Mesopotamian journal of Cybersecurity
Vol.4,No.2, **pp**. 74–87
DOI: https://doi.org/10.58496/MJCS/2024/008; ISSN: 2958-6542
https://mesopotamian.press/journals/index.php/cybersecurity

Research Article

LipPrint: Using Lip Movements as a Silent Password

Ahmed S. Ketab [1,2,*], Nidhal K. El Abbadi[3]

[1] *Department of Computer Science, Faculty of Computer Science and Mathematics, University of Kufa, Najaf, Iraq*

[2] *Directorate of Education of Thi-Qar, Department of Vocational Education, Thi-Qar, Iraq*

[3] *Computer Engineering Techniques, College of Engineering Techniques, Al-Mustaqbal University, Babylon, Iraq*

ARTICLE INFO

Article history

Received 03 May 2024
Accepted 02 Jul 2024
Published 23 Jul 2024

Keywords

Password

Lip movements

Visual speech recognition

Arabic digits dataset

Modified VGG16

ABSTRACT

Authentication and confidentiality are critical elements in many systems, and passwords are used to protect these systems. One way to enter the password into the system is via voice. The problem with this type of password is that people listen to it easily. The general purpose of this study was to make the password based on lip movement instead of sound. To address this problem, a visual speech recognition (VSR) system was proposed that detects speech from lip movement as a silent password. The proposed system consists of preprocessing the captured video via the proposed algorithms and feature extraction and classification of spoken digits via a modified VGG16. Fairness, robustness, and privacy are also considered in the proposed system. Research has focused on Arabic digits because there is little research on Arabic visual speech recognition. The reported results of the model show remarkable performance, with a validation accuracy of 96.74%. It was concluded that the proposed system, which relies on analysing lip movement instead of sound, increases security by preventing others from hearing the spoken password.

1. INTRODUCTION

Authentication and confidentiality are critical components of many systems. Strong passwords are essential for protecting these systems. However, traditional methods such as voice are vulnerable to interception during login, which poses security risks [1]. To address this problem, a promising solution involves silent password engines based on lip movement analysis. It is also known as lipreading, which is the skill of comprehending spoken words through visual cues that involve tongue and teeth movements without relying on sound. Automating this human capability of lip reading is known as visual speech recognition (VSR) [2]. The VSR has gained significant attention from researchers because of its ability to interpret visual signals. It has emerged as an alternative to audio speech recognition (ASR), especially in situations where audio is not possible or requires high-security measures [3]. Moreover, VSR may also increase the accuracy of ASR, as it is not affected by audio noise interference [4].

Most VSR research has focused on English, and there has been limited research on Arabic. Despite advancements in English speech recognition techniques, applying them to Arabic poses challenges because of their distinct characteristics. [5]. In Arabic, each letter can indeed be pronounced in three different ways depending on its position within a word or sentence. These variations in pronunciation are known as phonemes, and they introduce distinctive segmentation challenges in the analysis of spoken language [6].

The Arabic language boasts elaborate and significant origins, derivations, rules, amplitude, flexibility, and formation, features not commonly found in other languages. Its letters are distinguished by their articulation (exits) and sound by harmony. Lip reading could be a crucial topic that could be effectively utilized in human–computer interactions across

*Corresponding author. Email: ahmedsaudketab@utq.edu.iq

multiple languages. However, Arabic still requires extensive study by researchers, especially in comparison to other languages [7].

Compared with audio datasets, VSR datasets are scarce, especially in Arabic. Existing datasets often have low-quality recordings and may not align with modern video standards, prompting researchers to develop new datasets with higher-quality recordings [8].

In this research, due to the lack of available Arabic datasets, we contributed to creating a new dataset for Arabic numerals. Furthermore, a lip-reading system has been proposed. The proposed system uses several proposed algorithms aimed at enhancing the fairness and robustness of the proposed system.

In addition to the new dataset, this research contains three contributions:

- A new algorithm is proposed that aligns the mouth to ensure a consistent appearance across the dataset and solves the problem of head or camera movements. This makes it easier to extract meaningful features and helps in accurate lip analysis, thus enhancing the accuracy and robustness of the speech recognition system.
- A new lips masking algorithm has been proposed to precisely identify the mouth region by eliminating skin and facial hair. This algorithm reduces potential biases related to gender and skin tone variations, ensuring fairness in the system's outcomes. The mouth area should be determined more accurately, as most research defines the mouth as a rectangular shape that contains the skin and/or facial hair. Therefore, a new lip masking algorithm was proposed that removes skin and facial hair to reduce potential biases related to differences in sex and skin color, which ensures fairness in the system's results.
- A new frame selection algorithm has been proposed that deletes similar frames and selects the most representative frames of the video. This leads to improved feature extraction and more reliable speech recognition results.

This study aims to develop and implement a robust VSR system that uses lip movements instead of sounds, providing a silent, intuitive, and accessible authentication method to increase security. It introduces new algorithms designed to prioritize fairness and robustness while collecting a comprehensive dataset of lip movements from diverse individuals under various conditions. The proposed system is trained on this dataset, and its performance is evaluated and compared with that of other existing systems.

The significance and implications of using lip movements instead of sound are multifaceted and extend across various domains:

- Enhanced Security: Compared with traditional text-based passwords, lip movements are harder to replicate or steal, reducing vulnerability to common security threats such as phishing, keylogging, and shoulder surfing. Additionally, visual verification through lip movements makes it difficult for attackers to use audio recordings or mimic the password, offering a more secure alternative to voice recognition systems.
- Hands-Free Authentication: Users can authenticate without needing to type, remember complex passwords, or use their hands, which is particularly beneficial in situations where hands are occupied or dirty. Lip movements are a natural and intuitive form of interaction, potentially increasing user satisfaction and acceptance.
- Accessibility and Inclusivity: This study provides an alternative authentication method for users with physical disabilities or impairments that make typing or traditional biometric authentication challenging. It is useful in environments where silence is required or preferred, such as libraries, meetings, or public transportation.
- Multimodal Authentication Potential: Lip analysis can be combined with other biometric methods (e.g., facial recognition) to create a more reliable and robust multimodal authentication system.

The remaining sections of this paper are organized as follows: Section 2 focuses on related works. Section 3 presents our methodology for the proposed system. In section 4, our dataset is explained in detail. The proposed system was also tested on the dataset, and its results were presented and compared with those of other systems. Finally, section 5 presents the conclusions of the study.

