

POLİTEKNİK DERGİSİ

JOURNAL of POLYTECHNIC

ISSN: 1302-0900 (PRINT), ISSN: 2147-9429 (ONLINE)
URL: http://dergipark.org.tr/politeknik

Deep learning-based speech emotion recognition for iot edge devices: a comparative study

Nesnelerin interneti (iot) uç birimleri için derin öğrenme tabanlı konuşma duygusu tanıma: karşılaştırmalı bir çalışma

***Yazar(lar) (Author(s))**: Buket İŞLER[1]*

ORCID[1]: 0000-0002-9393-9564

To cite to this article: İşler B., "Deep Learning-Based Speech Emotion Recognition for Iot Edge Devices: A Comparative Study", ***Journal of Polytechnic***, 29(4):290419:1-13 (2026).

Bu makaleye şu şekilde atıfta bulunabilirsiniz: İşler B., "Deep Learning-Based Speech Emotion Recognition for Iot Edge Devices: A Comparative Study", ***Politeknik Dergisi***, 29(4):290419:1-13 (2026).

Erişim linki (To link to this article): http://dergipark.org.tr/politeknik/archive

DOI: 10.2339/politeknik.1729678

Deep Learning-Based Speech Emotion Recognition for IoT Edge Devices: A Comparative Study

Highlights

- *Three hybrid deep learning models (CNN–Dense, LSTM–CNN, Dense–LSTM) were proposed for speech-based emotion recognition.*
- *The CNN–Dense architecture demonstrated superior performance in recognising high-salience emotional states.*
- *Experimental results confirmed the feasibility of deploying the models in low-latency edge computing environments.*

Graphical Abstract

This study evaluates three hybrid deep learning models—CNN–Dense, LSTM–CNN, and Dense–LSTM—for speech emotion recognition in edge environments, focusing on their performance and latency.

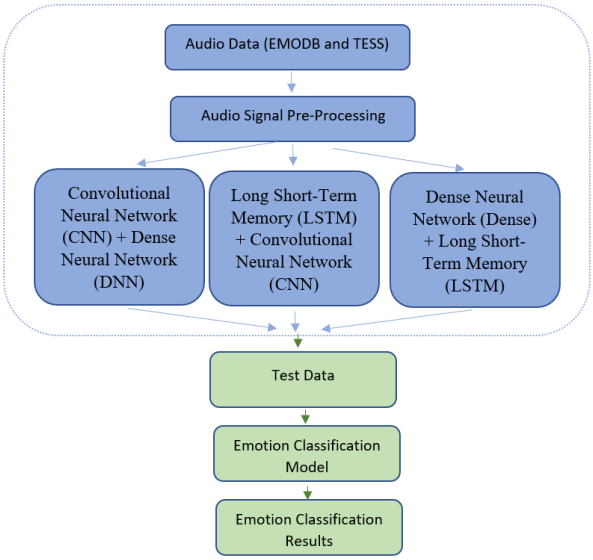


Figure. Hybrid Deep Learning Workflow for Speech Emotion Recognition

Aim

This study aims to compare three hybrid models in speech-based emotion recognition.

Design & Methodology

The CNN–Dense, LSTM–CNN, and Dense–LSTM models were applied using MFCC features extracted from the EMODB and TESS datasets

Originality

This study aims to expand on existing efforts by offering a comparative evaluation of hybrid models under simulated edge computing conditions.

Findings

CNN–Dense yielded the best results, while others showed balanced performance and low latency.

Conclusion

Hybrid models can be effectively deployed in edge computing environments.

Declaration of Ethical Standards

The author(s) of this article declare that the materials and methods used in this study do not require ethical committee permission and/or legal-special permission

Deep Learning-Based Speech Emotion Recognition for IoT Edge Devices: A Comparative Study

Araştırma Makalesi / Research Article

Buket İŞLER[1*]

[1]Mühendislik ve Doğa Bilimler Fakültesi, Yazılım Mühendisliği Bölümü, İstanbul Topkapı Üniversitesi, Türkiye

(Geliş/Received : 30.06.2025 ; Kabul/Accepted : 12.09.2025 ; Erken Görünüm/Early View : 28.09.2025)

ABSTRACT

With advancements in artificial intelligence (AI), especially in pattern recognition, emotion recognition (ER) systems have significantly progressed in identifying human emotions from speech, facial movements, and physiological signals. However, the rapid growth of Internet of Things (IoT)-based infrastructures has placed increasing demands on traditional cloud computing systems due to the high volume of transmitted data and the need for real-time responsiveness. Edge computing has emerged as a promising alternative by enabling localised processing, thus reducing latency and network dependency. This study investigates the classification performance of three hybrid deep learning (DL) models—Convolutional Neural Network–Dense Neural Network (CNN–Dense), Long Short-Term Memory–Convolutional Neural Network (LSTM–CNN), and Dense–Long Short-Term Memory (Dense–LSTM) —within a simulated edge computing environment. Experiments were conducted using two benchmark speech emotion datasets: the Toronto Emotional Speech Set (TESS) and the Berlin Emotional Speech Database (EMODB). MFCCs were used for feature extraction, and simulations were performed on an Amazon Web Services (AWS) virtual instance configured to reflect resource-limited edge conditions. Performance was evaluated using macro-averaged precision, recall, and F1-score metrics. Among the models, CNN–Dense achieved the highest F1-scores: 96% on TESS and 87% on EMODB, followed by LSTM–CNN and Dense–LSTM. These findings demonstrate the strong generalisability of the CNN–Dense model and its effectiveness in latency-sensitive, decentralised environments. Overall, the study highlights the practical value of hybrid DL models for emotion recognition and aims to demonstrate their applicability in decentralised IoT-based environments.

Keywords: Internet of things, speech emotion recognition, edge computing, machine learning.

Nesnelerin İnterneti (IoT) Uç Birimleri İçin Derin Öğrenme Tabanlı Konuşma Duygusu Tanıma: Karşılaştırmalı Bir Çalışma

ÖZ

Yapay zekâ (YZ), özellikle örüntü tanıma alanındaki gelişmelerle birlikte, duygu tanıma (ER) sistemleri; konuşma, yüz hareketleri ve fizyolojik sinyaller aracılığıyla insan duygularını tanıma konusunda önemli ilerlemeler kaydetmiştir. Bununla birlikte, Nesnelerin İnterneti (IoT) tabanlı altyapıların hızla büyümesi, iletilen veri hacminin artması ve gerçek zamanlı yanıt gereksinimi nedeniyle geleneksel bulut bilişim sistemleri üzerinde artan bir baskı oluşturmuştur. Bu soruna bir çözüm olarak, yerelleştirilmiş veri işleme imkânı sunan uç (edge) bilişim yaklaşımı, gecikmeyi azaltarak ve ağ bağımlılığını düşürerek umut verici bir alternatif hâline gelmiştir. Bu çalışma, üç hibrit derin öğrenme (DL) modelinin—Convolutional Neural Network–Dense Neural Network (CNN–Dense), Long Short-Term Memory–Convolutional Neural Network (LSTM–CNN) ve Dense–Long Short-Term Memory (Dense–LSTM) —uç bilişim ortamını simüle eden bir deneysel yapı altında sınıflandırma performanslarını incelemektedir. Deneyler, konuşmaya dayalı duygu tanıma alanında yaygın olarak kullanılan iki veri seti ile gerçekleştirilmiştir: Toronto Emotional Speech Set (TESS) ve Berlin Emotional Speech Database (EMODB). Özellik çıkarımı için Mel Frekans Kepstrum Katsayıları (MFCC) kullanılmış; simülasyonlar ise kısıtlı kaynaklara sahip uç bilişim koşullarını yansıtacak şekilde yapılandırılmış bir Amazon Web Services (AWS) sanal örneği üzerinde yürütülmüştür. Modellerin performansı; makro ortalamalı kesinlik (precision), geri çağırma (recall) ve F1-skoru ile değerlendirilmiştir. Bu kapsamda CNN–Dense modeli en yüksek F1 skorlarını elde etmiştir: TESS veri setinde %96, EMODB veri setinde ise %87. Diğer modeller ise sırasıyla LSTM–CNN ve Dense–LSTM olarak sonuçlanmıştır. Bulgular, CNN–Dense modelinin güçlü genellenebilirliğini ve gecikmeye duyarlı, merkeziyetsiz ortamlardaki etkinliğini ortaya koymaktadır. Genel olarak, bu çalışma, hibrit DL modellerinin duygu tanıma sürecindeki pratik değerine dikkat çekmekte ve bu modellerin merkeziyetsiz IoT tabanlı ortamlardaki uygulanabilirliğini göstermeyi amaçlamaktadır.

