Spectrum of Engineering Sciences
ISSN (e) 3007-3138 (p) 3007-312X Volume 3, Issue 5, 2025

FACIALFUSIONNET: A NOVEL TRIPLE-BRANCH CNN ENSEMBLE FOR DECEPTION DETECTION FROM FACIAL MICRO-EXPRESSIONS

Memoona Malik[1], Shabana Ramzan[1], Aqsa Mahmood[1], Hafiza Hamida Zulqarnain[1], Muhammad Faheem Mushtaq[*2], Wajahat Hussain [2]

[1]*Department of Computer Science & IT, Government Sadiq College Women University Bahawalpur, 63100, Pakistan*
[2]*Department of Data Science, The Islamia University of Bahawalpur, 63100, Pakistan*

memonabwp@gmail.com, shabana@gscwu.edu.pk, aqsa.mahmood@gscwu.edu.pk, 2605hhz@gmail.com, faheem.mushtaq@iub.edu.pk, jamwajahat@gmail.com

DOI: https://doi.org/10.5281/zenodo.23914518

Keywords

Deception Detection, Facial Expression Analysis, Deep Learning, Convolutional Neural Networks (CNN), FacialFusionNet Model, Feature Fusion, Transfer Learning, Grad-CAM, MU3D Dataset, Computer Vision.

Article History

Received on 12 April 2025
Accepted on26 May 2025
Published on28 May 2025

Copyright @Author

Corresponding Author: *
Shabana Ramzan

Abstract

Deception detection through facial expression analysis has emerged as a significant area of research owing to its wide-ranging applications in security, forensic investigation, and behavioral analysis. The research employs the Miami University Deception Detection Database (MU3D), which contains 320 videos, half truthful and half deceptive. A frame extraction algorithm is used to encode videos into sample images used to train Convolutional Neural Networks (CNN). To evaluate the performance of pretrained models in this context, three transfer learning architectures–ResNet50, MobileNetV2, and EfficientNetB0–were implemented and assessed individually. Building upon these baseline models, a hybrid architecture, termed FacialFusionNet, is proposed, which integrates ResNet50, MobileNetV2, and EfficientNetB0 through feature-level fusion to enhance overall classification performance. Experimental results demonstrate that ResNet50 and EfficientNetB0 individually achieve a classification accuracy of 96%, while MobileNetV2 attains 95%. Notably, the proposed FacialFusionNet model outperforms all individual baseline models, achieving a classification accuracy of 99%, along with superior precision, recall, and F1-score. Furthermore, Grad-CAM is employed to visualize the salient facial regions influencing model predictions, thereby enhancing the interpretability of the proposed framework. The experimental findings suggest that the proposed approach effectively captures subtle facial expression cues, offering a robust and efficient solution for automated deception detection.

1. INTRODUCTION

Deception is one of the most basic forms of communication and it is crucial that it is accurately detected, with implications for security, psychology, and human–computer interaction. With the growing use of video communication in daily life and the increasing sophistication of deceptive activity, there is an urgent need for effective, automatic and non-invasive tools for detecting deception. Facial expressions are one of the most promising cues among the different types of behavioral signals that are available, because they are less consciously controlled and may express emotions and psychologies that are opposite to the face's verbal representation [1, 2]. Physiological measurement techniques, such as polygraphy (measuring physiological responses including heart rate, skin conductance and respiration to determine the presence or absence of truthfulness) have been used in the past to detect deception. Polygraph techniques are widely applied in

forensic and investigative settings, but they are intrusive, labor-intensive, and highly dependent on specialization and trained examiners, and have been criticized many times as unreliable in real-world applications [3]. In addition to forensic use, deception detection can be of great benefit in a wide variety of areas, such as criminal investigations, border security, online transaction verification, recruitment interviews, and in daily interpersonal interactions. If the deceptive activity is not accurately detected in these situations, it could result in severe consequences, including wrongful convictions, financial abuse, and security breaches [4, 5]. As video mediated interactions are increasingly common, research has increasingly turned to computational methods that can directly analyze the facial behavior from video data. Despite the progress made, current deceptive detection techniques have a number of shortcomings. Some methods rely on handcrafted features for micro-expression detection which cannot capture the subtle and fleeting nature of micro-expression, and others rely on a single deep learning model that is effective, but fails to fully capture the complementary strengths of various network architectures. In addition, many of the current methods are black-box models, lacking in interpretability for security and forensic applications due to the need for transparency and trust. Existing deception detection systems are limited in their robustness, generalizability and applicability to real-world scenarios, thus calling for an approach that is both accurate, efficient and interpretable. This work presents an automated deception detection framework based on deep learning algorithm which is assessed on the MU3D dataset by analyzing facial expressions. An end-to-end processing pipeline is designed, which includes video frame extraction, facial preprocessing, data augmentation, and dataset balancing to overcome class imbalance. To set up base lines, three popular CNN models, ResNet50, EfficientNetB0, and MobileNetV2 are implemented and evaluated on their own based on transfer learning. Additionally, we propose a novel triple-branch CNN ensemble architecture named FacialFusionNet that combines ResNet50, EfficientNetB0 and MobileNetV2 by feature-level fusion and effectively utilizes deep hierarchical feature learning and efficiency to boost the overall classification performance. To further increase transparency of the model, Grad-CAM is used for facial region visualization by identifying the face regions that are most crucial for model decision-making; this helps in providing interpretable results of the model decision-making process and increases the trust in the proposed system. The key contributions of this study are summarized as follows:

1. To design and implement a complete deep learning pipeline for facial expression-based deception detection, including frame extraction, preprocessing, augmentation, and class balancing on the MU3D dataset.
2. To evaluate three pretrained CNN architectures ResNet50, MobileNetV2, and EfficientNetB0 individually via transfer learning, establishing strong baseline benchmarks for the task.
3. To propose FacialFusionNet, a novel triple-branch CNN ensemble architecture that fuses features extracted from ResNet50, MobileNetV2, and EfficientNetB0 at the feature level, achieving superior classification accuracy, precision, recall, and F1-score compared to individual baseline models.
4. To incorporate Grad-CAM-based visualization to interpret model predictions, identifying salient facial regions associated with deceptive behavior and enhancing the interpretability of the proposed framework.
5. To demonstrate that the proposed approach effectively captures subtle facial micro-expression cues, offering a robust, efficient, and interpretable solution for automated deception detection with practical applicability in real-world scenarios.

The remainder of this paper is organized as follows. Section 2 reviews related work on deception detection and facial expression analysis. Section 3 describes the proposed methodology, including the dataset, preprocessing pipeline, baseline models, and the FacialFusionNet architecture.

RELATED WORK

The initial studies on deception detection were mostly psychological and physiological in nature. The most popular physiological method is the polygraph, which is based on the assumption that deceptive intent is indicated by changes in autonomic nervous system

functioning (heart rate variability, skin conductance, respiration patterns, etc.) [3].

These techniques were historically important, but they do need direct physical contact with the subject in question, special equipment and trained examiners, and their scientific validity has been questioned many times in the past because they sometimes do not perform accurately from person to person and place to place. Meanwhile, behavioral scientists investigated verbal and non-verbal information such as speech patterns, gaze aversion, body posture, and gesture to look for signs of deceptive intent [6, 7].

These early psychological experiments formed the basis of early psychological theories of the involuntary leakage of hidden emotions through the body and face, which would later be adopted by computational methods. Most of these studies were however observational or they used the manual coding of trained human annotators, which restricted their scalability and applicability to automated, real time detection systems. As evidence began to accumulate that subtle facial movements or micro-expressions of a few seconds are difficult to control consciously, subsequent study focused on facial expression as one of the key modalities for deception detection [2]. Research projects included in this category usually use facial action coding systems or handmade spatiotemporal features to measure the small movements of the muscles that occur when emotional states are hidden. These techniques showed that facial cues contain discriminative information for lie detection, but they are based on manually-designed features and are thus sensitive to variations in illumination, head pose and video quality, and lack the ability to transfer to different populations and recording environments. However, with the advent of deep learning, there is now a possibility of learning features from data, which can go much further than handcrafted features. Recently, the convolutional neural network (CNN) and Recurrent architecture have been exploited to detect spatial and temporal facial features directly from video frames, which has boosted the detection accuracy over the traditional method [8]. Recent approaches have suggested hybrid and ensemble deep learning architectures using several pathways for feature extraction, or integrating face information with other modalities like audio or physiological signals to further improve the detection performance [9].