2. RELATED WORKS

Research in the field of VSR for the Arabic language is significantly less extensive than that for the English language. Machine Learning (ML) and Deep Learning (DL) are powerful techniques for feature extraction and prediction, utilized in various fields including security [9-13]. They have also been employed to understand spoken Arabic words and phrases. To achieve optimal results, these techniques require abundant, high-quality, and diverse data for training [14]. From 2019 to 2024, several noteworthy research efforts have contributed to the advancement of Arabic VSR technology.

In 2019, Elrefaei, Alhassan, and Omar introduced the Arabic visual speech dataset (AVSD), which comprises 1100 videos containing 10 words spoken by 22 Arabic speakers. Initially, they manually cropped and labelled mouth regions. Their model is machine learning-based, where they used discrete cosine transform (DCT) for feature extraction and a support vector machine (SVM) for recognition. The accuracy of visual speech recognition reached 70.09% [8]. However, manual cropping of mouth regions can be considered a disadvantage because of the time and effort involved.

In the next year, Mohamad Ezz et al. suggested a hybrid voting framework for silent password recognition via lip movement analysis. The framework combines three techniques for automatic visual feature extraction: SURF, HoG, and Haar. These features are separately fed into a hidden Markov model (HMM), resulting in three outputs that are integrated via a voting scheme. The dataset used was a self-created digit dataset consisting of 2000 videos, achieving an accuracy of 96.2% [1].

Nadia H. Alsulami collected an Arabic visual dataset in 2021. The dataset contains 2400 records of Arabic digits and 960 records of Arabic phrases from 24 native speakers. They employed facial landmark detection from the Dlib toolkit to detect faces and crop mouth regions. These cropped mouth regions were used to produce concatenated frame images (CFI) that represent the utterance sequence in a single image. Compared with VGG-19 without batch normalization, VGG-19 with batch normalization achieved greater accuracy: 94% for digit recognition and 97% for phrase recognition [15]. In the same year, Doaa Sami Khafaga introduced a real-time Arabic viseme recognition system using the SAVE dataset for Arabic visemes. The system comprises visual feature extraction and deep CNN classification, achieving the best accuracy of 95.3% [16].

The next year, Dweik et al. proposed a lipreading system capable of distinguishing ten Arabic words via a dataset of 1051 videos captured from various participants. After preprocessing, the input video is converted to grayscale image frames, which separately enter the CNN, TD-CNN-BiLSTM, and TD-CNN-LSTM. A voting model combines the results of these models, achieving a testing accuracy of 82.84% [17].

In the 2023 timeframe, two research efforts emerged. The first one was introduced by N. F. Aljohani and E. S. Jaha. They created a new Arabic lip-reading dataset comprising 10,490 videos, including single letters, disjoined letters, and words from the Quran, captured from three different viewpoints. They trained a deep learning CNN model, achieving accuracies of 83.3%, 80.5%, and 77.5% for words, disjoined letters, and single letters, respectively. However, this study has limitations in applying models trained on frontal view data to nonfrontal viewpoints[7]. Khalil I. Alsaif et al. utilized a database containing 1080 images from five participants to track the geometrical features of lips for Arabic word utterance recognition [18]. The method calculates various geometric features of the mouth region, but the results lack evaluation criteria.

The latest research conducted this year was by Ali Baaloul et al. in 2024. They created a robust Arabic audio–visual dataset with 1,383 videos of 10 daily communication phrases spoken by 9 native Arabic speakers. Their dataset is publicly available on GitHub. They proposed a CNN model that compared favourably with other models. The proposed CNN models and the Vision Transformer (ViT) model with transfer learning achieved validation accuracies of 90% and nearly 98%, respectively [19].

Research on Arabic speech recognition through lip movement has faced challenges because of the scarcity of available datasets. As a result, most researchers create their own datasets. These datasets include images or videos of isolated mouth regions, as they perform preprocessing. These preprocessed formats of the dataset, although they reduce time and effort for researchers, may cause limitations in their research and make them unable to apply their algorithms. Therefore, formatting the dataset before preprocessing is very important because it allows researchers to perform preprocessing, such as determining the number of frames. The GRID [20] is an example of this format of an English dataset, where participants appear with full faces.

To overcome these limitations, our dataset is presented in two versions: the raw version and the preprocessed version. The raw version contains videos of individuals speaking with full faces, whereas the processed version includes images of isolated mouths. This allows researchers to choose the appropriate version according to their needs and empowers them to apply their algorithms and preprocessing.

The mouth area was also defined more precisely than in these studies. The mouth is often cropped into a rectangular shape. We noticed that this area also contains parts unrelated to speech, such as facial hair and skin tone. Therefore, these parts were removed via the proposed lips masking algorithm, reducing potential biases related to gender and skin tone variations and ensuring fairness in the system's outcomes. We also selected fewer frames than did other studies to represent the number of frames. The proposed frame selection algorithm identifies the 16 most representative frames from a video. This addressed the issue of varying video lengths and differences in speaking speed.

In previous studies, either the lips were not aligned at all or the face was aligned. In this study, the lips were aligned according to the characteristics of the mouth. Lip alignment ensures a consistent appearance across the dataset and makes it easier to extract meaningful features for more accurate analysis. Lip alignment also eliminates the need for rotation augmentation.

3. PROPOSED SYSFTEM

In this section, we present our proposed speech recognition system. The first part of this system is preprocessing. The most important preprocessing algorithms used were facial detection, the proposed lip alignment, the proposed lip masking, and the proposed frame selection. The second part is related to feature extraction and classification. A VGG16 model is used for this purpose, but its top fully connected layers are modified to better suit the recognition of ten Arabic digits in visual speech.

3.1 Preprocessing

Preprocessing plays an important role in the development and performance of each recognition system. This is an important part of our proposed system. The preprocessing stages are aimed at minimizing data that are not related to speech (face and lip detection and lip masking), correcting unintended movements of the speaker (lip alignment), and determining the frames that enter the model by removing the redundant frames (frame selection). Figure 1 illustrates the preprocessing part of the proposed lip-reading system.

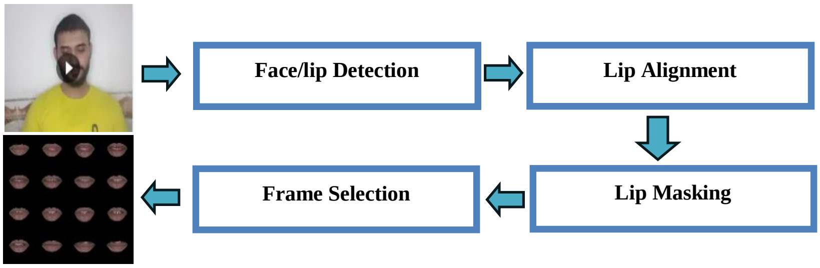


Fig. 1. Preprocessing steps.

