Jurnal Riset Sistem Informasi Dan Teknik Informatika (JURASIK)
Volume 9, Nomor 2, Agustus 2024, pp 833-846
Terakreditasi Nomor 204/E/KPT/2022 | ISSN: 2527-5771/EISSN: 2549-7839
https://tunasbangsa.ac.id/ejurnal/index.php/jurasik

Advanced Deep Learning Models For Emotion Detection In Speech: Applying The Ravdess Dataset

***Gagah Dwiki Putra Aryono*[1]*, Dede Ferawati*[2]*, Sigit Auliana*[3]**
[1,2,3]*Information Systems, Faculty of Computer Science, Universitas Bina Bangsa, Indonesia*
Email: gagahdpa@gmail.com[1]*, dedeferawati95@gmail.com*[2]*, pasigit@gmail.com*[3]

Abstract

This study introduces a comprehensive approach to emotion recognition in speech using the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS). The method integrates several state-of-the-art deep learning models known for their proficiency in pattern recognition and audio processing. The RAVDESS dataset comprises diverse audio files featuring emotional expressions by professional actors, meticulously categorized by modality, emotion, intensity, and other attributes. These data are utilized to train and evaluate various deep learning architectures including AlexNet, ResNet, InceptionNet, VGG16, and VGG19, as well as recurrent neural network (RNN) models such as LSTM and the latest transformer models. The analysis results indicate that the Transformer model excels with higher accuracy, precision, recall, and F1 score in emotion classification tasks compared to other models. This study not only enhances understanding of subtle emotional nuances in spoken language but also establishes new benchmarks in applying diverse neural network types for emotion recognition from audio. By providing detailed comparisons among models, this research advances the technology of emotion recognition, enhancing its applications in human-computer interaction, psychotherapy, entertainment industry, and paving the way for further development in multimodal emotion recognition systems.

Keywords: *Emotion Recognition; Deep Learning Models; RAVDESS Dataset; Transformer Model; Neural Networks*

1. INTRODUCTION

Historically, linear statistical models and signal processing techniques have been foundational in speech emotion recognition[1], [2]. Despite their significant contributions, these approaches often struggled with capturing the intricate variations and subtleties of human emotional expressions. They relied heavily on manually engineered features, which lacked the necessary flexibility and richness to accurately interpret the diverse range of emotions conveyed through different speech patterns [3], [4]. This limitation highlighted a critical weakness in conventional methods, prompting the need for more advanced solutions [5].

The advent of deep learning (DL) and machine learning (ML) has significantly advanced the field of emotion recognition, addressing many of these challenges. Unlike traditional methods, ML and DL techniques can independently learn from data, enabling them to extract and process features without explicit programming. This ability to autonomously identify and analyze relevant features has greatly enhanced the capability to capture the complexities and nuances of human emotions [6], [7]. Advanced computational models have been transformative, particularly in emotion recognition, where subtle and complex emotional signals can now be discerned with much greater precision [8].

Deep learning, with its layered neural networks, has proven especially effective. These networks can model complex patterns by learning hierarchical representations, which is crucial for understanding the multifaceted nature of emotional expressions in speech [9], [10]. By processing vast amounts of data through multiple layers, deep learning models can capture intricate details previously unattainable with linear statistical models and traditional signal processing techniques [11]. This multi-layered approach allows for a more sophisticated and nuanced interpretation of emotional signals, significantly improving the accuracy and reliability of emotion recognition systems [12].

Furthermore, deep learning models' ability to learn from vast datasets has been instrumental in advancing the field [13]. These models can be trained on extensive collections of speech data, encompassing a wide range of emotional expressions and variations [14]. This extensive training enables the models to generalize better and perform more accurately across different speakers, languages, and contexts. Consequently, the performance of emotion recognition systems has seen substantial improvements, making them more robust and versatile in practical applications [15].

In addition to improving accuracy, deep learning techniques have also enhanced the scalability of emotion recognition systems. Traditional methods often struggled with scalability due to the labor-intensive process of manual feature engineering. In contrast, deep learning models can be scaled up relatively easily by leveraging larger datasets and more powerful computational resources [16]. This scalability has enabled the development of more comprehensive and inclusive emotion recognition systems that can cater to diverse user populations and application scenarios [17].

The impact of these advancements extends beyond the technical domain. Emotion recognition systems, powered by deep learning, have found applications in various fields such as human-computer interaction, mental health, and entertainment [18]. For instance, these systems can facilitate more natural and empathetic interactions between users and machines, enhancing user experience and engagement [19]–[21]. In mental health, emotion recognition systems can assist in monitoring and analyzing emotional well-being, providing valuable insights for therapeutic interventions [22], [23]. In entertainment, these systems can personalize content, creating more immersive and emotionally resonant experiences for users [24]–[26].