Transfer learning, especially, has been applied successfully to deception detection tasks which often are based on small labeled datasets, in which the pretrained CNN architectures originally trained on large scale image were fine-tuned to extract discriminative facial representations with less training overheads [8, 9]. Yet, many current deep learning-based techniques use only a single-branch architecture, which might not fully reflect the complementary advantages (e.g., deep hierarchy feature extraction versus low computational complexity from the design of the network). Moreover, most of these models are black-box predictors with only limited interpretability about which facial regions play the largest role in a particular classification decision, an area of the field that has only recently started to be tackled with methods such as Grad-CAM or saliency mapping. The development of deception detection methods that rely on deep learning has long been linked to the availability of annotated video databases of facial movements that are genuine or deceptive. The Bag of Lies dataset contains multimodal recordings consisting of eye gaze, EEG and facial video collected in controlled deception scenarios that allow multimodal deception detection research [10].

The Miami University Deception Detection (MU3D) dataset has been established as a reliable benchmark for deception detection studies using facial expression [11] and has been used in various studies, for example, in [15] for training a deep neural network to classify

facial expressions for deception detection. Although these datasets have enabled the shift from lab-based studies to more robust and data-driven deep learning methods, the limited scale of publicly available deception datasets (when compared with large-scale datasets for general-purpose vision tasks) still presents some issues concerning class imbalance and model generalization. In addition to the deception detection literature, ensemble and hybrid deep learning architectures have been successful in various detection tasks, such as acoustic fault detection based on mel-frequency cepstral features [12], electrical fault classification in power transmission [13], medical image classification based on radiographs and mammograms [14] [15], and security-related threat and fraud detection [16] [17].

In parallel, deep architectures are also being used to infer emotional and opinion states from user-generated content, such as attention-based models [18] and emotion classifiers based on domain-specific corpora [19] and [20]. Similar trade-offs between detection accuracy and model interpretability have also appeared in other security-related scenarios, e.g., detecting network intrusions and denial-of-service attacks [21], [22] and in broader Internet of Things security scenarios [23]. Considering these studies, a conclusion can be drawn that the integration of complementary learners is generally better than any individual model that the principle applied by FacialFusionNet does in the field of facial micro-expression analysis. Practically speaking, deception detection systems are rarely used in a vacuum, but rather are part of larger security, sensing, and communication networks. In related work in this area, lightweight cipher design and cryptographic protection mechanisms for protecting transmitted data have been discussed [24], [25], [26], [27], wireless network monitoring using rogue access point detection and localization [28] and sensor-based measurement and instrumentation systems [29] have been explored, as has the design of the ciphers and other energy-efficient security measures for IoT deployments [30]. These contributions all come from various levels of the technology stack that are being studied in the present work, but all have in common the need for automated systems to be accurate, efficient and secure after they are deployed in operational settings. All the literature reviewed show clearly that the shift from physiological and behavioral to facial expression based deep learning approaches is evident. There, however, are three major gaps that must be addressed. First, most current deep learning models use only a single backbone architecture, which results in the loss of many complementary strengths of multiple pretrained models. Second, there are not systematic dataset balancing and augmentation methods proposed for many of the methods that attempt to tackle class imbalance issues in many deception detection databases like MU3D. Thirdly, much less research has been done in the area of model interpretability, which is essential for certain security-and-forensic-related applications where end-users need to trust and verify the automated predictions made by models. To tackle these limitations together, the present study proposes a triple-branch CNN ensemble which combines ResNet50, EfficientNetB0, and MobileNetV2 at the feature level combined with a systematic preprocessing and balancing pipeline, coupled with a Grad-CAM based Interpretability analysis. Table 1 shows a summary of literature reviewed.

Table 1. Summary of related work.

Study	Method / Approach	Dataset / Input	Key Contribution	Limitation
[1] (2023)	Survey of AI techniques	Multiple datasets	Comprehensive review of deception detection methods	Lacks experimental validation

Study	Method / Approach	Dataset / Input	Key Contribution	Limitation
[31] (2022)	Lightweight CNN (mobile backbone)	Image-based	Efficient model for real-time tasks	Slight accuracy trade-off
[8] (2025)	Micro-expression + feature selection	Facial expressions	Improved detection using feature selection	Needs high-quality data
[9] (2024)	Hybrid deep neural network	Facial micro-expressions	Improved accuracy using hybrid model	Increased complexity
[32] (2023)	Interpretable deep learning model	Facial expressions	Added explainability	Limited generalization
[33] (2023)	ML + physiological signals	Multimodal	Combined facial and pulse data	Requires sensors
[34] (2022)	Facial muscle analysis	Facial movements	Effective facial cue detection	Not fully automated

2. PROPOSED METHODOLOGY

The proposed system presents a deep learning-based architecture for automatic lie detection using facial expressions extracted from a sequence of videos. The methodology converts raw video information into discriminative spatial features and categorizes them as truthful or deceptive. The overall framework consists of dataset preparation, preprocessing, model development, training, evaluation, and explainability, as shown in Figure 1.

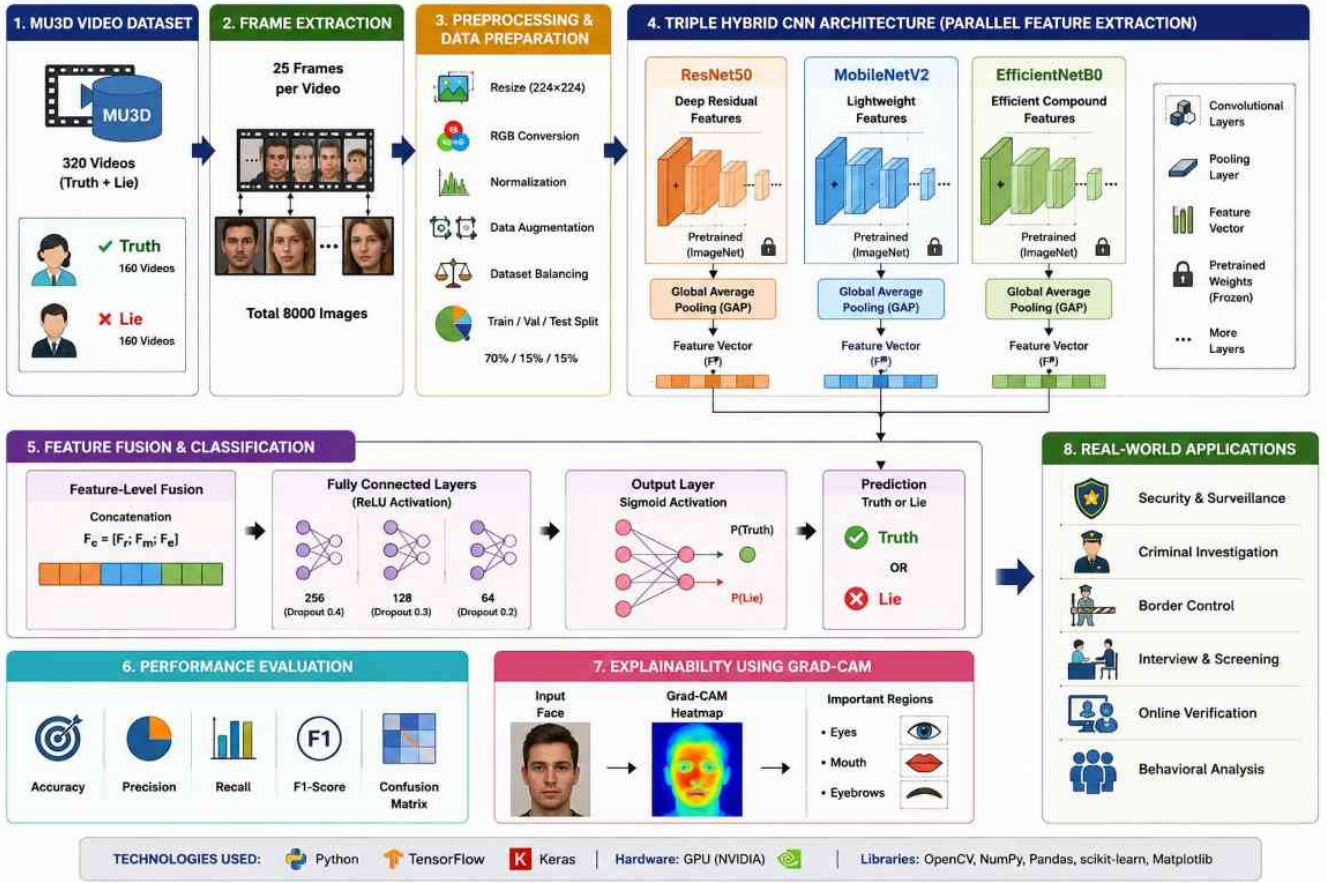


Figure 1. Architecture of the proposed system.