3.1.1 Face and Lip Detection

This step aims to remove visual data that are not relevant to speech, such as the background, speaker's hair, nose, etc., that do not contribute to speech production. However, the mouth is responsible for producing sound and speech, so only this area is cropped. In other words, lip detection aims to remove irrelevant facial features, which helps to identify speech more accurately. For accurate lip detection, an accurate facial detection algorithm is needed. For this purpose, the MediaPipe face mesh model developed by Google has been used [21]. It predicts 468 facial landmarks that represent the detected face. These facial landmarks include the lip landmarks that are drawn to determine the mouth (see Figure 3(a)).

3.1.2 Proposed Lip Alignment Algorithm

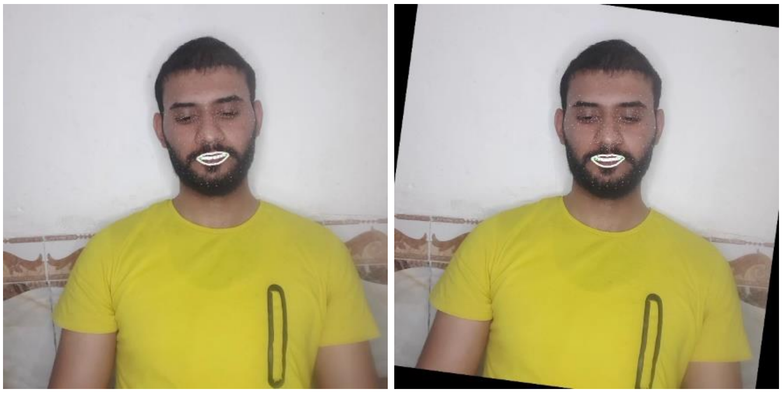
The purpose of alignment in visual speech processing is to ensure that the mouth region is consistently positioned across the video. This helps standardize the appearance of the mouth across the dataset, making it easier to extract meaningful features and facilitating accurate analysis and recognition of speech. Alignment also helps in removing variations due to differences in head movements or camera angles, thus improving the robustness of the speech recognition system. Lip alignment leads to automatic facial alignment. The lip alignment process aims to correct unintended movements of the speaker. In Figure 3, the speaker’s head is slightly tilted. The proposed lip alignment algorithm removes this tilt. As mentioned, the facial mesh model predicts 468 facial landmarks. The proposed lip alignment algorithm takes advantage of two of them, 183 and 308. These two landmarks are on either side of the edge of the mouth and are colored green for illustration in Figure 3. The horizontal angle between these two landmarks was subsequently calculated. Finally, the frame was rotated by the calculated angle. The steps of the proposed lip alignment are shown in Algorithm 1.

Algorithm 1 Proposed lip alignment

```
Input: A frame
Output: Lip aligned frame
1: Load the MediaPipe Face Mesh model, detect facial landmarks, and then draw the lips landmarks for the given frame.
2: Identify the landmarks with indices 183 and 308, which correspond to the corners of the mouth. Figure 2 (a) illustrates the application of steps 1 and 2.
3: Retrieve the x and y coordinates of the identified landmarks 183 and 308.
```

- Let the coordinates of landmark 183 be (x1, y1).
- Let the coordinates of landmark 308 be (x2, y2).

4: Calculate the angle θ between the two points (x1, y1) and (x2, y2).
5: Rotate the frame by the calculated angle θ to align the lips horizontally, as shown in Figure 2(b).



(a) (b)

Fig. 2. (a) Face and lips detection. (b) Lips alignment.

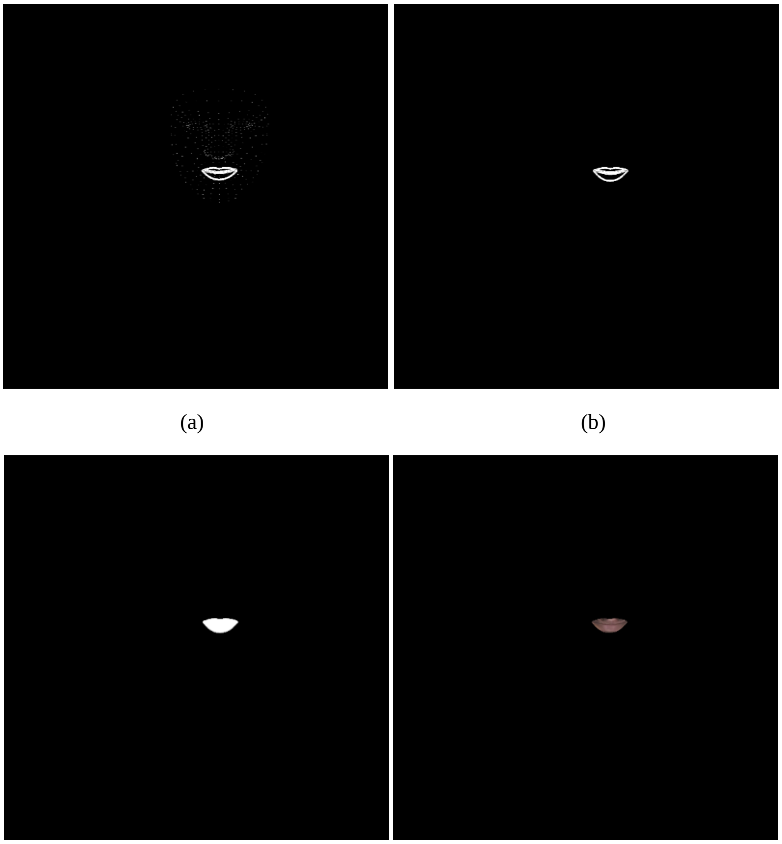
3.1.3 Proposed Lip Masking Algorithm

The precise delineation of the region of interest (ROI) plays a crucial role in system performance. Many research endeavors typically define the ROI as a rectangular area that covers the mouth region. However, this area may include irrelevant information, such as facial hair and variations in skin tone, which can affect the speech recognition process. To address this issue, we suggest a lips masking algorithm that precisely identifies the mouth region. By eliminating skin and facial hair, the algorithm reduces potential biases related to gender and skin tone variations, thereby ensuring fairness in the system's outcomes. After accurate delineation of the lip region, the cropping operation is performed so that the mouth is in the center of the cropped box, with dimensions of 200 × 150 pixels.

The steps of the proposed lip masking algorithm are shown in Algorithm 2.

