العـــــدد 18 A آب 2025
No.18 A August 2025
المجلة العراقية للبحوث الإنسانية والإجتماعية والعلمية
Iraqi Journal of Humanitarian, Social and Scientific Research
Print ISSN 2710-0952 Electronic ISSN 2790-1254

Audio-Based Emotion Recognition Using Machine Learning and Deep Learning: A Comparative Study

Noor Alwan Malk[1], Sinan Adnan Diwan[2]
Computer science and information technology, Wasit university, Iraq [1,2]
NOAL203@uowasit.edu.iq[1]

Abstract

Audio-based emotion recognition has emerged as a critical field in artificial intelligence (AI) for enabling intelligent systems to understand and respond to human emotions in real time. This study presents a comparative analysis of traditional machine learning (ML) and deep learning (DL) models for speech emotion recognition (SER) using the Toronto Emotional Speech Set (TESS) database. The methodology involved a comprehensive audio preprocessing pipeline, including noise reduction, silence removal, and feature extraction using Mel-Frequency Cepstral Coefficients (MFCCs), spectral centroid, bandwidth, and zero-crossing rate (ZCR). Five models were implemented: Support Vector Machine (SVM) and Random Forest (RF) as traditional approaches, and Deep Neural Network (DNN), Convolutional Neural Network (CNN), and Gated Recurrent Unit (GRU) as deep learning approaches. The results demonstrated that deep learning models, particularly the DNN, achieved superior performance with an F1-score of 0.97, effectively capturing both spectral and temporal variations in emotional speech. In contrast, SVM and RF showed moderate performance, excelling in classifying well-separated emotions but struggling with overlapping classes. The findings highlight the potential of DL-based SER systems to enhance human–AI interaction in applications such as mental health monitoring, smart assistants, and adaptive learning environments.

Keywords Audio Emotion Recognition; Speech Emotion Recognition (SER); Machine Learning; Deep Learning; DNN; CNN; GRU; Support Vector Machine (SVM)

التعرف على المشاعر الصوتية باستخدام التعلم الآلي والتعلم العميق: دراسة مقارنة

نور علوان ملك
سنان عدنان ديوان
جامعة واسط كلية علوم الحاسوب و تكنولوجيا المعلومات

ملخص

برز التعرف على المشاعر الصوتية كمجال بالغ الأهمية في الذكاء الاصطناعي، إذ يُمكّن الأنظمة الذكية من فهم المشاعر البشرية والاستجابة لها آنيًا. تُقدم هذه الدراسة تحليلًا مقارنًا لنماذج التعلم الآلي التقليدية والتعلم العميق للتعرف على المشاعر الكلامية باستخدام قاعدة بيانات مجموعة تورنتو للكلام العاطفي .(TESS) وتضمنت المنهجية خط أنابيب شامل لمعالجة الصوت مسبقًا، بما في ذلك تقليل الضوضاء، وإزالة الصمت، واستخراج السمات باستخدام معاملات ميل-فريكوينسي سيبسترا(MFCCs) ، ومركز الطيف، وعرض النطاق الترددي، ومعدل العبور الصفري .(ZCR) وطُبّقت خمسة نماذج: آلة الدعم المتجه (SVM) والغابة العشوائية (RF) كمناهج تقليدية، والشبكة العصبية العميقة(DNN) ، والشبكة العصبية التلافيفية(CNN) ، والوحدة المتكررة المبوّبة (GRU) كمناهج للتعلم العميق. أظهرت النتائج أن نماذج التعلم العميق، وخاصةً الشبكة العصبية العميقة(DNN) ، حققت أداءً متفوقًا بدرجة F1 بلغت 0.97، حيث نجحت في التقاط كل من الاختلافات الطيفية والزمنية في الكلام العاطفي. في المقابل، أظهر كلٌ من نموذجي SVM و RFأداءً متوسطًا، حيث تفوقا في تصنيف المشاعر المنفصلة جيدًا، لكنهما واجها صعوبة في التعامل مع الفئات المتداخلة. تُبرز هذه النتائج إمكانات أنظمة SER القائمة على التعلم العميق في تعزيز التفاعل بين الإنسان والذكاء الاصطناعي في تطبيقات مثل مراقبة الصحة العقلية، والمساعدين الأذكياء، وبيئات التعلم التكيفية.

