Knowledge Engineering and Data Science (KEDS)
Vol 7, No 2, December 2024, pp. 152–163

pISSN 2597-4602
eISSN 2597-4637

Constructing Qur'an Recitation Classification using Alexnet Algorithm

Harits Ar Rosyid [a,1*], Dzulkifli Abdullah [a,2], Mohammed S. Alqahtani [b,3]

[a]Departmen of Electrical Engineering, Universitas Negeri Malang
Jl. Semarang no. 05, Malang 65145, Indonesia
[b]Mechanical Engineering Department, College of Engineering, King Saud University
Riyadh 11451, Saudi Arabia
[1] harits.ar.ft@um.ac.id*; [2] dzulkifli.abdullah.1905356@students.um.ac.id; [3] malqahtanib@ksu.edu.sa
* corresponding author

ARTICLE INFO

Article history:
Received 20 September 2024
Revised 03 November 2024
Accepted 15 December 2024
Published online 24 December 2024

Keywords:
Classification
Speech Recognition
CNN
AlexNet
Al-Qur'an

ABSTRACT

The growing demands for accurate and efficient methods in the Qur'an recitation classification highlight the limitations of existing models, particularly in assisting the memorization process. This study aims to address these challenges by implementing the AlexNet Convolutional Neural Network architecture, widely recognized for its effectiveness in image classification, to classify the Qur'an recitations using the Mel-Frequency Cepstral Coefficient (MFCC) as the feature extraction method. The research involves several stages, including data collection, preprocessing (audio segmentation by verse), data augmentation, feature extraction, and classification using the AlexNet architecture, followed by performance evaluation. Key results demonstrate that the combination of MFCC and AlexNet yields promising accuracy in classifying Surah Al-Ikhlas recitations, suggesting its potential application for automatic reading correction. This approach significantly improves over traditional methods, contributing to more effective tools for Qur'an memorization assistance. Future work could explore its application in other significant improvement contexts and address potential challenges related to varying audio quality.

I. Introduction

Memorizing the Qur'an involves recalling its verses, which are then recited in front of a teacher or guide to ensure proper retention and pronunciation [1]. *Hafidz*, individuals who memorize the Qur'an, employ various methods to facilitate this process. The common memorization techniques include intensive programs such as the 30-Day Qur'an Memorization Camp, which is designed to enhance students' retention capabilities [2]. Another approach is the "One Day One Ayah (ODOA)" method, which encourages daily practice [3]. Additionally, the *Tikrar* Method, the oldest and most widely used strategy, relies on the repetitive recitation of verses to aid memorization [4]. However, the implementation of these methods often faces significant challenges. Learners frequently experience fatigue, sleepiness, and boredom due to demanding schedules during the intensive programs.

Distractions such as a preference for leisure activities over memorization diminish enthusiasm. Other obstacles include the extensive time required for oral evaluations and inconsistent memorization progress without direct guidance from teachers, or in this context, they usually call them-*ustadz* [5][6]. These issues underscore the urgent need for automated evaluation tools leveraging speech recognition technology integrated with machine learning algorithms.

Several studies have investigated the automated Qur'an recitation evaluation. For example, an Android-based application utilizing speech recognition technology has been developed to support the memorization of several Surahs, including *Al-Ikhlas*, *An-Naas*, and *Al-Kautsar* [7]. Similarly, speech recognition systems have been utilized to evaluate *Iqro's* recitations, addressing challenges in recognizing *hijaiyah* characters [8]. These studies highlight the potential for automated recitation evaluations through speech recognition. However, they did not delve deeply into the algorithms used. Another study also utilized Mel-Frequency Cepstral Coefficient (MFCC) feature extraction for checking Qur'an recitations [9]. Although MFCC effectively captured Qur'an recital audio features, its classification accuracy averaged only 51.8%, primarily due to the absence of advanced deep-

https://doi.org/10.17977/um018v7i22024p152-163

learning techniques [10]. These findings suggest the need for more sophisticated approaches, particularly those leveraging deep learning, to enhance the accuracy of automated Qur'an recitation evaluation.

Recent advancements in machine learning, particularly in deep learning, provide opportunities to improve upon these existing limitations. Convolutional Neural Networks (CNNs), a class of deep learning algorithms known for their exceptional performance in image and audio recognition tasks, have demonstrated significant potential in this area [11][12]. Among various CNN architectures, AlexNet stands out due to its balance of accuracy and computational efficiency [13]. Previous studies have highlighted AlexNet's superior performance, achieving accuracy rates of 98.41% to 99.14% in tasks such as classifying Arabic letter pronunciation [14]. This performance surpasses other CNN architectures, such as DCNN and ResNet, making it an ideal candidate for evaluating Qur'an recitations [15]. Given these challenges, selecting an appropriate deep learning model becomes crucial to improving the accuracy and efficiency of automated Qur'an recitation evaluation.

The superiority of AlexNet over other models like ResNet and Google Speech API lies in its simplicity and efficiency in feature extraction and classification [16]. While ResNet excels in tasks requiring deeper network architectures, it often entails higher computational demands and longer training times [17]. In contrast, AlexNet provides an optimally balanced trade-off between computational efficiency and accuracy, particularly for datasets with limited size [18]. Moreover, while Google Speech API has been widely applied for general speech recognition tasks, it lacks customization and fine-tuning capabilities for domain-specific tasks, such as the evaluation of Qur'an recitations [19]. These limitations highlight the need for an optimized approach that combines efficient feature extraction with a robust classification model to enhance Qur'an recitation evaluation.

Therefore, this study proposes a novel approach by integrating AlexNet with MFCC for feature extraction to enhance the accuracy of audio classification in Qur'an recitations, with a specific focus on *Surah Al-Ikhlas* as a case study. By inputting Al-Qur'an recital audio into the MFCC for feature extraction and processing it with the AlexNet algorithm, this research aims to significantly improve the classification accuracy of Qur'an recitations [7]. This methodology does not only address the technical limitations of previous systems but also provides a scalable solution for wider application in Qur'anic education.

Furthermore, advancements in speech recognition technology further enable the development of robust systems capable of accommodating diverse accents and recitation styles. For instance, integrating data augmentation techniques and transfer learning could further enhance model performance, making the system more adaptable to various recitation contexts [20][21]. The integration of this system into educational platforms holds the potential to revolutionize the way Qur'an memorization is taught, evaluated, and supported on a global scale.

By combining AlexNet and MFCC, this study introduces a groundbreaking approach to the limitations of earlier methods, paving the way for the automatic evaluation of Qur'an recitations. This proposed system is expected to contribute significantly to the field of Qur'anic studies and the development of AI-based educational tools. Beyond its technical advancements, it offers a practical framework for improving the accessibility and quality of Qur'anic education, promoting inclusivity and efficiency for learners worldwide. Ultimately, this integration not only enhances recitation assessment accuracy but also sets a foundation for future advancements in AI-driven religious education.