Algorithm 2 Proposed lip masking

Input: (Aligned) frame
Output: Lip-masked frame
1: Input the aligned frame from the previous lips’ alignment algorithm (Figure 2 (b)).
2: Create a black mask image with the same dimensions as the input frame.
3: Apply the Face Mesh model to the input frame to detect facial landmarks.
4: Project the detected facial landmarks onto the black mask image and then connect the lip landmarks (Figure 3 (a)).
5: Convert the black mask to a binary image.
6: Perform a morphological opening operation on the binary image to remove small white regions and smooth the lip area (Figure 3 (b)).
7: Find the black regions. The opened mask may have three distinct black regions: the background, upper lip, and lower lip.
8: Convert all black regions except the largest black region (background) to white to create the final mask (Figure 3 (c)).
9: Perform multiplication between the input frame and the final mask to output the masked lips (Figure 3 (d)).



(a) (b)

(c) (d)

Fig. 3. (a) Face and lips detection. (b) Lips alignment. (c) Final mask. (d) Lips masked.

3.1.4 Proposed Frame Selection Algorithm

The third proposed algorithm is frame selection, which addresses the issue of varying lengths of recorded videos and differences in speaking speed. The frame selection algorithm aims to choose the most representative frames from the video by eliminating redundant frames. The input to the proposed lips masking algorithm is a video containing many frames, whereas the output is a fixed and small number of frames (16 frames in this study). This process ensures that the model receives a manageable set of frames that accurately represent the video. This improves the quality of feature extraction, resulting in more reliable recognition results. The proposed algorithm uses the single-scale structural similarity index (SSIM). To calculate the SSIM between two frames, n, and m, equation (1) is applied [22]:

$$\mathrm{SSIM}(n, \mathrm{m}) = \frac{(2\mathrm{A_n A_m}+\mathrm{S_1})(2\mathrm{A_{nm}}+\mathrm{S_2})}{(\mathrm{A_n^2}+\mathrm{A_m^2}+\mathrm{S_1})(\mathrm{V_n^2}+\mathrm{V_m^2}+\mathrm{S_2})} \qquad (1)$$

where:

- A_n and A_m are the average pixel intensities of images n and m, respectively.
- V_n and V_m are the variances of the pixel intensities in images n and m, respectively.
- V_{nm} is the covariance of the pixel intensities between frames n and m.
- S_1 and S_2 are constants used to stabilize the division with a weak denominator.

Algorithm 3 explains the proposed frame selection:

Algorithm 3 Proposed frame selection

```
Input: Lip-masked frames (video)
Output: desired number of frames
1: Input the lips-masked video as input.
2: Determine the desired number of frames, k.
```

```
3: Set the threshold value for the structural similarity index (SSIM) at 0.9.
4: Initialize an empty list to store selected frames.
5: Designate the first frame of the video as the selected frame.
6: Initialize counter F to 0 to count the selected frames.
7: Put the next frame of the video as the current frame.
8: Convert both the selected and current frames to grayscale.
9: Compute the SSIM between them.
10: If the computed SSIM is less than the threshold value:
        • Increment the counter variable F by 1.
        • Append the selected frame to the list.
        • Update the selected frame to the current frame.
11: If the current frame is not the last frame in the video, proceed to step 7.
12: Calculate the number of frames to be deleted, M, by subtracting k from F.
13: Remove M frames from the list using frame deletion rule in (2).
14: Output the list (k frames).
```

Note that the output of the proposed lips masking algorithm is a fixed and small number of frames (k). Here, this number is 16 and can be changed as needed.

Initially, the first frame is taken as the selected frame, and the second frame is taken as the current frame. Then, the SSIM of the selected frame with the current frames is computed. Here, the SSIM represents the similarity ratio between two frames. If the SSIM is greater than the threshold, the current frame is cancelled, and the next frame is taken as the current frame. However, if the SSIM is less than the threshold, add 1 to the counter of the selected frames (F); the selected frame is saved, and the selected frame is updated to the current frame. This process continues until the last frame.

Here, we obtained only the most different frames, and it was not necessary to obtain only 16 frames. It may provide 16 or more or fewer frames. To solve this problem, first, the threshold is carefully chosen so that it is not given fewer than 16 frames. This step leads us to obtain 16 frames or more. Second, the number of additional frames is calculated, and this number of frames is removed to obtain the 16 most different frames that represent the input video as follows:

$$\text{Frame deletion rule} = [\,2\text{i}, \text{F} - \text{j},\ \ldots\,]\ \text{continues for M elements} \qquad (2)$$

where:

- i= 1, 2, 3, 4, …
- j= 1, 3, 5, 7, …
- F is the total number of frames obtained by the SSIM.
- M= F-k is the number of additional frames to delete.
- k=16 is the number of desired frames.

For more explanation, let us take an example. Suppose that the total number of frames obtained by the SSIM on a video is 20. First, we compute the number of additional frames to delete according to the following: F= 20, k=16, and M=F-k= 20--16=4; thus, 4 frames must be deleted. The frame deletion rule shown in Table 1 is then applied to determine the position of the frames to delete.

TABLE I. FRAME DELETION RULE WITH F= 20

m	i & j	2i	F-j	Position of deletion
1	i=1	2*1=2	-	2
2	j=1	-	20-1=19	19
3	i=2	2*2=4	-	4
4	j=3	-	20-3=17	17

Therefore, the frames located at positions 2, 19, 4, and 17 are deleted.

3.1.5 Combined Image

After 16 frames are obtained, these frames are combined into one image. The resulting image is produced by arranging all four frames in one row. Therefore, four rows are needed. Since the dimensions of each frame are 200 × 150, the final image is 800 × 600. When the dataset is loaded into the model, it is resized to 224× 224. After resizing, the pixel values of the images are normalized by dividing each pixel value by 255, scaling the values to the range [0, 1]. Figure 5 shows a sample of the combined image.

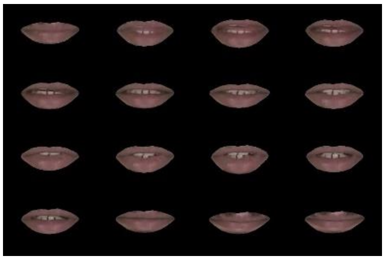


Fig. 4. Combined image

3.1 Feature Extraction and Classification

The combined image is now prepared for feature extraction and classification. We utilized the VGG16 [23] model with pretrained, frozen weights from ImageNet. The fully connected layers at the top of the VGG16 model were excluded. Instead, custom dense layers were added, including a flattened layer followed by a batch normalization layer, two dense layers (the first with 512 units, the next with 256 units) with ReLU activation, another batch normalization layer, and a final dense layer (10 units) with softmax activation. The modified VGG16 architecture is illustrated in Fig. 6.

