BAREKENG: Journal of Mathematics and Its Applications
September 2026 Volume 20 Issue 3 Page 2561-2574
P-ISSN: 1978-7227 E-ISSN: 2615-3017
https://doi.org/10.30598/barekengvol20iss3pp2561-2574

# OPTIMIZATION OF LIE DETECTION WITH DEEP LEARNING APPROACH USING FUSION METHOD

**Dewi Kusumawati** [1*], **Fitriyanti Andi Masse** [2], **Wulan** [3]

[1,2,3] *Informatics Departement, STMIK Bina Mulia Palu*
*Jln. Letjen Soeprapto, Palu, 94112, Indonesia*

*Corresponding author's e-mail: * **dewikusumawati@binamulia.ac.id***

***Article Info***

***Article History:***

*Received: 15th September 2025*
*Revised: 14th January 2026*
*Accepted: 17th March 2026*
*Published:8th April 2026*

***Keywords:***

*Accuracy;*
*Clasification;*
*Evaluation;*
*Fusion;*
*Lie detection.*

***ABSTRACT***

*In lie detection, early fusion methods that combine information from multiple modalities, such as images and sounds, are used. To improve performance, a lie detection system is designed using mean fusion techniques. The feature extraction method, which uses Optical Flow (OF) and GaussianBlur, uses image data as input. This process generates facial feature change data as numeric values, enabling more efficient processing and allowing the model to be trained quickly and effectively. Evaluation of the model with accuracy, precision, recall, and F1 score using 10 (Fold) cross-validation using a Convolutional Neural Network (CNN) architecture to find features associated with lying in visual content. At the same time, voice signals are studied through voice signal processing and voice feature extraction methods using Mel Frequency Cepstral Coefficients (MFCC) feature extraction and classification using Mel Frequency Cepstral Coefficients (LSTM). The purpose of this process is to discover lying patterns through the audio module. The mean fusion model combines the decisions of multiple lie detection models for each modality, enabling the system to leverage the strengths of each modality to create a broader feature representation. The dataset used contains images, and voice is used for performance evaluation. This dataset can show various lying situations and contexts. The experimental results show that the fusion method using the mean fusion model achieves a lie detection accuracy of 99% and an F1-Score of 0.99. In the context of lying, this research helps develop a more comprehensive and reliable lie-detection system model. The main contribution of this work is a measurable multimodal fusion strategy that integrates pupil-based facial landmarks and temporal voice features, yielding an accuracy improvement of over 14% compared to unimodal baselines.*



***How to cite this article:***

D. Kusumawati, F. A. Masse and Wulan., "OPTIMIZATION OF LIE DETECTION WITH DEEP LEARNING APPROACH USING FUSION METHOD", *BAREKENG: J. Math. & App.*, vol. 20, no. 3, pp. 2561-2574, Sep, 2026.


Journal homepage: *https://ojs3.unpatti.ac.id/index.php/barekeng/*
Journal e-mail: *barekeng.math@yahoo.com; barekeng.journal@mail.unpatti.ac.id*

**Research Article · Open Access**

## 1. INTRODUCTION

In daily communication, people usually decide whether to lie or tell the truth on their own accord [1]. This is also the case during a trial to determine whether the defendant is guilty of the charges brought against them, or whether their testimony is true or false in general. In this process, judges may pay attention to non-verbal cues, such as the facial expressions of those giving testimony [2]. Likewise, verbal cues affect how strong the voice signal produced when lying is; each word in a lying statement can show changes in the voice signal [3]- [4]. Lie detection can be a complementary tool for forensic investigators to enhance their decisions regarding the credibility of statements or lie detection[5]. Polygraph is a traditional method widely used to detect lies based on physiological changes in a person, such as skin conductivity, heart rate, blood pressure, breathing, and pulse [6]. Eye-tracking technology, which records eye movements, has attracted attention as a viable alternative[7]. Different eye movement markers are used in various lie detection technologies, and the accuracy obtained is 78%[8]. Changes in pupil size are also an indicator of emotional arousal and medical conditions. The sympathetic system controls pupillary dilation by activating the radial dilator muscle, where decreased sympathetic activity will cause a decrease in pupil diameter. Conversely, the parasympathetic system regulates pupillary constriction through the iris sphincter muscle as a reflex response to light, with efferent pathways originating from the Edinger-Westphal complex in the oculomotor nucleus [9]. Subjects who lied showed greater pupil dilation, more frequent speech interruptions, higher voice pitch, and higher blink rates [10]. CNN is a method for classifying images [11], which consists of an input layer, a hidden layer and an output layer[12]. In the CNN-based lie detection method, the processed data is an image extracted from a video using the VGG 16 and VGG 19 architectures [13]. Health Domain Multimodal Rumor Detection Neural Network that combines various fusion techniques to detect health-related rumors with high accuracy [14].

The study [1] used a relatively small sample size of 32 respondents, which may limit the range of detectable differences between truth and Lie. The sample size is not large enough to generalize broadly [2], limiting participants' natural expression and potentially affecting the authenticity of their responses. Polygraph deception detection can be manipulated by individuals skilled enough to maintain normal physiological functioning, resulting in low accuracy and inadmissibility in court [6]. Lie detection systems that focus solely on eye movements and gaze achieve only 78% accuracy using the SVM method, and are limited in their multidimensional analysis, making it difficult to distinguish between honest and false behavior. Dilated pupils are an indication of lying, while non-dilated pupils are considered neutral [9], but many other factors can cause pupil dilation, such as lighting, emotional state, or medical conditions, so accurate data is needed to see changes that occur in the pupil area to determine whether someone is honest or lying. Using a single dataset type (video converted to images) can limit data variation and model generalization. Processing and combining image data from multiple sources can increase computational complexity and require greater computational resources. The aim of this research is to address the problems identified in previously conducted research and to develop a lie detection model that uses a combination of facial and voice features, employing a fusion method. The datasets used for model generalization are heterogeneous, sourced from public datasets and from datasets built with scenarios we have designed. Multimodal fusion (combining image and voice) has the advantage of leveraging complementary information from both modalities, increasing robustness to noise in one modality, and allowing the model to learn correlations across multiple modalities. In this research, we implemented a fusion strategy to integrate visual and acoustic representations, enabling the system to utilize temporal and spatial features simultaneously. The multimodal approach yielded better and more stable performance than models that used only image or audio modalities. The contribution of this research is a CNN-based lie-detection classification model built on MobileNet that converts image data into numerical features to reduce computational time and capture temporal changes in each video frame. The second contribution is the use of late fusion to combine facial and voice features, making the model more robust at detecting lies across modalities. Apart from the fusion strategy, we use a CNN mobilenet network for face with feature extraction using optical flow and pupil diameter change algorithms, and an LSTM for voice with feature extraction using MFCC.