Anahtar Kelimeler: Nesnelerin interneti, konuşma duygu tanıma, uç bilişim, makine öğrenmesi

1. INTRODUCTION

Emotion recognition (ER) has become a multidisciplinary research field that aims to interpret human affective states by analysing behavioural and physiological indicators, such as facial expressions, vocal tone, gestures, and biometric signals. Its practical implementations span interactive systems, clinical diagnostics, surveillance technologies, and human–machine interaction, all of which benefit from emotionally adaptive responses and enhanced user experience [1–3]. With the rise of artificial intelligence, particularly deep learning techniques, the performance of

*Sorumlu Yazar (Corresponding Author)
e-posta : buketisler @topkapi.edu.tr

ER systems has improved significantly. Among these techniques, Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks have emerged as leading architectures due to their respective strengths in capturing spatial and temporal patterns in multimodal emotional datasets [4–7]. In parallel, the proliferation of edge computing and the Internet of Things (IoT) has enabled the deployment of ER applications in environments requiring real-time responsiveness and privacy preservation. While IoT provides seamless connectivity across distributed systems, edge computing shifts computational load from centralised cloud servers to decentralised local nodes, reducing latency and communication overhead [9–12]. These infrastructures have become integral to affect-aware systems in fields such as ambient assisted living, smart healthcare, and adaptive automation [13,14]. However, designing ER systems suitable for such decentralised contexts introduces challenges in handling heterogeneous data, computational constraints, and low-latency requirements. Solutions such as Domain-Specific Modelling Languages (DSMLs) and semantic architectures like Emotional Knowledge Graphs (EmoKGs) have been proposed to address these challenges [15–17]. In the speech domain, the abundance of labelled corpora has allowed the application of complex embedding techniques such as Bag-of-Audio-Words (BoAW) and recurrent models. LSTMs, in particular, have demonstrated robustness in noisy acoustic conditions [18,19], and hardware-efficient versions such as spiking LSTMs have been developed for deployment on energy-constrained devices [20]. The use of densely connected layers and synthetic data augmentation has further enhanced model generalisability [21–24].

Hybrid architectures that combine CNNs, LSTMs, GRUs, or dense layers have shown potential in capturing the intricate structure of emotional speech signals [25–27]. Nevertheless, much of the existing research has remained focused on evaluations under cloud-based or laboratory conditions. Recent studies in related fields also confirm the importance of hybrid and optimisation-driven approaches for affective computing tasks. For example, Donuk et al., [28] proposed a BPSO–SVM-based deep feature selection method for facial emotion recognition, demonstrating that optimisation-based hybrid pipelines can enhance classification accuracy. Similarly, Demirel and Çam [29] explored sentiment analysis on large-scale Turkish social media datasets to investigate fluctuations in cryptocurrency transactions, highlighting the broader applicability of emotion and sentiment recognition in dynamic environments. In the speech domain, a real-time laboratory application for voice-controlled in-car devices [30] has shown the feasibility of integrating speech recognition with human–machine interfaces, underscoring the need for low-latency solutions in edge-inspired scenarios. In contrast to these applications, this study aims to evaluate the performance and latency characteristics of three hybrid models—CNN–Dense, LSTM–CNN, and Dense–LSTM—under simulated edge conditions using two widely adopted speech emotion datasets: the Toronto Emotional Speech Set (TESS) and the Berlin Emotional Speech Database (EMODB). TESS provides standardised recordings in seven emotional categories, articulated by female speakers from different age groups, whereas EMODB contains authentic expressions of emotion in German, recorded under controlled acoustic conditions. Together, these datasets offer a robust experimental ground for evaluating model generalisability and performance. To extract perceptually meaningful features, Mel-Frequency Cepstral Coefficients (MFCCs) were employed. The selected models were trained and tested within a virtualised Amazon Web Services (AWS) environment configured to emulate edge-like conditions. Due to limited access to physical edge hardware, the study adopted an AWS t3.medium instance (2 vCPUs, 4 GB RAM), which—although more powerful than typical embedded edge devices such as Raspberry Pi boards or low-power microcontrollers—was selected to simulate key operational constraints, including decentralised computation, limited communication overhead, and low-latency inference. The limitations of this simulation, particularly in replicating hardware constraints, are acknowledged and further discussed in the conclusion section. The hypothesis guiding this study is that hybrid models combining convolutional and recurrent layers can achieve superior performance compared to standalone CNN or LSTM models. This superiority stems from the dual capability of hybrid models to exploit both local feature extraction (via CNNs) and long-range temporal dependencies (via LSTMs), which is particularly advantageous for recognising emotion in continuous speech data. These advantages are expected to improve accuracy and responsiveness in latency-sensitive deployment scenarios typical of IoT edge environments.

2. INTERNET OF THINGS (IOT) AND EMOTIONAL INTELLIGENCE

2.1. IoT

The notion of the Internet of Things (IoT) was first introduced in 1999 through studies conducted at the Auto-ID Laboratory of the Massachusetts Institute of Technology, providing an early formal framework for the concept. Its international prominence increased following the International Telecommunication Union's (ITU) 2005 report titled The Internet of Things [31]. Essentially, IoT denotes the smooth integration and coordination of physical devices—such as smartphones, sensors, and embedded systems—via various technological infrastructures to perform specific functions [32]. It is commonly regarded as a unified platform that enables autonomous communication among interconnected devices in intelligent environments. Al-Fuqaha et al. [33] describe IoT as a system composed of sensor-equipped physical entities and software, working together to facilitate data sharing,

intelligent operations, and decentralised decision-making. Such systems are typically supported by continuous wireless communication technologies, including Bluetooth, Wi-Fi, ZigBee, Wireless Sensor Networks (WSNs), Low Power Wide Area Networks (LPWANs), and mobile broadband, all of which are used for real-time data transmission and analysis. This level of connectivity enables remote control and monitoring of physical environments, thereby contributing to improved quality of life. Recent forecasts suggest that over 50 billion devices—including mobile phones, computers, sensors, and gaming consoles—will be incorporated into the global internet infrastructure via diverse technologies such as RFID and WSNs [34]. This expansive digital convergence is anticipated to transform various sectors by promoting automation, operational efficiency, and enriched user experiences. A typical IoT system is structured in a layered architecture that enables modularity and scalability across diverse application domains. At the foundational level lies the perception layer, which comprises intelligent sensors and RFID modules tasked with detecting environmental stimuli and gathering raw data for further processing [35]. The network layer follows, establishing communication pathways for the transmission of this data between devices and broader infrastructural systems, thereby ensuring robust and reliable connectivity [36]. The service layer then processes and manages the data, delivering context-specific functionalities tailored to application requirements. At the top, the application layer—also referred to as the interface layer—enables end-users or software systems to interact with the IoT environment. This stratified architecture supports the development of intelligent, adaptive, and context-aware applications capable of real-time interaction, thus facilitating integration into advanced domains such as emotional intelligence systems and smart environments.