II. Method

The research method stages in this study are shown in Figure 1. Audio data recording is collected utilizing purposive sampling through TDI-BBQ, the Qur'an recitation guidance program, at Universitas Negeri Malang. The recordings were categorized into three classes: Class A for recitations with perfect and fluent pronunciation, Class B for those with good but not flawless pronunciation, and Class C for less accurate recitations. Thirty audio samples were collected for each category, comprising 15 male and 15 female voice recordings, resulting in a total of 90 data samples.

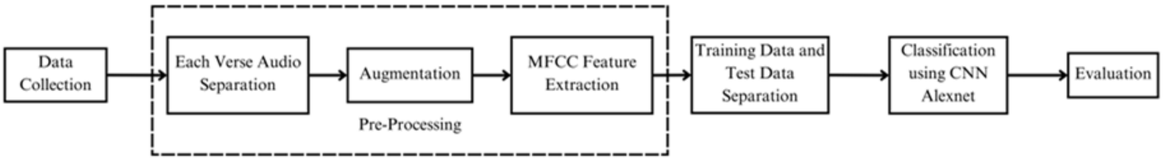

Data Collection
Each Verse Audio Separation
Augmentation
MFCC Feature Extraction
Pre-Processing
Training Data and Test Data Separation
Classification using CNN Alexnet
Evaluation

Fig. 1. Method flowchart

The audio recordings were segmented into five parts: Bismillah, *Ayah* 1, *Ayah* 2, *Ayah* 3, and *Ayah* 4 to create a verse-specific dataset. This segmentation was essential for developing a verse-by-verse training model. During the augmentation stage, the segmented data were enhanced using Audacity software and the Librosa library, following these steps:

A. *Tempo Adjustment*

The tempo adjustment was applied using the time stretch function from Librosa [22], increasing the tempo of the audio by 50% to introduce variations in the rhythm of the recitation. This adjustment reflects the natural diversity in recitation speeds among different reciters, enhancing the model's ability to generalize to faster-paced recitations.

B. *Pitch Adjustment*

Pitch adjustment was conducted using the pitch shift method from the same library, adjusting the pitch by ±30% to simulate variations in vocal frequencies. This ensures the dataset reflects a broader range of vocal characteristics, enabling the model to handle differences between male and female voices as well as natural pitch variations in human recitation.

C. *Adding White Noise*

White noise with an amplitude of 0.02 was added to the audio signal using a Gaussian distribution to emulate real-world conditions where background noise may be present [23]. Through these steps, the dataset size was expanded from 90 samples to 2,250 samples, incorporating diverse variations in tempo, pitch, and background noise. A comparative evaluation of models trained on the original dataset and the augmented dataset demonstrated a marked performance improvement. The results showed a significant improvement in accuracy (from 85% to 93%) and F1-score (from 0.82 to 0.91), highlighting the effectiveness of the augmentation techniques.

The augmented dataset was subsequently utilized in the feature extraction stage, where MFCC was calculated. This process converted the audio into 2D arrays consisting of 13-dimensional features and 13 delta coefficients. The extracted features were stored in CSV format and partitioned into training and testing datasets in an 80:20 ratio, ensuring adequate data for model training and evaluation. These augmentation and preparation steps were critical for training the AlexNet-based model, enabling it to perform reliably across diverse recitation styles and environments. This is consistent with findings from recent studies that highlight the importance of data variability in enhancing the generalization capabilities of deep learning models [24][25]. The augmented dataset was integral to the subsequent feature extraction and classification stages, contributing to the development of a robust system for evaluating recitations of *Surah Al-Ikhlas*.

The following pseudocode illustrates the audio augmentation process, including tempo adjustment, pitch shifting, and noise injection, to enhance data variability for training machine learning models on Qur'an recitation datasets. Recent studies have emphasized the critical impact of dataset size and quality on the generalization capabilities and reliability of deep learning models, particularly CNN architectures like AlexNet. Limited data restricts the ability of models to effectively learn diverse patterns, increasing the risk of overfitting and reducing performance on unseen data [26][27]. Illustrates the audio augmentation process, including tempo adjustment, pitch shifting, and noise injection, to increase data variability for training a machine learning model on a Qur'an recitation dataset presented in Pseudocode 1. These processes were implemented in three following stages.

PSEUDOCODE 1. Add audio data

```
    BEGIN AugmentAudio(file_path, output_path)

        # Step 1: Load the audio file
        audio = LoadAudio(file_path)

        # Step 2: Apply tempo adjustment
        tempo_adjusted_audio = AdjustTempo(audio, factor=1.5)

        # Step 3: Apply pitch adjustment
        pitch_up_audio = AdjustPitch(tempo_adjusted_audio, semitones=+1)
        pitch_down_audio = AdjustPitch(tempo_adjusted_audio, semitones=-3)

        # Step 4: Add white noise
        noisy_audio_up = AddWhiteNoise(pitch_up_audio, amplitude=0.02)
        noisy_audio_down = AddWhiteNoise(pitch_down_audio, amplitude=0.02)

        # Step 5: Save augmented audio files
        SaveAudio(noisy_audio_up, output_path + "-up.mp3")
        SaveAudio(noisy_audio_down, output_path + "-down.mp3")

    END AugmentAudio


    FUNCTION LoadAudio(file_path):
        RETURN audio_data

    FUNCTION AdjustTempo(audio, factor):
        # Modify the playback speed
        RETURN tempo_adjusted_audio

    FUNCTION AdjustPitch(audio, semitones):
        # Shift the pitch by the specified number of semitones
        RETURN pitch_adjusted_audio

    FUNCTION AddWhiteNoise(audio, amplitude):
        # Generate white noise and add it to the audio signal
        noise = GenerateNoise(length=audio.length, amplitude=amplitude)
        noisy_audio = Combine(audio, noise)
        RETURN noisy_audio

    FUNCTION SaveAudio(audio, file_path):
        # Save the processed audio to the specified path
        RETURN

 END
```

D. *Feature Extraction*

CNNs inherently require input in the form of 2D arrays, such as images or structured table data, for effective feature extraction and classification [12][28]. Thus, converting audio into arrays through a coefficient computation process applying MFCC is necessary. This study extracted 26 features, comprising 13-dimensional features and 13 delta coefficients, from the audio signals. The resulting data was stored in a CSV format.