3.1 Dataset Description

This study uses the publicly available Miami University Deception Detection Database (MU3D) [11], which is designed for the analysis of deception. It consists of 320 video clips of individuals saying the truthful and untruthful words. The sample consists of 80 participants that are evenly split by gender and race (20 Black females, 20 Black males, 20 White females, and 20 White males). Each participant delivers four videos, half of them true and half false, half of them positive and half negative. Table 2 gives a summary of the data.

Table 2. MU3D dataset summary.

Attribute	Description
Total videos	320
Participants	80
Truth videos	160
Lie videos	160

The dataset is suitable for deception detection tasks based on facial expression because all videos are recorded under controlled conditions and labelled with the help of trained assistants.

3.2 Frame Extraction

Since the MU3D data set is made up of video sequences, a frame extraction procedure is applied to convert the time dependent data to image based representations that can be used by models in deep learning. For every video, 25 frames are evenly sampled throughout the clip in order to capture significant facial expressions while eliminating redundancy. The extracted frames is resized to some desired fixed resolution, and stored in different folders based on their label, truth and lie. This method ensures a well-structured and balanced dataset for model training. The frame extraction process is summarized in the following table (Table 3).

Table 3. Frame extraction summary.

Category	Videos	Frames per video	Total frames
Truth	160	25	4,000
Lie	160	25	4,000
Total	**320**	–	**8,000**

The resulting image data is passed to the subsequent preprocessing and deep learning stages.

3.3 Data Preprocessing

Subsequent to the extraction, every frame of the video undergoes various preprocessing stages. Firstly, the images are resized to an agreed dimension of either 160×160 pixels or 224×224 pixels depending on the architecture of the model used. The images are transformed into their RGB form, and then, their pixel values are normalized via specific preprocessing functions developed for the model to subtract the average pixel value and assist the pretrained networks to function better.

3.4 Data Splitting

Using stratified sampling, the dataset is divided into training, validation, and testing datasets to maintain the constant balance of class distribution at 50:50. The data used for training contains 70% of the dataset (5,600 frames) used for training the model, the validation dataset contains 15% (1,200 frames) for hyperparameter tuning and the early processes of stopping the training, and the testing dataset also contains 15% (1,200 frames). This separation makes sure that the testing data is not leaked to the training or validation data, thus producing a true result.

Table 4. Data distribution across training, validation, and test splits.

Split	Percentage	Total frames	Truth	Lie
Training	70%	5,600	2,800	2,800
Validation	15%	1,200	600	600
Testing	15%	1,200	600	600
Total	**100%**	**8,000**	**4,000**	**4,000**

3.5 Data Balancing and Augmentation

By applying image augmentation techniques in this study, the dataset gets enlarged and the quality of the training models is improved. Some techniques employed in this study include random rotations, horizontal flipping, zooming, and small translations which allow the model to learn invariant features and decrease overfitting. The dataset is class-balanced, meaning there are the same number of honest and dishonest samples in the MU3D database. Furthermore, stratified sampling is used to keep the classes balanced when dividing the dataset into training, validation, and testing sets. With this method, the models will not be biased towards any of the classes and will be able to learn the discriminative features easily.

3.6 Deep Learning Models

Three recent Convolutional Neural Network (CNN) architectures were used to perform image classification: EfficientNetB0, MobileNetV2, and ResNet-50. The choice of these models was based on their high performance, scalability, and applicability in both accuracy-critical and resource-constrained environments.

3.6.1 EfficientNetB0

EfficientNetB0 is a compact yet high-precision CNN model that applies a compound scaling strategy, uniformly scaling network depth, width, and resolution.

Input layer. The EfficientNetB0 model takes a colour image that has been resampled to a standard resolution of 224 × 224 × 3, as given in Equation (1). This layer standardizes the input dimensions of the network so that it remains compatible with the pretrained weights.

$$\mathbf{X} \in \mathbb{R}^{224\times224\times3} \tag{1}$$

Convolutional feature extraction. EfficientNetB0 extracts hierarchical visual features through a series of convolutional and mobile inverted bottleneck (MBConv) blocks. The architecture employs compound scaling to control depth, width, and resolution so that feature learning requires fewer parameters. These blocks capture both low-level features (edges, textures) and high-level semantic features. The convolutional backbone is frozen in a manner that preserves the features learned from ImageNet, as shown in Equation (2).

$$\mathbf{F} = f_{\text{EfficientNetB0}}(\mathbf{X}), \quad \mathbf{F} \in \mathbb{R}^{7\times7\times1280} \tag{2}$$

Global average pooling layer. The global average pooling layer reduces each feature map to a single scalar. This keeps the number of parameters in the model very low relative to fully connected layers and enhances generalization by limiting overfitting, as expressed in Equation (3).

$$g_k = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} F_{i,j,k}, \quad \mathbf{g} \in \mathbb{R}^C \tag{3}$$

Output classification layer. Binary classification is performed using a fully connected dense layer with sigmoid activation. This layer converts the extracted feature vector into a probability score representing the likelihood that the input belongs to the positive class, as expressed in Equation (4).

$$\hat{y} = \sigma(z) = \frac{1}{1 + e^{-z}} \tag{4}$$

3.6.2 ResNet-50

ResNet-50 incorporates residual learning through skip connections, which facilitate the network to make use of identity mappings. This mechanism is crucial in overcoming the issue of vanishing gradients and facilitates the training of deeper networks. Another advantage of the ResNet-50 system is its strong expressive ability which provides it with sufficient capabilities for extraction of advanced image features required for classification.

Input layer. The ResNet-50 model takes an image of size 224 × 224 × 3. This standardized input format makes the reuse of ImageNet pretrained weights possible, as given in Equation (5).

$$\mathbf{X} \in \mathbb{R}^{224\times224\times3} \tag{5}$$

Residual convolutional blocks. ResNet-50 is composed of stacked residual blocks containing convolutional layers and identity shortcut connections. Such residual connections enable gradients to pass directly through the network, making the training of deep architectures effective. The model learns increasingly complex feature representations across its 50 layers. The frozen convolutional layers are used as a fixed feature extractor, as expressed in Equations (6), (7), and (8).

$$\mathbf{y} = \mathcal{F}(\mathbf{x}, \mathbf{W}) + \mathbf{x} \tag{6}$$

$$\mathbf{F} = f_{\text{ResNet50}}(\mathbf{X}) \tag{7}$$

$$\mathbf{F} \in \mathbb{R}^{H\times W\times C} \tag{8}$$

Global average pooling layer. The global average pooling layer compresses the spatial features into a compact vector. This layer replaces the traditional fully connected layers and simplifies the model while retaining the discriminative information, as expressed in Equation (9).

$$g_k = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} F_{i,j,k}, \quad \mathbf{g} \in \mathbb{R}^C \tag{9}$$

Classification layer. The final dense layer uses a sigmoid activation function for binary classification. It returns the probability of the expected class of the input image, as given in Equations (10) and (11).

$$z = \mathbf{w}^T \mathbf{g} + b \tag{10}$$

$$\hat{y} = \sigma(z) \tag{11}$$

3.6.3 MobileNetV2

MobileNetV2 is optimized for resource-constrained environments. It uses depthwise separable convolutions and inverted residual blocks with linear bottlenecks to keep computational and memory requirements considerably low.