## 2. RESEARCH METHODS

Fig. 1 shows a block diagram of the implementation of this research model, starting with video input, then searching for key points based on pupil changes, and then voice input. The facial and voice data are

processed separately to determine the performance of each feature, which will ultimately be combined with a fusion model.

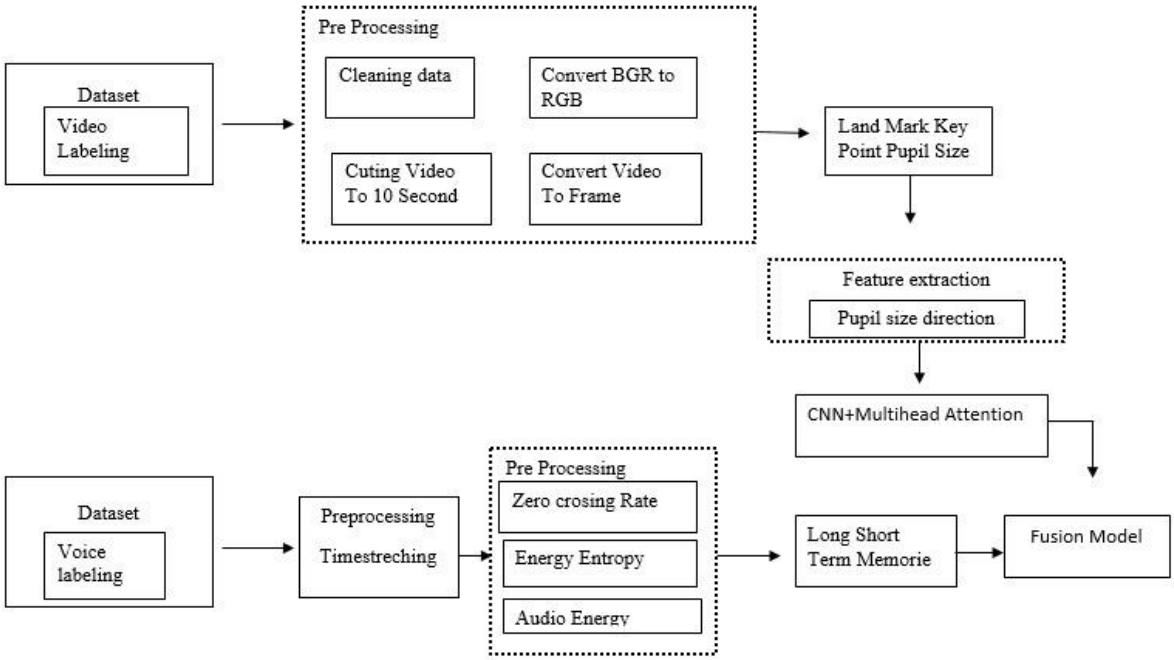


**Figure 1**. Research Process Design

## 2.1 Data Collection and Ground Truth Annotation

The present study combined covert examination monitoring with structured post-examination interviews. First, the examination room was equipped with multiple synchronized CCTV cameras, including top- and frontal-view configurations, to record students' behavior in real time from different angles. All cameras were synchronized using a unified time reference at least one hour prior to each session to ensure temporal consistency across all recordings. During the examination, several students were observed engaging in collaborative behavior while answering questions, as captured on CCTV footage. After the examination, students identified as having collaborated with others were individually invited to participate in structured interviews. In addition, a comparable number of students who completed the examination independently were recruited as a control group. In total, 25 students participated in the interview sessions (20 males and 5 females from the collaboration group, and 25 students from the independent-working group). Each interview lasted approximately 5–10 minutes and was recorded with a concealed webcam to minimize behavioral alteration from awareness of being observed.

The interview recordings were manually aligned with the corresponding CCTV footage. Because the interview questions explicitly referred to events that occurred during the examination, each student's verbal responses could be temporally matched to the relevant segments in the CCTV recordings. For instance, when a student discussed a specific examination question, the corresponding event was located within the CCTV timeline by matching timestamps and identifiable behaviors. This cross-checking process enabled a direct comparison between interview statements and objectively recorded examination behaviors, despite differences in camera viewpoints. Timestamp synchronization and visible student identifiers facilitated this alignment, ensuring that each claim could be verified against objective video evidence. To ensure high data quality, all recording equipment was tested and calibrated prior to each session. Camera focus, resolution, and audio levels were verified, and recordings were conducted continuously without intentional interruptions. Video data inherently support repeated review, thereby enhancing reproducibility. In line with best practices for behavioral recording, multiple cameras were deployed to capture different aspects of the scene, with overlapping camera views to record shared areas simultaneously from multiple viewpoints. Trained psychologists and coders were instructed to interpret partial or occluded views accordingly. Recording consistency was periodically assessed by replaying random samples to ensure that no frames were missing. As a result, the video data provides a reliable and verifiable record of both examination and interview sessions.

The study participants were undergraduate students aged approximately 18–24 years. In total, recordings were obtained from 36 male and 21 female students. Data were partitioned subject-wise, with all recordings from a given participant grouped and assigned exclusively to either the training set (approximately 80% of subjects) or the test set (20%). This subject-independent split ensured that no individual's facial or behavioral patterns appeared in both sets, thereby reducing the risk of overfitting to subject-specific characteristics. The final label distribution in the dataset consisted of 20 truthful cases and 5 lying cases, as determined through cross-verification of CCTV evidence and interview responses. In addition to data from the examination scenario, the dataset included recordings from investigative sessions involving suspects who were evaluated using polygraph examinations. These recordings provided supplementary data in which the validity of truthful and lying labels was supported by polygraph-based assessments. Our research also used public datasets as a comparison to examine changes in facial features between honest and lying labels.