The dataset file was divided into five categories, *Ayah* 1, *Ayah* 2, *Ayah* 3, and *Ayah* 4 to accommodate the training of five scenario models. The data obtained from MFCC is split into training and test data with a ratio of 80:20, where 80% is used for training and 20% for testing. The training data is used to train the model, while the test data is used to evaluate the trained model.

E. *Model Training*

The AlexNet architecture, consisting of five convolutional layers and three fully connected layers, was implemented for classification. This architecture offers a robust framework for feature extraction and prediction [29]. Therefore, the hyperparameter adjustments, such as batch size and the number of epochs, were fine-tuned. The batch size determines the amount of data processed in one epoch, with experiments conducted using sizes of 8, 16, 24, and 50, which refers to the number of training iterations. In this study, the model was trained over 60 epochs. During each epoch, the training data was propagated through all layers of the model. Following the processing of all samples, the model's weights and biases were updated using backpropagation to enhance learning and prediction accuracy.

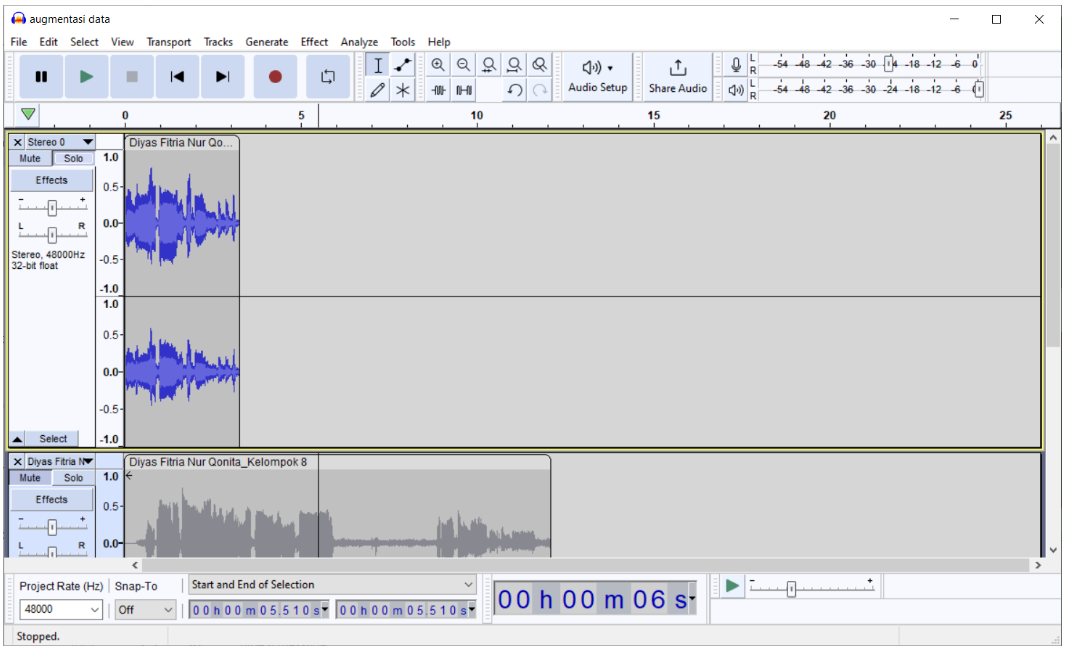
F. Evaluation

The evaluation phase involved testing the best-performing model from each scenario using the reserved 20% test dataset. Metric parameters used for evaluation included accuracy, precision, recall, and F1-score. The objective was to identify the model that achieved the highest performance in recognizing each verse of *Surah Al-Ikhlas*.

III. Result and Discussion

The data for this study were collected using purposive sampling from students enrolled in the Islamic Religious Education course and participants of the TDI-BBQ program at Universitas Negeri Malang. A total of 90 students participated, divided equally into three groups: 30 in Class A, 30 in Class B, and 30 in Class C. Each class was further stratified by gender, with 15 male and 15 female students in each group. Audio recordings of *Surah Al-Ikhlas* were obtained through the TDI-BBQ program administrators, resulting in 90 recordings.

The collected audio recordings were subsequently segmented into five parts: (1) *Bismillah*; (2) *Ayah* 1; (3) *Ayah* 2; (4) *Ayah* 3; and (5) *Ayah* 4. This segmentation was performed using Audacity software and carefully adjusted to match the length of each verse. Each segmented audio file was exported individually, creating a verse-specific dataset. The process of splitting the audio recordings into separate verses is illustrated in Figure 2.



Stereo, 48000Hz
32-bit float

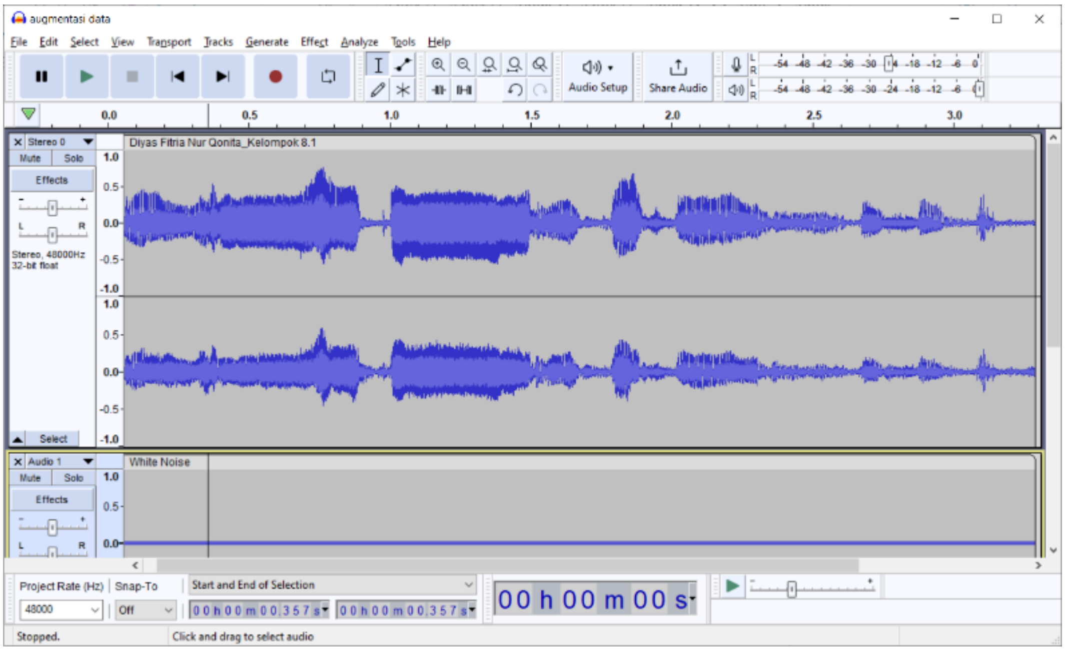
Project Rate (Hz) 48000 | Snap-To Off | Start and End of Selection 00 h 00 m 05.510 s 00 h 00 m 05.510 s

00 h 00 m 06 s

Stopped.