Input layer. MobileNetV2 receives input images scaled to 224 × 224 × 3. This provides compatibility with the pretrained network and ensures uniform feature extraction.

Depthwise separable **convolution and inverted residual blocks.** MobileNetV2 uses depthwise separable convolutions to lower computational cost by separating spatial and channel processing. The inverted residual blocks with linear bottlenecks expand the feature dimensionality to facilitate effective learning and then project it back to a lower-dimensional space. Information flow is preserved through skip connections, which enhance feature reuse. Training of the convolutional backbone is frozen so as to reduce overfitting, as expressed in Equations (12) to (16).

$$\mathbf{Y}_{dw} = \mathbf{X} * \mathbf{K}_{dw} \tag{12}$$

$$\mathbf{Y}_{pw} = \mathbf{Y}_{dw} * \mathbf{K}_{pw} \tag{13}$$

$$\mathbf{Z} = \mathrm{ReLU6}\left(\mathbf{X} * \mathbf{W}_e\right) \tag{14}$$

$$\mathbf{F} = \mathbf{Z} * \mathbf{W}_p \tag{15}$$

$$\mathbf{Y} = \mathbf{F} + \mathbf{X} \tag{16}$$

Global average pooling layer. The global average pooling layer reduces each feature map to a single value, which minimizes the number of parameters and improves generalization. This keeps the model lightweight and deployable in resource-constrained settings.

Dropout layer. After pooling, a dropout layer is applied in order to randomly disable neurons during training. This regularization method is used to avoid overfitting and enhance the robustness of the model, as expressed in Equation (17).

$$\mathbf{g}' = \mathbf{g} \odot \mathbf{m}, \quad m \sim \mathrm{Bernoulli}(1-p) \tag{17}$$

Classification layer. The final binary classification output is produced by a sigmoid-activated dense layer, which gives the likelihood of the input image belonging to the target class, as given in Equations (18) and (19).

$$z = \mathbf{w}^T \mathbf{g}' + b \tag{18}$$

$$\hat{y} = \sigma(z) \tag{19}$$

3.7 FacialFusionNet Model Architecture

Based on ResNet50, EfficientNetB0, and MobileNetV2, the FacialFusionNet deep learning model is proposed to improve feature representation and classification performance. All the three networks run in parallel on the same input image, and the learned features of all of them are fused to make the final prediction. The triple-branch design is based on the complementary capabilities of each backbone: ResNet50 for deep residual representations, EfficientNetB0 for efficient scaling and MobileNetV2 for lightweight feature extraction.

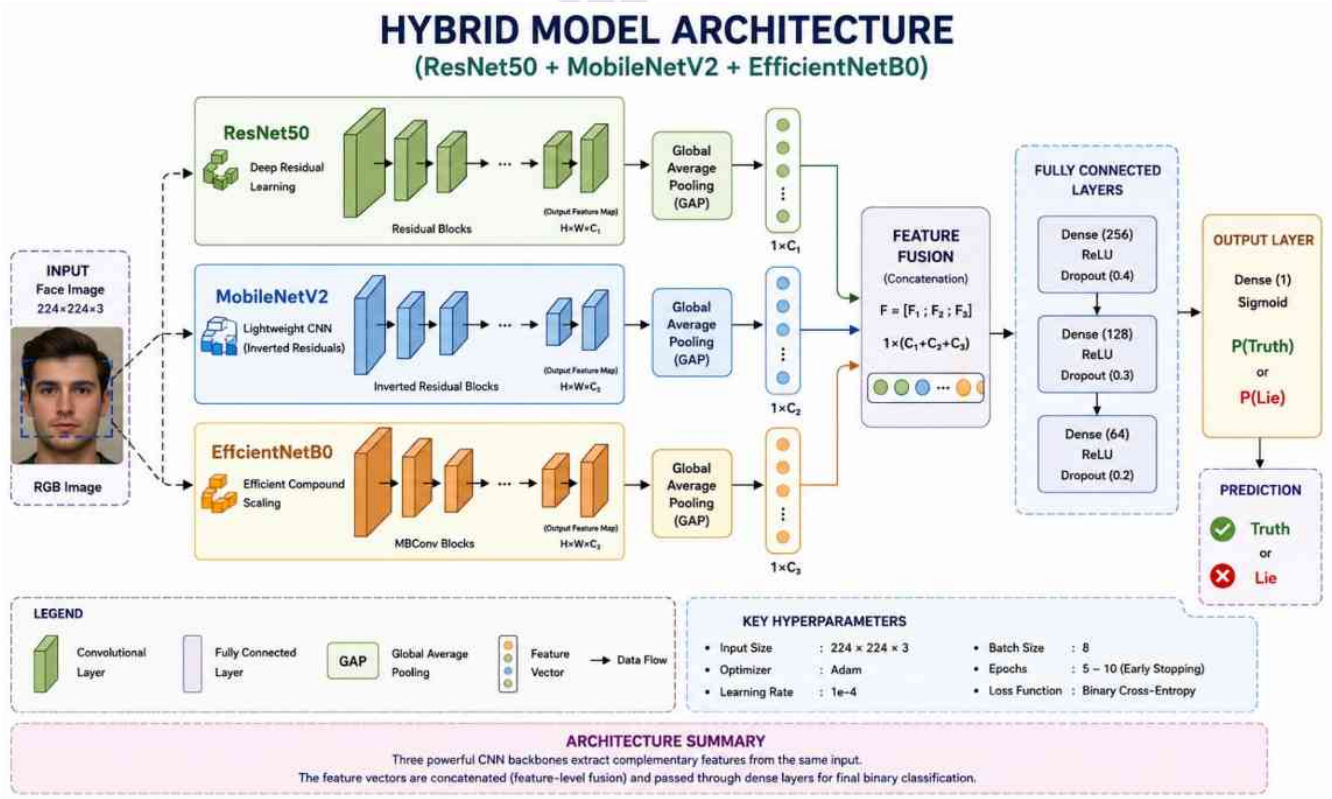


Figure 2. FacialFusionNet model architecture with ResNet50, MobileNetV2, and EfficientNetB0 branches.

3.7.1 Input Layer

The input layer receives RGB images resized to a fixed dimension, as shown in Equation (20).

$$X \in \mathbb{R}^{224\times224\times3} \tag{20}$$

3.7.2 ResNet50 Feature Extraction Layer

The first branch uses ResNet50 as a feature extractor, as expressed in Equation (21). Only the deeper layers are fine-tuned (the last 40 layers are trainable), while earlier layers remain frozen to retain pretrained ImageNet knowledge.

$$F_r = f_{\text{ResNet50}}(X) \tag{21}$$

3.7.3 MobileNetV2 Feature Extraction Layer

The second branch employs MobileNetV2 for efficient feature extraction, as expressed in Equation (22). This lightweight network captures complementary features at lower computational cost.

$$F_m = f_{\text{MobileNetV2}}(X) \tag{22}$$

3.7.4 EfficientNetB0 Feature Extraction Layer

The third branch utilizes EfficientNetB0 to extract high-performance features, as expressed in Equation (23). EfficientNetB0 balances depth, width, and resolution to achieve superior representation quality.

$$F_e = f_{\text{EfficientNetB0}}(X) \tag{23}$$

3.7.5 Global Average Pooling Layer

To reduce spatial dimensions and convert the feature maps into vectors, global average pooling (GAP) is applied to each branch, as expressed in Equation (24).

$$\begin{aligned} G_r &= \text{GAP}(F_r) \\ G_m &= \text{GAP}(F_m) \\ G_e &= \text{GAP}(F_e) \end{aligned} \tag{24}$$

3.7.6 Feature Fusion Layer

The pooled features from all three branches are concatenated to form a unified representation, as defined in Equation (25). This triple fusion combines deep, efficient, and lightweight features into a single comprehensive feature vector.

$$F_c = [G_r;\, G_m;\, G_e] \tag{25}$$

3.7.7 Fully Connected Layers

The fused feature vector is passed through dense layers to learn complex patterns. The progressive dimensionality reduction is shown in Equations (26) to (28), where ϕ denotes the ReLU activation function. Dropout regularization is applied after each dense layer to prevent overfitting, with rates of 0.4, 0.3, and 0.2, respectively.

$$H_1 = \phi\left(W_1 F_c + b_1\right), \quad \dim = 256 \tag{26}$$

$$H_2 = \phi(W_2 H_1 + b_2), \quad \dim = 128 \tag{27}$$

$$H_3 = \phi(W_3 H_2 + b_3), \quad \dim = 64 \tag{28}$$

3.7.8 Output Layer

The final classification is performed using a sigmoid activation, as defined in Equation (29), where the output lies in the interval [0, 1] and represents the probability of the input being truthful (class 1).

$$\hat{y} = \sigma(W_4 H_3 + b_4) \tag{29}$$

3.7.9 Loss Function

Binary cross-entropy, shown in Equation (30), is used to optimize the model.