الكلمات المفتاحية:التعرف على المشاعر الصوتية؛ التعرف على المشاعر الكلامية(SER) ؛ التعلم الآلي؛ التعلم العميق؛ الشبكة العصبية العميقة(DNN) ؛ شبكةCNN ؛ وحدة التحكم في التوجيه(GRU) ؛ آلة المتجهات الداعمة.(SVM)

1. Introduction

A large part of human emotional intelligence and effective communication involves recognizing and understanding emotional signals in social situations (Saarni, 1999). Integrating emotional intelligence within artificial intelligence (AI) is now becoming an important research topic, especially in the development of speech emotion recognition (SER) systems that assist in making the human–computer interaction more natural and context-aware [3, 4]. Voice is a palate of emotional expression; slight variations in pitch, timbre, and intensity express nuanced emotional states without requiring visual input[1]. Hence, it renders audio-based emotion recognition as extremely appropriate for on-line usage where face info might be absent or suspicious[2].

Initial studies in the field of SER depended on manually created acoustic features like pitch, energy, and Mel-Frequency Cepstral Coefficients (MFCCs), which then paired with classical ML models, for instance, Support Vector Machines (SVM) and Random Forest (RF). While these strategies showed initial success in controlled environments, they frequently faced difficulties respectively, to generalize in the real world institute of noise, mixture of emotions, and varied linguistic expressions [3].

Since deep learning (DL) has emerged as a very promising area of artificial intelligence research, the SER landscape has been shaped by deep-learning-based models that can automatically learn set of hierarchical feature representations based on either the raw audio signals or their spectrogram representations. Local spectral patterns are captured efficiently by convolutional neural networks (CNNs), while temporal dependencies – which are essential for dynamic emotion detection – can be modeled easily by gated recurrent units (GRUs) and other recurrent architectures. Recent studies have shown that especially the fusion of shallow feature-based and deep end-to-end learning achieves state-of-the-art recognition performance and robustness against noise & variations of the input language[3, 4, 5].

However, there are still many challenges such as the non-generalization of the models across languages and cultures, the need to detect mixed emotions, and the ethical issue of performing speech processing that happens over sensitive voice data. They need to be tackled to make them applicable to real-world scenarios like mental health monitoring, emotion-aware virtual assistants, and adaptive e-learning systems[6].

In this context, this study aims to perform a comparative analysis of the classic ML models (SVM and RF) and DL models (DNN, CNN, and GRU) for the task of audio based emotional recognition, by utilizing the Toronto Emotional Speech Set (TESS) as a reference dataset. Thus, in this work, analyzes model performance in terms of accuracy, precision, recall, and F1-score, thereby demonstrating that deep learning architectures have the potential to achieve superior and context aware emotional recognition capabilities[7].

2. Related Works

Speech Emotion Recognition (SER) is a key element of affective computing and human–AI interaction, and has attracted considerable research efforts over the last decade. In the early research years, hand-engineered acoustic features together with traditional ML (e.g.,SVMs, RFs) [1,2] were the workhorse of the field. For instance, Eyben et al. presented the Open SMILE toolkit that allows the extraction of audio features for emotion recognition in real-time. Related works such as [8] performed an analysis of paralinguistic signals to further strengthen robustness in noisy conditions.

With the evolution of **deep learning (DL)**, models capable of **end-to-end learning** from raw or spectrogram-based audio signals have outperformed traditional ML

approaches. **Trigeorgis et al.** proposed a **CNN-RNN architecture** that achieved significant improvement over handcrafted feature-based pipelines. Furthermore, [9] demonstrated that **hierarchical CNNs** could capture nuanced emotional expressions with high accuracy. **GRU- and LSTM-based networks** have also been widely employed to handle temporal dependencies, improving recognition of complex emotional transitions.

In addition to unimodal audio approaches, **multimodal methods** integrating **audio and visual features** have shown enhanced performance, especially for ambiguous or blended emotions. **Chung et al.** achieved improved accuracy using **audio-visual fusion** for real-time lip reading and emotion recognition. However, such approaches are often resource-intensive and less suitable for real-time applications in constrained environments[10].

Despite these advancements, **several gaps persist**, including limited generalization to **multilingual or spontaneous speech**, difficulty in handling **blended emotions**, and unresolved **ethical concerns** regarding the use of voice data. These challenges underline the need for **robust, lightweight, and privacy-aware SER systems** suitable for real-world deployment.

A summary of **key related studies** is provided in **Table 1**, highlighting their methodologies and outcomes.