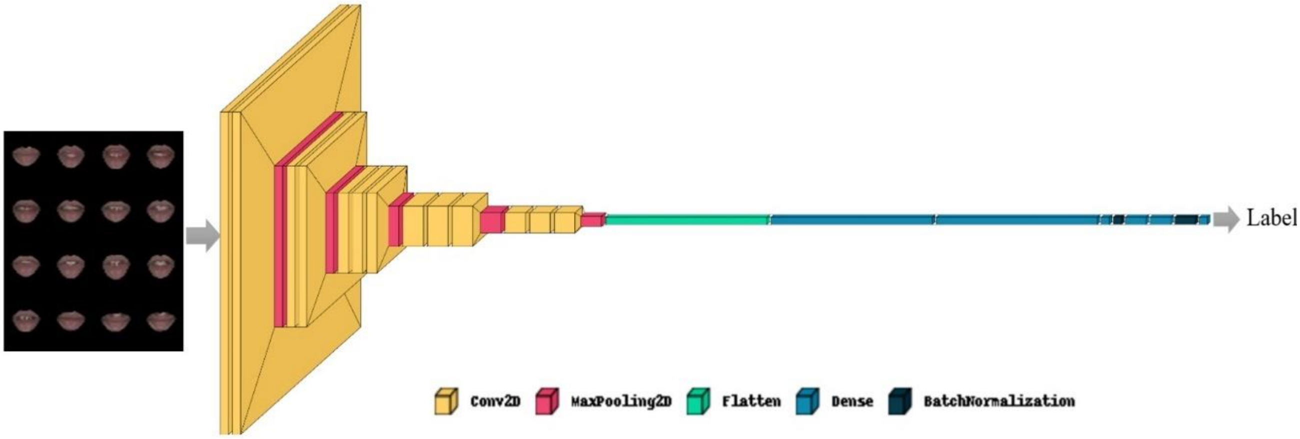


Fig. 5. VGG16 architecture

4. EXPERIMENTS AND RESULTS

4.1 New Arabic Digits Dataset

Public VSR datasets for the English language, such as GRID [20], OuluVS2 [24], CUAVE [25] and BANCA [26], are readily available. However, the availability of Arabic datasets is limited. Therefore, we gathered our Arabic digit dataset. Table 2 provides a list of Arabic digits along with their details.

TABLE II. ARABIC DIGITS

Arabic digit	Arabic spelling	English digit
صفر	Sefer	Zero
واحد	Wahed	One
اثنين	Ethnein	Two
ثلاثة	Thalatha	Three
اربعة	Arbaa	Four
خمسة	Khamsa	Five
ستة	Setah	Six
سبعة	Sabaa	Seven
ثمانية	Thamania	Eight
تسعة	Tesaa	Nine

The dataset is specifically designed for isolated words, where each word begins and ends with a period of silence. The videos were recorded by the Galaxy Note 8 selfie camera in several places with different backgrounds and light conditions.

The dataset encompasses recordings from Arabic-speaking volunteers spanning various age groups, including children, youth, and adults aged 8-60 years. Both genders were represented among the participants, with male participants including individuals with and without mustache/beard. The inclusion of children introduces additional challenges to the dataset due to the inherent variability in children's speech, including issues related to pronunciation and verbal fluency. In total, the dataset comprises 40 participants, including 32 males and 8 females. Each digit was repeated three times. However, some videos that violated the speaking conditions were excluded, namely, videos in which the mouth was closed before and after speaking and some videos in which lip movement habits were not related to speech. In the end, there were 963 videos. Figure 7 illustrates some of the participants.

Fig. 6. Some participants of Arabic digits dataset

Robustness is a crucial aspect of trustworthy AI and is defined as a system's ability to function correctly despite invalid inputs or stressful environmental conditions. According to the EU's ethical guidelines and IEEE's definition, robustness is essential for building AI systems that users can rely on [27]. In this dataset, participants from all age groups, including children, were included. Children's speech may contain pronunciation errors, which may be considered invalid inputs or stressful environmental conditions. A database of diverse age groups, including children, exposes the system to a wide range of input variations, enhancing its robustness and ensuring its ability to perform reliably in real-world scenarios.

Privacy is also an important consideration in AI systems. Users are more likely to trust systems that protect their personal data and identity. To protect privacy, sensitive and nonsensitive data must be distinguished during the preprocessing and modelling stages [27]. Therefore, two versions of the dataset are presented. The first version contains videos of participants with their full faces. We consider this version to be sensitive. This version is shared with researchers only after they accept agreements and publish licences that protect the rights of participants. The second version is the preprocessed dataset, which contains images of mouths. This version is publicly available because it does not contain sensitive data.

Fairness is a fundamental principle in AI systems, ensuring that they do not discriminate against any individual or group. In this study, several measures were taken to promote fairness [27]. First, the same individuals were used across all age groups for each class, and the same number of words was used for each person. This ensures that the data are evenly distributed among different age groups, reducing the risk of bias. Second, the videos were recorded under various lighting conditions and backgrounds, with a similar number of videos recorded for each class. This balanced dataset helps mitigate bias caused by environmental factors.

Fairness is a basic principle in AI systems and aims to prevent discrimination against any individual or group [27]. In this study, several considerations were taken to enhance fairness. First, the same individuals were used across all age groups for each class. This ensures that the data are evenly distributed between different classes and avoids bias. Second, the videos were recorded under different lighting conditions and backgrounds. This mitigates bias caused by environmental factors. Third, to avoid bias arising due to skin color or gender, the proposed mask algorithm was used, which subtracts the color of the skin, beard, and mustaches.

4.2 Evaluation Metrics

The proposed system was evaluated via four important metrics. Accuracy assesses overall correctness, precision evaluates the accuracy of positive predictions, recall measures the model’s ability to capture all positive instances, and the F1 score offers a nuanced understanding of the classification model’s performance. See (3) [28][29], (4) [30], (5) [31] and (6) [32].

$$Accuracy = {}^{[TP + TN]}\!\Big/_{[TP + FP + TN + TP]} \qquad (3)$$

$$Precision(P) = {}^{TP}\!\big/_{[TP + FP]} \qquad (4)$$

$$Recall(R) = {TP}/{[TP + FN]} \qquad (5)$$

$$F1_{Score} = 2 * {[P * R]}/{[P + R]} \qquad (6)$$

4.3 Results

All the experiments were conducted in the Google Colab Pro environment with 51 GB of RAM, 16 GB of GPU, and 166.8 GB of disk space. The programming language used in this environment is Python. The dataset has been uploaded to Google Drive. The dataset was divided into 90% for training and 10% for validation. The model is trained via the Adam optimizer with the default learning rate for 20 epochs and a batch size of 32. Flip-flop augmentation is utilized to increase the diversity of training data artificially. It mirrored the frames along the horizontal axis. Through the augmentation process, 1,183 combined images were created, bringing the total training data to 2,146 combined images. The model achieved a training loss of 0.009 with a training accuracy of 99.9%, indicating strong performance on the training data. For the validation set, a loss of 0.965 and an accuracy of 96.74% were obtained, demonstrating effective generalizability to new data. Figure 7 illustrates the accuracy and loss of training and validation of the proposed lip-reading system.