$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N}\left[y_i \log(\hat{y}_i) + (1-y_i)\log(1-\hat{y}_i)\right] \tag{30}$$

3.7.10 Hyperparameter Settings

A summary of the hyper parameters used in this study can be seen in Table 5. In this table are the important hyper parameters for training including the size of the input image, optimizers parameters, the rate of learning, size of batch, epochs, and regularization methods. The hyper parameters are well-defined, because of that, the experiment can provide good results.

Table 5. Hyperparameters for the FacialFusionNet model.

Parameter	Value
Input size	224 × 224 × 3
Backbone models	ResNet50 + MobileNetV2 + EfficientNetB0
Trainable layers	Last 40 layers per backbone
Pooling	Global average pooling
Dense layers	256, 128, 64
Activation function	ReLU
Output activation	Sigmoid
Optimizer	Adam
Learning rate	1×10^{-4}
Batch size	8
Epochs	5–10 (with early stopping)
Dropout rates	0.4, 0.3, 0.2
Loss function	Binary cross-entropy (Equation 30)
Callbacks	EarlyStopping (patience = 3), ReduceLROnPlateau

3.8 Performance Evaluation

Standard classification metrics—accuracy, precision, recall, F1-score, and the confusion matrix—are used to evaluate the performance of the proposed models. These measures provide an overall assessment of how well each model differentiates between honest and dishonest classes.

3.9 Explainability Using Grad-CAM

Gradient-weighted Class Activation Mapping (Grad-CAM) is used to make the deep learning models more interpretable. The method offers visual explanations that highlight the parts of the input image that most affect the decision of the model. Grad-CAM creates a heatmap that is superimposed on the original facial image, indicating which facial regions have the greatest influence on the classification. In deception detection, the model usually places particular emphasis on the eyes, mouth, and eyebrows. This visualization assists in comprehending the decision-making process of the model and makes the predictions more trustworthy. It also enables verification that the model is learning salient facial features rather than irrelevant background characteristics.

3. RESULTS AND DISCUSSION

The performance of the proposed system is evaluated using multiple deep learning models, including ResNet50, MobileNetV2, EfficientNetB0, and the proposed FacialFusionNet model. The evaluation is conducted on the MU3D test dataset using standard classification metrics.

4.1 Frame Extraction and Data Distribution

Frames obtained from the MU3D dataset undergo visual checks to ensure the validity of the preprocessing pipeline. In order to extract a good representation of facial expressions that take place in the video over time, the video content is transformed into frames. Frames are kept separately for truth and lie labels, ensuring label consistency in the dataset. The frames extracted are of a quality suitable for establishing deep learning models and are displayed in Figure 3. The balance of the dataset is analyzed to facilitate fair training. A chart that represents the number of samples of the truth and lie classes in training, validation, and test data splits is shown in

Figure 4. This is done to confirm that there is no bias in model training.

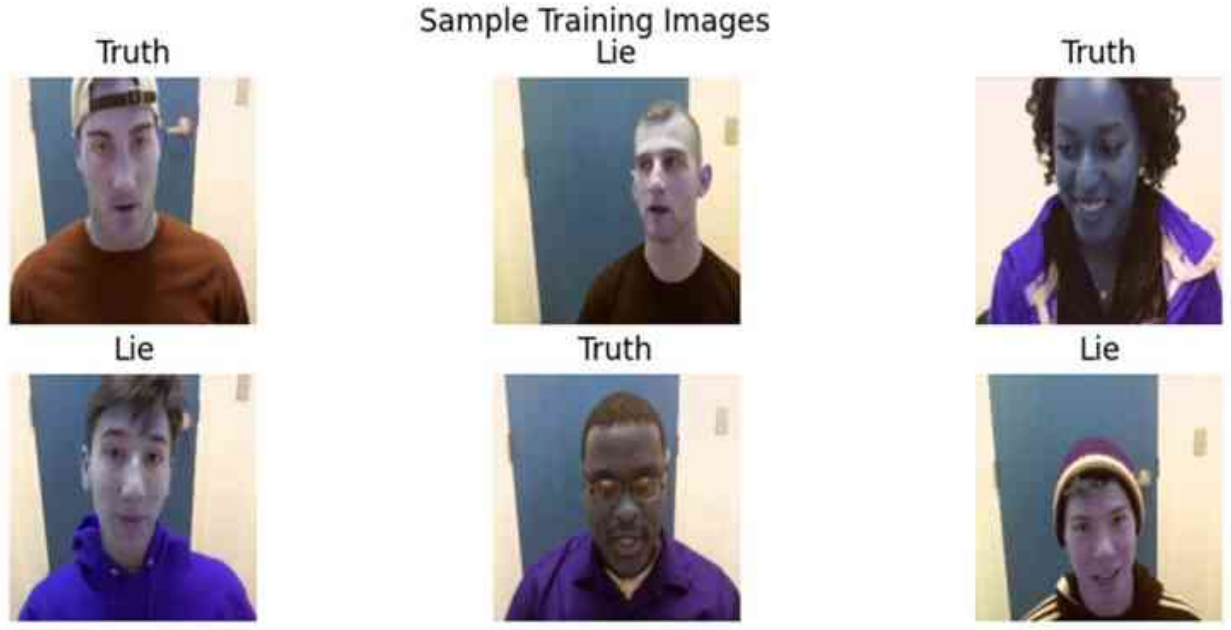


Figure 3. Extracted frames of the MU3D dataset.

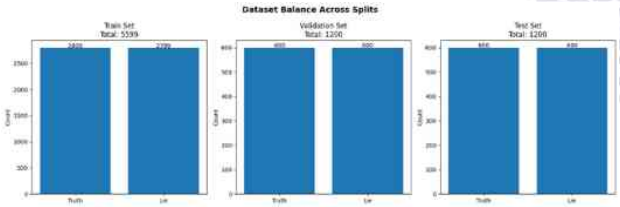


Figure 4. Dataset distribution across training, validation, and test splits.

4.2 Model-wise Performance Evaluation

Each model is evaluated separately using confusion matrices and classification reports. The results highlight differences in performance across architectures.

4.2.1 ResNet50 Results

ResNet50 exhibits effective performance in terms of identifying truthful and untruthful expressions. The evaluation against test dataset shows excellent and consistent performance for both classes. The results show that the model achieves the accuracy of 96%. Such an outcome indicates the successful feature learning and generalization by the model. The two classes (lie and truth) are classified similarly, hence showing that the model does not show significant biased performance towards either class. The classification report can be found in Table 6 while the confusion matrix can be found in Figure 5.

Table 6. Classification report of the ResNet50 model.

Class	F1-Score	Precision	Recall
Lie	96%	97%	96%
Truth	96%	96%	97%
Accuracy	**96%**	–	–
Macro avg	96%	96%	96%
Weighted avg	96%	96%	96%

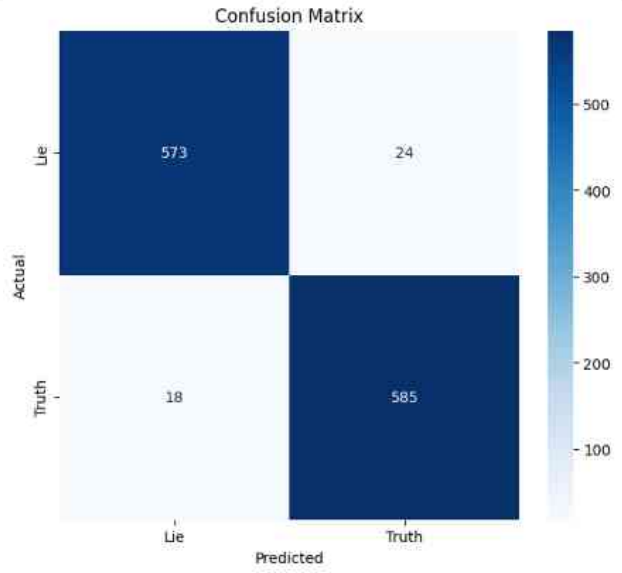


Figure 5. Confusion matrix of the ResNet50 model.