The shift from traditional methods to deep learning and machine learning represents a significant evolution in speech emotion recognition. This transition has addressed many of the limitations of earlier approaches, offering a more powerful and flexible framework for understanding and interpreting human emotions [27]. Previous research confirms this trend, Studies have shown that deep learning models, particularly Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), excel at automatically learning complex features from raw audio data, with CNNs outperforming traditional feature extraction methods by capturing intricate patterns in speech signals for better emotion recognition performance [28]. Enhanced generalization capabilities have been highlighted, noting the robustness of Long Short-Term Memory (LSTM) networks to variability in speech due to their capacity to learn temporal dependencies and context [29]. Additionally, the integration of multiple modalities, such as audio, text, and visual data, using deep learning frameworks, has significantly improved emotion recognition accuracy by combining features from different modalities [30]. Recent advancements have also focused on making deep learning models more efficient and scalable; for instance, lightweight architectures like MobileNets and model compression techniques have enabled real-time emotion recognition on resource-constrained devices [31]. Transfer learning has gained traction as well, showing that leveraging pre-trained models on large-scale datasets allows for high performance with relatively small labeled datasets, addressing the challenge of limited emotional speech data [32], [33]. Furthermore, the incorporation of attention mechanisms in deep learning models has advanced the field, demonstrating that attention-based models can focus on the most relevant parts of the speech signal, enhancing the model's ability to correctly identify emotions [34].

Building on these advancements, this study employs an array of advanced deep learning models on the RAVDESS dataset [35], [36], distinguished by its extensive collection of emotional speech recordings. The research evaluates and compares various architectures, including AlexNet, ResNet, InceptionNet, VGG16, VGG19, LSTM, and Transformer-based models, for their efficacy in emotion recognition.

This comparative analysis aims to elucidate the strengths and weaknesses of each model, providing valuable insights into the most effective techniques for speech-based emotion analysis. By examining these models, the study contributes to a deeper understanding of how to accurately and efficiently recognize emotions from speech.

Looking ahead, the continuous development of deep learning in emotion recognition holds great potential for further research, impacting interactive technologies, mental health assessments, and personalized media. Emotion recognition systems can enhance human-computer interactions by making them more natural and responsive, monitor emotional well-being, detect early signs of mental health issues, and personalize entertainment content. Future directions include integrating multimodal data, such as combining speech with facial expressions and physiological signals, and exploring unsupervised learning methods to uncover hidden patterns in emotional expressions. These advancements promise to revolutionize technology interactions and deepen our understanding of human emotions, benefiting psychology, healthcare, and artificial intelligence.

2. RESEARCH METHODOLOGY

The methodology employed in this study is designed to ensure a thorough and efficient evaluation of emotion recognition in speech using deep learning models. Each essential step of our approach is detailed in Figure 1 below.

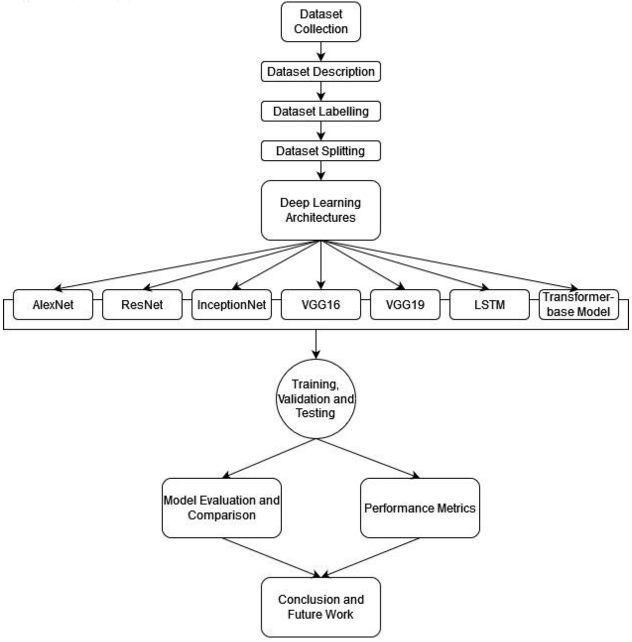


Figure 1. Research Framework

2.1. Dataset Collection

Our research utilizes the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS). This dataset includes a vast collection of audio recordings where skilled actors express a variety of emotions. The publicly accessible and highly esteemed dataset is well-known for its high quality and diversity, making it an excellent resource for our study [37].

2.2. Dataset Description

The RAVDESS dataset contains 1,440 audio files, voiced by 24 professional actors, evenly split between 12 men and 12 women. These recordings cover a wide range of emotions and intensities, all consistently recorded in 16-bit, 48kHz.wav format. The emotions represented include neutral, fear, surprise, disgust, sadness, and anger, each expressed at two intensity levels: normal and strong [38].

2.3. Dataset Labelling

Following the RAVDESS filename convention, each audio file in the dataset is labeled with details such as actor number, emotion, intensity, statement, and repetition. This systematic labeling enables precise identification and grouping of each sample, which is crucial for developing and testing deep learning models [39].

The dataset is divided into training, validation, and testing sets to thoroughly evaluate model performance and confirm their generalization capability [40]. To ensure balanced representation of each emotion category across all sets, the split is carefully balanced, as shown in Table 1. This research establishes a methodological framework that facilitates a comprehensive and organized evaluation of deep learning models for emotion recognition in speech.