2.2.Emotional Intelligence and IoT

Emotional Intelligence (EI) is widely recognised as a crucial factor in personal well-being, social competence, and performance. It is linked to constructs like empathy, self-esteem, and life satisfaction. With digital transformation, Emotional AI (AEI) has emerged, enabling machines to detect and simulate human emotions through advances in machine learning and affective computing. AI systems can now interpret emotional cues in real time, enhancing user interaction. In this context, IoT devices such as smart mirrors and wearables gather biometric data, which is analysed by cloud-based AI models to monitor emotional states and provide personalised feedback, supporting mental well-being in line with diagnostic standards like the DSM.

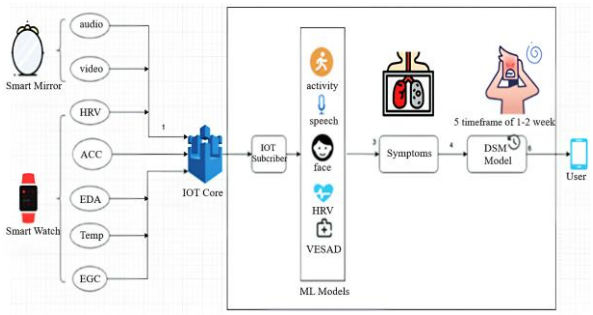


Figure 1. IoT-Based Artificial Emotional Intelligence Architecture

3. MATERIAL AND METHODOLOGY

3.1. Dataset

In the present research, SER was implemented through a combination of multiple deep learning architectures, with empirical evaluations conducted on two benchmark corpora: EMODB and TESS, both widely recognised in affective computing research. The EMODB dataset, developed by the Berlin Technical University, consists of professional German speech recordings from ten actors (five male and five female) who expressed seven emotions: anger, boredom, disgust, fear, happiness, sadness, and neutrality. In total, the database contains 535 utterances, each recorded under carefully designed conditions to balance natural expressiveness with acoustic clarity. The recordings were sampled at 16 kHz with 16-bit resolution and stored in WAV format. Each file is annotated with both the speaker identifier and the corresponding emotional category, enabling supervised learning and cross-corpus validation. The TESS dataset comprises high-quality English speech recordings produced by two female actors—one younger (approximately 26 years old) and one older (approximately 64 years old). These speakers vocalised a standardised set of 200 semantically neutral sentences derived from the North Wind and the Sun narrative. Each sentence was recorded in seven discrete emotional states: neutral, happiness, sadness, anger, fear, disgust, and surprise. The dataset consists of approximately 2800 speech samples (200 utterances × 7 emotions × 2 speakers), all captured in a controlled studio environment. The audio files were digitised with a 44.1 kHz sampling rate and 16-bit resolution, stored in WAV format, and explicitly labelled with speaker and emotional category information, facilitating supervised learning and reliable evaluation.

3.2 Methodology

The architecture introduced in this study (shown in Figure 2) is structured around a cloud–edge integration model that simulates low-latency operational settings for the primary purpose of comparing the classification performance of different hybrid emotion recognition models based on speech input. The system workflow is divided into three key stages: data management, pre-processing at the edge, and training and evaluation of models in the cloud. Initially, the speech samples from the TESS and the EMODB—each comprising utterances

labelled with seven core emotions—were uploaded to Amazon Simple Storage Service (S3). S3 served as a reliable and scalable cloud-based storage solution, ensuring organised access to audio data across all phases of the experimental workflow. Before training begins, raw audio signals are subjected to a pre-processing phase. In this phase, MFCCs are computed to extract spectral features that reflect the perceptual characteristics of human hearing. This transformation standardises the input data, optimising it for deep learning-based classification. Although such operations are typically handled in the cloud, this research incorporates an Edge Runtime environment to locally perform MFCC extraction and segment the input signals. This design strategy minimises communication latency by reducing the need to send raw audio to remote servers and aligns with practical deployment conditions where responsiveness is critical. After local processing, the extracted features are securely transferred to the cloud infrastructure via AWS IoT Core. To facilitate communication, a publish–subscribe architecture using an IoT Device Gateway enables asynchronous and bidirectional data exchange between edge nodes and cloud services. This approach not only ensures scalable transmission of data but also reserves cloud-based computing power for more intensive tasks such as model training. The experimental setup in this study focused on the implementation and comparative evaluation of hybrid deep learning architectures. Three distinct configurations—namely CNN–Dense, LSTM–CNN, and Dense–LSTM—were developed and assessed within the cloud-based Amazon SageMaker platform, which facilitated model training and deployment under controlled conditions. Each architecture was deliberately assembled to exploit the specific strengths of its constituent layers: convolutional components were tasked with extracting spatially localised features from MFCC-derived inputs; recurrent units were responsible for encoding temporal dependencies inherent in speech sequences; and fully connected layers aimed to abstract higher-order emotional patterns. The design rationale centred on enhancing the model's ability to encapsulate the multifaceted structure of affective speech. Evaluation was performed using standardised classification indicators, including overall accuracy, precision, recall, and F1-score, to ensure a comprehensive assessment of predictive capability. Upon completing hyperparameter optimisation, the models were implemented within a simulated edge computing environment using AWS services, specifically a t3.medium instance (2 vCPUs, 4 GB RAM). Although this configuration is more powerful than many embedded edge devices (e.g., Raspberry Pi or microcontrollers), it was selected to approximate edge resource limitations and latency-aware inference. To maintain system consistency and synchronisation, AWS IoT Shadow services were used to create virtual instances of physical edge devices. These services supported device state tracking, remote configuration, and synchronisation. Furthermore, AWS Device Management and AWS Device Defender were integrated to ensure the operational reliability, security, and stability of the end-to-end emotion recognition system.

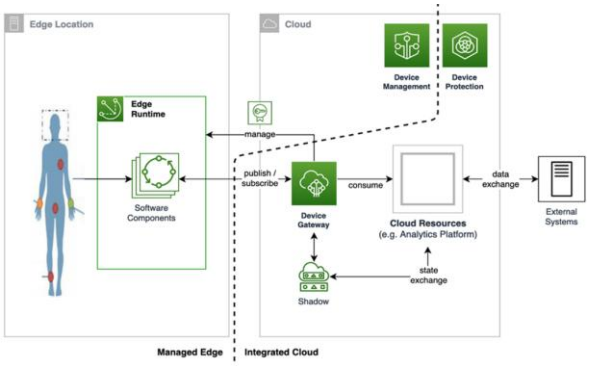


Figure 2. Emotion Detection Architecture

3.2.1. Modelling

Artificial neural networks are widely used in Speech Emotion Recognition (SER) for capturing complex acoustic patterns. This study developed three hybrid deep learning models—CNN–Dense, LSTM–CNN, and Dense–LSTM—to combine the strengths of convolutional, recurrent, and dense layers. CNNs extract spectro-temporal features, LSTMs capture temporal dependencies, and dense layers handle abstraction and classification. These models were evaluated in a simulated edge computing environment to explore their performance across emotion datasets and identify architectures suitable for real-time IoT applications.