Each interview recording was analyzed by a licensed psychologist with expertise in behavioral analysis. The psychologist reviewed both visual and audio streams to identify changes in facial expressions and variations in vocal tone that may indicate lying behavior. The analysis followed an established facial behavior coding scheme. As annotations were performed by a single expert, we acknowledge the potential for subjectivity. To mitigate this limitation, expert judgments were verified against CCTV recordings; any discrepancies between interview statements and recorded examination behaviors were resolved through joint re-examination of both video sources. In future work, multiple independent raters will be involved to enable explicit measurement of inter-rater agreement (e.g., using Cohen's kappa). In the present study, the psychologist's annotations served as the primary labels, while CCTV recordings functioned as an objective ground truth reference for verifying the veracity of each claim. This annotation strategy is consistent with prior lie detection research, which has employed manual annotation of facial action units, gaze direction, head movements, and mouth motion dynamics. Building on these approaches, the present study further incorporates key facial and verbal cues extracted from each interview.

### 2.2 Pre Processing

The preprocessing stages, as shown in Figure 1, are a crucial part of the data analysis process, aiming to clean and prepare raw data for further use. Data used for analysis must be accurate, complete, and consistent before processing. Here are some of the main functions of data preprocessing:

1. Data Cleaning: Eliminating or correcting corrupted, incomplete, or inaccurate data. This includes filling in missing values, correcting entry errors, and eliminating duplicate data.
2. Resolution and Frame Format Unification
3. Video Cropping
4. BGR to RGB Conversion.
5. Using a time-stretching approach.

### 2.3 Landmark

Facial landmarks are important points on the human face used to recognize and indicate the location of facial and hand features [15]. For this study, as shown in Figure 1, we will search for landmarks in the pupil area. To find facial landmarks using the facemesh from MediaPipe, in this study, the landmarks found for the pupil area are as follows:

LEFT_PUPIL = [474, 475, 476, 477]

RIGHT_PUPIL = [469, 470, 471, 472].

### 2.4 Feature Extraction

In this research, we developed a calculation to monitor changes in pupil size from frame to frame. This change is determined based on a set threshold. The pupil is considered enlarged based on the set threshold and the average change is calculated. Based on the points identified in the previous step within the facial landmarks of the face mesh, we then calculate the pupil circle to determine the pupil size in each frame. Fig. 2 below is the pseudocode. To find the change value:

// For each pupil (left and right):

```
(l_cx, l_cy) is the center of the left circle
l_radius is the radius of the left circle
(r_cx, r_cy) is the center of the right circle
r_radius is the radius of the right circle
Convert (l_cx, l_cy) to an integer array and store it as center_left
Convert (r_cx, r_cy) to an integer array and store it as center_right
Draw a circle on the frame with the center at center_left and radius l_radius
Draw a circle on the frame with the center at center_right and radius r_radius
```

**Figure 2**. Pseudocode to Look for Changes in Pupil Size

Voice Extraction Using MFCC

1. Pre emphasis to lower noise and enhance speech signal clarity.
2. Windowing: A series of tiny, overlapping frames are created from the audio signal. To lessen edge effects, a window (such as a Hamming window) is multiplied by each audio frame. The frames in this study were separated at 4-second intervals.
3. Next, a Fourier transform is used to determine each time frame's power spectrum.
4. To accommodate the properties of human hearing, the power spectrum is subsequently transformed to the Mel scale.
5. Cepstral coefficients to depict the speech signal's properties at each time interval.

Fig. 3 below is a source fragment in the MFCC feature extraction process, with a central coefficient of 20.

```
Audio extraction using MFCC
import librosa
import numpy as np
# Load audio file
audio_path = 'path_to_audio_file.wav'
audio, sr = librosa.load(audio_path, sr=None)
# Extract MFCC while preserving a specific number of cepstral coefficients
n_mfcc = 20
mfccs = librosa.feature.mfcc(y=audio, sr=sr, n_mfcc=n_mfcc)
```

**Figure 3**. Source Sound Feature Extraction Using MFCC

1. Zero-Crossing Rate

   The zero-crossing rate (ZCR) is the frequency at which the signal's amplitude changes. This is the absolute value of the sign of the nth sample minus the sign of the (n-1)th sample, divided by twice the number of samples. The sign of the nth sample is 1 if the sample is positive and -1 if the sample is negative.

$$ZCR = \frac{\sum_{n-1}^{N} |\, sgn\, x(n) - sgn\, x(n-1)\,|}{2N}. \quad (1)$$

   Sgn x(n) = Sign of x(n), equals 1 if x(n) is positive and -1 if x(n) is negative. N = total number of samples in the audio chunk.

2. Audio energy

$$(i) = \frac{1}{W_L} \sum_{n=1}^{W_L} |x_i\,(n)|^2. \quad (2)$$

3. Entropy of energy

   Captures drastic changes whether smooth or drastic to the audio data.

$$e_j = \frac{E_{sub\ Frame\ j}}{E_{short\ Frame\ i}}. \quad (3)$$

### 2.5 CNN

With the advancement of biotechnology, the most frequently proposed model is the convolutional neural network. The entire input space is locally filtered by neurons arranged in an orderly manner to provide a comprehensive understanding of the image within the field of view. Convolutional neural networks can perform in-depth feature extraction on the input image.

For image processing and visual pattern recognition, based on Fig. 4, CNN is a type of artificial neural network architecture consisting of convolutional layers designed to automatically extract hierarchical features from images [16]. By integrating attention layers, CNN models can focus attention on important areas of features discovered by convolutional layers. This can help CNN models extract more important information from images by adding attention layers, allowing the model to process information in parallel across multiple images [17]. The input data consists of visual data of the eye area, specifically the pupil. Keypoints obtained from landmarks capture changes in pupil size and movement. The input data is processed via convolution to extract spatial features from the pupil, enabling the model to recognize patterns in pupil shape and movement. The convolution output is flattened into a 1-dimensional vector. Multihead attention is added to emphasize important features for lie-based and truth-based classification. Processing will be performed with a batch size of 64 and a sigmoid layer for lie-based or truth-based classification. This process is shown in Fig. 4.