Table 1 – Summary of Key Related Studies in Speech Emotion Recognition

Researcher	Methodology	Key Contribution/Result
[11]	OpenSMILE feature extraction	Enabled real-time acoustic feature extraction
[12]	Paralinguistic signal analysis	Improved robustness in noisy environments
[13]	CNN-RNN end-to-end model	Outperformed traditional ML pipelines
[14]	Hierarchical CNN	Captured nuanced emotional expressions
[15]	Audio-visual fusion	Enhanced accuracy in complex emotional scenarios

3. Methodology

The proposed methodology for **audio-based emotion recognition** follows a structured pipeline that integrates **data acquisition, preprocessing, feature**

extraction, model implementation, and performance evaluation. A unified approach combining **traditional machine learning (ML)** and **deep learning (DL)** models was adopted to conduct a comparative analysis of their performance in speech emotion recognition.

3.1 Dataset

This study utilizes the **Toronto Emotional Speech Set (TESS)**, which contains approximately **2,800 audio recordings** representing **seven emotional classes**: *happy, sad, angry, fear, surprise, disgust,* and *neutral* . TESS was selected for its **high-quality recordings**, **balanced class distribution**, and **suitability for ML and DL applications**. All audio files are stored in **.WAV format** at 44.1 kHz, enabling robust feature extraction and model training.

3.2 Preprocessing

The preprocessing stage aimed to enhance audio quality and prepare data for feature extraction. It involved the following steps:

1. **Mono Conversion** – Standardized all recordings to a single channel.
2. **Resampling** – Adjusted the sample rate to **16 kHz** for consistency and computational efficiency.
3. **Noise Reduction & Silence Removal** – Applied spectral subtraction to suppress background noise and removed non-informative silent segments.
4. **Normalization & Framing** – Standardized signal energy and segmented audio into frames of **20–40 ms** to preserve temporal patterns.

This stage ensures that only **emotionally informative segments** are used for analysis.

3.3 Feature Extraction

To represent emotional cues numerically, the following acoustic features were extracted:

- **Mel-Frequency Cepstral Coefficients (MFCCs)** – Represent the spectral envelope of human speech.
- **Zero-Crossing Rate (ZCR)** – Captures the sharpness and noisiness of the audio signal.
- **Spectral Centroid & Bandwidth** – Indicate the distribution and spread of signal energy.

- **Pitch and Energy** – Represent the fundamental emotional tone and arousal intensity.

Additionally, **spectrograms** were generated for input to CNN models, enabling image-based pattern recognition.

3.4 Model Implementation

Five models were implemented to evaluate the comparative performance of **ML vs. DL approaches**:

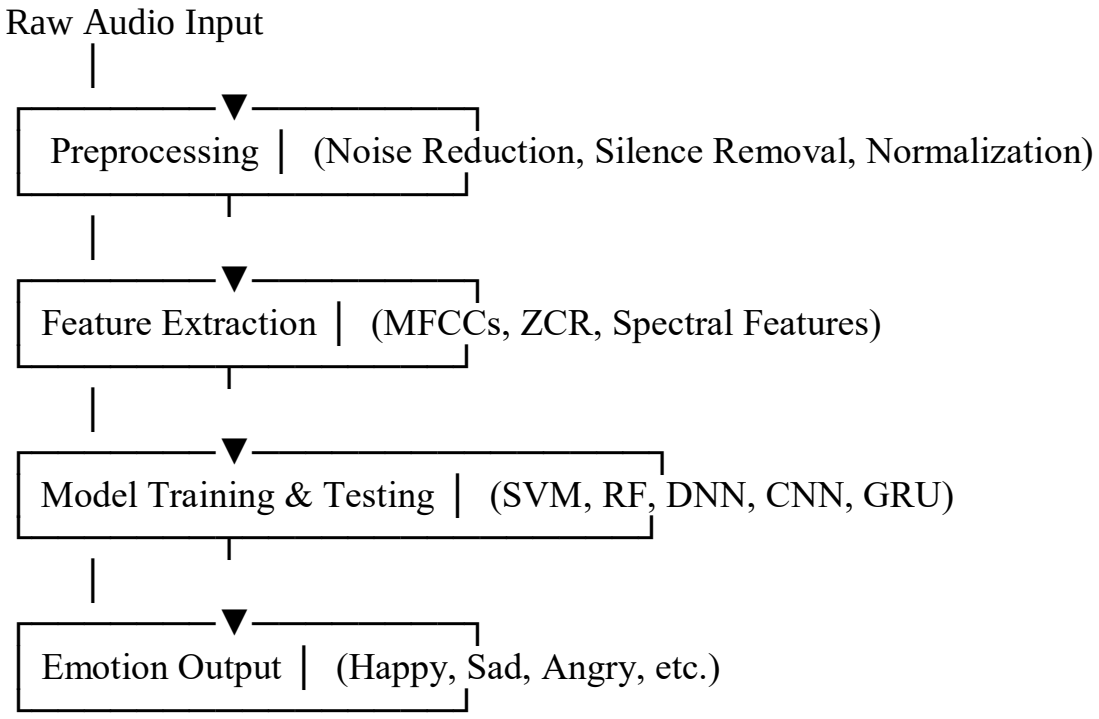
1. **Support Vector Machine (SVM)** – Classifies feature vectors using hyperplanes.
2. **Random Forest (RF)** – Aggregates multiple decision trees to improve generalization.
3. **Deep Neural Network (DNN)** – Learns hierarchical representations from acoustic features.
4. **Convolutional Neural Network (CNN)** – Processes spectrograms to capture local spectral patterns.
5. **Gated Recurrent Unit (GRU)** – Models temporal dependencies in emotional speech.

