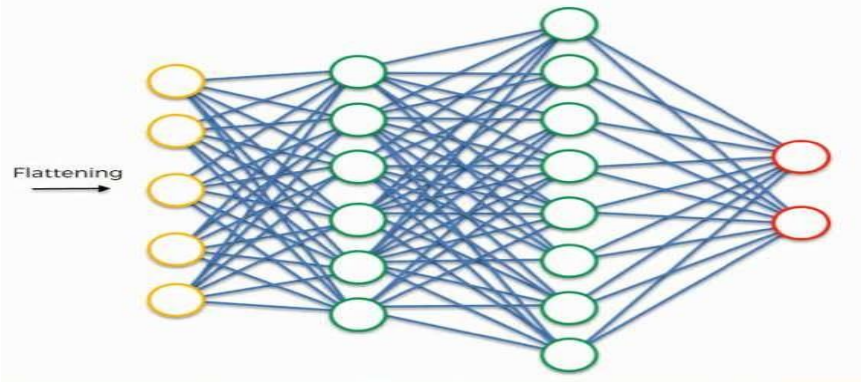


Figure 13. Fully Connected Layer Representation

3.3.Performance Parameters

The proposed system assessment utilizes four performance metrics which include accuracy together with precision and recall and F1-score assessment. Each metric is explained below:

Precision (PR) stands for how many correct positive outcomes exist among all positive results the system labels. The model's performance in identifying correct positive instances matters especially when false positive mistakes have significant financial or operational costs.

$$PR = \frac{TP}{TP + FP} \quad (4)$$

The calculation includes TP representing correctly identified positives while FP stands for mistaken positive predictions among actual negative examples.

Recall (RE) is the ability of a model to find all existing positive cases which professionals also call Sensitivity. The measure becomes vital for reducing wrong negative identifications in critical situations.

$$RE = \frac{TP}{TP + FN} \quad (5)$$

The cases which should have been positive but were incorrectly labeled negative fall under the category of FN (False Negatives).

F1-score calculates its value as the harmonic mean between precision and recall. The F1-score evaluates system performance through both precision and recall metrics therefore it proves practical when class groups are unbalanced or false positive errors and false negative errors bear comparable weights.

$$F1 - Score = 2 \times \frac{PR \times RE}{PR + RE} \quad (6)$$

Accuracy (AC) encompasses the total number of correct classifications from both positive and negative outcomes. It is defined as:

$$AC = \frac{TP + TN}{TP + FN + TN + FN} \quad (7)$$

The correct identification of negative instances is called TN (True Negatives).

In order to achieve high precision rates along with low recall the model makes cautious decisions about positive predictions thus decreasing false positives but potentially missing some valid positives. The model reaches high recall values by identifying most positive instances although it results in numerous false positives. F1-score becomes essential because it balances precision values with recall results. The accuracy metric works well for balanced datasets although it proves ineffective within unbalanced datasets.

4. Results And Discussion

4.1.Experimental Assessment, Implementation Tool And Data Acquisition

The experimental analysis evaluates a CNN model which functions as an emotion identifier through seven basic emotions (neutral, angry, sad, pleased, surprise, fear) and as a gender user (man, woman). Validation metrics along with accuracy determine model performance and training occurs with 80% of the dataset followed by testing on the remaining 20%.

4.1.1. Tool And Language Selection For Implementation

Python's Deep Learning Library enables implementation of real-time testing using a Deep Neural Network (DNN) model with one input layer followed by multiple hidden layers. The facial emotion recognition and classification system uses the CNN model.

4.1.2. Data Acquisition

Actionable dataset selection stands at the center of this research since it trains and evaluates the experimental model. The system uses various learning patterns to achieve effective emotion and gender classification. Three influential datasets named FER-2013 and Cohn-Kanade (CK+ and Karolinska Directed Emotional Faces (KDEF collectively serve this study for evaluating model performance at depth.

4.1.3. Benchmark Datasets For Gender And Emotion Recognition

The FER-2013 dataset includes 33,000 grayscale facial images that group emotions into seven classes starting with neutral and including pleased and anger as well as sadness and surprise and fearing and disgust. A computer system automatically aligns all faces into one position during registration. The proposed technique undergoes evaluation using a testing portion which amounts to 20% of the data while training occurs on 80%.

The Cohn-Kande (CK +) database contains 993 grayscale face pictures that measure 640×490 pixels and have limited background scenery stored in JPEG file format. The Karolinska Directed Emotional Faces (KDEF) database developed by Lindquist, Flykt, and Ohman in 1998 contains 490 JPEG images measuring 72 pixels with seven emotional expressions of 70 individuals consisting of 35 women and 35 men.

The IMDB database includes 4,000 images which are 128×128 pixels in dimension while it automatically divides 20% of data for validation purposes. The total number of training samples becomes 4000 while validation samples amount to 800. The proposed system uses these database collections to achieve time-sensitive gender and emotional identification.

Figure 14. Example Images From The FER-2013 Dataset.

The FER-2013 dataset contains images which are shown in Figure 14. Table 1 demonstrates the performance outcome of our model by showing the confusion matrix for detecting the seven emotions present in the FER-2013 dataset including Angry, Disgust, Fear, Happy, Neutral, Sad and Surprise categories.