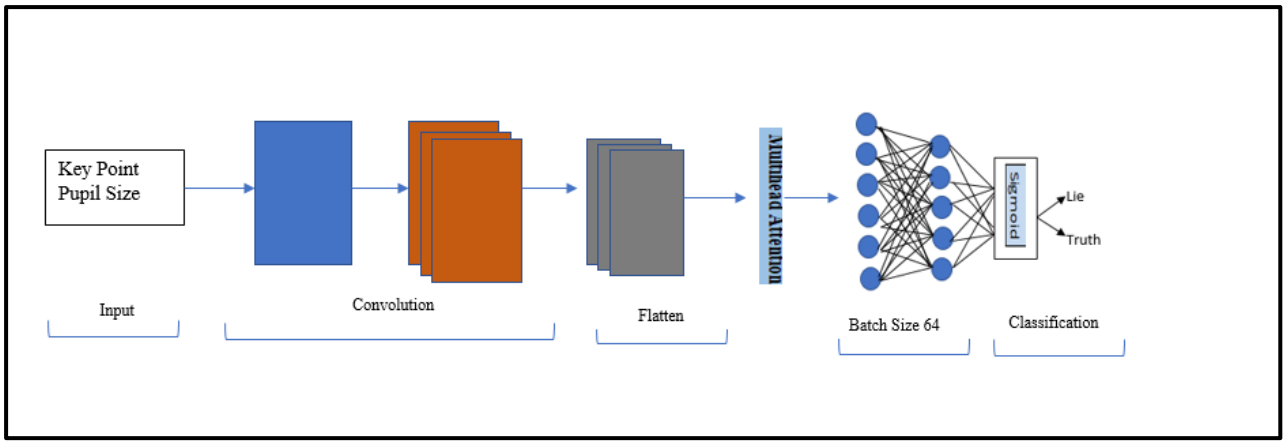


**Figure 4**. CNN Architecture Used

### 2.6 Attention

Multihead attention is an attention layer added to the CNN in this study. This technique allows the model to calculate the relationship between each pair of elements in the input sequence in a more complex way than regular spatial attention [18], as seen in Fig. 4. Linear Projection: for each feature point in the input X, project X into three different feature spaces: Q=XWQ, K=XWK, and V=XWV, where WQ, WV, and WK are weight matrices.

### 2.7 Long Short-Term Memory (LSTM)

The Long Short-Term Memory (LSTM) algorithm is a type of Recurrent Neural Network (RNN) architecture commonly used in deep learning applications [19]. LSTM neural networks are used to classify, process, and make predictions based on input via audio and LSTM processing based on time series data, because there may be unknown duration gaps between important events in the time series.

### 2.8 Fusion Model

Multimodal fusion refers to the process of gathering and integrating information from multiple modalities, which enhances performance relative to relying on a single modality [20]. Fusion can typically be performed at the input stage (early fusion), at the decision stage (late fusion), or at an intermediate level. In this work, a late fusion approach is applied, where each modality is independently processed for facial features using a Convolutional Neural Network (CNN) and for voice features using an LSTM, before their respective scores are integrated at the final stage.

### 2.9 Evaluation Model

During the experiment, analysis will be conducted by designing several test scenarios to obtain results that meet the research objectives. Evaluation is the stage of measuring the performance of the resulting model. The evaluation will be carried out using a confusion matrix to determine the success rate, including the calculation of accuracy, precision, recall, and the F1 score [21]. The receiver operating characteristic (ROC) area under the curve (AUC) serves as an essential accuracy measure in voice detection, representing the balance between sensitivity (the rate of true positives) and specificity (the rate of true negatives) [22] .

$$Accuracy = \frac{TP + TN}{TP + TN + FN + FP} X100\% , \quad (4)$$

$$Precision = \frac{TP}{TP + FP} X100\%, \quad (5)$$

$$Recall = \frac{TP}{TP + FN} X100\%, \quad (6)$$

$$F1 = 2\,X \frac{Precision * Recall}{Precision + Recall} X100\%, \quad (7)$$

$$AUC = \frac{1 + Recall - FPR}{2} X100\%. \quad (8)$$

## 3. RESULTS AND DISCUSSION

Model evaluation was conducted using a subject-independent data split, with 80% of subjects allocated to the training set and the remaining 20% to the test set. Several data partitioning configurations were examined, as summarized in Table 2. The scenario-based dataset consisted of 25 interview video samples, including 20 truthful samples and 5 lying samples. Additional video data were obtained from investigative recordings integrated with polygraph examinations, providing complementary labeled instances of truthful and lying behavior. Model performance was evaluated using standard classification metrics, namely Accuracy (ACC), Precision (P), Recall (R), and F1-score (F1). All metrics were computed based on the confusion matrix to mitigate potential bias arising from class imbalance.

Each interview video in the dataset has an average duration of approximately 10 minutes. To balance temporal resolution and computational efficiency, frame extraction was performed at 1 frame per second (1 FPS). Given a total duration of 600 seconds, this strategy resulted in 600 frames per interview video. Each extracted frame was subsequently processed using MediaPipe Face Mesh to obtain a numerical facial representation. MediaPipe Face Mesh detects 468 three-dimensional facial landmark points, each represented by Cartesian coordinates (x, y, z). Consequently, each frame yields 1,404 numerical feature values (468×3). An example of pupil-related facial landmarks is illustrated in Fig. 5. Under this scheme, a single 10-minute interview video generates a total of 842,400 numerical facial feature values (600 frames ×1,404 features). For the complete dataset comprising 25 interview videos, this process yields 15,000 frames and an aggregate of 21,060,000 numerical facial feature values.

This frame-based facial landmark representation results in a numerically large and temporally rich dataset. The use of Long Short-Term Memory (LSTM) networks is therefore well-suited to capture the temporal dynamics of facial expressions and supports more stable quantitative modeling of behavioral patterns associated with lying responses. Examples of the extracted numerical representations are provided in Table 1.

Fig. 5 is an overlay visualization of frames 1 and 2 of the right and left eyes. We collected samples from public datasets to examine changes in pupil diameter during truthfulness and lying. Colored landmarks indicate changes in pupil position detected by the pipe tool. A cropping process is used to study the right- and left-eye pupil movement patterns in more detail, which are psychological indicators of whether someone is telling the truth or lying. The pipeline generates 478 three-dimensional landmarks, consisting of 468 facial points derived from the MediaPipe Face Mesh, with the eye-region landmarks as shown in Fig. 5, refined for greater precision.