Fig. 2. Splitting audio by verse using Audacity

The augmentation stage was carried out manually using Audacity and Jupyter with the Librosa library on audio recordings segmented into verses from Classes A, B, and C. This process involved adding white noise (0.02), pitch-up +30%, pitch-down -30%, and tempo +50% to introduce variations in the audio. The augmented audio files were exported in WAV format. The augmentation produced more audio data to a total of 2250, with 450 samples per verse across five parts. These processes are shown in Figure 3.



Diyas Fitria Nur Qonita_Kelompok 8.1

Stereo, 48000Hz
32-bit float

White Noise

Project Rate (Hz) 48000

Snap-To Off

Start and End of Selection 00 h 00 m 00.357 s 00 h 00 m 00.357 s

Stopped.

Click and drag to select audio

Fig. 3. Audio augmentation with Audacity

Once augmentation was completed, the augmented audio was processed using MFCC feature extraction. This step produced 26 features for each audio sample, including 13 feature coefficients and 13 delta coefficients. The extracted features were exported as arrays and stored in CSV files, each labelled to represent the respective class and verse. The resulting dataset from feature extraction was grouped into 5 CSV files based on the five verses to conduct training and produce five models.

After preprocessing and feature extraction, the dataset was split into training and testing subsets with an 80:20 ratio. The training data was used to train the models, while the testing data was reserved for evaluation. Furthermore, the next step is to input the datasets into the CNN algorithm. In this process, the classification is divided into five scenarios by training five models according to verses, starting from bismillah, *Ayah* 1, *Ayah* 2, *Ayah* 3, and *Ayah* 4. Each scenario utilized the dataset prepared for the respective verse.

For the Bismillah model scenario, the dataset consisted of 450 samples categorized into three classes: bismillahA, bismillahB, and bismillahC. With an 80:20 split, 360 samples were allocated for training and 90 for testing. The model was trained five times to determine the best performance in each trial. The model training results obtained are shown in Figure 4.

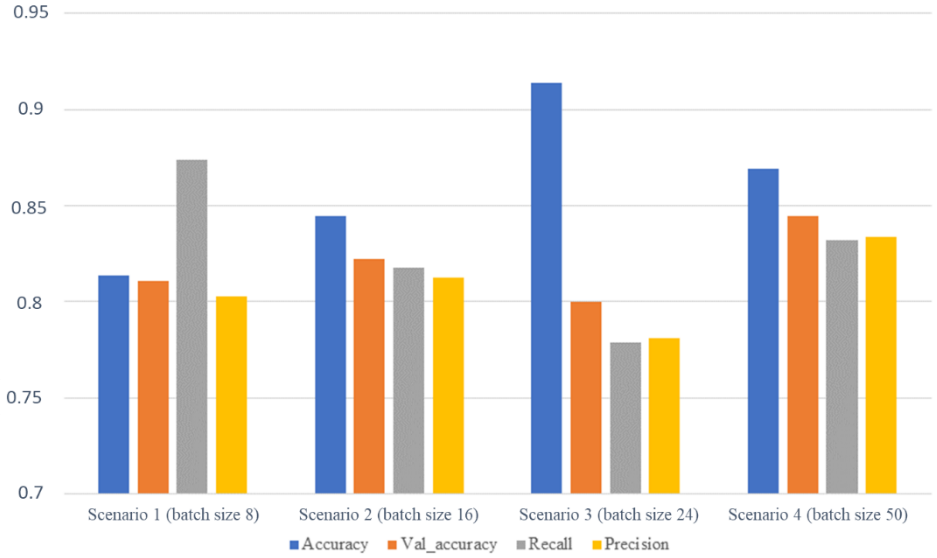


Fig. 4. Graph of *Bismillah* model training results

In the *Ayah* 1 model scenario, the dataset consisted of 450 samples distributed across three classes: *Ayah*1A, *Ayah*1B, and *Ayah*1C. The data was split into training (360 samples) and testing (90 samples) subsets in an 80:20 ratio. The model was trained five times to determine the best results from each trial. The training results for the *Ayah* 1 model are presented in Figure 5.

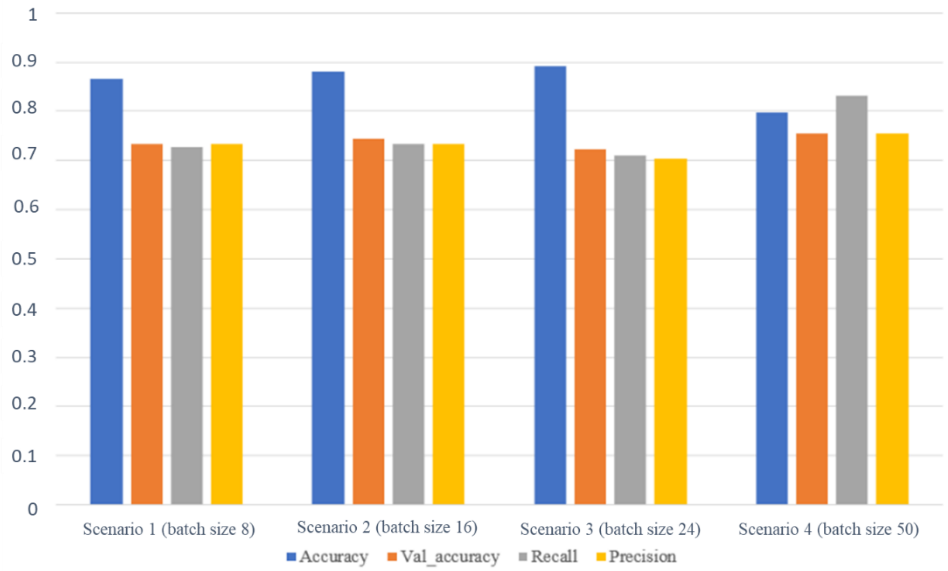


Fig. 5. Graph of the training results of the *Ayah* 1 model

Similarly, in the *Ayah* 2 model scenario, the dataset also included 450 samples divided into three classes: *Ayah*2A, *Ayah*2B, and *Ayah*2C. The same 80:20 split was applied, allocating 360 samples for training and 90 for testing. The model underwent five training iterations to achieve optimal performance in each trial. The training results for the *Ayah* 2 model are presented in Figure 6.