Table 1. Confusion Matrix Of The FER-2013 Dataset For Facial Emotion Classification

Emotion	Angry	Disgust	Fear	Happy	Neutral	Sad	Surprise
Angry	90	0	0	1	1	8	0
Disgust	0	85	6.69	8.31	0	0	0
Fear	2.14	0	99	0	1	5	0
Happy	0	0	0	90.31	0	1	1
Neutral	0	0	0	0	98	2	0
Sad	0	0	1	0	0	99	0
Surprise	0	0	0	2.31	0	7.69	90
Avg Accuracy				93			

According to Table 1 the proposed model operates at an average 93% accuracy level and specifically performs at 99% precision in detecting Fear and Sadness while also attaining 98% precision in identifying Neutral emotions. Recognition performances for Happy (90.31%) match those of Angry (90%). The emotion Disgust shows the least accurate recognition at 85% because evaluators mistake it for Fear (6.69%) and Happy (8.31%) most frequently. The detection of Surprise produces incorrect emotions in 2.31% of cases as Happy and 7.69% of times as Sad while Sad is incorrectly identified as Fear in 5% of instances and as Angry in 8% of occurrences. The need exists to improve feature extraction techniques since the current dataset for Disgust fails to differentiate its signatures properly. Building the model's accuracy in recognizing smallest distinctions between emotions would boost its effectiveness when deployed for facial emotion identification.

The Cohn-Kanade (CK+) dataset contains these example images that appear in Figure 15. The evaluation of Cohn-Kanade (CK+) uses Table 2 to show how the model identifies seven face emotions including Angry, Disgust, Fear, Happy, Neutral, Sad, and Surprise.

Figure 1. Example Images From The Cohn-Kanade (CK+) Dataset.

Table 2. Confusion Matrix Of The Cohn-Kanade (CK+) Dataset For Classifying Seven Facial Emotions.

Emotion	Angry	Disgust	Fear	Happy	Neutral	Sad	Surprise
Angry	87	1	3	0	0	9	0
Disgust	0	89.50	0	0	0	0	10.50
Fear	10	0	88.60	0	0	0	2.40
Happy	0	0	0	93.78	0	0	6.22
Neutral	0	0	0	0	95.89	0	4.11
Sad	0	0	3.50	0	1.22	94.28	0
Surprise	0	0	2.22	2	0	0	95.78
Avg Accuracy				92			

The proposed model generates a 92% average accuracy according to the data presented in Table 2 which demonstrates superior classification capabilities. The detection accuracy surpasses 93% for each of the four expressions including Happy at 93.78%, Neutral at 95.89%, Sad at 94.28% and Surprise at 95.78%. The classifications of Disgust at 89.5% and Fear at 88.6% were successful though Fear was occasionally mistaken for either Angry at 10% or Surprise at 2.4%. The occurrence of mistaken anger emotions happens in Sad facial expressions (9%) as well as Fear (3%). Disgust receives incorrect classification as Surprise by 10.50% while Surprise mistakes happen as Happy by 6.22% and Neutral by 4.11%. The incorrect classifications of Fear and Disgust and Angry face emotions show the need for enhanced strategies to extract features and classify emotional responses. The model's performance strength will increase when corrected errors which will enhance its capability to operate effectively in security operations and psychological studies and human-computer interaction applications.

An illustration of KDEF dataset images appears in Figure 16. The KDEF dataset confusion matrix in Table 3 demonstrates how the model performed in identifying seven facial emotions between Angry, Disgust, Fear, Happy, Neutral, Sad, and Surprise.

Figure 16. Example Images From The KDEF Dataset

Table 3. Confusion Matrix Of The KDEF Dataset For Classifying Seven Facial Emotions.

Emotion	Angry	Disgust	Fear	Happy	Neutral	Sad	Surprise
Angry	77.78	11.11	0	5.56	0	0	5.56
Disgust	0	81.50	0	2.85	4.62	11.03	0
Fear	0	7.29	85.86	2.44	3	2.66	0
Happy	0	0	0	93.33	1.67	0	5
Neutral	0	0	0	1.33	95	0	3.67
Sad	0	0	5.33	0	2.67	93	0
Surprise	0	0	1	8	1.56	0	90.44
Avg Accuracy				88			

The proposed model exhibits 88% average accuracy while delivering specific high classification rates for Happy at 93.33% and for Neutral at 95% and Sad at 93% according to Table 3. The classification accuracy for Fear reaches 85.86% despite its minor overlap with Disgust which amounts to 7.29%. Anger and Disgust both show misidentification as Happy according to the participants (11.11% and 5.56% respectively) whereas Disgust occasionally gets mixed up with Sad (11.03%). The Surprise category achieves 90.44% precision yet fails to distinguish correctly between it and Happy emotions in 8% of cases. The facial features used for identifying Disgust and Fear display similar characteristics thus causing the classification system to mistake between these expressions. An improved model for real-world facial emotion identification demands better feature detection technology combined with a broadened dataset for clearer emotion distinction.

The IMDB dataset contains several images which are displayed in Figure 17. The gender identification performance of the model based on IMDB dataset appears in Table 4 through its confusion matrix for Male and Female group recognition.

Figure 17. Example Images From The IMDB Dataset.

Table 4. Confusion Matrix For Gender Classification In The IMDB Dataset.