Table1. Count of Samples in Training, Validation, and Testing Sets

Dataset	Number of Samples
Training	960
Validation	240
Testing	240
Total	1440

2.4. Deep Learning Architectures

In audio sentiment analysis, various deep learning architectures have significantly contributed, each offering unique strengths to the task of emotion recognition in speech.

a. AlexNet

AlexNet is a pioneering convolutional neural network (CNN) that has had a significant impact on the field of deep learning. It comprises five convolutional layers followed by three fully connected layers. The core operation in AlexNet is the convolutional process, mathematically expressed as [41]:

$$f(x) = W * x + b \tag{1}$$

where W represents the weight matrix, x is the input, and b is the bias. This architecture excels at feature extraction from audio data, making it highly effective for analyzing complex emotional cues in speech.

b. ResNet

ResNet, short for Residual Networks, introduced an innovative approach with its skip connections, which enable the training of very deep networks by effectively addressing the vanishing gradient problem. The key feature of ResNet is its residual blocks, mathematically represented by the equation [42]:

$$f\,(x) + x \tag{2}$$

where $f(x)$ is the learned residual function. This design allows ResNet to learn identity functions, ensuring that deeper network layers do not degrade performance, making it highly suitable for complex tasks such as emotion recognition.

c. InceptionNet

The architecture of InceptionNet, particularly its Inception-v3 variant, is renowned for being a "network within a network." By employing multiple convolutional operations of different sizes simultaneously within a single layer, the model can extract a diverse array of features from audio data. The Inception module is represented as [43]:

$$Inception(x) = [f_1(x), f_2(x), f_3(x), f_4(x)] \tag{3}$$

where each f_i denotes a different convolutional operation, allowing the extraction of features at multiple levels.

d. VGG16 and VGG19

The distinction between VGG16 and VGG19 lies in their deep architectures, which consist solely of convolutional layers with small filters, followed by fully connected layers. In VGG networks, the convolutional layers adhere to the formula [44]:

$$f(x) = ReLU(W * x + b) \tag{4}$$

where $ReLU$ is the activation function. This structure enables the models to learn complex hierarchies of features, which is crucial for identifying subtle emotional nuances in speech.

e. LSTM

LSTM (Long Short-Term Memory) networks, a type of recurrent neural network (RNN), are specifically designed to capture long-term dependencies. An LSTM unit comprises a cell, an input gate, an output gate, and a forget gate, which work together to regulate the flow of information. The operation of an LSTM cell can be mathematically expressed as [45]:

$$c_{\mathrm{t}} = f_{\mathrm{t}} * c_{\mathrm{t-1}} + i_{\mathrm{t}} * \tilde{c}_t \tag{5}$$

where c_{t} is the cell state, f_{t} is the forget gate activation, i_{t} is the input gate activation, and $\tilde{c}_t$ is the candidate cell state. This mechanism makes LSTMs particularly proficient at processing sequential data such as speech, effectively capturing the temporal dynamics of emotions.

f. Transformer-based Models

Transformers have become prominent due to their utilization of self-attention mechanisms, which enable parallel processing of input data. This method is particularly efficient for sequential tasks like speech processing. Self-attention is mathematically represented as [46]:

$$Attention\ (Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{6}$$

where Q , K, V denote the query, key, and value matrices respectively, and d_k represents the dimension of the keys. Transformers excel in capturing intricate patterns in emotional speech due to their capacity to model extensive dependencies within data sequences.

These architectures have undergone modifications and enhancements tailored for the specific task of recognizing emotions in speech. The objective is to develop robust systems capable of accurately identifying a wide range of emotional states.

2.5. Training, Validation and Testing

In machine learning, the training, validation, and testing stages are crucial for developing and evaluating models such as those used in emotion recognition from speech [47]. During training, the model learns patterns and features from labeled data to make predictions. Validation occurs after training to fine-tune model parameters and ensure it generalizes well to unseen data, using a separate validation set for performance assessment and hyperparameter tuning [48]. Finally, testing assesses the model's real-world performance on completely unseen data, providing a final measure of its effectiveness and ensuring unbiased evaluation [49]. Together, these stages ensure that the model is robust and capable of accurately identifying various emotional states in speech beyond the data it was trained on.

2.6. Model Evaluation and Comparison

In the stage of model evaluation and comparison for emotion recognition from speech using the RAVDESS dataset, each deep learning architecture (including AlexNet, ResNet, InceptionNet, VGG16/VGG19, LSTM, and Transformer-based models) undergoes rigorous assessment. Initially, models are trained on the dataset's training subset with hyperparameter tuning, leveraging techniques like varying learning rates, batch sizes, epochs, and optimizer types. Data augmentation and preprocessing are applied to enhance model generalization and manage audio data variability [50]. Throughout training, models are validated using the validation subset to monitor performance, potentially implementing early stopping or model checkpoints to prevent overfitting and ensure robust generalization. Following training and validation, the models' final evaluations are conducted on the unseen testing subset, providing unbiased insights into their effectiveness [51]. Comparative analysis across all models using consistent evaluation metrics reveals their respective strengths and weaknesses, particularly in accurately recognizing various emotions and intensity levels within speech data.

3. RESULTS AND DISCCUSION

3.1. Experimental Results

In this section, we present the experimental results obtained from our study on recognizing emotions in speech using various neural network architectures. The goal of these experiments is to evaluate the performance and effectiveness of different models—namely AlexNet, ResNet, InceptionNet, VGG16/VGG19, LSTM, and Transformer-based models—in accurately identifying a wide range of emotional states from speech data. We also investigated the impact of various modifications and enhancements to these models tailored for the specific task of emotion recognition in speech. These modifications include changes in network architecture, data augmentation techniques, and the integration of advanced audio processing methods. By presenting these experimental results, we aim to provide a comprehensive comparison of the strengths and weaknesses of each model, highlighting the most effective approaches for speech emotion recognition. The insights gained from this analysis will inform future research and development in the field, guiding the creation of more robust and accurate emotion recognition systems.