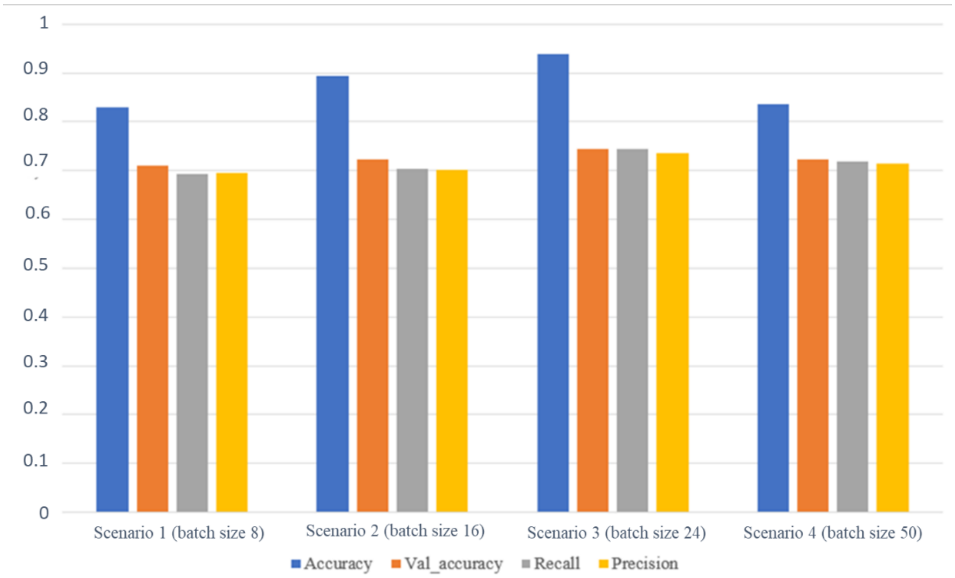


Fig. 6. Graph of the training results of the *Ayah* 2 model

In the *Ayah* 3 model scenario, the dataset again comprised 450 samples, categorized into three classes: *Ayah*3A, *Ayah*3B, and *Ayah*3C. The data distribution and training procedure mirrored the previous scenarios, with 360 training samples and 90 testing samples. This training model was trained 5 times to get the best results from each trial. The training results for the *Ayah* 3 model are presented in Figure 7.

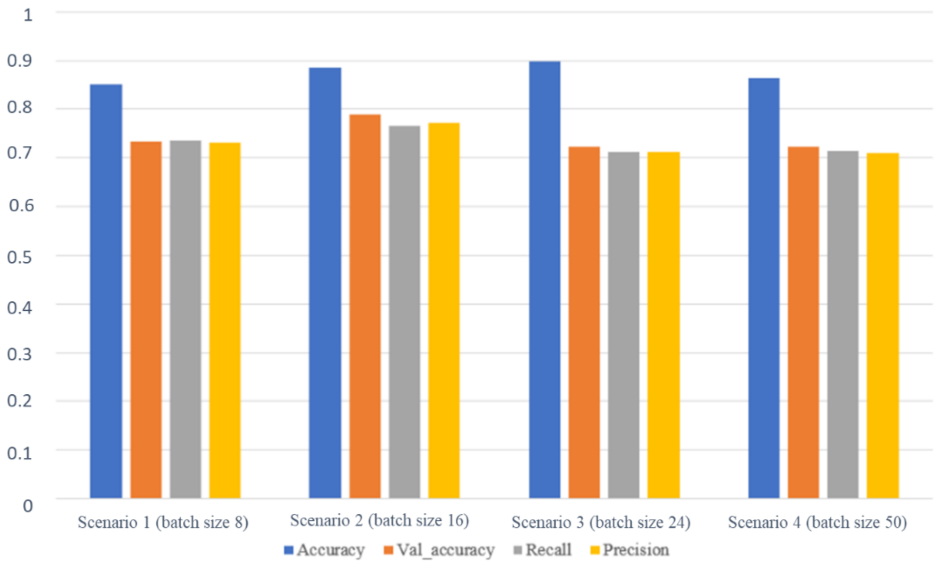


Fig. 7. Graph of the training results of the *Ayah* 3 model

Lastly, the *Ayah* 4 model scenario utilized a dataset of 450 samples grouped into three classes: *Ayah*4A, *Ayah*4B, and *Ayah*4C. As in the other scenarios, the data was divided into 360 training samples and 90 testing samples. The model was trained five times to refine its performance. The training results for the *Ayah* 4 model are presented in Figure 8.

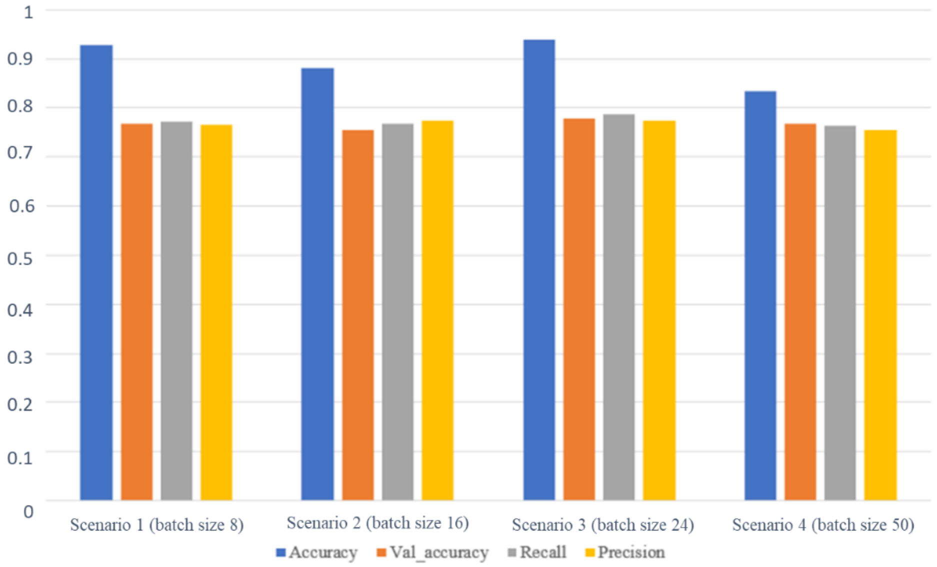


Fig. 8. Graph of the training results of the *Ayah* 4 model

After completing all scenarios and experiments by adjusting hyperparameters and running each model five times, the best accuracy and validation accuracy (val_accuracy) for each scenario were determined. It is presented in Table 1.