Gender	Male	Female
Male	96	04
Female	8	92
Avg Accuracy	94%	

The proposed model reaches 94% accuracy in its classification performance according to Table 4. The proposed model recognizes male subjects accurately 96% of the time but mistakes 4% of subjects as females. The model correctly identifies 92% of female subjects in the data set yet it assigns incorrect gender labeling to 8% of those subjects as male. Some errors in the model classification exist due to gender-related facial characteristic overlaps while conducting gender identification. The model misclassifies subjects because of inconsistent lighting and different facial perspectives as well as physical traits which do not cleanly match traditional gender categories. The performance of this model can be enhanced by adding images from different age ranges and facial types and various image conditions which will improve its capabilities to generalize while decreasing misclassification errors so it can function effectively during real-world gender identification situations.

4.2.Experimental Analysis And Performance of Model Classification

The proposed experiments utilize publicly available data to achieve maximum accuracy at an acceptable level of user experience. Overall system performance depends on the number and quality of images drawn from FER-2013, Cohn-Kande (CK+), Karolinska Directed Emotional Faces (KDEF), and IMDB datasets. The framework proves its effectiveness through the demonstrated results.

4.2.1. Performance Analysis Of The Fine-Tuned 4-Layer CNN Model

Figure 18 depicts the training and validation loss of the 4-layer fine-tuned CNN model over 100 epochs. Initially, both losses are high, indicating poor performance. As training progresses, the losses decrease rapidly, signifying improved learning. By around 90 epochs, they stabilize, showing model convergence. The small gap between training and validation loss suggests good generalization with minimal overfitting.

4.2.2. Performance Analysis Of The Fine-Tuned 4-Layer CNN Model

The training and validation loss patterns of the 4-layer fine-tuned CNN model are presented through Figure 18 spanning 100 epochs. During the initial phase of training the system demonstrates poor performance because the losses remain at a high level. The learning of better concepts happens quickly during the training period as both loss variables reduce substantially. The model reaches convergence point by approximately 90 epochs. Both training and validation loss display limited variation showing that the model has accurately generalized while maintaining minimal overfitting.

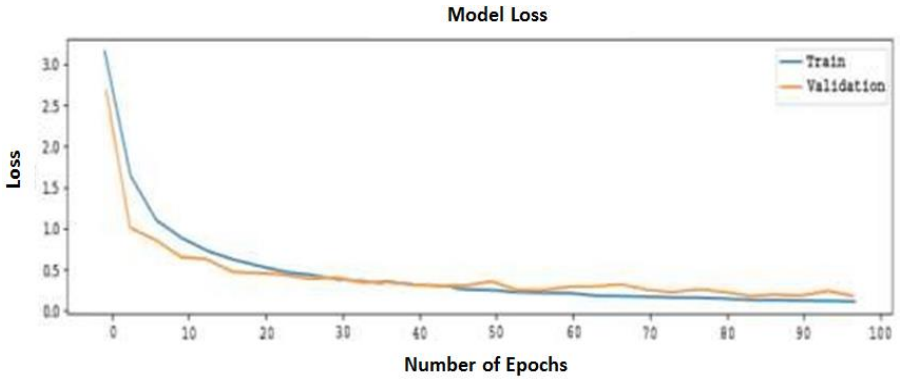


Figure 18. Training And Validation Loss Of The 4-Layer CNN Model Over 100 Epochs.

Figure 19 illustrates how the fine-tuned CNN model performs regarding training and validation accuracy throughout its execution period.

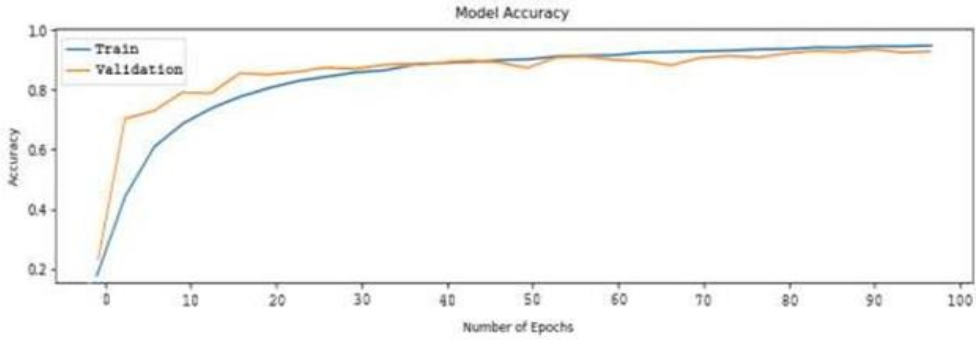


Figure 19. Training And Validation Accuracy Of The 4-Layer Fine-Tuned CNN Model Over 100 Epochs

The learning performance is evident from the models' steady increase of accuracy through every epoch run. The model reaches 93% training precision along with 90% validation precision at epoch 100 which demonstrates appropriate generalization properties.

The model reaches a better accuracy level than conventional approaches when used for testing the recognition of angry, happy, and surprise emotions. Although it shows weakness in processing both disgust and fear emotions specifically. The new model system enhances its capacity to detect both gender and emotions during real-time operations.

Upon training, the refined 4-layer CNN model is saved in HDF5 format. It reaches 93% training accuracy and 90% validation accuracy after 100 epochs (Figure 19), and accuracy increases stepwise over time. In testing, the model performs better than conventional methods in identifying angry, happy, and surprise emotions. Its accuracy in disgust and fear, however, is relatively lower. The model also exhibits improved real-time gender and emotion recognition functions, rendering it applicable in real-world applications.

4.2.3. Training And Validation Performance Of Gender Classification

Figure 20 shows layer 2 accuracy as a trained CNN model with fine-tuning during training and validation (0 to 1).