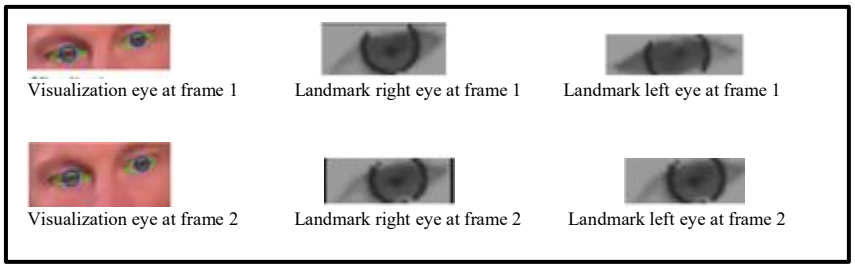


**Figure 5**. Illustrates Eye Landmarks

In this study, we performed feature extraction before processing with a CNN, converting facial images to numerical form so the model could recognize changes in detail. The values processed in feature extraction were the key points of the eyes, marked as in Fig. 2, obtained from the facemesh in the media pipe using the left and right pupil indices. Calculating circles for the left and right pupil points.

Based on the feature extraction, the numerical values of the change in pupil size for each video and frame are shown in Table 1.

**Table 1**. Results Of Feature Extraction Are Displayed in Numerical Form

| Video ID | Frame | Face | Eyes | Pupil Size | Status |
|---|---|---|---|---|---|
| 1 | 2 | 4.07E-13 | 3.80E-14 | 12.50578 | 1 |
| 1 | 3 | 4.23E-13 | 3.65E-14 | 12.39159 | 1 |
| 1 | 4 | 4.28E-13 | 3.59E-14 | 12.34786 | 1 |
| 1 | 5 | 3.83E-13 | 3.40E-14 | 12.38846 | 1 |
| 1 | 6 | 4.81E-13 | 4.11E-14 | 12.51274 | 1 |
| 1 | 7 | 5.38E-13 | 4.66E-14 | 12.51363 | 1 |
| 2 | 2 | 3.19E-13 | 3.10E-14 | 12.5001 | 1 |
| 2 | 3 | 3.45E-13 | 3.55E-14 | 12.3777 | 1 |
| 2 | 4 | 3.26E-13 | 2.97E-14 | 12.6685 | 1 |
| 2 | 5 | 3.47E-13 | 3.23E-14 | 12.7514 | 1 |
| 2 | 6 | 3.76E-13 | 3.91E-14 | 12.55233 | 1 |
| 2 | 7 | 4.07E-13 | 3.80E-14 | 12.45352 | 1 |

After obtaining the feature extraction value from the pupil, the next step is to extract voice features separately using MFCC with 20 coefficients in the librosa library, and to add the parameters zero-crossing rate, energy entropy, and audio energy. The results of the MFCC feature extraction process, with several parameters discussed previously, produce the voice extraction value shown in Fig. 6.

```
MFCCs: [[-6.5296918e+02 -6.5296918e+02 -6.2765710e+02 ... -4.7214276e+02
  -4.6454364e+02 -4.7412494e+02]
 [ 0.0000000e+00  0.0000000e+00  3.3230656e+01 ...  1.1159120e+02
   1.1942374e+02  1.1024862e+02]
 [ 0.0000000e+00  0.0000000e+00  2.7413689e+01 ...  2.1445797e+01
   2.3924566e+01  1.8422380e+01]
 ...
 [ 0.0000000e+00  0.0000000e+00 -2.0221109e+00 ...  5.1052737e+00
   5.8169746e+00  3.7081342e+00]
 [ 0.0000000e+00  0.0000000e+00 -1.9016300e+00 ...  6.0476738e-01
   1.1187967e+00  4.9811441e-01]
 [ 0.0000000e+00  0.0000000e+00 -1.4188266e+00 ...  4.4069872e+00
  -2.5808897e+00 -4.4594641e+00]]
```

**Figure 6**. Values Obtained from Voice Extraction

The feature extraction values from the face and voice were processed separately: the face using a CNN and the voice using an LSTM. Both deep learning models were trained with the specifications listed in Table 2. The following table contains information on the dataset type, data partitioning method, number of training epochs, number of classes, learning rate, optimizer, and activation function used in both deep learning models.

Table 2. Experimental In Model Testing

| Dataset type | Audio |
|---|---|
| Split data | 80:20; 60:40; 50:50 |
| Epoch | 50;100;150;200;250;300 |
| Class | 2 class (Lie, truth) |
| Rates of learning tested | 0.01;0.001;0.0001 |
| Optimization Algoritm applied | ADAM; Adadelta; Rmsprop |
| Activation | Relu; Tanh; Sigmoid |
| Cross validation | K=1;3;5;7;10 |

Table 3. Comparison of Fold Values for CNN-LSTM Models

| Fold | Accuracy (%) | Precision | Recall | F1-score | AUC |
|---|---|---|---|---|---|
| 1 | 87.2 | 0.86 | 0.83 | 0.84 | 0.89 |
| 3 | 86.9 | 0.85 | 0.82 | 0.83 | 0.88 |
| 5 | 87.8 | 0.87 | 0.84 | 0.85 | 0.9 |
| 7 | 87.5 | 0.86 | 0.84 | 0.85 | 0.89 |
| 10 | 99.1 | 0.98 | 0.99 | 0.99 | 0.98 |

Model performance was evaluated using 10-fold cross-validation to assess model stability and generalization. The evaluation results for each fold are presented in Table 3, which includes Accuracy, Precision, Recall, F1-score, and AUC metrics. Although the average model accuracy was around 89.6%, Fold-10 demonstrated the highest performance, reaching 99% accuracy. This improvement demonstrates that, under certain subject-separation configurations, the CNN-LSTM model can effectively learn temporal patterns of facial expressions.