Table 1. Best results from each scenario

Hyperparameter		**Scenario**	**Accuracy**	**Val_accuracy**
Batch Size	**Epoch**			
50	60	*Bismillah* Model	86,94%	84,44%
50	60	*Ayah* 1 Model	79,72%	75,56%
24	60	*Ayah* 2 Model	93,89%	74,44%
16	60	*Ayah* 3 Model	88,61%	78,89%
24	60	*Ayah* 4 Model	93,89%	77,78%

The summarized results of accuracy and validation accuracy reinforce that optimal performance varies across scenarios, influenced by the dataset's complexity and the model's parameter configurations. The best-performing model from each scenario was further evaluated using a confusion matrix. This evaluation involved testing data previously separated during the dataset preparation stage. The testing data was input into the model to predict labels, which were then compared with the true labels to assess the model's prediction accuracy. The evaluation metrics derived from the confusion matrix included accuracy, precision, recall, and F1 score, providing a comprehensive assessment of the model's performance. The results of the confusion matrix are shown in Figure 9.

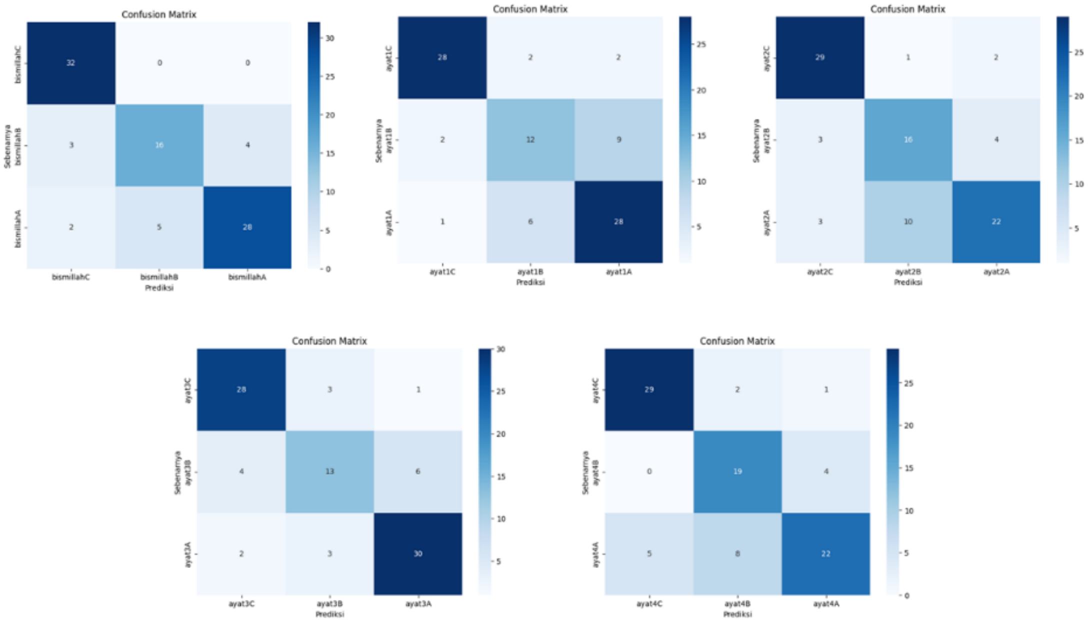


Fig. 9. Confusion matrix evaluation of the best results

Table 2 shows the performance evaluations of models trained on individual segments of *Surah Al-Ikhlas*. Each model's performance is assessed using four key metrics: Accuracy, Precision, Recall, and F1-Score. These metrics provide a comprehensive overview of how effectively each model classifies recitations of the corresponding segment. Key observations from the table include the following results.

Table 2. Model performance evaluation

Scenario	Accuracy	Precision	Recall	F1-Score
Bismillah Model	86,94%	83,39%	83,18%	83,01%
Ayah 1 Model	79,72%	74,03%	73,22%	73,45%
Ayah 2 Model	93,89%	73,56%	74,34%	73,46%
Ayah 3 Model	88,61%	77,28%	76,58%	76,69%
Ayah 4 Model	93,89%	77,42%	78,70%	77,30%

The models trained on *Ayah* 2 and *Ayah* 4 achieved the highest accuracy at 93.89%, making them the best-performing scenarios. This superior performance is likely due to the distinct phonetic features in these verses, which facilitate easier classification. High precision (77.42%) and recall (78.70%) further demonstrate the models' ability to classify positive instances while minimizing false negatives correctly. This observation aligns with [30], which emphasizes that datasets with clear phonetic distinctions often lead to better classification outcomes in audio recognition tasks.

In contrast, the model trained on *Ayah* 1 exhibited the lowest accuracy at 79.72%. The relatively lower precision (74.03%) and recall (73.22%) suggest challenges in accurately distinguishing recitations for this segment. This difficulty may be attributed to the phonetic similarities of *Ayah* 1 with other classes, increasing the complexity of classification. Prior research [25] suggests that overlapping audio features, such as shared phonemes, can significantly degrade model performance, particularly when the dataset size is limited.

The model trained on *Bismillah* achieved an accuracy of 86.94%, with balanced precision (83.39%) and recall (83.18%). Despite moderate performance, the F1-Score of 83.01% reflects the model's ability to generalize well across the dataset. The distinct rhythm and tonal features of *Bismillah* likely contributed to reducing feature overlap with other segments, which reduced feature overlap with other segments.

The F1-Score, which balances precision and recall, varied across models, highlighting the influence of verse-specific characteristics. Scores ranged from 73.45% (*Ayah* 1) to 83.01% (*Bismillah*). These results highlight the role of phonetic clarity in improving classification robustness. As discovered by [26], audio models trained on datasets with high intra-class variability tend to achieve lower F1 scores due to increased false negatives. Therefore, advanced preprocessing techniques, such as phonetic-specific adjustments, as demonstrated in [30], could enhance feature distinctiveness and improve classification accuracy.

The use of data augmentation techniques tempo adjustment, pitch shifting, and white noise addition—significantly enhanced the model's performance. Without augmentation, the original dataset of 90 recordings would have likely caused overfitting, leading to suboptimal accuracy. The augmentation expanded the dataset to 2,250 samples, introducing variations in rhythm, pitch, and background noise. This process allowed the model to learn more robust representations of Qur'an recitations. This is aligned with findings in [31][32] which emphasized the importance of dataset diversity in mitigating overfitting.

The results also emphasize the role of hyperparameter tuning. The models with batch sizes of 24 and 16 consistently performed better, particularly for *Ayah* 2 and *Ayah* 4. In addition, the smaller batch sizes introduced stochasticity during training, which enhanced generalization [33].