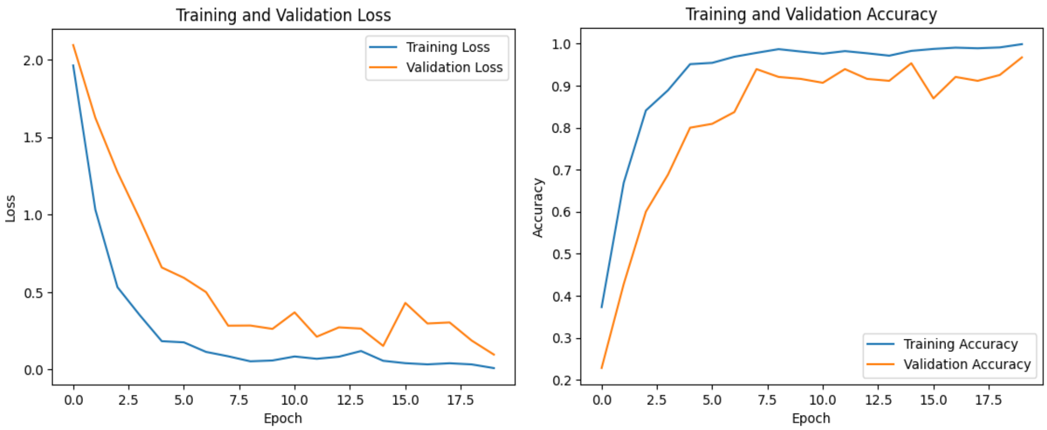


Fig. 7. Loss and accuracy of the proposed system

Both the training and validation losses decrease significantly over the epochs, indicating that the model is learning and improving its performance. After approximately 5 epochs, the loss stabilizes, with the validation loss occasionally spiking but remaining low overall. The training loss decreases steadily, whereas the validation loss also decreases with minor fluctuations, suggesting that the model is generalizing well without overfitting. On the other hand, the accuracy for both training and validation increases significantly over the epochs. After approximately 5 epochs, both accuracies stabilize, with the training accuracy being slightly higher than the validation accuracy. The high validation accuracy indicates that the model performs well on unseen data, and the minimal gap between training and validation accuracy suggests good generalizability.

To obtain a clearer picture, the confusion matrix illustrating the proposed system's performance is shown in Figure 8. Then, a classification report can be extracted from the confusion matrix, as shown in Figure 9. It contains the precision, recall, and F1 score of each class.

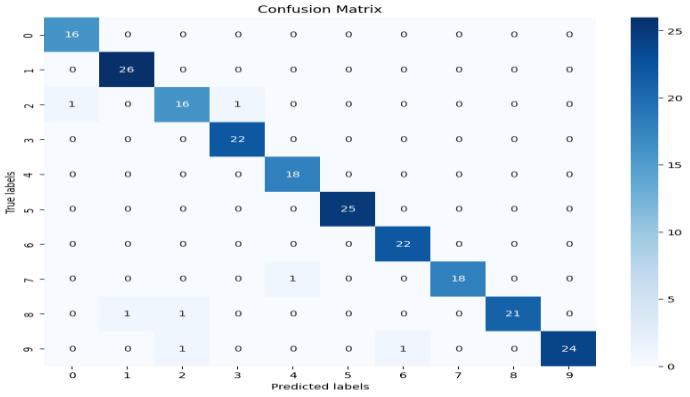


Fig. 8. Confusion matrix of the proposed system

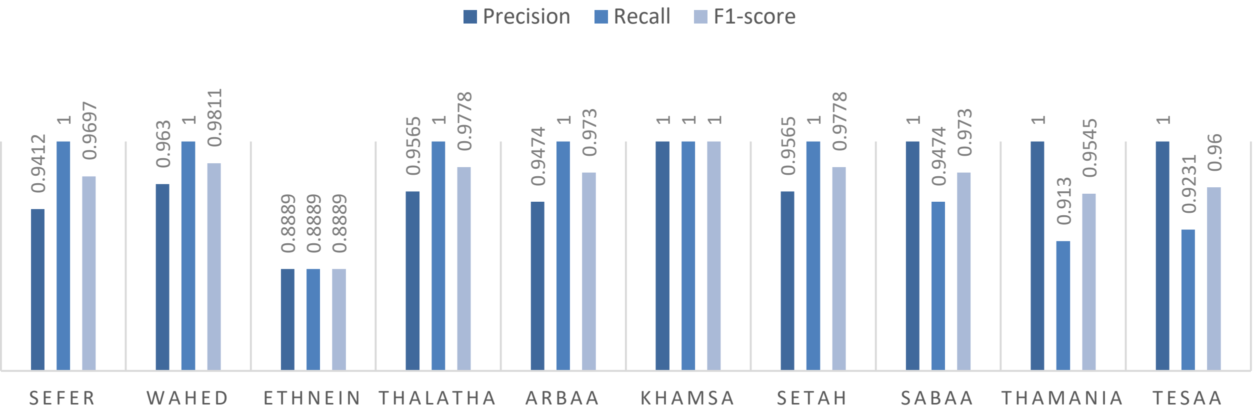


Fig. 9. Classification report

The proposed system demonstrates exceptional performance across most Arabic digits, with near-perfect precision, recall, and F1 scores for digits sefer, thalatha, arbaa, and setah. The perfect scores for khamsa indicate that the model can flawlessly classify this digit. The system shows slightly lower performance for ethnein. Overall, high metrics for most digits reflect the robustness and reliability of the proposed system in classifying Arabic digits.

4.4 Comparisons

The proposed system was compared with three Arabic lip-reading models via our new Arabic digit dataset. To the best of our knowledge, these are the only three models that have utilized the combined image style to feed their models into the Arabic lip-reading system to date. Table 3 compares the performance of these researchers' techniques in terms of accuracy, precision, recall, and F1 score.

TABLE III. MODEL COMPARISON

Researchers	**Technique**	**Accuracy**	**Precision**	**Recall**	**F1-Score**
Ali et al. [1]	CNN + SoftMax	0.9070	0.9155	0.9070	0.9030
Nadia et al [5]	VGG19 + SoftMax	0.9302	0.9738	0.8651	0.9327
Lamiaa et al. [8]	DCT+SVM	0.8046	0.8082	0.8046	0.8050
Proposed system	Modified VGG16 + SoftMax	0.9674	0.981	0.9628	0.9656

Compared with other techniques, the proposed system demonstrates superior performance. This suggests that the modifications made to the VGG16 architecture effectively enhance the model's ability to classify correctly and handle both false positives and false negatives more efficiently.