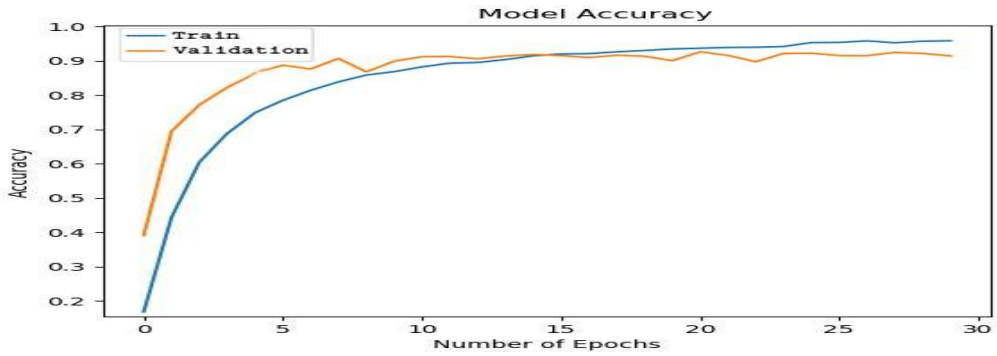


Figure 20. Training And Validation Accuracy Of The Fine-Tuned CNN Model For Gender Classification Over 30 Epochs.

The x-axis represents epochs, and red is used for validation accuracy and blue for training accuracy. Validation accuracy begins at 40% initially, while training begins at 10%. Accuracy oscillates over periods, and it reaches 90% validation and 94% training accuracy at 30 epochs.

Figure 20 shows the training and validation accuracy curves for gender classification. Initially, the training accuracy (blue) begins at approximately 10%, and the validation accuracy (orange) begins at around 40%. Both accuracies rise through subsequent epochs and level off at 94% for training and 90% for validation, reflecting the learning trajectory and convergence of the model.

Figure 21 shows six different facial emotions. Faces are bounded within green rectangles with labels placed above. This is the first role of the suggested model, to identify faces and put labels according to identified emotions.

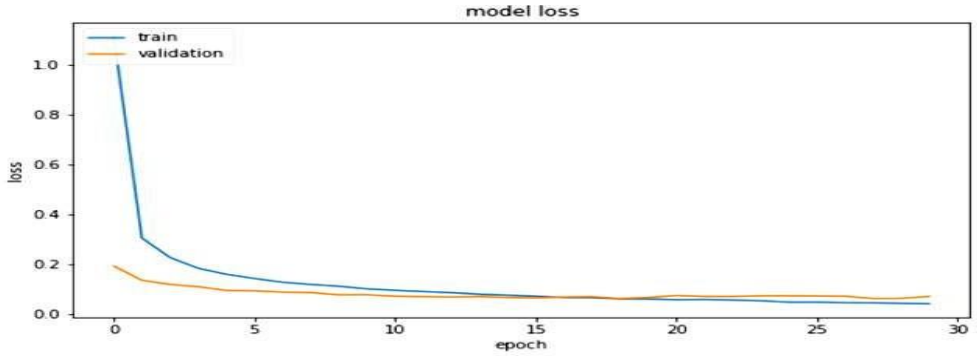


Figure 21. Training And Validation Loss Of The Fine-Tuned CNN Model For Gender Classification Over 30 Epochs

Figure 21 shows the respective loss curves. At the beginning, the training and validation losses are both high, indicating poor model performance. During training, the losses subsequently reduce dramatically, with minimal variation, indicating effective learning. By the last 30 epochs, the loss settles at around zero, indicating effective training with little overfitting.

4.2.4. Visual Analysis Of Real-Time Facial Emotion Recognition

This section illustrates the visual results from running real-time facial emotion recognition using FER-2013, CK+, and KDEF datasets. The emotional detection model identifies happy sad and fear and angry plus neutral and surprised emotions however it struggles with mistakes between fear and disgust and

fear and sadness as shown in Figure 22 ((a), (b), (c), and (d)). The refined model promotes better feature extraction that enhances both, model interpretability along with performance reliability.

Several images display real-time facial emotion recognition results in Figure 22 (a), (b), (c) and (d). The method successfully identifies emotional states although it partially misinterprets close facial expressions since developers must focus on developing this part further. The model encounters generalization problems because overlapping features across different conditions together with dataset variations produce classification mistakes.