Training and validation accuracy trends, as shown in Figure 4 to Figure 8, reveal distinct patterns. The *Bismillah* model converged consistently with a training accuracy of 86.94% and a slightly lower validation accuracy of 84.44%. This indicates potential room for optimization, such as exploring advanced augmentation techniques or refining feature extraction methods. In addition, the models for *Ayah* 2 and *Ayah* 4 displayed stable accuracy improvements, reflecting the phonetic uniqueness of these verses. However, the model for *Ayah* 1 showed fluctuating accuracy, indicating difficulties in learning overlapping features. Incorporating phonetic-specific preprocessing, as proposed in [34], could mitigate these challenges and improve performance.

These evaluations show that the integration of MFCC with AlexNet represents a substantial improvement over traditional methods, such as the Google Speech API, which lacked the customization needed for domain-specific tasks. Earlier MFCC-based models achieved an average accuracy of 51.8% [35], while the current approach delivers significantly higher accuracies across all scenarios. This underscores the effectiveness of combining advanced feature extraction techniques with deep learning architectures for Qur'an recitation classification.

IV. Conclusion

This study utilized the AlexNet architecture CNN algorithm to develop a speech recognition model in classifying memorized recitations of *Surah Al-Ikhlas*. This approach integrates audio augmentation preprocessing and MFCC feature extraction. The MFCC-extracted features are trained on the AlexNet CNN algorithm to produce five models: Bismillah, *Ayah* 1, *Ayah* 2, *Ayah* 3, and *Ayah* 4. These models are designed for practical application in assessing Qur'an memorization.

The results indicate different ranges of accuracy and validation accuracy (val_accuracy) produced for the five models. For the Bismillah model, the accuracy ranged between 81.39% and 91.39%, with val_accuracy between 80% and 84.44%. For the *Ayah* 1 model, the accuracy ranged from 79.72% to 89.17%, with val_accuracy between 72.22% and 75.66%. The *Ayah* 2 model achieved accuracy between 83.06% and 93.89%, with val_accuracy ranging from 71.11% to 74.44%. The *Ayah* 3 model showed accuracy between 85% and 89.72%, with val_accuracy from 72.22% to 78.89%. Lastly, the *Ayah* 4 model recorded accuracy between 83.33% and 93.89%, with val_accuracy ranging from 75.56% to 77.78%. All experiments were conducted with hyperparameters such as batch sizes of 8, 16, 24, and 50 and an epoch count of 60. These results confirm that the combination of MFCC and the AlexNet CNN architecture is highly capable of performing speech recognition classification on *Surah Al-Ikhlas* audio.

This study highlights the significant contribution of the AlexNet-based model integrated with MFCC in the area of Qur'an recitation classification. By employing advanced audio augmentation techniques and leveraging deep learning algorithms, the model achieves remarkable accuracy, particularly for verses with distinct phonetic features, such as *Ayah* 2 and *Ayah* 4. This study not only advances methodologies for evaluating Qur'an recitations but also establishes a benchmark for future research in this domain.

The practical applications of the developed model are substantial. It can serve as an automated tool for assessing Qur'an memorization, reducing reliance on manual evaluation while ensuring accuracy and consistency. Moreover, the model's robustness in handling varied audio conditions makes it ideal for deployment in real-world educational settings, such as Islamic schools and Qur'an learning centres.

Future studies could focus on expanding the dataset to include diverse recitation styles and exploring alternative deep-learning architectures to enhance model performance further. Additionally, integrating the model into user-friendly mobile or web-based applications would broaden its accessibility and utility. This future study would underscore the importance of combining innovative feature extraction methods with tailored neural network architectures, significantly advancing the capabilities of Qur'anic recitation classification systems. It sets a foundation for more sophisticated tools that can support and improve Qur'an learning and memorization processes globally.

Declarations

Author contribution

All authors contributed equally as the main contributor of this paper. All authors read and approved the final paper.

Conflict of interest

The authors declare no known conflict of financial interest or personal relationships that could have appeared to influence the work reported in this paper.

Additional information

Reprints and permission information are available at http://journal2.um.ac.id/index.php/keds.

Publisher's Note: Department of Electrical Engineering and Informatics - Universitas Negeri Malang remains neutral with regard to jurisdictional claims and institutional affiliations.

References

[1] N. M. S. A. Nik Abdullah, F. S. Mohd Sabbri, and R. A. Muhammad Isa, "Tahfiz Students' Experiences in Memorizing the Qur'an: Unveiling Their Motivating Factors and Challenges," IIUM J. Educ. Stud., vol. 9, no. 2, pp. 42–63, Jun. 2021.

[2] A. B. Baried and M. Hannase, "Sufis And Women: The Study of Women's Sufis In The Western World," Refleksi, vol. 21, no. 1, Oct. 2022.

[3] R. Sari, S. Sakban, and D. Deprizon, "The Effect Of Application Of The ODOA (One Day One Verse) Method On The Ability To Memorize The Al-Qruan Of Class IV Students In Memorizing Surah Al- Bayyinah At Muhammdiyah 03 Unggunlan Pekanbaru Primary School," Kalijaga J. Penelit. Multidisiplin Mhs., vol. 1, no. 4, pp. 127–134, Aug. 2024.

[4] A. M. Diponegoro, I. H. Khotimah, and F. S. Setiawan, "Implementation of the Tikrar Method in BTQ (Guidance for Tahfidz Al Qur'an) Learning at Madrasah Ibtidaiyah," MUDARRISA J. Kaji. Pendidik. Islam, vol. 16, no. 2, pp. 269–283, Dec. 2024.

[5] A. Makrus and L. Usriyah, "Teacher Strategies in Enhancing Quranic Memorization and Psychological Implications for Quranic Memorizers: A Study at Mukhtar Syafa'at Banyuwangi's Distinguished Junior High School," IJIE Int. J. Islam. Educ., vol. 2, no. 1, pp. 13–28, Jun. 2023.

[6] N. Naufalita and R. Sari, "Understanding Anxiety among Students Who Memorize the Qur'an," J. Psikol. Integr., vol. 12, no. 1, pp. 66–82, Jun. 2024.

[7] R. Hadiyansah and R. Andamira, "Convolutional Neural Network (CNN) for Detecting Al-Qur'an Reciting and Memorizing," Khazanah J. Relig. Technol., vol. 1, no. 2, pp. 44–48, Dec. 2023.

[8] A. Asroni, K. R. Ku-Mahamud, C. Damarjati, and H. B. Slamat, "Arabic speech classification method based on padding and deep learning neural network," Baghdad Sci. J., vol. 18, no. 2 (Suppl.), p. 925, 2021.

[9] G. Samara, E. Al-Daoud, N. Swerki, and D. Alzu'bi, "The Recognition of Holy Qur'an Reciters Using the MFCCs' Technique and Deep Learning," Adv. Multimed., vol. 2023, pp. 1–14, Mar. 2023.