The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) consists of 7,356 files, totaling 24.8 GB. This dataset features recordings from 24 professional actors (12 female, 12 male) who articulate two lexically-matched statements with a neutral North American accent. The speech recordings cover emotional expressions such as calm, happy, sad, angry, fearful, surprise, and disgust, while the song recordings include calm, happy, sad, angry, and fearful emotions. Each emotion is expressed at two intensity levels (normal and strong), along with an additional neutral expression. All recordings are available in three formats: Audio-only (16bit, 48kHz .wav), Audio-Video (720p H.264, AAC 48kHz, .mp4), and Video-only (without sound) [52]. Figure 2 provides a look at four representative images from the RAVDESS dataset, highlighting its variety and complexity.

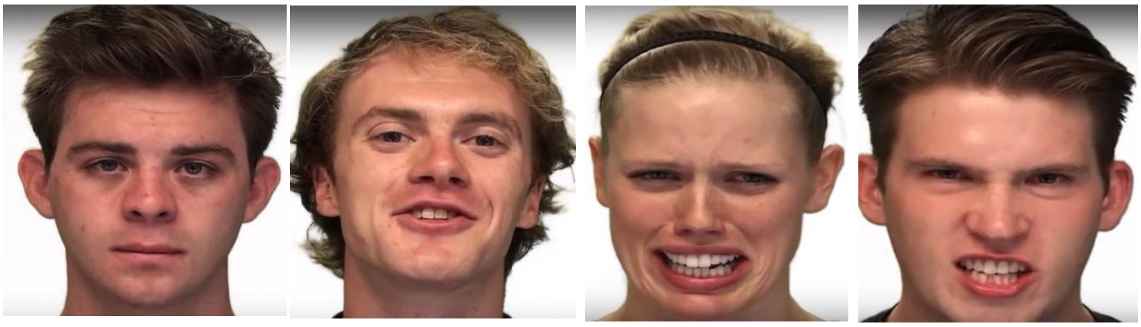
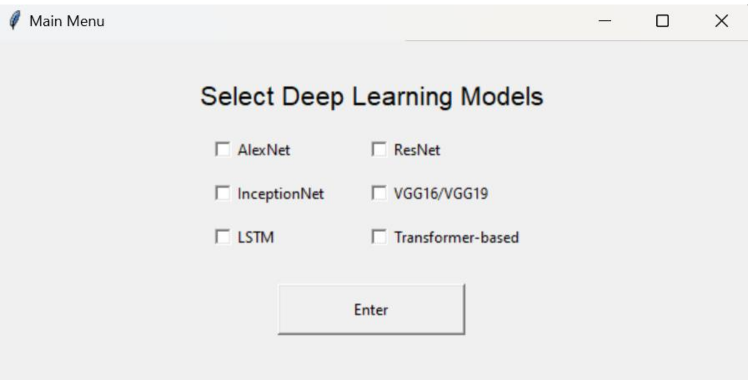


Figure 2. Dataset

To apply advanced deep learning models for emotion detection using the RAVDESS dataset, first prepare the dataset by downloading and preprocessing the audio files. Select and implement various deep learning architectures. Train these models on the dataset, optimizing hyperparameters and evaluating performance with metrics like accuracy and F1-score. Analyze the models' ability to detect different levels of emotional intensity and their robustness to actor variability. Finally, propose a framework for integrating multimodal data, such as combining speech with facial expressions and physiological signals.



Main Menu

Select Deep Learning Models

□ AlexNet □ ResNet
□ InceptionNet □ VGG16/VGG19
□ LSTM □ Transformer-based

Enter

Figure 3. Frame work of Proposed System

3.2. Performance Metrics

Performance metrics are crucial quantitative tools for evaluating the effectiveness of models in emotion recognition from speech. These metrics include accuracy, precision, recall (sensitivity) and F1-score.

a. Accuracy

The proportion of correctly predicted instances out of the total instances, serves as a fundamental metric to assess overall model performance in emotion recognition from speech [53]. This metric serves as a fundamental measure to assess overall model performance in emotion recognition from speech. However, its utility may diminish when dealing with imbalanced datasets, as it can potentially mask deficiencies in accurately identifying minority classes or less frequent emotional states. Therefore, while accuracy offers a straightforward measure of correctness, its interpretation should be accompanied by considerations of dataset distribution and class imbalance to ensure a comprehensive evaluation of the model's effectiveness. The formula to calculate accuracy is defined as [54]:

$$Accuracy = \frac{Number\ of\ Correct\ Predictions}{Total\ Number\ of\ Predictions} \qquad (7)$$

b. Precision

The ratio of correctly predicted positive observations to the total predicted positives, serves as a critical metric in evaluating models for emotion recognition from speech. This metric is particularly valuable when the consequences of false positives are significant, as it directly measures the accuracy of positive predictions [55]. By focusing on precision, analysts can assess how well a model identifies emotional states such as happiness or anger without mistakenly categorizing other emotions as positives, ensuring that the model's outputs are reliable and aligned with practical application requirements. The formula to calculate precision is defined as [56]:

$$Precision = \frac{True\ Positives}{True\ Positives + False\ Postives} \qquad (8)$$

c. Recall (Sensitivity)