4.2.2 MobileNetV2 Results

As presented in Table 7 and Figure 6, the mobileNetV2 achieves an accuracy of 95%, which is an efficient performance with reduced computational complexity. Overall, the model is well balanced between the two classes, with slightly reduced recall values suggesting that occasional misclassifications occur with subtle deceptive expressions. Even though it is lightweight, MobileNetV2 yields stable performance for resource-constrained environments.

Table 7. Classification report of the MobileNetV2 model.

Class	F1-Score	Precision	Recall
Lie	94%	95%	94%
Truth	95%	94%	95%
Accuracy	**95%**	–	–
Macro avg	95%	95%	95%
Weighted avg	95%	95%	95%

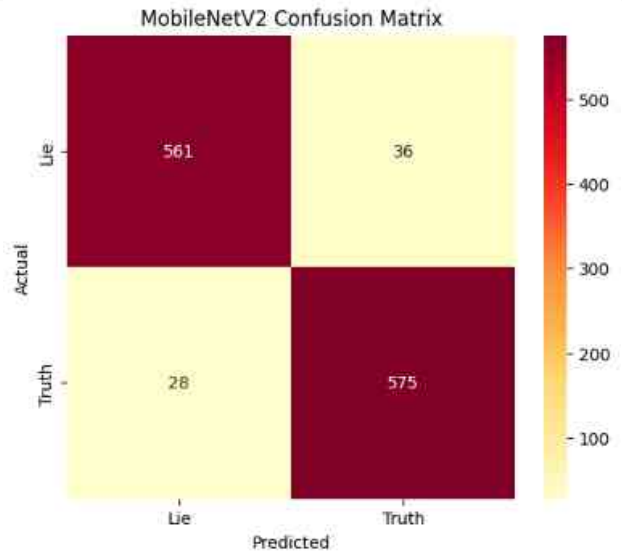


Figure 6. Confusion matrix of the MobileNetV2 model.

4.2.3 EfficientNetB0 Results

The confusion matrix shown in Figure 7 and the accuracy results from Table 8 demonstrate that EfficientNetB0 provides a good balance between accuracy and efficiency, with an accuracy of 96%. The model has better generalization performance than MobileNetV2 and slightly better stability for classification. It adopts a compound scale transformation, which achieves better feature representation than common CNN architectures.

Table 8. Classification report of the EfficientNetB0 model.

Class	F1-Score	Precision	Recall
Lie	96%	96%	95%
Truth	96%	95%	96%
Accuracy	**96%**	–	–
Macro avg	96%	96%	96%
Weighted avg	96%	96%	96%

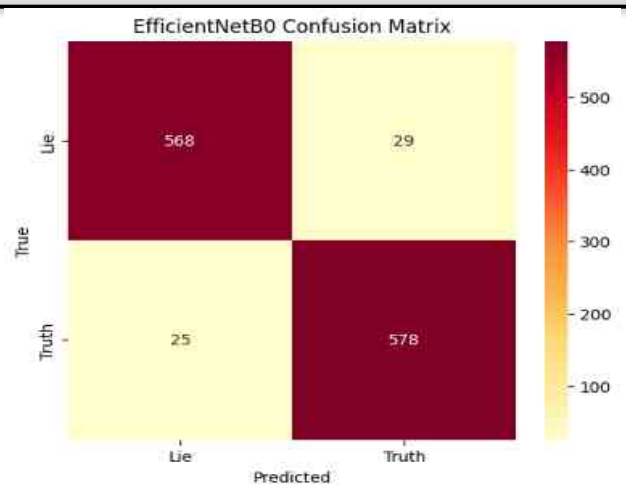


Figure 7. Confusion matrix of the EfficientNetB0 model.

4.3 FacialFusionNet Model Results

The accuracy of proposed FacialFusionNet model is best of all individual architecture with 99%. The F1-scores of 99% in both classes for lie and truth represent high balanced and reliable classification performance as seen in Table 9 and the confusion matrix in Figure 8. The model's better performance is said to be due to feature-level fusion of ResNet50, EfficientNetB0 and MobileNetV2, enabling it to learn deep semantic features and light spatial representations. This combination helps to substantially mitigate misclassification errors and improve robustness.

Table 9. Classification report of the proposed FacialFusionNet model on the MU3D test dataset.

Class	F1-Score	Precision	Recall
Lie	99%	99%	99%
Truth	99%	99%	99%
Accuracy	**99%**	–	–
Macro avg	99%	99%	99%
Weighted avg	99%	99%	99%

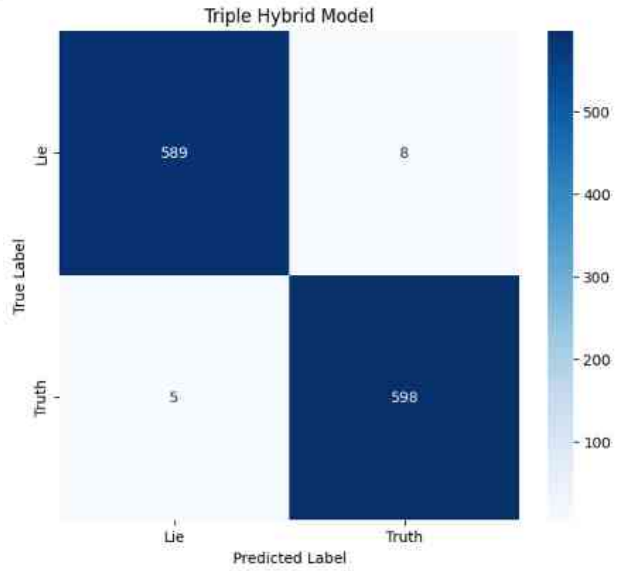


Figure 8. Confusion matrix of the proposed FacialFusionNet model on the MU3D test dataset.

4.4 Comparative Analysis

A comparative evaluation shows that the hybrid model consistently outperforms the individual architectures in terms of accuracy and F1-score. This confirms that feature fusion improves the ability of the model to capture subtle facial expression variations associated with deception.

4.5 Grad-CAM Explainability

Grad-CAM visualizations are used to interpret model predictions. The heatmaps indicate that the models focus on key facial regions such as the eyes, mouth, and eyebrows. The hybrid model shows more consistent and refined attention, indicating better feature learning.

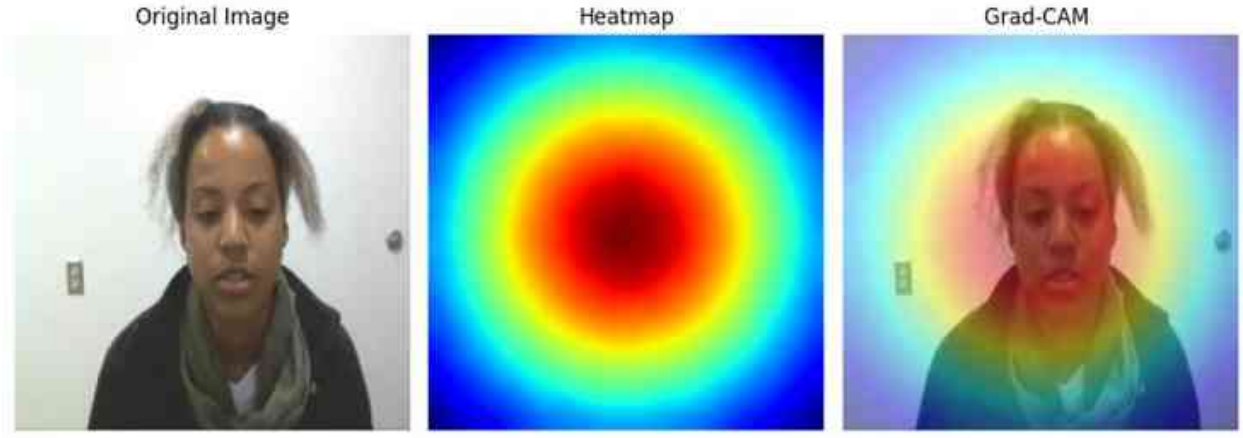


Figure 9. Grad-CAM heatmap of the MU3D test dataset.