[10] Z. Li, B. Chen, S. Wu, M. Su, J. M. Chen, and B. Xu, "Deep learning for urban land use category classification: A review and experimental assessment," Remote Sens. Environ., vol. 311, p. 114290, Sep. 2024.

[11] L. Nanni, G. Maguolo, S. Brahnam, and M. Paci, "An Ensemble of Convolutional Neural Networks for Audio Classification," Appl. Sci., vol. 11, no. 13, p. 5796, Jun. 2021.

[12] M. M. Islam, S. Nooruddin, F. Karray, and G. Muhammad, "Human activity recognition using tools of convolutional neural networks: A state of the art review, data sets, challenges, and future prospects," Comput. Biol. Med., vol. 149, p. 106060, Oct. 2022.

[13] A. Ullah, H. Elahi, Z. Sun, A. Khatoon, and I. Ahmad, "Comparative Analysis of AlexNet, ResNet18 and SqueezeNet with Diverse Modification and Arduous Implementation," Arab. J. Sci. Eng., vol. 47, no. 2, pp. 2397–2417, Feb. 2022.

[14] A. Asif, H. Mukhtar, F. Alqadheeb, H. F. Ahmad, and A. Alhumam, "An Approach for Pronunciation Classification of Classical Arabic Phonemes Using Deep Learning," Appl. Sci., vol. 12, no. 1, p. 238, Dec. 2021.

[15] D. Jaganathan, S. Balsubramaniam, V. Sureshkumar, and S. Dhanasekaran, "Concatenated Modified LeNet Approach for Classifying Pneumonia Images," J. Pers. Med., vol. 14, no. 3, p. 328, Mar. 2024.

[16] S. Sharma and K. Guleria, "Deep Learning Models for Image Classification: Comparison and Applications," in 2022 2nd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE), Apr. 2022, pp. 1733–1738.

[17] M. Razavi, S. Mavaddati, and H. Koohi, "ResNet deep models and transfer learning technique for classification and quality detection of rice cultivars," Expert Syst. Appl., vol. 247, p. 123276, Aug. 2024.

[18] A. A. Masaoodi, H. I. Shahadi, and H. H. Abbas, "Eye Movement Recognition: Exploring Trade-Offs in Deep Learning Approaches with Development," 2024, pp. 238–251.

[19] R. Sobti, K. Guleria, and V. Kadyan, "Comprehensive literature review on children automatic speech recognition system, acoustic linguistic mismatch approaches and challenges," Multimed. Tools Appl., vol. 83, no. 35, pp. 81933–81995, Mar. 2024.

[20] S. P. Jakkaladiki and F. Maly, "Integrating hybrid transfer learning with attention-enhanced deep learning models to improve breast cancer diagnosis," PeerJ Comput. Sci., vol. 10, p. e1850, Feb. 2024.

[21] H. Kheddar, Y. Himeur, S. Al-Maadeed, A. Amira, and F. Bensaali, "Deep transfer learning for automatic speech recognition: Towards better generalization," Knowledge-Based Syst., vol. 277, p. 110851, Oct. 2023.

[22] A. Abeysinghe, S. Tohmuang, J. L. Davy, and M. Fard, "Data augmentation on convolutional neural networks to classify mechanical noise," Appl. Acoust., vol. 203, p. 109209, Feb. 2023.

[23] W. N. Manamperi, T. D. Abhayapala, P. N. Samarasinghe, and J. (Aimee) Zhang, "Drone audition: Audio signal enhancement from drone embedded microphones using multichannel Wiener filtering and Gaussian-mixture based post-filtering," Appl. Acoust., vol. 216, p. 109818, Jan. 2024.

[24] S. Ali et al., "Assessing generalisability of deep learning-based polyp detection and segmentation methods through a computer vision challenge," Sci. Rep., vol. 14, no. 1, p. 2032, Jan. 2024.

[25] P. Papadimitroulas et al., "Artificial intelligence: Deep learning in oncological radiomics and challenges of interpretability and data harmonization," Phys. Medica, vol. 83, pp. 108–121, Mar. 2021.

[26] C. Aliferis and G. Simon, "Overfitting, Underfitting and General Model Overconfidence and Under-Performance Pitfalls and Best Practices in Machine Learning and AI," 2024, pp. 477–524.

[27] M. M. Bejani and M. Ghatee, "A systematic review on overfitting control in shallow and deep neural networks," Artif. Intell. Rev., vol. 54, no. 8, pp. 6391–6438, Dec. 2021.

[28] T. Kattenborn, J. Leitloff, F. Schiefer, and S. Hinz, "Review on Convolutional Neural Networks (CNN) in vegetation remote sensing," ISPRS J. Photogramm. Remote Sens., vol. 173, pp. 24–49, Mar. 2021.

[29] S. B. Akbar, K. Thanupillai, and S. Sundararaj, "Combining the advantages of AlexNet convolutional deep neural network optimized with anopheles search algorithm based feature extraction and random forest classifier for COVID-19 classification," Concurr. Comput. Pract. Exp., vol. 34, no. 15, Jul. 2022.

[30] H. Aldarmaki, A. Ullah, S. Ram, and N. Zaki, "Unsupervised Automatic Speech Recognition: A review," Speech Commun., vol. 139, pp. 76–91, Apr. 2022.

[31] T. Islam, M. S. Hafiz, J. R. Jim, M. M. Kabir, and M. F. Mridha, "A systematic review of deep learning data augmentation in medical imaging: Recent advances and future research directions," Healthc. Anal., vol. 5, p. 100340, Jun. 2024.

[32] A. Gracia Moisés, I. Vitoria Pascual, J. J. Imas González, and C. Ruiz Zamarreño, "Data Augmentation Techniques for Machine Learning Applied to Optical Spectroscopy Datasets in Agrifood Applications: A Comprehensive Review," Sensors, vol. 23, no. 20, p. 8562, Oct. 2023.

[33] M. Segu, A. Tonioni, and F. Tombari, "Batch normalization embeddings for deep domain generalization," Pattern Recognit., vol. 135, p. 109115, Mar. 2023.

[34] Z. Liu, Z. Huang, L. Wang, and P. Zhang, "A Pronunciation Prior Assisted Vowel Reduction Detection Framework with Multi-Stream Attention Method," Appl. Sci., vol. 11, no. 18, p. 8321, Sep. 2021.

[35] L. Syafa'ah, R. Prasetyono, and H. Hariyady, "Enhancing Qur'anic Recitation Experience with CNN and MFCC Features for Emotion Identification," Kinet. Game Technol. Inf. Syst. Comput. Network, Comput. Electron. Control, May 2024.