In the realm of digit datasets, two significant studies have addressed the Arabic digit dataset (as mentioned in related work). Mohamad Ezz et al. reported a model accuracy of 96.2%, whereas Nadia H. Alsulami achieved an accuracy of 94%. Compared with these studies, the proposed system outperforms their results, with an accuracy of 96.74%.

The performance of the proposed system was also studied in comparison with that of well-known models such as Inception v3 [33], MobileNet v2 [34], DenseNet121 [35], NASNetMobile [36], and Xception [37]. The pretrained networks initialize the base model with pretrained weights, freeze its layers, and then add flattened and dense layers on top. Table 4 lists the performance metrics of these models.

TABLE IV. PRETRAINED MODEL COMPARISON

DL approach	**Inception v3**	**MobileNetV2**	**DenseNet121**	**NASNetMobile**	**Xception**	**Proposed system**
Accuracy	0.8698	0.8698	0.9116	0.8233	0.8884	0.9674
Precision	0.8976	0.8738	0.9194	0.8365	0.8967	0.9670
Recall	0.8558	0.8698	0.9023	0.8093	0.8884	0.9535
F1Score	0.8663	0.8707	0.9139	0.8233	0.8888	0.9587

Compared with existing deep learning approaches, the proposed system shows superior performance, highlighting the success of architectural modifications. DenseNet121 also performs well, surpassing other deep learning approaches except for the proposed system. In contrast, NASNetMobile has the lowest performance among the evaluated models.
Upon comparison, it is evident that our proposed system outperforms the other models, with an accuracy of 96.74%. Additionally, the proposed system exhibited excellent precision, recall, and F1 score values, indicating its reliability, which is an important aspect of trustworthy AI.

5. CONCLUSIONS

This study presents a novel lip-reading system for silent passwords that addresses the vulnerabilities of voice passwords. The proposed system enhances security by using lip movements, which are harder to replicate, reducing risks such as phishing and keylogging.
The system comprises three main stages: preprocessing, feature extraction, and classification. In the preprocessing stage, three new algorithms are introduced. The lip alignment algorithm solves head or camera movement issues by aligning the mouth using two landmarks at the corners of the mouth. The lip masking algorithm removes extraneous elements such as skin, mustaches, and beards to define the mouth region more accurately, ensuring fairness by reducing biases related to gender and skin tone variations. The frame selection algorithm selects 16 representative frames from each video, addressing variations in video length and speaking speed for improved feature extraction. Feature extraction and classification are based on a modified VGG16 model.
A new dataset focusing on Arabic digits, comprising 963 videos from 40 Arabic speakers of various age groups, was developed. The system achieved remarkable performance, with a validation accuracy of 96.74%, outperforming existing Arabic models and pretrained models in all performance metrics. Despite these successes, the study's limitation is the exclusive use of Arabic digits, which do not cover the full range of real-world password components such as alphabets and words. Future research should address this limitation to further enhance the system's applicability. The results underscore the system's reliability and potential for integration with other biometric methods for a more secure authentication system.

Conflicts of interest

The authors declare that they have no conflicts of interest.

Funding

The acknowledgements section of the paper does not mention any financial support from institutions or sponsors.

Acknowledgement

The author acknowledges all participants in the Arabic digit dataset.

Data availability

The Arabic digit dataset is available in Kaggle at the following link: Arabic Digits Dataset (kaggle.com)

References

[1] M. Ezz, A. M. Mostafa, and A. A. Nasr, "A silent password recognition framework based on lip analysis," *IEEE Access*, vol. 8, pp. 55354–55371, 2020, doi: 10.1109/ACCESS.2020.2982359.

[2] M. Hao, M. Mamut, N. Yadikar, A. Aysa, and K. Ubul, "A survey of research on lipreading technology," *IEEE Access*, vol. 8, pp. 204518–204544, 2020, doi: 10.1109/ACCESS.2020.3036865.

[3] J. S. Chung and A. Zisserman, "Learning to lip read words by watching videos," *Computer Vision and Image Understanding*, vol. 173, pp. 76–85, 2018.

[4] V. Estellers and J.-P. Thiran, "Multi-pose lipreading and audio-visual speech recognition," *EURASIP Journal on Advances in Signal Processing*, vol. 2012, pp. 1–23, 2012.

[5] N. S. Ghadban, J. Alkheir, and M. Saii, "Performance evaluation hybrid visual speech recognition features on Arabic isolated words," *International Journal of Computer Science Trends and Technology (IJCST)*, vol. 5, no. 5, pp. 47–50, 2017.

[6] A. Al-Ghanim, A.-O. Nourah, R. Al-Haidary, S. Al-Zeer, S. Altammami, and H. A. Mahmoud, "I See What You Say (ISWYS): Arabic lip reading system," in *2013 International Conference on Current Trends in Information Technology (CTIT)*, IEEE, Dec. 2013, pp. 11–17. doi: 10.1109/CTIT.2013.6749470.

[7] N. F. Aljohani and E. S. Jaha, “Visual Lip-Reading for Quranic Arabic Alphabets and Words Using Deep Learning.,” *Computer Systems Science & Engineering*, vol. 46, no. 3, 2023.

[8] L. A. Elrefaei, T. Q. Alhassan, and S. S. Omar, “An Arabic visual dataset for visual speech recognition,” *Procedia Computer Science*, vol. 163, pp. 400–409, 2019, doi: 10.1016/j.procs.2019.12.122.

[9] “Machine Learning-Based Detection of Smartphone Malware: Challenges and Solutions,” *Mesopotamian Journal of Cyber Security*, pp. 134–157, Aug. 2023, doi: 10.58496/MJCS/2023/017.

[10] S. S. M, H. D, and R. R. Vallem, “Cyber Security System Based on Machine Learning Using Logistic Decision Support Vector,” *Mesopotamian Journal of Cyber Security*, pp. 64–72, Mar. 2023, doi: 10.58496/MJCS/2023/011.

[11] J. J. Hephzipah, R. R. Vallem, M. S. Sheela, and G. Dhanalakshmi, “An efficient cyber security system based on flow-based anomaly detection using Artificial neural network,” *Mesopotamian Journal of Cyber Security*, pp. 48–56, Mar. 2023, doi: 10.58496/MJCS/2023/009.