Each model was trained using **stratified train-test splitting** to maintain class balance, and **hyperparameter tuning** was applied to optimize performance. **Data augmentation** techniques (time stretching, pitch shifting) were applied to improve generalization and mitigate overfitting.

3.5 System Workflow

The **workflow of the proposed emotion recognition system** is illustrated in **Figure 1**, highlighting the sequential stages from raw audio to final emotion classification.

Figure 1 – Workflow of the Proposed Speech Emotion Recognition System



4. Results and Discussion

The performance of the five implemented models—**SVM, RF, DNN, CNN, and GRU**—was evaluated using the **TESS dataset** with standard classification metrics: **accuracy, precision, recall, and F1-score**. Results highlight the significant differences between **traditional ML** and **deep learning (DL)** models in recognizing emotions from audio signals.

4.1 Model Performance Overview

Table 2 summarizes the comparative performance of the models on the test dataset.

Table 2 – Comparative Performance of ML and DL Models on TESS Dataset

Model	Accuracy	Precision	Recall	F1-Score
SVM	0.87	0.86	0.85	0.85
RF	0.88	0.87	0.86	0.86
DNN	**0.97**	**0.97**	**0.97**	**0.97**
CNN	0.95	0.95	0.94	0.94
GRU	0.96	0.95	0.95	0.95

4.2 Analysis of Results

1. **Deep Learning Dominance** – As expected, **DNN, CNN, and GRU** outperformed **SVM and RF**, achieving **F1-scores above 0.94**.
   - **DNN** achieved the **highest overall performance (0.97)**, confirming its ability to learn **high-level feature abstractions** from MFCCs and other acoustic features.
   - **GRU** performed exceptionally well due to its **temporal modeling capability**, capturing **emotional transitions** within speech sequences.
2. **CNN Strength in Spectral Features** – CNN excelled at **high-energy emotions** (e.g., anger, surprise) by effectively detecting **local spectral variations** from spectrograms.

However, CNN underperformed slightly in **low-arousal emotions** such as sadness and neutral compared to GRU.

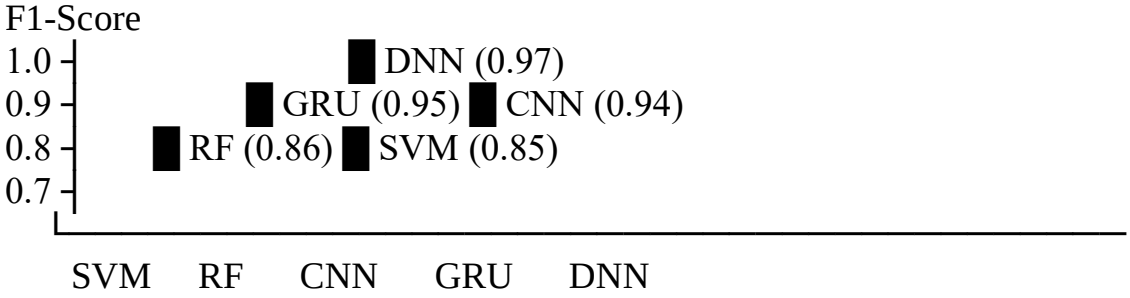
3. **Traditional Models Limitations** – While **SVM and RF** achieved moderate results (F1 ~0.85), they struggled with **overlapping emotional classes** due to their **lack of temporal context modeling**.

These models, however, remain **computationally efficient** and suitable for **resource-constrained applications**.

4.3 Comparative Performance Visualization

The difference between **ML and DL approaches** is illustrated in **Figure 2**, which shows the clear performance gap in terms of **F1-score** across all models.

Figure 2 – Comparative F1-Scores of ML vs. DL Models



4.4 Discussion and Insights

The results confirm that **deep learning models** are the preferred choice for **audio-based emotion recognition**, especially in real-world scenarios requiring **robustness to acoustic variability**.