4.6 Discussion

The experimental results prove that deep learning models are highly effective in detecting deception based on facial expressions. Of the tested models, the proposed hybrid architecture delivers the highest performance with an accuracy of 99%, which exceeds that of ResNet50 and EfficientNetB0 (96%) and MobileNetV2 (95%).

The high performance of the hybrid model can be explained by the fact that it integrates complementary

feature representations. ResNet50 extracts deep hierarchical features, while MobileNetV2 adds lightweight and efficient spatial features and EfficientNetB0 contributes compound-scaled representations. Their combination leads to greater discrimination between truthful and deceptive expressions, especially in the difficult case of minute changes in facial expression.

The results of ResNet50 and EfficientNetB0 are similar, which suggests that deeper architectures and compound scaling approaches are both effective in feature extraction. Both models, however, have minor limitations in resolving fine-grained differences compared with the hybrid approach. MobileNetV2 is computationally efficient but attains a lower accuracy level because of its reduced model complexity.

Grad-CAM visualizations further illustrate these results by revealing that all models centre their attention on pertinent parts of the face, including the eyes, the mouth, and the eyebrows. The hybrid model generates more stable and localized attention maps, suggesting better feature learning and more reliable decision making.

4.7 Ablation Study

To analyze the contribution of the different components in the proposed framework, an ablation study is conducted by comparing the individual models with the hybrid architecture, as shown in Table 10.

Table 10. Ablation study results.

Model	Accuracy
MobileNetV2	95%
ResNet50	96%
EfficientNetB0	96%
FacialFusionNet model	**99%**

The results highlight the following observations:

- Removing feature fusion (i.e., using single models) results in reduced performance.
- The combination of ResNet50, EfficientNetB0, and MobileNetV2 improves accuracy by 2–3%, confirming the effectiveness of feature-level fusion.
- Lightweight models alone are less effective in capturing complex facial cues compared with deeper or hybrid architectures.
- The FacialFusionNet model consistently achieves higher precision and recall across both classes, indicating improved robustness.

4.8 State-of-the-Art Comparison

Recent studies published between 2022 and 2025 have focused on automated deception detection using facial cues, micro-expressions, and hybrid deep learning architectures, as shown in Table 11.

Table 11. State-of-the-art comparison.

Study / Year	Approach / Modality	Accuracy	Comparison with current work
[34] (2022)	Facial muscle movement analysis	87%	Limited computational modelling; the current work uses hybrid CNN fusion.
[33] (2023)	Facial + physiological data	88%	Multimodal approach, but lower accuracy and no multi-stream feature fusion.
[32] (2023)	Deep learning on facial cues	90%	Single-stream CNN model; the current work employs hybrid fusion for richer feature extraction.

Study / Year	Approach / Modality	Accuracy	Comparison with current work
[9] (2024)	Hybrid deep learning with micro-expressions	91%	Comparable performance; the current work additionally uses triple-branch fusion and Grad-CAM visualization.
[35] (2025)	Micro-expression + gaze + pose + feature selection	92%	Robust multimodal features; the current work achieves improved performance using hybrid CNN embeddings.
Our work	**Facial expression-based FacialFusionNet CNN (ResNet50 + EfficientNetB0 + MobileNetV2) using extracted video frames with Grad-CAM**	**99%**	**Utilizes video-to-frame transformation and hybrid feature fusion to capture spatial facial cues with enhanced interpretability.**

IEER
Institute for Excellence in Education & Research

4. CONCLUSION

In this paper, a deep-learning model for lie detection via facial expressions extracted from the MU3D dataset was introduced. In this process, an end-to-end pipeline was designed and implemented, which included video frame extraction, video frame pre-processing, data augmentation, and dataset balancing. Three CNN architectures (ResNet50, EfficientNetB0, and MobileNetV2) were developed and evaluated. Furthermore, FacialFusionNet architecture, combining ResNet50, EfficientNetB0, and MobileNetV2, was introduced to benefit both the deep feature extraction and computational efficiency. The experimental results revealed that the proposed FacialFusionNet model achieves better accuracy, precision, recall and F1-score than the individual models. This improved performance underlines the effectiveness of combining several architectures to capture the complex facial patterns related to deceptive behaviour. Moreover, the model was made more explainable and reliable by using Grad-CAM to provide insight into the decision making process of the model, which highlighted important parts of the face including the eyes and mouth. To conclude, this study validates the feasibility of automated deception detection with deep learning techniques with respect to facial expressions. The proposed approach is well suited to real-world implementation as it provides a trade-off between accuracy, efficiency, and interpretability. However, additional improvements can be realized with the incorporation of multimodal information, temporal component and improved attention mechanisms. Further studies along these lines can result in more powerful and scalable lie detection systems.

REFERENCES

[1] H. Alaskar, Z. Sbaï, W. Khan, A. Hussain, and A. Alrawais, Intelligent techniques for deception detection: A survey and critical study. *Soft Computing, 27*(7), 3581–3600, 2023.

[2] M. Zloteanu, Reconsidering facial expressions and deception detection. *Handbook of Facial Expression and Emotion, 3*, 238–284, 2020.

[3] F. Herbig, The 'psychology' of polygraph: Engendering differential salience—Concerns and caveats. *Journal of Psychological Research, 2*(2), 1–10, 2020.

[4] J. Sánchez-Monedero and L. Dencik, The politics of deceptive borders: 'Biomarkers of deceit' and the case of iBorderCtrl. *Information, Communication & Society, 25*(3), 413–430, 2022.

[5] K. Mattes, V. Popova, and J. R. Evans, Deception detection in politics: Can voters tell when politicians are lying? *Political Behavior, 45*(1), 395–418, 2023.

[6] J. Weimann and J. Brosig-Koch, *Methods in experimental economics*. Springer, 2019.

[7] U. Fischbacher and F. Föllmi-Heusi, Lies in disguise: An experimental study on cheating. *Journal of the European Economic Association, 11*(3), 525–547, 2013.

[8] J. Li, H. Chen, and M. Zhao, Deception detection based on micro-expression and feature selection methods. *EURASIP Journal on Image and Video Processing, 2025*(12), 1–17, 2025.

[9] A. Singh, R. Kumar, and V. Gupta, A study on the accuracy of micro-expression based deception detection with hybrid deep neural network models. *European Journal of Electrical and Computer Engineering, 12*(1), 50–63, 2024.

[10] V. Gupta, M. Agarwal, M. Arora, T. Chakraborty, R. Singh, and M. Vatsa, Bag-of-Lies: A multimodal dataset for deception detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2019.

[11] E. P. Lloyd, J. C. Deska, K. Hugenberg, A. R. McConnell, B. Humphrey, and J. W. Kunstman, Miami University Deception Detection Database. *Behavior Research Methods, 51*(1), 429–439, 2019.

[12] M. Ali, M. F. Mushtaq, U. Akram, N. A. Samee, M. M. Jamjoom, and I. Ashraf, Detecting vehicle mechanical defects using an ensemble deep learning model

with Mel frequency cepstral coefficients from acoustic data. *CMES–Computer Modeling in Engineering & Sciences, 145*(2), 1863–1901, 2025.

[13] U. Farooq, S. N. Khosa, U. Akram, and M. F. Mushtaq, Intelligent electrical fault detection and classification using ensemble deep learning architecture. *Spectrum of Engineering Sciences, 4*(5), 2474–2493, 2026.

[14] S. N. Khosa, N. Kiran, H. Arshad, and U. Akram, Deep learning-based bone age growth disease detection (BAGDD) using RSNA radiographs. *International Journal of Advanced Computer Science and Applications, 16*(6), 2025.