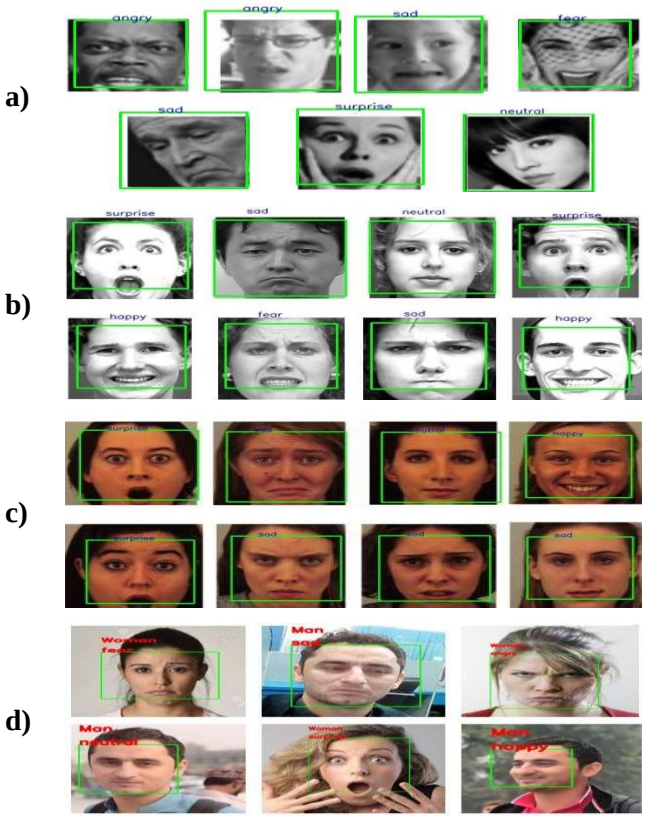


Figure 22.Images (A), (B), (C), And (D) Show The Visual Results Of The Proposed Framework On The FER-2013, CK+, And KDEF Datasets, Including Real-Time Scenarios.

The fourth-layer fine-tuned model enhances feature representation quality thus improving recognition dependability. The current version of the model functions properly under controlled circumstances although it requires more improvement to function effectively in complex real-world settings.

Among the essential challenges for the model are data examples that are uncertain and lighting issues and facial expression inconsistency. Three enhancements for accuracy improvement should be implemented in future studies: additional training data, advanced deep learning systems equipped with attention mechanisms and improved face-specific region extraction techniques. Domain adaptation technologies utilized on the model will improve its data flexibility and strengthen its performance when processing real-time applications.

4.2.5. Evaluative Comparative Analysis of The Proposed Work With Existing Research

Loss minimization and accuracy enhancement determine the main evaluation metrics in this research where the proposed CNN model receives attention for its evaluation against existing leadership

standards in the field. The proposed methodology leads to enhanced performance outcomes in deep learning-based gender and emotion recognition models as presented in Table 5.

Table 5.Comparative Analysis of Deep Learning-Based Gender and Emotion Recognition Models.

Authors	Techniques	Dataset	Accuracy (%)
Lahariya, Abhinav et-al. [47].	CNN	FER-2013	68
Çavşi Zaim et al. [48]	CNN	IMDB-WIKI	90
Yanç, Ibrahim el al. [49]	CNN	FER-2013, CK+, and KDEF	72
Proposed work	4 layer CNN	FER-2013	93
	2 layer CNN	IMDB	94

Several CNN-based models have been applied in the studies of facial emotion recognition, but they work at different levels on different datasets. Lahariya et al. [47] used a CNN-based model on the FER-2013 dataset with a 68% accuracy level, which indicates limitations in processing complex facial expressions. Çavşi Zaim et al [48] achieved 90% accuracy on the IMDB-WIKI dataset, indicating improved performance; however, their model lacks a multi-layer optimization strategy. Yanç et al. [49] used FER-2013, CK+, and KDEF databases to train their CNN model with 72% accuracy, which is significantly lower than the proposed approach.

As compared to this, the proposed CNN model achieves an accuracy of 93% in FER-2013 using a 4-layer CNN structure and 94% in IMDB using a 2-layer CNN structure. These results reflect the robustness of the proposed deep learning technique, which outperforms existing techniques by up to 26% in some cases.

Multiple significant results become apparent through the analysis process. The proposed CNN model achieves superior accuracy when compared to previous research studies on both FER-2013 and IMDB datasets. A 4-layer CNN structure improves feature extraction capabilities, which leads to enhanced identification of gender along with emotions attaining a high accuracy rate of 93% compared to previous FER-2013 models. The 2-layer CNN model demonstrates 94% accuracy when working on IMDB while using fewer layers than Çavşi Zaim et al.'s method thus proving the effectiveness of this proposed architecture. The utilization of CNN-based approaches faces multiple performance constraints especially during processing of FER-2013 and CK+ and KDEF datasets. The proposed model performs effectively to resolve these problems.

New optimized methods in real-time gender and emotion recognition constitute the main contributions of this study. The research develops a deep learning architecture that includes four CNN layers for FER-2013 alongside two CNN layers for IMDB resulting in high performance with optimized computational benefits. The method shows better performance through experiments that show superiority to traditional CNN-based approaches on various benchmark datasets to demonstrate its ability to generalize across different applications. The optimized CNN layers enhance emotion recognition by extracting precise facial characteristics which lead to better discrimination of emotions like fear and disgust as well as sadness thereby lowering recognition mistakes. The proposed CNN model outperforms deep learning models with its efficient computational capacity since it strikes a balance between depth and computational efficiency which makes it appropriate for real-time applications.

Multiple optimization methods implemented in this proposed CNN system improve its operational performance. Both the 4-layer and 2-layer CNN models have different feature extraction capabilities where the first uses layer-wise optimization to improve hierarchical detection and the second makes gender pattern identifications producing higher accuracy with minimized layers. The proposed model applies data augmentation and regularization techniques with dropout layers and batch normalization

alongside dataset augmentation to stop overfitting so the model can successfully apply to new unseen information. Transformed hyperparameters within this model ensure better convergence speed as well as higher accuracy through their optimized learning rates and activation functions and kernel sizes. A combination of Adam and RMSprop optimizers enables loss minimization and adaptive learning during the gradient descent process thus minimizing classification errors and achieving better stability during training.