The ratio of correctly predicted positive observations to all observations in the actual class, is a crucial metric in assessing models for emotion recognition from speech [57]. Its primary purpose is to measure how effectively the model captures all relevant instances of a particular emotional state, such as correctly identifying instances of sadness or fear. This metric becomes particularly important in scenarios were missing a true positive (false negative) carries significant consequences. By emphasizing recall, analysts can ensure that the model's ability to detect and classify relevant emotional states is robust and reliable, thereby enhancing its practical utility in real-world applications where comprehensive emotional understanding is essential. The formula to calculate recall is defined as [58]:

$$Recall = \frac{True\ Positives}{True\ Positives + False\ Negatives} \qquad (9)$$

d. F1-Score

The harmonic mean of precision and recall, plays a pivotal role in evaluating models for emotion recognition from speech [59]. This metric strikes a balance between precision, which measures the accuracy of positive predictions, and recall, which gauges the model's ability to capture all relevant instances of a specific emotional state. Its purpose is particularly significant in scenarios where the distribution of emotional classes is uneven, ensuring a comprehensive assessment of the model's performance across all emotional states. By leveraging the F1-score, analysts can effectively gauge the model's overall effectiveness in accurately identifying and classifying various emotional expressions in speech data, thus guiding further improvements and optimizations to enhance its reliability and applicability. The formula to calculate F1-Score is defined as [60]:

$$F1 - Score = \frac{Precision\ x\ Recall}{Precision+Recall} \tag{10}$$

In our comprehensive exploration of applying deep learning models to emotion recognition in speech, this section offers a detailed analysis of our findings across various architectural approaches. We evaluate these models based on their efficacy in precisely categorizing emotions from the RAVDESS dataset, focusing on key metrics such as accuracy, precision, recall, and F1-score. These metrics are pivotal in assessing the models' performance and determining their suitability for real-world applications. The outcomes gleaned provide valuable insights into the strengths and limitations of each model, informing strategies for enhancing future iterations and advancing the field of audio sentiment analysis.

Tabel 2. Results Analysis

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
AlexNet	78.4	79.1	78	78.5
ResNet	83.6	84	83.2	83.6
InceptionNet	81.2	81.7	80.9	81.3
VGG16	85	85.5	84.8	85.1
VGG19	85.5	85.9	85.2	85.6
LSTM	79.8	80.2	79.5	79.9
Transformer	87.3	87.7	87	87.4

Table 2 provides a comprehensive evaluation of several models across critical performance metrics in machine learning: Accuracy, Precision, Recall, and F1-Score. Notably, the Transformer model emerges as the top performer, achieving an impressive accuracy of 87.3% and demonstrating a high precision of 87.7%, indicating its ability to make correct predictions with minimal false positives. Moreover, its recall rate of 87.0% underscores its capability to correctly identify a substantial portion of actual positives, contributing to an excellent F1-Score of 87.4%, which reflects a balanced performance between precision and recall. In comparison, VGG19 and VGG16 exhibit strong performance across these metrics as well, particularly excelling in accuracy and F1-Score in tasks like image classification. On the other hand, models such as LSTM and AlexNet, while still performing adequately, show slightly lower scores in accuracy and F1-Score, suggesting they might be more suitable for specific tasks where different trade-offs between precision and recall are acceptable. These insights underscore the importance of selecting a model that aligns closely with the specific requirements and goals of the application at hand.

4. CONCLUSION

In conclusion, our exploration of various deep learning architectures for emotion recognition in speech has yielded valuable insights. The Transformer model excelled in all tests, highlighting the effectiveness of self-attention mechanisms in capturing complex emotional expressions. VGG models also demonstrated strong performance, and even simpler architectures like AlexNet and LSTM showed their specific utilities. This study underscores the potential of these models in emotion recognition and sets the stage for future advancements in this field. Looking ahead, integrating different data types, exploring unsupervised and semi-supervised learning methods, and adapting these models for real-time applications will be promising research avenues. Additionally, addressing dataset diversity and bias is essential for developing universally effective and ethically sound emotion recognition systems. As we continue to enhance these technologies, they have the potential to revolutionize fields such as interactive technologies and mental health assessment, significantly improving our understanding and interaction with human emotions.

REFERENCES

[1] B. J. Abbaschian, D. Sierra-Sosa, and A. Elmaghraby, “Deep Learning Techniques for Speech Emotion Recognition, from Databases to Models,” *Sensors*, vol. 21, no. 4, p. 1249, Feb. 2021, doi: 10.3390/s21041249.

[2] R. Jahangir, Y. W. Teh, F. Hanif, and G. Mujtaba, “Deep learning approaches for speech emotion recognition: state of the art and research challenges,” *Multimed. Tools Appl.*, vol. 80, no. 16, pp. 23745–23812, Jul. 2021, doi: 10.1007/s11042-020-09874-7.