[15] B. Mughal, and A. Buriro, Classification of breast lesions in combination with metamorphic segmentation and saliency feature block. In *International Conference on Intelligent Technologies and Applications* (pp. 573–580). Springer, 2019.

[16] T. B. Tariq, S. N. Khosa, M. Z. Hadi, T. Kiran, M. Mansab, U. Akram, and M. F. Mushtaq, Intelligent cyber security framework for threat detection using ensemble learning techniques. *Journal of Computing & Biomedical Informatics, 8*(2), 2025.

[17] M. Mansab, T. B. Tariq, and U. Akram, An improved fraud detection system using the LDX ensemble machine learning technique. *Technical Journal, 30*(2), 31–39, 2025.

[18] M. Ali, U. Akram, D. S. Hassan, N. A. Samee, and I. Ashraf, Improving prediction accuracy of emotions from X (Twitter) post opinions with enhanced hierarchical attention model. *International Journal of Computational Intelligence Systems, 19*(1), 203, 2026.

[19] M. M. Iqbal, W. Mehmood, S. Ramzan, and M. F. Mushtaq, Leveraging deep learning for sentiment analysis: Insights from academic and general feedback across dual datasets. *Multimedia Tools and Applications*, 1–27, 2025.

[20] M. Ahmad, N. Wahid, A. Mehmood, G. S. Choi, R. A. Hamid, M. F. Mushtaq, and S. S. Saharuddin, Mental health app reviews analyzer (MHARA) using logistic regression and tri-gram. In *International Conference on Soft Computing and Data Mining* (pp. 274–282). Springer, 2019.

[21] A. A. Ghali, R. Ahmad, and H. Alhussian, A framework for mitigating DDoS and DoS attacks in IoT environment using hybrid approach. *Electronics, 10*(11), 1282, 2021.

[22] A. Srivastava, S. Tiwari, B. S. Rawat, and S. A. Dhondiyal, DDoS attacks detection in IoT networks using Naive Bayes and Random Forest. In *2024 3rd International Conference on Applied Artificial Intelligence and Computing (ICAAIC)* (pp. 1609–1612). IEEE, 2024.

[23] A. A. Ghali, M. H. Abdullah, A. G. Jaafar, and N. Talpur, A review: Internet of Things (IoT) security challenges. *Open International Journal of Informatics, 14*(1), 1–20, 2026.

[24] C. H. Yuet, C. J. Zhi, C. J. Kai, C. S. Fong, and A. A. Ghali, A new method CZ4 cipher: Your next innovation. *The Metascience, 2*(1), 42–63, 2024.

[25] A. Arwind, L. Chandra, C. K. Xuan, G. J. Qing, A. Vickram, L. Ravi, and A. A. Ghali, Custom cipher—Reverse Alpha Adding Cipher (RAAC). *The Metascience, 2*(1), 22–29, 2024.

[26] X. Ru, C. K. Yann, T. W. En, C. Ziqi, and A. A. Ghali, A novel SecurX algorithm. *The Metascience, 2*(1), 30–41, 2024.

[27] C. R. Genn, L. L. Chih, T. E. Qi, and A. A. Ghali, The Grand Diversion Cipher: A proposed information security protection mechanism and cryptography. *The Metascience, 2*(1), 64–75, 2024.

[28] C. M. Lee, B. K. Ang, S. L. Kok, A. A. Ghali, and N. A. Sabri, Rogue access point detection and tracking system using trilateration algorithm, n.d..

[29] H. U. Siddiqui, R. Brown, S. Dudley, and M. F. Mushtaq, Gas metering using optical sensor. In *2019 IEEE 6th International Conference on Engineering Technologies and Applied Sciences (ICETAS)* (pp. 1–5). IEEE, 2019.

[30] R. Kaban, Advancing energy-efficient technologies in the Internet of Things (IoT): A comprehensive analysis. *JCEIT: Journal of Computer Engineering and Information Technology, 1*(2), 2025.

[31] P. K. A. Vasu, J. Gabriel, J. Zhu, O. Tuzel, and A. Ranjan, *An improved one millisecond mobile backbone.* arXiv:2206.04040, 2022.

[32] FacialCueNet., FacialCueNet: Unmasking deception—An interpretable model for criminal interrogation using facial expressions. *Applied Intelligence, 53*(6), 7451–7465, 2023.

[33] X. Zhang, Y. Li, and H. Wang, Detecting deception using machine learning with facial expressions and pulse rate. *Journal of Intelligent Systems, 35*(2), 102–115, 2023.

[34] N. Alsharari, Q. Zhang, and R. Khan, Intentional-deception detection based on facial muscle movements in an interactive social context. *Journal of Behavioral Studies, 14*(3), 215–230, 2022.

[35] S. Yuan, Z. Shao, Z. Ma, T. Cao, H. Xing, Y. Liu, and Y. Cao, Deception detection based on micro-expression and feature selection methods. *EURASIP Journal on Image and Video Processing, 2025*(1), 8, 2025.

[36] D. Avola, L. Cinque, M. De Marsico, A. Fagioli, and G. L. Foresti, LieToMe: Preliminary study on hand gestures for deception detection via Fisher-LSTM. *Pattern Recognition Letters, 138*, 455–461, 2020.

[37] E. Brynjolfsson, C. Wang, and X. Zhang, The economics of IT and digitization: Eight questions for research. *MIS Quarterly, 45*(1), 473–477, 2021.

[38] N. R. Butski, J. R. E. Boostrom, and R. Bonds, An honest discussion about lies: How components of a seller's lies can impact customer relationship commitment. *Marketing Management Journal, 32*(1), 17, 2022.

[39] C. F. Camerer, A. Dreber, and M. Johannesson, Replication and other practices for improving scientific quality in experimental economics. In *Handbook of research methods and applications in experimental economics* (pp. 83–102). Edward Elgar Publishing, 2019.

[40] H.-W. Hsu, T.-Y. Wu, S. Wan, W. H. Wong, and C.-Y. Lee, QuatNet: Quaternion-based head pose estimation with multiregression loss. *IEEE Transactions on Multimedia, 21*(4), 1035–1046, 2019.

[41] B. Huang, R. Chen, W. Xu, and Q. Zhou, Improving head pose estimation using two-stage ensembles with top-k regression. *Image and Vision Computing, 93*, 103827, 2020.

[42] M. Junger, L. Koning, P. Hartel, and B. Veldkamp, In their own words: Deception detection by victims and near victims of fraud. *Frontiers in Psychology, 14*, 1135369, 2023.

[43] N. Ruiz, E. Chong, and J. M. Rehg, Fine-grained head pose estimation without keypoints. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops* (pp. 2155–2164), 2018.

[44] R. Saini and P. Rani, LDM: A systematic review on lie detection methodologies, 2022.

[45] M. Schepisi, G. Porciello, S. M. Aglioti, and M. S. Panasiti, Oculomotor behavior tracks the effect of ideological priming on deception. *Scientific Reports, 10*(1), 9555, 2020.

[46] J. B. Taylor, S. R. Beach, and N. Kontos, The therapeutic discharge: An approach to dealing with deceptive patients. *General Hospital Psychiatry, 46*, 74–78, 2017.

[47] P. Werner, A. Al-Hamadi, R. Niese, S. Walter, S. Gruss, and H. C. Traue, Towards pain monitoring: Facial expression, head pose, a new database, an automatic system and remaining challenges. In *Proceedings of the British Machine Vision Conference* (pp. 1–13), 2013.

[48] P. Werner, F. Saxen, and A. Al-Hamadi, Landmark based head pose estimation benchmark and method. In *IEEE International Conference on Image Processing (ICIP)* (pp. 3909–3913), 2017.

[49] Y. Xiong, H. J. Kim, and V. Singh, Mixed effects neural networks (MeNets) with applications to gaze estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 7743–7752), 2019.

[50] X. Zhang, Y. Sugano, M. Fritz, and A. Bulling, MPIIGaze: Real-world dataset and deep appearance-based gaze estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 41*(1), 162–175, 2017.

IEER
Institute for Excellence in Education & Research