The proposed CNN-based approach delivers excellent results in real-time gender and emotion recognition while reaching state-of-the-art achievements on FER-2013 and IMDB datasets. This research has brought together several key features which now make the CNN solution effective for real-time applications in human-computer interaction and surveillance and behavioral analysis. Research results show that the new proposed CNN model beats existing approaches thus establishing itself as a leading innovation in deep learning-based facial recognition.

Conclusion

This study presents a real-time system which uses deep learning to identify gender and emotions while resolving main issues in facial expression assessment. Two optimized CNN architectures combine within the proposed framework where a 2-layer CNN operates for gender classification and a 4-layer CNN executes FER tasks. The proposed model delivers unparalleled accuracy which outperforms previous systems through a 26% improvement as it reaches 94% accuracy for gender identification on IMDB and 93% accuracy for emotion detection using FER-2013. The system proves its effectiveness through multiple benchmark tests on both CK+ and KDEF datasets without showing signs of overfitting in its generalization performance. The CNN models operate efficiently to support real-time use together with Viola-Jones face detection and batch normalization which allows smooth processing sequences. The highlight of the 4-layer CNN framework lies in its identification of facial characteristics at different scales but the lightweight 2-layer CNN achieves high accuracy measurements efficiently. Due to its architecture the system can effectively serve human-computer interaction and surveillance and behavioral analysis applications. Despite its numerous advantages the model struggles with identifying disgust (85%) and fear (88.6%) emotions because their facial expressions share similarities. The system sometimes misidentifies gender when individuals have different lighting or head orientations (8% female gender errors are identified as male). The generalization of results encounters challenges because the available datasets lack sufficient diversity regarding age demographics and contain few examples of ethnicities and illumination variations.

Future research work will prioritize attention systems integration for better feature extraction since it will combine 3D CNNs and transformer-based models alongside GAN methods to generate synthetic data for diversified datasets. Lower power device optimization combined with voice and text modular features will enable the model to tackle affective computing applications together with healthcare needs plus robotics applications.

References

[1] Lapakko, David. "Three cheers for language: A closer examination of a widely cited study of nonverbal communication." Communication Education 46, no. 1 (1997): 63-67.

[2] Donato, Gianluca, Marian Stewart Bartlett, Joseph C. Hager, Paul Ekman, and Terrence J. Sejnowski. "Classifying facial actions." IEEE Transactions on pattern analysis and machine intelligence 21, no. 10 (1999): 974-989.

[3] Fasel, Beat, and Juergen Luettin. "Automatic facial expression analysis: a survey." Pattern recognition 36, no. 1 (2003): 259-275.

[4] Ekman, Paul. "Facial expressions of emotion: an old controversy and new findings." Philosophical transactions of the royal society of London. Series B: Biological Sciences 335, no. 1273 (1992): 63-69.

[5] Essa, Irfan A., and Alex P. Pentland. "Facial expression recognition using a dynamic model and motion energy." In Proceedings of IEEE International Conference on Computer Vision, pp. 360-367. IEEE, 1995.

[6] Turk, Matthew. "Perceptive media: machine perception and human computer interaction." CHINESE JOURNAL OF COMPUTERS-CHINESE EDITION- 23, no. 12 (2000): 1235-1244.

[7] Mather, George. Essentials of sensation and perception. Routledge, 2014.

[8] Senior, A., V. Vanhoucke, P. Nguyen, and T. Sainath. "Deep neural networks for acoustic modeling in speech recognition." IEEE Signal processing magazine (2012).

[9] Vinyals, Oriol, Alexander Toshev, Samy Bengio, and Dumitru Erhan. "A picture is worth a thousand (coherent) words: building a natural description of images." November. http://googleresearch. blogspot. co. uk/2014/11/a-picture-is-worththousand-coherent. html (2014).

[10] Picard, Rosalind W. Affective computing. MIT press, 2000.

[11] Pentland, Alex. "Social signal processing [exploratory DSP]." IEEE Signal Processing Magazine 24, no. 4 (2007): 108-111.

[12] Han, Jinsoo, Chang-Sic Choi, Wan-Ki Park, Ilwoo Lee, and Sang-Ha Kim. "Smart home energy management system including renewable energy based on ZigBee and PLC." IEEE Transactions on Consumer Electronics 60, no. 2 (2014): 198-202.

[13] Kim, Yelin, Honglak Lee, and Emily Mower Provost. "Deep learning for robust feature generation in audiovisual emotion recognition." In 2013 IEEE international conference on acoustics, speech and signal processing, pp. 3687-3691. IEEE, 2013.

[14] Sandbach, Georgia, Stefanos Zafeiriou, Maja Pantic, and Lijun Yin. "Static and dynamic 3D facial expression recognition: A comprehensive survey." Image and Vision Computing 30, no. 10 (2012): 683-697.

[15] Viola, Paul, and Michael J. Jones. "Robust real-time face detection." International journal of computer vision 57 (2004): 137-154.

[16] Calvo, Manuel G., and Daniel Lundqvist. "Facial expressions of emotion (KDEF): Identification under different display-duration conditions." Behavior research methods 40, no. 1 (2008): 109-115.

[17] Khan, Rizwan Ahmed, Alexandre Meyer, Hubert Konik, and Saida Bouakaz. "Framework for reliable, real-time facial expression recognition for low resolution images." Pattern Recognition Letters 34, no. 10 (2013): 1159-1168.

[18] Ekman, Paul, and Wallace V. Friesen. "Facial action coding system." Environmental Psychology & Nonverbal Behavior (1978).

[19] Damasio, Antonio R. Descartes' error. Random House, 2006.