CNN (Convolutional neural networks)

NNs are deep learning models widely applied in image and sound recognition [37–39]. Their strength lies in capturing spatial and partial temporal dependencies via convolutional filters, which extract local patterns and build higher-level features. Despite many variants, CNNs generally share a structure of convolutional layers for feature extraction, pooling for dimensionality reduction, and fully connected layers for classification [40, 41]. This modular design supports efficient, scalable learning, making CNNs a core element of modern AI. Figure 3 illustrates the CNN architecture used in this study, showing progression from raw input to classification through convolutional, pooling, and dense layers.

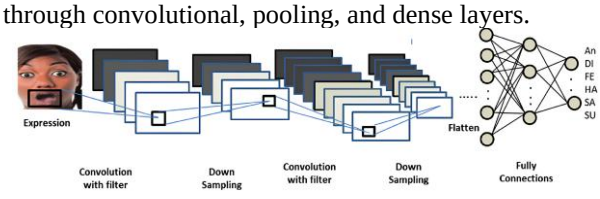


Figure 3. Typical CNN architecture for emotion recognition, comprising convolution, pooling, flattening, and fully connected layers. Classified emotions: An = Anger; DI = Disgust; FE = Fear; HA = Happiness; SA = Sadness; SU = Surprise (adapted from [41]).

LSTM (Long short-term memory)

LSTM models offer an innovative and scalable approach to learning from sequential data. In various applications, they have outperformed CNN and conventional RNN models. While RNNs are capable of processing short-term dependencies, they are inadequate in capturing long-term dependencies and dealing with inputs dispersed across time [42]. These limitations primarily stem from the vanishing gradient problem, which slows down or halts learning when weight values are small, and the exploding gradient problem, where large gradients destabilise the learning process [43-45]. LSTM models were specifically designed to overcome such issues through their unique architecture, enabling effective learning of long-term dependencies and providing a robust solution for tasks that involve the analysis of sequential information.

Dense neural networks

Dense neural networks**,** also referred to **as** fully connected architectures**,** are considered one of the foundational components of artificial intelligence and deep learning frameworks. In these architectures, each neuron in a given layer is fully connected to all neurons in the preceding layer, allowing for the modelling of abstract and complex relationships in high-dimensional input spaces [46]. As shown in Figure 4, a typical dense network comprises an input layer (x_1 to x_n), one or more hidden layers characterised by full interconnectivity (P_1 to P_n), and an output layer that produces the final prediction Y=$f = (x_1, \dots, x_n)$. During the training phase, model parameters such as weights and biases are updated iteratively using gradient-based optimisation algorithms to minimise the prediction error and enhance overall performance [47]. To introduce non-linearity and enable the learning of complex patterns, activation functions such as ReLU, sigmoid, or tanh are applied within each layer. Due to their architectural simplicity, flexibility, and strong generalisation capabilities**,** dense networks have been widely adopted in speech technology applications, including emotion recognition**,** speech command classification**,** and voice-based authentication systems [48]. In emotion recognition tasks, raw audio waveforms are typically pre-processed using feature extraction techniques like MFCCs. These features, which reflect perceptual characteristics of human hearing, are then fed into the dense layers for classification. The fully connected structure enables the model to capture subtle acoustic cues and higher-order emotional representations**,** thus enhancing its capacity to differentiate between nuanced affective states in speech [49]. This makes dense neural networks a suitable candidate for deployment in latency-sensitive, real-time IoT environments**,** particularly when integrated into hybrid deep learning architectures.

Audio feature extraction methods

MFCCs were chosen as the main audio features due to their ability to capture perceptually relevant spectral properties of speech with low computational cost. The extraction process segments the signal into overlapping Hamming-windowed frames, applies FFT, maps the spectra to the Mel scale, log-compresses, and reduces them through DCT to obtain cepstral coefficients. Prior studies show that MFCCs outperform alternative features in speech-emotion recognition in both accuracy and efficiency [50–52]. In this study, these coefficients were used in deep-learning models and were crucial for accurate emotion classification.

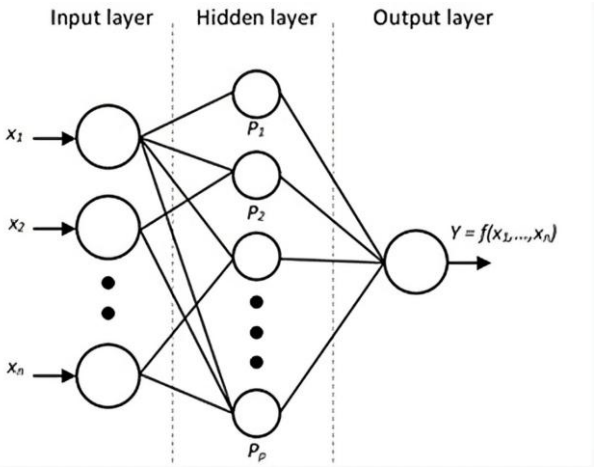


Figure 4. Sample dense neural network

4. EXPERIMENTAL RESULTS

4.1. Evaluation of the CNN-Dense Architecture

To capture spectral patterns in emotional speech, a hybrid deep learning architecture combining a CNN with fully connected Dense layers was implemented and evaluated on both the TESS and EMODB datasets. Each utterance was represented by 60 MFCCs. The datasets were partitioned using an 80:20 stratified split to ensure class balance, with 20% of the training data allocated for validation. The model architecture consisted of three sequential one-dimensional convolutional layers (64, 128, and 256 filters), each followed by batch normalisation, ReLU activation, max-pooling, and a dropout layer (rate: 0.25). The extracted features were flattened and passed through three dense layers (512, 256, and 128 units), each followed by batch normalisation, ReLU activation, and dropout (rate: 0.5). A final Softmax layer produced class probabilities. The model was trained using the Adam optimiser (learning rate = 0.001) and categorical cross-entropy loss with a batch size of 64 for up to 300 epochs. Early stopping (patience = 20) was used to prevent overfitting, and ReduceLROnPlateau and ModelCheckpoint were applied to improve training stability. To ensure robustness, the training and evaluation cycle was repeated across five independent runs with different random seeds. This repeated hold-out validation approach was adopted instead of k-fold cross-validation due to edge-device simulation constraints. As summarised in Table 1, the CNN–Dense model achieved an overall accuracy of 96% on the TESS dataset. The highest F1-scores were observed for disgust (0.98),

sadness (0.98), and anger (0.96), while the remaining classes—fear (0.95), neutral (0.96), happiness (0.94), and surprise (0.93) —also yielded high classification performance. To assess generalisability, the same model was evaluated on the EMODB dataset using identical architecture and training settings. An overall accuracy of 87% was achieved, with high F1-scores observed for disgust (0.94), sadness (0.90), anger (0.88), and boredom (0.88). Classification of anxiety/fear (0.77) was more challenging due to acoustic overlap. Nevertheless, macro and weighted F1-scores remained at 0.87, indicating stable cross-corpus performance. Latency evaluation conducted in a simulated edge environment (AWS t3.medium: 2 vCPUs, 4 GB RAM) revealed an average inference time of 58 ms per utterance, supporting the model's suitability for real-time, speech-based affective computing in resource-constrained scenarios.