[12] N. A. Bajao and J. Sarucam, “Threats Detection in the Internet of Things Using Convolutional neural networks, long short-term memory, and gated recurrent units,” *Mesopotamian Journal of Cyber Security*, pp. 22–29, Feb. 2023, doi: 10.58496/MJCS/2023/005.

[13] T. Muhammad and H. Ghafory, “SQL Injection Attack Detection Using Machine Learning Algorithm,” *Mesopotamian Journal of Cyber Security*, pp. 5–17, Feb. 2022, doi: 10.58496/MJCS/2022/002.

[14] M. A. Fadhel *et al.*, “Comprehensive Systematic Review of Information Fusion Methods in Smart Cities and Urban Environments,” *Information Fusion*, p. 102317, 2024. doi: https://doi.org/10.1016/j.inffus.2024.102317

[15] N. H. Alsulami, A. T. Jamal, and L. A. Elrefaei, “Deep learning-based approach for Arabic visual speech recognition,” *Computers, Materials & Continua*, vol. 71, no. 1, pp. 85–108, 2022, doi: https://doi.org/10.32604/cmc.2022.019450.

[16] D. Khafaga, “Novel algorithm utilizing deep learning for enhanced Arabic lip reading recognition,” *International Journal of Advanced Computer Science and Applications*, vol. 12, no. 11, pp. 811–816, 2021. doi: 10.14569/IJACSA.2021.0121192

[17] W. Dweik, S. Altorman, and S. Ashour, “Read my lips: artificial intelligence word-level arabic lipreading system,” *Egyptian Informatics Journal*, vol. 23, no. 4, pp. 1–12, Dec. 2022, doi: 10.1016/j.eij.2022.06.001.

[18] K. I. Alsaif and N. S. Allella, “Lips Reading Spoken Arabic Word Based on The Geometric Shape Features of The Lip,” pp. 624–634, 2023. doi: 10.32628/IJSRST2310164.

[19] A. Baaloul, N. Benblidia, F. Z. Reguieg, M. Bouakkaz, and H. Felouat, “An arabic visual speech recognition framework with CNN and vision transformers for lipreading,” *Multimedia Tools Application*, pp. 1–35, 2024, doi: https://doi.org/10.1007/s11042-024-18237-5.

[20] M. Cooke, J. Barker, S. Cunningham, and X. Shao, “An audio-visual corpus for speech perception and automatic speech recognition,” *The Journal of the Acoustical Society of America*, vol. 120, no. 5, pp. 2421–2424, 2006, doi: https://doi.org/10.1121/1.2229005.

[21] Ü. Atila and F. Sabaz, “Turkish lip-reading using Bi-LSTM and deep learning models,” *Engineering Science and Technology, an International Journal*, vol. 35, Nov. 2022, doi: 10.1016/j.jestch.2022.101206.

[22] M.-J. Chen and A. C. Bovik, “Fast structural similarity index algorithm,” *J Real Time Image Process*, vol. 6, no. 4, pp. 281–287, Dec. 2011, doi: 10.1007/s11554-010-0170-9.

[23] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” in *International Conference on Learning Representations*, 2014. doi: https://doi.org/10.48550/arXiv.1409.1556.

[24] I. Anina, Z. Zhou, G. Zhao, and M. Pietikäinen, “Ouluvs2: a multi-view audiovisual database for non-rigid mouth motion analysis,”*IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, IEEE, 2015, pp. 1–5. doi: 10.1109/FG.2015.7163155.

[25] E. K. Patterson, S. Gurbuz, Z. Tufekci, and J. N. Gowdy, “CUAVE: a new audio-visual database for multimodal human-computer interface research,”*IEEE International conference on acoustics, speech, and signal processing*, IEEE, 2002. doi: 10.1109/ICASSP.2002.5745028.

[26] E. Bailly-Bailliére *et al.*, “The BANCA database and evaluation protocol,” in *Audio-and Video-Based Biometric Person Authentication: 4th International Conference*, Springer, 2003, pp. 625–638.

[27] L. Alzubaidi *et al.*, “Towards risk-free trustworthy artificial intelligence: Significance and requirements,” *International Journal of Intelligent Systems*, vol. 2023, 2023. doi: https://doi.org/10.1155/2023/4459198

[28] M. A. Fadhel, L. Alzubaidi, Y. Gu, J. Santamaría, and Y. Duan, “Real-time diabetic foot ulcer classification based on deep learning & parallel hardware computational tools,” *Multimedia Tools Application*, pp. 1–26, 2024. doi: https://doi.org/10.1007/s11042-024-18304-x

[29] F. K. H. Mihna, M. A. Habeeb, Y. L. Khaleel, Y. H. Ali, and L. A. E. Al-saeedi, “Using Information Technology for Comprehensive Analysis and Prediction in Forensic Evidence,” *Mesopotamian Journal of CyberSecurity*, vol. 2024, pp. 4–16, Mar. 2024, doi: 10.58496/MJCS/2024/002.

[30] L. Alzubaidi *et al.*, “Reliable deep learning framework for the ground penetrating radar data to locate the horizontal variation in levee soil compaction,” *Engineering Application of Artificial Intelligence*, vol. 129, p. 107627, 2024. doi: https://doi.org/10.1016/j.engappai.2023.107627

[31] A. A. Saihood, M. A. Hasan, M. A. Fadhel, L. Alzubaid, A. Gupta, and Y. Gu, “Multiside graph neural network-based attention for local co-occurrence features fusion in lung nodule classification,” *Expert System with Applications*, vol. 252, p. 124149, 2024. doi: https://doi.org/10.1016/j.eswa.2024.124149

[32] L. Alzubaidi *et al.*, “MEFF–A model ensemble feature fusion approach for tackling adversarial attacks in medical imaging,” *Intelligent Systems with Applications*, vol. 22, p. 200355, 2024. doi: https://doi.org/10.1016/j.iswa.2024.200355

[33] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, “Rethinking the inception architecture for computer vision,” in *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, 2016, pp. 2818–2826. doi: https://doi.org/10.48550/arXiv.1512.00567

[34] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, “Mobilenetv2: Inverted residuals and linear bottlenecks,” in *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, 2018, pp. 4510–4520. doi: https://doi.org/10.48550/arXiv.1801.04381

[35] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, “Densely Connected Convolutional Networks,” in *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, 2017, pp. 4700–4708.

[36] B. Zoph, V. Vasudevan, J. Shlens, and Q. V Le, “Learning transferable architectures for scalable image recognition,” in *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, 2018, pp. 8697–8710. doi: https://doi.org/10.48550/arXiv.1707.07012

[37] F. Chollet, “Xception: deep learning with depthwise separable convolutions,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 1251–1258. doi: https://doi.org/10.48550/arXiv.1610.02357