[20] Johnson-Laird, Philip N., and Eldar Shafir. "The interaction between reasoning and decision making: An introduction." Cognition 49, no. 1-2 (1993): 1-9.

[21] Cohn, Jeffrey F., Zara Ambadar, and Paul Ekman. "Observer-based measurement of facial expression with the Facial Action Coding System." The handbook of emotion elicitation and assessment 1, no. 3 (2007): 203-221.
[22] Yacoob. "Computing spatio-temporal representations of human faces." In 1994 Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, pp. 70-75. IEEE, 1994.
[23] Choi, Hyun-Chul, and Se-Young Oh. "Realtime facial expression recognition using active appearance model and multilayer perceptron." In 2006 SICE-ICASE International Joint Conference, pp. 5924-5927. IEEE, 2006.
[24] Ghimire, Deepak, Joonwhoan Lee, Ze-Nian Li, and Sunghwan Jeong. "Recognition of facial expressions based on salient geometric features and support vector machines." Multimedia Tools and Applications 76 (2017): 7921-7946.
[25] Kim, D. J. "Facial expression recognition using ASM-based post-processing technique." Pattern Recognition and Image Analysis 26 (2016): 576-581.
[26] Zeng, Nianyin, Hong Zhang, Baoye Song, Weibo Liu, Yurong Li, and Abdullah M. Dobaie. "Facial expression recognition via learning deep sparse autoencoders." Neurocomputing 273 (2018): 643-649.
[27] Yang, Biao, Jin-Meng Cao, Da-Peng Jiang, and Ji-Dong Lv. "Facial expression recognition based on dual-feature fusion and improved random forest classifier." Multimedia Tools and Applications 77 (2018): 20477-20499.
[28] Li, Wei, Wen-jun Wu, Huai-min Wang, Xue-qi Cheng, Hua-jun Chen, Zhi-hua Zhou, and Rong Ding. "Crowd intelligence in AI 2.0 era." Frontiers of Information Technology & Electronic Engineering 18 (2017): 15-43.
[29] Zhang, Kaihao, Yongzhen Huang, Yong Du, and Liang Wang. "Facial expression recognition based on deep evolutional spatial-temporal networks." IEEE Transactions on Image Processing 26, no. 9 (2017): 4193-4203.
[30] Acevedo, Daniel, Pablo Negri, María Elena Buemi, Francisco Gómez Fernández, and Marta Mejail. "A simple geometric-based descriptor for facial expression recognition." In 2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017), pp. 802-808. IEEE, 2017.
[31] Al-agha, Lecturer Salwa A., P. H. H. Saleh, and P. R. F. Ghani. "Geometric-based feature extraction and classification for emotion expressions of 3D video film." Journal of Advances in Information Technology 8, no. 2 (2017).
[32] Liao, Shu, Wei Fan, Albert CS Chung, and Dit-Yan Yeung. "Facial expression recognition using advanced local binary patterns, tsallis entropies and global appearance features." In 2006 International Conference on Image Processing, pp. 665-668. IEEE, 2006.
[33] Tian, Y., and Shizhi Chen. "Understanding effects of image resolution for facial expression analysis." J Comput Vis Image Process (2012).
[34] Ahmed, Faisal, Hossain Bari, and Emam Hossain. "Person-independent facial expression recognition based on compound local binary pattern (CLBP)." Int. Arab J. Inf. Technol. 11, no. 2 (2014): 195-203.
[35] Happy, S. L., and Aurobinda Routray. "Automatic facial expression recognition using features of salient facial patches." IEEE transactions on Affective Computing 6, no. 1 (2014): 1-12.
[36] Malik, Fazal, Muhammad Suliman, Shehla Shaha, Muhammad Qasim Khan, and Abd Ur Rub. "Optimizing Pneumonia Diagnosis during COVID-19: Utilizing Random Forest for Accurate Classification and Effective Public Health Interventions." Journal of Computing & Biomedical Informatics 7, no. 01 (2024): 297-312.
[37] Malik, Fazal, Muhammad Suliman, Muhammad Qasim Khan, Noor Rahman, and Mohammad Khan. "Optimized XGBoost-based model for accurate detection and classification of COVID-19 pneumonia." Journal of Computing & Biomedical Informatics 7, no. 02 (2024).

[38] Suliman, Muhammad, Fazal Malik, Muhammad Qasim Khan, Ashraf Ullah, Noor Rahman, and Said Khalid Shah. "A Convolutional Neural Network (CNN) Based Framework for Enhanced Diagnosis and Classification of COVID-19 Pneumonia." VAWKUM Transactions on Computer Sciences 12, no. 2 (2024): 220-240.

[39] Suliman, Muhammad, Fazal Malik, Muhammad Qasim Khan, Irfan Ullah, and Abd Ur Rub. "Integrating data augmentation with AdaBoost for effective COVID-19 pneumonia classification." Journal of Computing & Biomedical Informatics 7, no. 01 (2024): 590-605.

[40] Khan, Muhammad Qasim, Fazal Malik, and Noor Rahman. "Optimized Sentiment Classification of Google Play Store App Ratings Using Advanced Machine Learning Models." VFAST Transactions on Software Engineering 12, no. 4 (2024): 252-266.

[41] Shah, Masroor, Fazal Malik, Muhammad Suliman, Noor Rahman, Irfan Ullah, Sana Ullah, Romaan Khan, and Salman Alam. "Dark Data in Accident Prediction: Using AdaBoost and Random Forest for Improved Accuracy." Journal of Computing & Biomedical Informatics 7, no. 02 (2024).