Table 1. Comparison of Emotion Recognition Performance (CNN-Dense)

Emotion	P-T	R-T	F1-T	P-E	R-E	F1-E
Anger	0.97	0.96	0.96	0.89	0.87	0.88
Disgust	0.98	0.97	0.98	1.00	0.89	0.94
Fear/ Anxiety	0.96	0.95	0.95	0.83	0.71	0.77
Happy/ Happiness	0.96	0.93	0.94	0.72	0.80	0.80
Neutral	0.97	0.96	0.96	0.82	0.88	0.85
Surprise	0.91	0.96	0.93	-	-	-
Sad/ Sadness	0.97	1.00	0.98	0.81	1.00	0.90
Boredom	-	-	-	0.88	0.88	0.88
Accuracy	-	-	0.96	-	-	0.87
Macro Avg	0.96	0.96	0.96	0.87	0.87	0.87
W. Avg	0.96	0.96	0.96	0.87	0.87	0.87

Note. *P = Precision; R = Recall; F1 = F1-Score; T = TESS dataset; E = EMODB dataset; W. Avg = Weighted Average; Macro Avg = Macro Average.*

Figure 5 presents the classification confusion matrices and learning curves of the CNN–Dense hybrid architecture for both the TESS and EMODB datasets, enabling a comparative visual assessment of class-wise performance in speech emotion recognition. On the TESS dataset, the model achieved high predictive accuracy across all emotional categories. Sadness (n = 80), disgust (n = 78), anger (n = 77), and neutral (n = 77) were classified with near-perfect precision, underscoring the model's ability to capture distinct acoustic patterns of high-salience emotions. Fear (n = 76) and pleasant surprise (n = 77) were also reliably detected, while happiness (n = 75) showed minor variability, with occasional misclassifications into the neutral category. Specific confusions—such as fear mislabelled as neutral (n = 3) and happiness as neutral (n = 4) —are consistent with known prosodic and spectral overlaps among low- to moderate-arousal affective states. The matrix's strong diagonal structure and uniformly high classification counts indicate balanced generalisation with no evident class-specific bias. In the EMODB dataset, the model maintained robust performance, albeit with slightly reduced overall accuracy, attributable to the corpus's more subtle and contextually nuanced emotional expressions. The most accurately classified emotions included anger (n = 29), boredom (n = 15), and neutral (n = 15), with sadness (n = 11) also demonstrating acceptable results. Anxiety/fear emerged as the most challenging class, showing considerable confusion with anger and disgust, aligning with previous findings that highlight acoustic and affective ambiguity among these categories. Additional misclassifications occurred between neutral and both boredom and happiness, reflecting typical overlaps in tone, intensity, and rhythm. Despite these challenges, the model accurately classified the majority of samples across all classes. These results affirm the CNN–Dense architecture's capacity to generalise over both clean and naturalistic speech conditions. The observed classification patterns are consistent with psychological models of affect and empirical research on emotion recognition, thereby supporting the model's suitability for real-world affective computing applications.

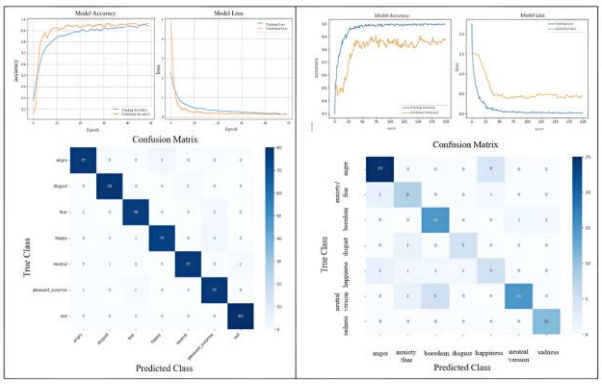


Figure 5. Classification Confusion Matrix for the CNN-Dense Architecture Using the TESS and EMODB

4.2. Evaluation of the LSTM-CNN Architecture

To capture the spectral and temporal dynamics of emotional speech, an LSTM–CNN hybrid architecture was developed and evaluated using both the TESS and EMODB datasets. Each utterance was represented by 60 MFCCs, extracted as input features. The datasets were partitioned using an 80:20 stratified split to maintain class balance, with 20% of the training data allocated for validation. The model architecture comprised three one-dimensional convolutional layers with 64, 128, and 256 filters, each followed by batch normalisation, ReLU activation, max-pooling, and dropout (rate: 0.25), allowing for robust spectral feature extraction. These representations were then passed to a stack of three LSTM layers (256, 128, and 64 units), with the first two configured to return full sequences. Batch normalisation and 0.25 dropout were applied to each recurrent layer to promote generalisation. The resulting sequence embeddings were processed by two fully connected layers (256 and 128 neurons), each followed by ReLU activation and 0.5 dropout. A final Softmax layer

produced class probabilities across the seven emotion categories defined in both the TESS and EMODB datasets. Model training was conducted using the Adam optimiser (initial learning rate = 0.001) and categorical cross-entropy loss. Learning stability was maintained via early stopping (patience = 20 epochs), as well as ReduceLROnPlateau and ModelCheckpoint callbacks. To ensure robustness, the training and evaluation process was repeated over five independent runs with different random seeds. Reported results represent mean ± standard deviation across these runs. On the TESS dataset, the LSTM–CNN model achieved an overall accuracy of 95.0 ± 0.4%, with high class-wise F1-scores: sadness (0.98), disgust (0.97), anger (0.96), fear (0.95), neutral (0.96), happiness (0.94), and surprise (0.93). These results confirm the model's effectiveness in learning both spectral and temporal emotion-related cues in structured speech. On the EMODB dataset, the model attained an overall accuracy of 74%, with macro and weighted F1-scores of 0.73, indicating moderate but variable performance across classes. Strong recognition was achieved for sadness (0.96), anger (0.89), and boredom (0.88), while lower scores were observed for neutral (0.50), disgust (0.67), and anxiety/fear (0.74), likely due to acoustic and affective overlaps. These results highlight the model's sensitivity to inter-class similarities in subtle or low-arousal affective expressions, particularly in more naturalistic speech recordings. Latency evaluation, conducted in a simulated edge environment using an AWS t3.medium instance (2 vCPUs, 4 GB RAM), revealed an average inference time of 71 ± 3 ms per utterance, indicating the model's feasibility for near-real-time edge deployment. In conclusion, the LSTM–CNN hybrid model effectively integrates spectral encoding with temporal sequence learning, demonstrating high performance on structured data (TESS) and generalisation capability on acoustically complex corpora (EMODB). The architecture shows particular promise for deployment in affective computing applications where both emotional salience and subtle variation must be accurately captured.

Figure 6 illustrates the training dynamics and class-wise prediction outcomes of the LSTM–CNN hybrid model on the TESS (left) and EMODB (right) datasets. The upper panels display training and validation accuracy/loss across epochs, while the lower panels present the corresponding confusion matrices. On the TESS dataset (left), the model exhibits stable and rapid convergence. Training and validation accuracies plateau around 95%, and both loss curves decline synchronously with no signs of overfitting. The associated confusion matrix confirms high discriminative performance across all emotion categories. Emotions such as sadness (n = 80), disgust (n = 78), anger (n = 76), and neutral (n = 77) were classified with near-perfect precision. Minor misclassifications were observed for fear and happy, particularly with confusion into the neutral category—an expected phenomenon due to shared low-arousal acoustic traits.