Table 4. Performance Comparison Between Baseline Models and the Proposed Fusion Model

| Model | Accuracy (%) | Precision | Recall | F1-score | AUC |
|---|---|---|---|---|---|
| CNN (Pupil-only) | 81.4 | 0.8 | 0.78 | 0.79 | 0.83 |
| LSTM (Voice-only) | 84.7 | 0.83 | 0.82 | 0.82 | 0.86 |
| **Proposed Fusion (CNN + LSTM)** | **99.1** | **0.98** | **0.99** | **0.99** | **0.98** |

The quantitative results presented in Table 4 show that the proposed fusion model consistently outperforms both unimodal baseline models. The CNN (pupil-only) model achieves 81.4% accuracy, indicating that static visual cues alone provide limited discriminative power for lie detection. The LSTM (voice only) model improves performance to 84.7%, highlighting the contribution of temporal vocal information. The proposed fusion model attains significantly higher performance, achieving an accuracy of 99.1% and an F1-score of 0.99. These improvements confirm that integrating pupil-based visual features with temporal vocal features enables the model to effectively capture complementary information from both modalities. Furthermore, the increase in AUC from 0.83 and 0.86 for the unimodal models to 0.98 for the fusion model demonstrates a substantial enhancement in discriminative capability achieved through multimodal fusion. Despite its high performance, these results were obtained in a controlled experiment and subject-independent validation, and further evaluation on larger and more diverse data sets is needed to assess generalizability.

To generate a decision, the classification results from the CNN implementation with an attention layer are fused with those of voice analysis using LSTM, using the mean fusion method to improve the accuracy of the lie detection model. This process uses voice and facial analysis. A holistic and comprehensive approach is provided through the use of architectures such as CNNs with attention layers and LSTMs. To utilize the strengths of each model and make a final decision, the test results show that the accuracy reaches 99%, indicating that the combined model can detect lies well and robustly.

Improving detection accuracy by using various neural network architectures, the Long Short Term Memory (LSTM) method in conjunction with the Convolutional Neural Network (CNN). A holistic approach to analyzing lie signatures is demonstrated by using optical flow to measure spatial changes and combining

the two multimodal features (voice and pupil size). We use an LSTM model for voice, and a CNN with an attention layer to focus on pupil size changes, using a fusion method. The formula for the fusion technique we developed is fusion_pred = (voice_pred_argmax + Pupil_pred_flaten)/2.

Suara_pred_argmax is the predicted result of the voice change model after taking argmax, so it has a dimension of (None, 2).

Pupil_pred_flaten is the predicted result of the pupil change model, with varying dimensions.

Fusion_pred is the final result of the mean fusion, which is the average value of the two predictions.

The sources for calling voice and face data in the fusion model are as follows:

```
from keras.models import load_model
# For example, load voice from dataset.npz
voice_data = np.load('dataset.npz')
voice _features = voice _data['data']
voice _labels = voice _data['labels']
#Load pupil from dataset.csv
pupil_data = pd.read_csv('dataset.csv')
pupil_features = pupil_data['Pupil Size']
pupil_labels = pupil_data['Status']
# Load voice model and pupil model
voice _model = load_model(' voice _model.h5', compile=False)
voice_model.compile(loss='categorical_crossentropy', optimizer='adam')
pupil_model = load_model('pupil_model.h5', compile=False)
pupil_model.compile(optimizer=tf.keras.optimizers.Adam(),
                    loss=tf.keras.losses.BinaryCrossentropy())
# Voice Data Processing
def perform_preprocessing_ voice (data):
    # Perform processing according to the needs of the sound model.
    processed_data = perform_actual_preprocessing_voice(data)
    return processed_data
# Processing functions according to the needs of the sound model
def perform_actual_preprocessing_voice(data):
    processed_data = data / 255 # example: normalization range in 0-1
    return processed_data
# Pemrosesan Pupil Data
def perform_preprocessing_pupil(data):
    # Perform processing according to the pupil model requirements.
    processed_data = perform_actual_preprocessing_pupil(data)
    return processed_data
# Processing functions according to pupil model requirements
def perform_actual_preprocessing_pupil(data):
    # For example, perform normalization or change dimensions.
    processed_data = data
    return processed_data
```

Table 5 presents the use of different models and features for lie detection, including eye movements, microfacial expressions, EEG, pupil diameter, and eye and visual features. Lie detection accuracy varied across models and features, ranging from 58% to 99%. However, our study demonstrated good performance with 99% accuracy using a mean fusion model that combines facial and voice features.

**Table 5**. Different Models and Compares Accuracy

| References | Model | Feature | Accuracy |
|---|---|---|---|
| [7] | Types of deception Software | Eye Movements | 86%-95% |
| [23] | Spatiotemporal, and clasification SVM, MKL, and RF | Face micro expression | 85% |
| [24] | Eeg | Eye blink | 95% |

| References | Model | Feature | Accuracy |
|---|---|---|---|
| [25] | Method Gabor Wavelet Transform and Decision Tree | Pupil diameter and eye | 95% |
| [26] | Fusion LBP, SVM, RF | Multimodal | 63% |
| [27] | Meta learning, adversarial learning | Visual | 58% |
| [28] | KNN | Voice | 97.3% |
| Ours Research | Fusion model (CNN attention + LSTM | Pupil and voice | 99% |

## 4. CONCLUSION

Multimodal fusion of pupil and voice features, using the proposed method, produces optimal results with an average accuracy of 99%. The combination of a CNN model for facial features and an LSTM with MFCC cepstral coefficients (20) for voice features, combined with a fusion method for high-accuracy detection, demonstrates the robustness of the model, achieving maximum accuracy. However, although these results show 99% accuracy, further development is needed to ensure the model's generalization and applicability to a wider dataset, varying lighting conditions, and voice quality, and to detect more complex situations. Overall, this method shows that the CNN-LSTM combination can build an excellent detection model by recognizing verbal (voice) and non-verbal (pupil) patterns. The results indicate that MediaPipe-based features can serve as an effective visual representation complementary to CNN-derived features. The integration of MediaPipe facial landmarks with an LSTM model enhances the model's ability to capture micro-expression dynamics in a more structured and interpretable manner. This approach opens opportunities for the development of hybrid feature fusion strategies, in which MediaPipe-based geometric facial features and CNN-based visual features are jointly leveraged to improve both the robustness and transparency of the lie detection model.