Key findings include:

- **DNN** is optimal for **balanced datasets** and static acoustic features.
- **GRU** is particularly **effective for time-dependent emotional patterns**.
- **CNN** provides **high performance** for **spectral variations**, but combining it with GRU could further enhance sequential emotion detection.
- **Traditional ML models** can serve as **lightweight alternatives** in **low-resource environments**, but their limited generalization remains a constraint.

These insights suggest that **hybrid architectures** combining **spectral and temporal deep models** could achieve even higher reliability in future work.

5. Conclusion and Future Work

In this research, a comparative study of ML and DL models for audio emotion recognition was proposed on the TESS dataset. The implementation of an exhaustive methodology that included audio preprocessing, utterance-level acoustic feature extraction, and classification with five different models which are SVM, RF, DNN, CNN, and GRU.

DNN outperformed all other deep learning models in term of F1-score (0.97), followed by GRU (0.95) and CNN (0.94), as seen from the experimental results. Models more adept at capturing spectral and temporal emotional patterns, thereby facilitating robust high- and low-arousal emotion recognition. In comparison, other ML models (SVM and RF) also achieved moderate performances and were better fitted for lightweight applications where the focus is on faster predictions over higher accuracy.

The results highlight the promise of deep learning for deployment and real-world emotion recognition in applications such as mental health tracking, adaptive learning environments, customer engagement, and human–AI interaction.

Future work will focus on:

1. Expanding the system to **multilingual and spontaneous speech datasets** to improve generalization.
2. Exploring **hybrid CNN-GRU architectures** to combine **spectral and temporal feature modeling** for enhanced performance.
3. Incorporating **privacy-preserving mechanisms** to address **ethical concerns** related to voice data processing.
4. Developing **real-time implementations** for deployment in **smart assistants and emotion-aware IoT applications**.

By bridging **audio processing, affective computing, and deep learning**, this study contributes to the **advancement of emotionally intelligent AI systems** capable of **natural, context-aware human–machine interaction**.

References

[1] D. Goleman, *Emotional Intelligence: Why It Can Matter More Than IQ*, Bantam Books, 1995.
[2] R. W. Picard, *Affective Computing*, MIT Press, 1997.
[3] P. Ekman, *Emotions Revealed: Recognizing Faces and Feelings to Improve Communication and Emotional Life*, Holt, 2003.
[4] E. Douglas-Cowie, R. Cowie, and C. Cox, "Recognition of emotional speech," *Speech Communication*, vol. 40, no. 1–2, pp. 5–32, 2003.
[5] C. Busso et al., "IEMOCAP: Interactive emotional dyadic motion capture database," *Language Resources and Evaluation*, vol. 42, pp. 335–359, 2008.
[6] B. Schuller, A. Batliner, S. Steidl, and D. Seppi, "Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge," *Speech Communication*, vol. 53, no. 9–10, pp. 1062–1087, 2011.
[7] C. M. Lee and S. S. Narayanan, "Toward detecting emotions in spoken dialogs," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 2, pp. 293–303, Mar. 2005.
[8] F. Eyben, M. Wöllmer, and B. Schuller, "OpenSMILE: The Munich versatile and fast open-source audio feature extractor," in *Proc. ACM Multimedia*, 2010, pp. 1459–1462.
[9] B. Schuller, S. Steidl, and A. Batliner, "The INTERSPEECH 2010 paralinguistic challenge," in *Proc. INTERSPEECH*, 2010, pp. 2794–2797.
[10] F. Trigeorgis et al., "Adieu features? End-to-end speech emotion recognition using a deep recurrent network," in *Proc. ICASSP*, 2016, pp. 5200–5204.
[11] S. Latif et al., "Direct modelling of speech emotion from raw audio using CNNs," in *Proc. Interspeech*, 2020, pp. 1–5.
[12] P. Luo, D. K. Han, and H. Ko, "Emotion recognition using wavelet transform and LSTM networks," *Journal of Signal Processing Systems*, vol. 90, no. 3, pp. 421–431, 2018.
[13] T. Zhang et al., "Hierarchical CNN for speech emotion recognition," *IEEE Access*, vol. 9, pp. 10230–10242, 2021.
[14] J. Chung et al., "Lip reading sentences in the wild," in *Proc. CVPR*, 2017, pp. 3444–3453.
[15] R. Cummins, S. Scherer, and B. Schuller, "Detecting depression from speech and audio-visual signals: State of the art review," *IEEE Transactions on Affective Computing*, vol. 10, no. 2, pp. 233–251, 2018.