Table 2. Comparison of Emotion Recognition Performance (LSTM-CNN)

Emotion	P-T	R-T	F1-T	P-E	R-E	F1-E
Anger	0.97	0.96	0.96	0.82	0.97	0.89
Disgust	0.97	0.96	0.97	0.80	0.57	0.67
Fear/ Anxiety	0.96	0.95	0.95	0.88	0.64	0.74
Happy/ Happiness	0.96	0.93	0.94	0.67	0.80	0.73
Neutral	0.97	0.95	0.96	0.64	0.41	0.50
Surprise	0.91	0.96	0.93	-	-	-
Sad/ Sadness	0.97	1.00	0.98	0.92	1.00	0.96
Boredom	-	-	-	0.55	0.71	0.62
Accuracy	-	-	0.95	-	-	0.74
Macro Avg	0.95	0.95	0.95	0.75	0.73	0.73
W. Avg	0.95	0.95	0.95	0.74	0.74	0.73

Note. *P = Precision; R = Recall; F1 = F1-Score; T = TESS dataset; E = EMODB dataset; W. Avg = Weighted Average; Macro Avg = Macro Average.*

Despite these overlaps, the confusion matrix remains strongly diagonal, indicating high per-class accuracy and balanced model behaviour. In contrast, performance on the EMODB dataset (right) reveals greater variability. While the training accuracy continues to improve steadily, the validation accuracy plateaus at a lower level and exhibits fluctuations. The validation loss increases in later epochs, suggesting mild overfitting. The confusion matrix reflects this trend, with accurate predictions for salient emotions such as anger (n = 20), sadness (n = 10), and boredom (n = 7), but notable misclassifications for anxiety/fear, neutral, and disgust. Specifically, anxiety/fear shows confusion with anger and disgust, whereas neutral is frequently misclassified as boredom or happiness, which reflects the known acoustic similarity and affective ambiguity among these classes. Overall, the figure demonstrates the LSTM–CNN model's capacity to generalise effectively on the structured and well-separated TESS dataset, while revealing performance limitations when applied to more naturalistic and acoustically overlapping data such as EMODB.

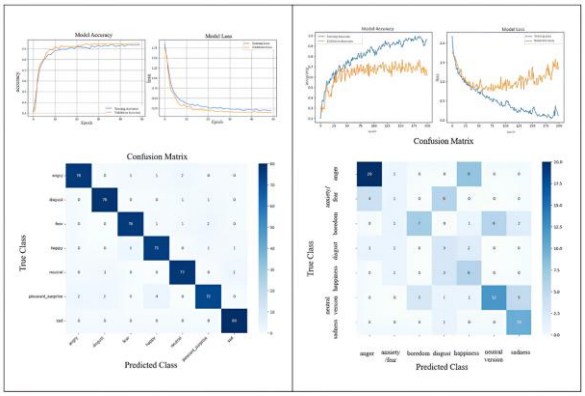


Figure 6. Classification Confusion Matrix for the LSTM-CNN Architecture Using the TESS and EMODB

4.3. Evaluation of the Dense-LSTM Architecture

To capture the sequential dynamics of affective speech, a deep Dense–LSTM hybrid architecture was constructed and comprehensively evaluated using the TESS and EMODB datasets. The input representation consisted of 60 MFCCs extracted per utterance, preserving both fine-grained spectral and temporal characteristics. The dataset was partitioned using an 80:20 stratified split, ensuring proportional class distribution across training and test subsets. Additionally, 20% of the training data was allocated for validation purposes. For both datasets, this corresponded to approximately 896 training, 224 test, and 179 validation samples in the case of TESS, and approximately 360, 90, and 72 samples, respectively, for EMODB. The architecture consisted of four sequential LSTM layers with progressively decreasing units: 512, 256, 128, and 64. The first three layers employed return_sequences=True to maintain full temporal resolution across time steps. Each LSTM block was followed by batch normalisation and dropout (rate: 0.25) to promote generalisation and suppress overfitting. The final LSTM output was passed to three fully connected Dense layers with 512, 256, and 128 neurons, each followed by ReLU activation, batch normalisation, and a dropout rate of 0.5 to enable abstract feature integration and robust representation learning. A final Softmax layer produced class probabilities over the seven emotional categories defined in both the TESS and EMODB datasets. Model training was performed using the Adam optimiser (initial learning rate = 0.001) and categorical cross-entropy loss function. The training was conducted over a maximum of 300 epochs with early stopping applied after 20 consecutive epochs of no improvement in validation loss. Convergence was further stabilised using the ReduceLROnPlateau (patience = 10, factor = 0.5) and ModelCheckpoint callbacks (monitoring validation accuracy). A batch size of 64 was used for both datasets. Training dynamics—including accuracy, loss, and learning rate—were logged across all epochs. To ensure robustness against random initialisation and sample selection variability, the complete training–evaluation pipeline was repeated five times with different random seeds. The reported performance metrics reflect the mean scores across these five independent runs. Latency analysis was conducted on an AWS t3.medium virtual instance (2 vCPUs, 4 GB RAM), yielding an average inference time of 84 ± 2 milliseconds per utterance, indicating moderate responsiveness suitable for edge-oriented deployment scenarios with limited real-time constraints. As shown in Table 3, the Dense–LSTM model achieved a mean accuracy of 93% on the TESS dataset. Class-wise performance was strong across all emotions: disgust (F1 = 0.94), sadness (0.93), fear (0.93), anger (0.94), happiness (0.94), and surprise (0.92), with neutral achieving an F1-score of 0.90. These results confirm the model's ability to accurately differentiate both high- and low-arousal emotions in structured acoustic conditions. On the more complex EMODB dataset, the model reached a mean accuracy of 79%, with relatively high F1-scores for sadness (0.91), boredom (0.86), and anger (0.79). However, classification performance for disgust (0.53) and anxiety/fear (0.57) was lower, reflecting the acoustic ambiguity inherent in these categories. The macro-averaged F1-score was 0.76, while the weighted average reached 0.78, indicating moderate generalisation capacity under acoustically diverse and emotionally subtle conditions. In summary, the Dense–LSTM model demonstrates strong performance on structured datasets like TESS, with moderate adaptability to less structured corpora such as EMODB. The architecture's balanced design—integrating deep temporal modelling with dense feature abstraction—proves effective for both salient and nuanced emotion recognition tasks, making it a viable candidate for speech-based affective computing in real-world applications.