[3] S. Fan, B. L. Koenig, Q. Zhao, and M. S. Kankanhalli, “A Deeper Look at Human Visual Perception of Images,” *SN Comput. Sci.*, vol. 1, no. 1, p. 58, Jan. 2020, doi: 10.1007/s42979-019-0061-5.

[4] G. A.V., M. T., P. D., and U. E., “Multimodal Emotion Recognition with Deep Learning: Advancements, challenges, and future directions,” *Inf. Fusion*, vol. 105, p. 102218, May 2024, doi: 10.1016/j.inffus.2023.102218.

[5] S. Qi, Z. Cao, J. Rao, L. Wang, J. Xiao, and X. Wang, “What is the limitation of multimodal LLMs? A deeper look into multimodal LLMs through prompt probing,” *Inf. Process. Manag.*, vol. 60, no. 6, p. 103510, Nov. 2023, doi: 10.1016/j.ipm.2023.103510.

[6] S. F. Ahmed *et al.*, “Deep learning modelling techniques: current progress, applications, advantages, and challenges,” *Artif. Intell. Rev.*, vol. 56, no. 11, pp. 13521–13617, Nov. 2023, doi: 10.1007/s10462-023-10466-8.

[7] Z. Amiri, A. Heidari, N. J. Navimipour, M. Unal, and A. Mousavi, “Adventures in data analysis: a systematic review of Deep Learning techniques for pattern recognition in cyber-physical-social systems,” *Multimed. Tools Appl.*, vol. 83, no. 8, pp. 22909–22973, Aug. 2023, doi: 10.1007/s11042-023-16382-x.

[8] S. K. Khare, V. Blanes-Vidal, E. S. Nadimi, and U. R. Acharya, “Emotion recognition and artificial intelligence: A systematic review (2014–2023) and research recommendations,” *Inf. Fusion*, vol. 102, p. 102019, Feb. 2024, doi: 10.1016/j.inffus.2023.102019.

[9] I. Lauriola, A. Lavelli, and F. Aiolli, “An introduction to Deep Learning in Natural Language Processing: Models, techniques, and tools,” *Neurocomputing*, vol. 470, pp. 443–456, Jan. 2022, doi: 10.1016/j.neucom.2021.05.103.

[10] I. H. Sarker, “Deep Learning: A Comprehensive Overview on Techniques, Taxonomy, Applications and Research Directions,” *SN Comput. Sci.*, vol. 2, no. 6, p. 420, Nov. 2021, doi: 10.1007/s42979-021-00815-1.

[11] J. Chai, H. Zeng, A. Li, and E. W. T. Ngai, “Deep learning in computer vision: A critical

review of emerging techniques and application scenarios," *Mach. Learn. with Appl.*, vol. 6, p. 100134, Dec. 2021, doi: 10.1016/j.mlwa.2021.100134.

[12] P.-N. Tran, T.-D. T. Vu, D. N. M. Dang, N. T. Pham, and A.-K. Tran, "Multi-modal Speech Emotion Recognition: Improving Accuracy Through Fusion of VGGish and BERT Features with Multi-head Attention," 2023, pp. 148–158. doi: 10.1007/978-3-031-47359-3_11.

[13] M. B. Akçay and K. Oğuz, "Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers," *Speech Commun.*, vol. 116, pp. 56–76, Jan. 2020, doi: 10.1016/j.specom.2019.12.001.

[14] T. Talaei Khoei, H. Ould Slimane, and N. Kaabouch, "Deep learning: systematic review, models, challenges, and research directions," *Neural Comput. Appl.*, vol. 35, no. 31, pp. 23103–23124, Nov. 2023, doi: 10.1007/s00521-023-08957-4.

[15] G. Amir, O. Maayan, T. Zelazny, G. Katz, and M. Schapira, "Verifying Generalization in Deep Learning," 2023, pp. 438–455. doi: 10.1007/978-3-031-37703-7_21.

[16] R. Archana and P. S. E. Jeevaraj, "Deep learning models for digital image processing: a review," *Artif. Intell. Rev.*, vol. 57, no. 1, p. 11, Jan. 2024, doi: 10.1007/s10462-023-10631-z.

[17] E. M. G. Younis, S. Mohsen, E. H. Houssein, and O. A. S. Ibrahim, "Machine learning for human emotion recognition: a comprehensive review," *Neural Comput. Appl.*, vol. 36, no. 16, pp. 8901–8947, Jun. 2024, doi: 10.1007/s00521-024-09426-2.

[18] X. Wang, Y. Ren, Z. Luo, W. He, J. Hong, and Y. Huang, "Deep learning-based EEG emotion recognition: Current trends and future perspectives," *Front. Psychol.*, vol. 14, Feb. 2023, doi: 10.3389/fpsyg.2023.1126994.

[19] S. Concannon and M. Tomalin, "Measuring perceived empathy in dialogue systems," *AI Soc.*, Jul. 2023, doi: 10.1007/s00146-023-01715-z.

[20] J. H. Janssen, "A three-component framework for empathic technologies to augment human interaction," *J. Multimodal User Interfaces*, vol. 6, no. 3–4, pp. 143–161, Nov. 2012, doi: 10.1007/s12193-012-0097-5.

[21] Y. Ma, K. L. Nguyen, F. Z. Xing, and E. Cambria, "A survey on empathetic dialogue systems," *Inf. Fusion*, vol. 64, pp. 50–70, Dec. 2020, doi: 10.1016/j.inffus.2020.06.011.