## Author Contributions

Dewi Kusumawati: Conceptualization, Data Collection, Formal Analysis, Research Implementation, and Writing-Initial Draft. Fitriyanti Andi Masse: Conceptual Design, Data Management, Formal Analysis, Methodology Development, Resource Provision, Validation, and Writing, Reviewing, and Editing-Manuscript. Wulan: Methodology, Data Collection, Formal Analysis, Writing, and Reviewing-Editing. All authors discussed the results and contributed to the final manuscript.

## Funding Statement

This research received financial support from the Ministry of Education, Culture, Research, and Technology of the Republic of Indonesia based on Decree No. 0419/C3/DT.05.00/2025 dated May 22, 2025, through Master Contract No.137/C3/DT.05.00/PL/2025 dated May 28, 2025, Sub-Contract No.879/LL16/AL.04/2025, and LPPM STMIK Bina Mulia Contract No. 37/LPPM/STMIK-BMP/VI/2025 dated June 2, 2025. The author expresses his gratitude to the Ministry for this funding. He also expresses his deepest gratitude to the Laboratory Team of the Department of Informatics Engineering, STMIK Bina Mulia Palu, and LPPM for their valuable technical contributions and ongoing assistance throughout the research.

## Acknowledgment

We would like to thank the Head of the Identification Section of the Central Sulawesi Regional Police for providing guidance, assistance with research data, and equipment. We also extend our gratitude to the Forensic Laboratory team for their assistance in completing our research.

## Declarations

No conflicting interests are disclosed by the authors.

## Declaration of Generative AI and AI-assisted technologies

ChatGPT was utilized only to improve the readability and grammatical structure of the manuscript. No AI tool was used to generate or alter the research data, methodology, results, or interpretations. All content was verified by the authors for accuracy and consistency with the study.

## REFERENCES


[1] K. Ask, S. Calderon, and E. Mac Giolla, "HUMAN LIE-DETECTION PERFORMANCE: DOES RANDOM ASSIGNMENT VERSUS SELF-SELECTION OF LIARS AND TRUTH-TELLERS MATTER?," *J. Appl. Res. Mem. Cogn.*, vol. 9, no. 1, pp. 128–136, 2020, doi: https://doi.org/10.1016/j.jarmac.2019.10.002

[2] A. Vrij and M. Hartwig, "DECEPTION AND LIE DETECTION IN THE COURTROOM: THE EFFECT OF DEFENDANTS WEARING MEDICAL FACE MASKS," *J. Appl. Res. Mem. Cogn.*, vol. 10, no. 3, pp. 392–399, 2021, doi: https://doi.org/10.1016/j.jarmac.2021.06.002

[3] Z. Yang *et al.*, "TOPIC AUDIOLIZATION: A MODEL FOR RUMOR DETECTION INSPIRED BY LIE DETECTION TECHNOLOGY," *Inf. Process. Manag.*, vol. 61, no. 1, p. 103563, 2024, doi: https://doi.org/10.1016/j.ipm.2023.103563

[4] H. J. Holm, "TRUTH AND LIE DETECTION IN BLUFFING," *J. Econ. Behav. Organ.*, vol. 76, no. 2, pp. 318–324, 2010, doi: https://doi.org/10.1016/j.jebo.2010.06.003

[5] Á. Escolà-Gascón, "NEW TECHNIQUES TO MEASURE LIE DETECTION USING COVID-19 FAKE NEWS AND THE MULTIVARIABLE MULTIAXIAL SUGGESTIBILITY INVENTORY-2 (MMSI-2)," *Comput. Hum. Behav. Reports*, vol. 3, no. December 2021, doi: https://doi.org/10.1016/j.chbr.2020.100049

[6] M. Rahmani, F. Mohajelin, K. Nastaran, and S. and S. D. Sheykhivand, "AN AUTOMATIC LIE DETECTION MODEL USING EEG SIGNALS BASED ON THE COMBINATION OF TYPE 2 FUZZY SETS AND DEEP GRAPH," *Biomed. Signal Process. Heal. Monit. Based Sensors*, 2024, doi: https://doi.org/10.3390/s24113598

[7] Y. V. Bessonova and A. A. Oboznov, "EYE MOVEMENTS AND LIE DETECTION," *Adv. Intell. Syst. Comput.*, vol. 722, pp. 149–155, 2018, doi: https://doi.org/10.1007/978-3-319-73888-8_25

[8] W. Khan, K. Crockett, J. O'Shea, A. Hussain, and B. M. Khan, "DECEPTION IN THE EYES OF DECEIVER: A COMPUTER VISION AND MACHINE LEARNING BASED AUTOMATED DECEPTION DETECTION," *Expert Syst. Appl.*, vol. 169, no.1, p. 14341, November 2021, doi: https://doi.org/10.1016/j.eswa.2020.114341

[9] F. V. Nurçin, E. Imanov, A. Işin, and D. U. Ozsahin, "LIE DETECTION ON PUPIL SIZE BY BACK PROPAGATION NEURAL NETWORK," *Procedia Comput. Sci.*, vol. 120, no. 2017, pp. 417–421, 2017, doi: https://doi.org/10.1016/j.procs.2017.11.258

[10] B. M. D. Miron Zuckerman Robert Rosenthal, *Verbal And Nonverbal Communication Of Deception*, vol. 2, no. 2. 1956.

[11] S. D. H. Permana, G. Saputra, B. Arifitama, Yaddarabullah, W. Caesarendra, and R. Rahim, "CLASSIFICATION OF BIRD SOUNDS AS AN EARLY WARNING METHOD OF FOREST FIRES USING CONVOLUTIONAL NEURAL NETWORK (CNN) ALGORITHM," *J. King Saud Univ. - Comput. Inf. Sci.*, vol. 34, no. 7, pp. 4345–4357, 2022, doi: https://doi.org/10.1016/j.jksuci.2021.04.013

[12] S. Arooj, S. Altaf, S. Ahmad, H. Mahmoud, and A. S. N. Mohamed, "ENHANCING SIGN LANGUAGE RECOGNITION USING CNN AND SIFT: A CASE STUDY ON PAKISTAN SIGN LANGUAGE," *J. King Saud Univ. - Comput. Inf. Sci.*, vol. 36, no. 2, p. 101934, 2024. https://doi.org/10.1016/j.jksuci.2024.101934