Table 3. Comparison of Emotion Recognition Performance (Dense- LSTM)

Emotion	P-T	R-T	F1-T	P-E	R-E	F1-E
Anger	0.96	0.93	0.94	0.80	0.78	0.79
Disgust	0.94	0.93	0.94	0.55	0.52	0.53
Fear / Anxiety	0.92	0.95	0.93	0.60	0.55	0.57
Happy / Happiness	0.97	0.92	0.94	0.78	0.76	0.77
Neutral	0.90	0.91	0.90	0.70	0.85	0.77
Surprise	0.92	0.92	0.92	-	-	-
Sad / Sadness	0.91	0.96	0.93	0.88	0.94	0.91
Boredom	-	-	-	0.95	0.78	0.86
Accuracy	-	-	0.93	-	-	0.79
Macro Avg	0.93	0.93	0.93	0.76	0.76	0.76
Weighted Avg	0.93	0.93	0.93	0.79	0.79	0.78

Note. *P = Precision; R = Recall; F1 = F1-Score; T = TESS dataset; E = EMODB dataset; W. Avg = Weighted Average; Macro Avg = Macro Average.*

Figure 7 presents the training trajectories and confusion matrices of the Dense–LSTM model across both TESS (right) and EMODB (left) datasets. These visualisations provide further insight into the learning stability and class-wise classification behaviour of the model. On the TESS dataset, the training and validation accuracy curves show a steady convergence, with final validation accuracy exceeding 93%. The corresponding loss values exhibit a smooth downward trend, with no significant divergence between training and validation loss—indicating minimal overfitting. The confusion matrix demonstrates that the model achieved near-perfect classification across all seven emotion categories. Particularly, anger, disgust, and sadness were correctly classified with minimal confusion (n = 75, 75, and 77, respectively), while fear, happy, and surprise also yielded strong performance. Minor misclassifications were observed, such as happy instances occasionally predicted

as neutral and a small number of pleasant surprise samples misclassified as happy or sad. Nevertheless, overall performance was highly balanced across both high- and low-arousal states. In contrast, performance on the EMODB dataset displayed greater variability. The training accuracy steadily increased and plateaued near 80%, while validation accuracy exhibited fluctuations, with the validation loss remaining consistently above the training loss—suggesting a degree of overfitting, likely due to the dataset's acoustic complexity and limited size. The confusion matrix reveals that while anger, sadness, and boredom were classified with acceptable precision (n = 28, 11, and 12, respectively), substantial confusion occurred among anxiety/fear, neutral, and disgust. For example, neutral samples were often misclassified as boredom, happiness, or sadness, indicating overlapping acoustic features in these classes. This highlights the model's sensitivity to inter-class ambiguity and subtle affective variations present in more naturalistic recordings. In summary, Figure 7 confirms that the Dense–LSTM model generalised well on the structured TESS dataset, achieving high class-wise performance with stable convergence. However, it encountered challenges when applied to the more acoustically nuanced EMODB corpus, especially in differentiating low-arousal and contextually ambiguous emotions. These results reflect the trade-off between model complexity and data generalisability, further underscoring the importance of dataset characteristics in emotion recognition performance.

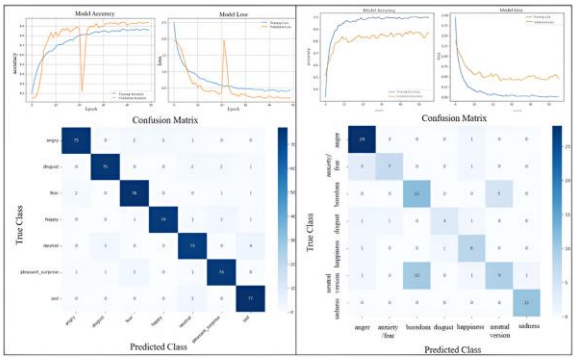


Figure 7. Classification Confusion Matrix for the Dense-LSTM Architecture Using the TESS and EMODB

5. DISCUSSION

This study has empirically examined and compared the performance of three hybrid deep learning models—CNN–Dense, LSTM–CNN, and Dense–LSTM—on two well-established emotional speech datasets: TESS and EMODB. By evaluating both classification performance and latency in a simulated edge computing setting, the investigation aimed to determine each model's suitability for real-time, resource-constrained environments. The CNN–Dense model emerged as the top-performing architecture, attaining the highest overall accuracy and F1-scores across both datasets. On TESS, it reached an overall accuracy of 96% with weighted and macro F1-scores of 0.96. EMODB results, while comparatively lower, remained acceptable, with 84% accuracy and 0.81 weighted F1-score. The model demonstrated high recognition rates for sadness (F1 = 0.98), anger (F1 = 0.96), and disgust (F1 = 0.97) in TESS. In EMODB, however, neutral (F1 = 0.82) and fear (F1 = 0.75) were classified with lower accuracy, possibly due to the subtler acoustic signatures of these low-arousal emotions. These findings highlight the CNN–Dense model's aptitude for detecting emotions characterised by pronounced spectral cues. Its average inference latency of 58 milliseconds further reinforces its feasibility in edge deployments requiring rapid responsiveness. The LSTM–CNN hybrid, integrating sequential feature modelling with spatial abstraction, yielded robust results across datasets. On TESS, it achieved 95% accuracy and high F1-scores for sadness (0.98), anger (0.96), fear (0.95), and disgust (0.97), while slightly lower scores were noted for surprise (F1 = 0.93). On EMODB, overall accuracy dropped to 74%, with more variable performance across emotion classes. Notably, high recognition was observed for sadness (0.96), boredom (0.88), and anger (0.89), whereas disgust (0.67) and neutral (0.50) remained challenging. These results align with prior studies indicating the strength of LSTM components in capturing dynamic temporal features (Wang et al., 2020 [56]; Zhao et al., 2022 [57]; Wan et al., 2023 [58], [59]), although increased model complexity and longer inference times may present limitations. Similar to these findings, Mishra et al. [60] also demonstrated that CNN–BiLSTM hybrids outperform standalone models, reinforcing the advantage of combining convolutional and recurrent modules for emotion recognition tasks. The observed latency of 71 milliseconds confirms that while suitable for edge deployment, recurrent models may introduce moderate computational delays relative to feedforward configurations. The Dense–LSTM model, which reverses the architecture by applying dense transformations prior to LSTM layers, yielded a balanced performance profile. On TESS, it recorded 93% accuracy with F1-scores above 0.93 for anger, disgust, and sadness. Lower scores for happy (0.94), neutral (0.90), and surprise (0.92) were still competitive. On EMODB, accuracy stood at 79%, with macro and weighted F1-scores of 0.76 and 0.78, respectively. The model performed well for sadness (0.91) and boredom (0.86), but struggled with disgust (0.53) and fear/anxiety (0.57). These results build upon prior literature reporting moderate performance of Dense and LSTM models individually (e.g., [21], [61–62]). Furthermore, earlier studies based on feature selection and shallow models (e.g., Liu et al. [54]) highlight the limits of traditional approaches, showing why hybrid deep networks such as Dense–LSTM can better capture the nuanced interplay between spectral and temporal cues. The model's inference time of 84 milliseconds, while close to the upper latency bound for edge environments, remains within tolerable limits for most IoT applications. To statistically validate the differences in performance, Pearson's chi-square test was conducted. The test revealed significant differences in error