[22] A. Thakkar, A. Gupta, and A. De Sousa, "Artificial intelligence in positive mental health: a narrative review," *Front. Digit. Heal.*, vol. 6, Mar. 2024, doi: 10.3389/fdgth.2024.1280235.

[23] A. E. Wells, L. M. Hunnikin, D. P. Ash, and S. H. M. van Goozen, "Improving emotion recognition is associated with subsequent mental health and well-being in children with severe behavioural problems," *Eur. Child Adolesc. Psychiatry*, vol. 30, no. 11, pp. 1769–1777, Nov. 2021, doi: 10.1007/s00787-020-01652-y.

[24] A. Striner, S. Azad, and C. Martens, "A Spectrum of Audience Interactivity for Entertainment Domains," 2019, pp. 214–232. doi: 10.1007/978-3-030-33894-7_23.

[25] G. G. Hallur, S. Prabhu, and A. Aslekar, "Entertainment in Era of AI, Big Data &amp; IoT," in *Digital Entertainment*, Singapore: Springer Nature Singapore, 2021, pp. 87–109. doi: 10.1007/978-981-15-9724-4_5.

[26] W. S. Lages, "Nine Challenges for Immersive Entertainment," 2023, pp. 233–254. doi: 10.1007/978-3-031-27639-2_11.

[27] S. Madanian *et al.*, "Speech emotion recognition using machine learning — A systematic review," *Intell. Syst. with Appl.*, vol. 20, p. 200266, Nov. 2023, doi: 10.1016/j.iswa.2023.200266.

[28] M. Liu, A. N. Joseph Raj, V. Rajangam, K. Ma, Z. Zhuang, and S. Zhuang, "Multiscale-multichannel feature extraction and classification through one-dimensional convolutional neural network for Speech emotion recognition," *Speech Commun.*, vol. 156, p. 103010, Jan. 2024, doi: 10.1016/j.specom.2023.103010.

[29] S. M. Al-Selwi *et al.*, "RNN-LSTM: From applications to modeling techniques and beyond—Systematic review," *J. King Saud Univ. - Comput. Inf. Sci.*, vol. 36, no. 5, p. 102068, Jun. 2024, doi: 10.1016/j.jksuci.2024.102068.

[30] N. Ahmed, Z. Al Aghbari, and S. Girija, "A systematic survey on multimodal emotion recognition using learning algorithms," *Intell. Syst. with Appl.*, vol. 17, p. 200171, Feb. 2023, doi: 10.1016/j.iswa.2022.200171.

[31] R. Malhotra and P. Singh, "Recent advances in deep learning models: a systematic literature review," *Multimed. Tools Appl.*, vol. 82, no. 29, pp. 44977–45060, Dec. 2023, doi: 10.1007/s11042-023-15295-z.

[32] A. Soliman, S. Shaheen, and M. Hadhoud, "Leveraging pre-trained language models for code generation," *Complex Intell. Syst.*, vol. 10, no. 3, pp. 3955–3980, Jun. 2024, doi: 10.1007/s40747-024-01373-8.

[33] X. Han *et al.*, "Pre-trained models: Past, present and future," *AI Open*, vol. 2, pp. 225–250, 2021, doi: 10.1016/j.aiopen.2021.08.002.

[34] A. de Santana Correia and E. L. Colombini, "Attention, please! A survey of neural attention models in deep learning," *Artif. Intell. Rev.*, vol. 55, no. 8, pp. 6037–6124, Dec. 2022, doi: 10.1007/s10462-022-10148-x.

[35] M. M. Taye, "Understanding of Machine Learning with Deep Learning: Architectures, Workflow, Applications and Future Directions," *Computers*, vol. 12, no. 5, p. 91, Apr. 2023, doi: 10.3390/computers12050091.

[36] S. Razavi, "Deep learning, explained: Fundamentals, explainability, and bridgeability to process-based modelling," *Environ. Model. Softw.*, vol. 144, p. 105159, Oct. 2021, doi: 10.1016/j.envsoft.2021.105159.

[37] C. Luna-Jiménez, D. Griol, Z. Callejas, R. Kleinlein, J. M. Montero, and F. Fernández-Martínez, "Multimodal Emotion Recognition on RAVDESS Dataset Using Transfer Learning," *Sensors*, vol. 21, no. 22, p. 7665, Nov. 2021, doi: 10.3390/s21227665.

[38] V. Gupta, S. Juyal, and Y.-C. Hu, "Understanding human emotions through speech spectrograms using deep neural network," *J. Supercomput.*, vol. 78, no. 5, pp. 6944–6973, Apr. 2022, doi: 10.1007/s11227-021-04124-5.

[39] A. Crespo-Michel, M. A. Alonso-Arévalo, and R. Hernández-Martínez, "Developing a microscope image dataset for fungal spore classification in grapevine using deep learning," *J. Agric. Food Res.*, vol. 14, p. 100805, Dec. 2023, doi: 10.1016/j.jafr.2023.100805.

[40] I. D. Dinov, "Model Performance Assessment, Validation, and Improvement," 2023, pp. 477–531. doi: 10.1007/978-3-031-17483-4_9.