[13] D. Kusumawati, A. A. Ilham, A. Achmad, and I. Nurtanio, "VGG-16 AND VGG-19 ARCHITECTURE MODELS IN LIE DETECTION USING IMAGE PROCESSING," in *Proceeding - 6th International Conference on Information Technology, Information Systems and Electrical Engineering: Applying Data Sciences and Artificial Intelligence Technologies for Environmental Sustainability, ICITISEE 2022*, pp. 340–345 2022, doi: https://doi.org/10.1109/ICITISEE57756.2022.10057748

[14] Y. Zhang and S. Huang, "JOURNAL OF KING SAUD UNIVERSITY - COMPUTER AND AUTOMATIC RUMOR RECOGNITION FOR PUBLIC HEALTH AND SAFETY : A STRATEGY COMBINING TOPIC CLASSIFICATION AND MULTI-DIMENSIONAL FEATURE FUSION," *J. King Saud Univ. - Comput. Inf. Sci.*, vol. 36, no. 5, p. 102087, 2024, doi: https://doi.org/10.1016/j.jksuci.2024.102087

[15] Y. Wu and Q. Ji, "FACIAL LANDMARK DETECTION: A LITERATURE SURVEY," *Int. J. Comput. Vis.*, vol. 127, no. 2, pp. 115–142, 2019, doi: https://doi.org/10.1007/s11263-018-1097-z

[16] J. Gu *et al.*, "RECENT ADVANCES IN CONVOLUTIONAL NEURAL NETWORKS," *Pattern Recognit.*, vol. 77, pp. 354–377, 2018, doi: https://doi.org/10.1016/j.patcog.2017.10.013

[17] X. Pan *et al.*, "ON THE INTEGRATION OF SELF-ATTENTION AND CONVOLUTION," *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 2022-June, pp. 805–815, 2022, doi: https://doi.org/10.1109/CVPR52688.2022.00089

[18] D. V. Sang and L. T. B. Cuong, "IMPROVING CRNN WITH EFFICIENTNET-LIKE FEATURE EXTRACTOR AND MULTI-HEAD ATENTION FOR TEXT RECOGNITION," *ACM Int. Conf. Proceeding Ser.*, pp. 285–290, 2019, doi: https://doi.org/10.1145/3368926.3369689

[19] B. Zohuri and S. Zadeh, "THE UTILITY OF ARTIFICIAL INTELLIGENCE FOR MOOD ANALYSIS, DEPRESSION DETECTION, AND SUICIDE RISK MANAGEMENT," *Journal of Health Science*. researchgate.net, 2020, [Online]. Available:https://www.researchgate.net/profile/Bahman Zohuri/publication/342448488_The_Utility_of_Artificial_Intelligence_for_Mood_Analysis_Depression_Detection_and_Suicide_Risk_Management/links/5ef4be6a45851550506f5125/The-Utility-of-Artificial-Intelligence-for-Mo.

[20] H. R. V. Joze, A. Shaban, M. L. Iuzzolino, and K. Koishida, "MMTM: MULTIMODAL TRANSFER MODULE FOR CNN FUSION," *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, pp. 13286–13296, 2020, doi: https://doi.org/10.1109/CVPR42600.2020.01330

[21] A. Luque, A. Carrasco, A. Martín, and A. de las Heras, "THE IMPACT OF CLASS IMBALANCE IN CLASSIFICATION PERFORMANCE METRICS BASED ON THE BINARY CONFUSION MATRIX," *Pattern Recognit.*, vol. 91, pp. 216–231, 2019, doi: https://doi.org/10.1016/j.patcog.2019.02.023

[22] A. A. Masrur Ahmed *et al.*, "DEEP LEARNING HYBRID MODEL WITH BORUTA-RANDOM FOREST OPTIMISER ALGORITHM FOR STREAMFLOW FORECASTING WITH CLIMATE MODE INDICES, RAINFALL, AND PERIODICITY," *J. Hydrol.*, vol. 599, no. May, p. 126350, 2021, doi: https://doi.org/10.1016/j.jhydrol.2021.126350

[23] M. Owayjan, A. Kashour, N. Al Haddad, M. Fadel, and G. Al Souki, "THE DESIGN AND DEVELOPMENT OF A LIE DETECTION SYSTEM USING FACIAL MICRO-EXPRESSIONS," *2012 2nd Int. Conf. Adv. Comput. Tools Eng. Appl. ACTEA 2012*, pp. 33–38, 2012, doi: https://doi.org/10.1109/ICTEA.2012.6462897

[24] J. Immanuel, A. Joshua, and S. Thomas George, "A STUDY ON USING BLINK PARAMETERS FROM EEG DATA FOR LIE DETECTION," in *2018 International Conference on Computer Communication and Informatics, ICCCI 2018*, pp. 1–5, 2018, doi: https://doi.org/10.1109/ICCCI.2018.8441238

[25] Z. Labibah, M. Nasrun, and C. Setianingsih, "LIE DETECTOR WITH THE ANALYSIS OF THE CHANGE OF DIAMETER PUPIL AND THE," *2018 IEEE Int. Conf. Internet Things Intell. Syst. Lie*, pp. 214–220, 2018, doi: https://doi.org/10.1109/IOTAIS.2018.8600918

[26] V. Gupta, M. Agarwal, M. Arora, T. Chakraborty, R. Singh, and M. Vatsa, "BAG-OF-LIES: A MULTIMODAL DATASET FOR DECEPTION DETECTION," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, 2019, vol.1 No.II pp. 83–90,June 2019, doi: https://doi.org/10.1109/CVPRW.2019.00016

[27] M. Ding, A. Zhao, Z. Lu, T. Xiang, and J. R. Wen, "FACE-FOCUSED CROSS-STREAM NETWORK FOR DECEPTION DETECTION IN VIDEOS," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2019, vol.1,No.2 , pp. 7794–7803,June 2019, doi: https://doi.org/10.1109/CVPR.2019.00799

[28] F. M. Talaat, "EXPLAINABLE ENHANCED RECURRENT NEURAL NETWORK FOR LIE DETECTION USING VOICE STRESS ANALYSIS," *Multimed. Tools Appl.*, no. 0123456789, 2023, doi: https://doi.org/10.1007/s11042-023-16769-w