[42] Malik, Fazal, Muhammad Suliman, Muhammad Qasim Khan, Noor Rahman, Khairullah Khan, and Muhammad Khan. "Optimizing malicious website detection with the XGBoost machine learning approach." Journal of Computing & Biomedical Informatics 7, no. 02 (2024).

[43] Malik, Fazal Malik Fazal, Muhammad Suliman Suliman, Irfan ullah Irfan, Shehla Shah Shehla, and Asiya Bibi Asiya. "Enhancing Cyber Security: A Holistic Strategy for Advanced Malicious Website Prediction Using AdaBoost Algorithm." Lahore Garrison University Research Journal of Computer Science and Information Technology 8, no. 3 (2024).

[44] Malik, F., A. U. Rahman, A. Ullah, R. Hussain, M. Javed, & S. Ullah. (2024). Optimizing Malicious Website Detection Through Comparative Analysis of Machine Learning Techniques. Pakistan Journal of Scientific Research, 4(1(Suppl.), 147–161. https://doi.org/10.57041/vol4iss1(Suppl.)pp147-16.

[45] Al-Sumaidaee, Saadoon AM, Mohammed AM Abdullah, Raid Rafi Omar Al-Nima, Satnam Singh Dlay, and Jonathon A. Chambers. "Multi-gradient features and elongated quinary pattern encoding for image-based facial expression recognition." Pattern recognition 71 (2017): 249-263.

[46] Lekdioui, Khadija, Rochdi Messoussi, Yassine Ruichek, Youness Chaabi, and Raja Touahni. "Facial decomposition for expression recognition using texture/shape descriptors and SVM classifier." Signal Processing: Image Communication 58 (2017): 300-312.

[47] Lahariya, Abhinav, Varsha Singh, and Uma Shanker Tiwary. "Real-time emotion and gender classification using ensemble CNN." arXiv preprint arXiv:2111.07746 (2021).

[48] Çavşi Zaim, Hande, Metin Yılmaz, and Esra Nergis Yolaçan. "Design of gender recognition system using quantum-based deep learning." Neural Computing and Applications 36, no. 4 (2024): 1997-2014.

[49] Yanç, Ibrahim, Aykan İpek, and Selma Yilmazyildiz Kayaarma. "Facial Emotion Recognition for Imitation in Human-Robot Interaction." In 2024 9th International Conference on Computer Science and Engineering (UBMK), pp. 654-659. IEEE, 2024.

[50] Mlakar, Uroš, Iztok Fister, Janez Brest, and Božidar Potočnik. "Multi-objective differential evolution for feature selection in facial expression recognition systems." Expert Systems with Applications 89 (2017): 129-137.

[51] Sun, Yaxin, and Guihua Wen. "Cognitive facial expression recognition with constrained dimensionality reduction." Neurocomputing 230 (2017): 397-408.

[52] Sun, Zhe, Zheng-Ping Hu, Meng Wang, and Shu-Huan Zhao. "Discriminative feature learning-based pixel difference representation for facial expression recognition." IET Computer Vision 11, no. 8 (2017): 675-682.

[53] Ding, Yuanyuan, Qin Zhao, Baoqing Li, and Xiaobing Yuan. "Facial expression recognition from image sequence based on LBP and Taylor expansion." IEEE Access 5 (2017): 19409-19419.

[54] Kumar, Sunil, Manas Kamal Bhuyan, and Biplab Ketan Chakraborty. "Extraction of informative regions of a face for facial expression recognition." IET Computer Vision 10, no. 6 (2016): 567-576.
[55] Lyons, Michael J., Julien Budynek, Andre Plante, and Shigeru Akamatsu. "Classifying facial attributes using a 2-d gabor wavelet representation and discriminant analysis." In Proceedings fourth IEEE international conference on automatic face and gesture recognition (Cat. No. PR00580), pp. 202-207. IEEE, 2000.
[56] Valstar, Michel François, Ioannis Patras, and Maja Pantic. "Facial action unit detection using probabilistic actively learned support vector machines on tracked facial point data." In 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)-Workshops, pp. 76-76. IEEE, 2005.
[57] Chen, Junkai, Zenghai Chen, Zheru Chi, and Hong Fu. "Facial expression recognition based on facial components detection and hog features." In International workshops on electrical and computer engineering subfields, pp. 884-888. 2014.
[58] Muhammad, K., Hussain, T. and Baik, S.W., 2020. Efficient CNN based summarization of surveillance videos for resource-constrained devices. Pattern Recognition Letters, 130, pp.370-375.
[59] Ioffe, Sergey, and Christian Szegedy. "Batch normalization: Accelerating deep network training by reducing internal covariate shift." In International conference on machine learning, pp. 448-456. pmlr, 2015.
[60] Lipton, Zachary C. "The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery." Queue 16, no. 3 (2018): 31-57.
[61] Kar, Santu, and Kumar Neeraj Jha. "Assessing criticality of construction materials for prioritizing their procurement using ANP-TOPSIS." International Journal of Construction Management 22, no. 10 (2022): 1852-1862.
[62] Pham, Vu, Théodore Bluche, Christopher Kermorvant, and Jérôme Louradour. "Dropout improves recurrent neural networks for handwriting recognition." In 2014 14th international conference on frontiers in handwriting recognition, pp. 285-290. IEEE, 2014.