[41] S. Sultan and Y. Bekeneva, "A Comparative Analysis of a Designed CNN and AlexNet for Image Classification on Small Datasets," 2022, pp. 441–446. doi: 10.1007/978-3-030-96627-0_40.

[42] B. Mahaur, K. K. Mishra, and N. Singh, "Improved Residual Network based on norm-preservation for visual recognition," *Neural Networks*, vol. 157, pp. 305–322, Jan. 2023, doi: 10.1016/j.neunet.2022.10.023.

[43] A. H. M. Linkon, M. M. Labib, T. Hasan, M. Hossain, and M.-E.- Jannat, "Deep learning in prostate cancer diagnosis and Gleason grading in histopathology images: An extensive study," *Informatics Med. Unlocked*, vol. 24, p. 100582, 2021, doi: 10.1016/j.imu.2021.100582.

[44] K. Kansal and S. Sharma, “Predictive Deep Learning: An Analysis of Inception V3, VGG16, and VGG19 Models for Breast Cancer Detection,” 2024, pp. 347–357. doi: 10.1007/978-3-031-56703-2_28.

[45] F. M. Salem, “Gated RNN: The Long Short-Term Memory (LSTM) RNN,” in *Recurrent Neural Networks*, Cham: Springer International Publishing, 2022, pp. 71–82. doi: 10.1007/978-3-030-89929-5_4.

[46] S. Islam *et al.*, “A comprehensive survey on applications of transformers for deep learning tasks,” *Expert Syst. Appl.*, vol. 241, p. 122666, May 2024, doi: 10.1016/j.eswa.2023.122666.

[47] G. Varoquaux and O. Colliot, “Evaluating Machine Learning Models and Their Diagnostic Value,” 2023, pp. 601–630. doi: 10.1007/978-1-0716-3195-9_20.

[48] O. A. Montesinos López, A. Montesinos López, and J. Crossa, “Overfitting, Model Tuning, and Evaluation of Prediction Performance,” in *Multivariate Statistical Machine Learning Methods for Genomic Prediction*, Cham: Springer International Publishing, 2022, pp. 109–139. doi: 10.1007/978-3-030-89010-0_4.

[49] B. M. Turner, B. U. Forstmann, and M. Steyvers, “Assessing Model Performance with Generalization Tests,” 2019, pp. 39–51. doi: 10.1007/978-3-030-03688-1_3.

[50] N. Le *et al.*, “K-Fold Cross-Validation: An Effective Hyperparameter Tuning Technique in Machine Learning on GNSS Time Series for Movement Forecast,” 2024, pp. 377–382. doi: 10.1007/978-3-031-43218-7_88.

[51] J. M. Zhang, M. Harman, B. Guedj, E. T. Barr, and J. Shawe-Taylor, “Model validation using mutated training labels: An exploratory study,” *Neurocomputing*, vol. 539, p. 126116, Jun. 2023, doi: 10.1016/j.neucom.2023.02.042.

[52] S. R. Livingstone and F. A. Russo, “The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English,” *PLoS One*, vol. 13, no. 5, p. e0196391, May 2018, doi: 10.1371/journal.pone.0196391.

[53] O. Rainio, J. Teuho, and R. Klén, “Evaluation metrics and statistical tests for machine learning,” *Sci. Rep.*, vol. 14, no. 1, pp. 1–14, 2024, doi: 10.1038/s41598-024-56706-x.

[54] F. Saeik *et al.*, “Task offloading in Edge and Cloud Computing: A survey on mathematical, artificial intelligence and control theory solutions,” *Comput. Networks*, vol. 195, p. 108177, Aug. 2021, doi: 10.1016/j.comnet.2021.108177.

[55] S. Taraji *et al.*, “Novel Machine Learning Algorithms for Prediction of Treatment Decisions in Adult Patients With Class III Malocclusion,” *J. Oral Maxillofac. Surg.*, vol. 81, no. 11, pp. 1391–1402, Nov. 2023, doi: 10.1016/j.joms.2023.07.137.

[56] J. Roessler and P. A. Gloor, “Measuring happiness increases happiness,” *J. Comput. Soc. Sci.*, vol. 4, no. 1, pp. 123–146, May 2021, doi: 10.1007/s42001-020-00069-6.

[57] D. Liu, “The effectiveness of three-way classification with interpretable perspective,” *Inf. Sci. (Ny).*, vol. 567, pp. 237–255, Aug. 2021, doi: 10.1016/j.ins.2021.03.030.

[58] B. Gao *et al.*, “Enhancing anomaly detection accuracy and interpretability in low-quality and class imbalanced data: A comprehensive approach,” *Appl. Energy*, vol. 353, p. 122157, Jan. 2024, doi: 10.1016/j.apenergy.2023.122157.

[59] E. M. Saoudi, J. Jaafari, and S. J. Andaloussi, “Advancing human action recognition: A hybrid approach using attention-based LSTM and 3D CNN,” *Sci. African*, vol. 21, p. e01796, Sep. 2023, doi: 10.1016/j.sciaf.2023.e01796.

[60] G. Naidu, T. Zuva, and E. M. Sibanda, “A Review of Evaluation Metrics in Machine Learning Algorithms,” 2023, pp. 15–25. doi: 10.1007/978-3-031-35314-7_2.