distributions among the three models ($p < 0.05$), corroborating that architectural configurations meaningfully affect classification outcomes. Despite the superiority of CNN–Dense in structured datasets such as TESS, its reduced performance on EMODB suggests limitations in generalisability. EMODB's higher acoustic variability and broader speaker diversity introduce classification ambiguity, particularly for low- to medium-arousal emotions such as neutral and fear. These limitations align with findings in recent literature, where cross-corpus speech emotion recognition models trained on specific datasets often experience marked performance degradation when applied to unseen corpora [63]. Domain adaptation and transfer learning techniques have therefore been increasingly proposed as effective solutions to mitigate such generalisation gaps [64]. As also highlighted in recent augmentation-focused work [53], performance robustness can be further improved by enriching training sets with synthetic or transformed samples to counteract dataset imbalance. While MFCCs offer a reliable baseline representation, enhancing their discriminative power through fusion with attention mechanisms (e.g., multiscale area attention) or transformer-based architectures may further improve performance robustness across variable acoustic domains. Additionally, transformer-based approaches (e.g., Wav2Vec2, HuBERT, SpeechT5), which have recently demonstrated superior performance in speech processing tasks, were not included in this study. Future research may compare the current hybrid models against these state-of-the-art frameworks to assess trade-offs in accuracy, latency, and model complexity. While all simulations were conducted in a software-based environment using a t3.medium AWS instance, the physical limitations of real-world edge devices—such as temperature fluctuations, power constraints, and energy consumption—were not incorporated. As noted in existing literature [65], hardware-level constraints may significantly impact model behaviour. Future work should explore model deployment on physical edge devices to evaluate robustness under such conditions. In addition to the simulation-based findings of this study, previous research has demonstrated the feasibility of deploying speech emotion recognition models on real embedded devices. Jetson Nano, equipped with a quad-core ARM processor and 4 GB RAM, has frequently been used as a reference platform for edge inference due to its low-power and small-form-factor characteristics [66]. Furthermore, lightweight architectures such as EdgeSpeechNets have achieved substantial reductions in model size (up to 7.8×), computational operations (36×), and latency (10×) when tested on mobile-class devices [67,68]. Similarly, TinySpeech has been proposed as an ultra-compact alternative, showing 507× fewer parameters and 2028× lower storage needs compared to conventional models [69]. These studies underscore the importance of assessing and reporting system-level constraints—such as RAM usage, parameter footprint, and inference latency—when developing deep learning models intended for real-world embedded deployment. Although the use of a t3.medium AWS instance (2 vCPUs, 4 GB RAM) exceeds the capabilities of typical edge devices such as Raspberry Pi 4 (4-core ARM Cortex-A72, 1–8 GB RAM), it was selected to balance feasibility and reproducibility in the absence of access to physical hardware. This instance provides a moderate baseline for latency evaluation, even if it cannot fully replicate thermal and power limitations of embedded platforms. In conclusion, the findings of this study contribute to both theoretical understanding and practical application of hybrid deep learning architectures in speech emotion recognition. CNN–Dense emerged as a high-performing, low-latency solution for controlled environments, while LSTM-based variants offer nuanced modelling for temporally complex or naturalistic settings. The observed dataset-specific performance differences, combined with evidence from both shallow-learning [54] and augmentation-focused studies [53], as well as hybrid deep learning approaches [59], emphasise the necessity of adaptable, resilient architectures for real-world affective computing applications.

6. CONCLUSION

In this study, a comparative evaluation of three hybrid deep learning architectures—CNN–Dense, LSTM–CNN, and Dense–LSTM—was carried out within a simulated edge computing environment for speech-based emotion recognition. The TESS and EMODB datasets were utilised to assess classification accuracy, F1-scores, and inference latency under latency-sensitive and resource-constrained conditions. Among the tested models, CNN–Dense achieved the highest performance (96% accuracy, 0.96 F1-score, 58 ms latency), indicating its suitability for real-time deployment in edge settings. While LSTM–CNN and Dense–LSTM also demonstrated acceptable accuracy, slightly longer inference times were observed. The current experimental design was limited to software-level simulation using AWS t3.medium instances. As a result, the physical and operational characteristics of actual embedded hardware—such as thermal variability, energy constraints, and latency under environmental stress—could not be fully captured. These factors have been recognised as limitations and are intended to be addressed in future work through direct implementation on platforms such as Raspberry Pi or Jetson Nano. Although the inclusion of EMODB alongside TESS provided a degree of acoustic and demographic variability, the generalisability of the findings remains limited. Therefore, it is envisaged that additional datasets such as RAVDESS and IEMOCAP will be incorporated in subsequent studies to enhance emotional diversity and contextual realism. Furthermore, transformer-based architectures (e.g., Wav2Vec2, HuBERT, SpeechT5), which were beyond the present study's scope, are expected to offer improvements in temporal and contextual emotion modelling. Future directions will also include the integration of attention mechanisms, the

development of emotion-specific spectral descriptors, and the optimisation of models for deployment on embedded edge devices.

DECLARATION OF ETHICAL STANDARDS

We declare that the materials and methods they use in their work do not require ethical committee approval and/or legal-specific permission.

AUTHORS' CONTRIBUTIONS

Buket İŞLER: Investigation; Methodology; Visualization; Writing - original draft; Supervision; Validation; Writing - review & editing.

CONFLICT OF INTEREST

There is no conflict of interest in this study.

REFERENCES

[1] Bonomi, F., Milito, R., Zhu, J., Addepalli, S., "Fog computing and its role in the internet of things" In Proceedings of the first edition of the MCC workshop on Mobile cloud computing (pp. 13-16), (2012).

[2] Al-Turjman, F.,"Intelligence and security in big 5G-oriented IoNT: An overview", ***Future Generation Computer Systems***, 102, 357-368, (2022).

[3] Mavaddati, S.," Voice-based age, gender, and language recognition based on ResNet deep model and transfer learning in spectro-temporal domain" ***Neurocomputing***, 580, 127429, (2024).

[4] Nonis, F., Dagnes, N., Marcolin, F., Vezzetti, E., "3D approaches and challenges in facial expression recognition algorithms-a literature review" ***Applied Sciences***, 9(18), 3904 ,(2019).

[5] Vasdev, D., Gupta, V., Shubham, S., Chaudhary, A., Jain, N., Salimi, M., Ahmadian, A.," Periapical dental X-ray image classification using deep neural networks", ***Annals of Operations Research***, (2022).

[6] Yang, Z., Li, Z., Zhou, S., Zhang, L., Serikawa, S., "Speech emotion recognition based on multi-feature speed rate and LSTM", ***Neurocomputing***, 601, 128177, (2024).

[7] Pitaloka, D. A., Wulandari, A., Basaruddin, T., Liliana, D. Y.," Enhancing CNN with preprocessing stage in automatic emotion recognition" ***Procedia computer science***, 116, 523-529, (2017).

[8] Ng, H. W., Nguyen, V. D., Vonikakis, V., Winkler, S.," Deep learning for emotion recognition on small datasets using transfer learning" In Proceedings of the 2015 ACM on international conference on multimodal interaction (pp. 443-449), (2015).

[9] Varshney, P., Simmhan, Y.," Demystifying fog computing: Characterizing architectures, applications and abstractions" In 2017 IEEE 1st international conference on fog and edge computing (ICFEC), 115-124., IEEE, (2017).

[10] Shi, W., Cao, J., Zhang, Q., Li, Y., Xu, L.," Edge computing: Vision and challenges", ***IEEE internet of things journal***, 3(5), 637-646. (2016).

[11] Kavitha, T., Malini, S., Senbagavalli, G., "Deep learning for emotion recognition", ***In Handbook of Research on Computer Vision and Image Processing in the Deep Learning Era,*** 56-91. IGI Global., (2023).

[12] Gastaldo, P., Ragusa, E., Dosen, S., Palmieri, F., "Special Issue on integration of machine learning and edge computing for next generation of smart wearable systems", ***Future Generation Computer Systems***, 107574, (2024).

[13] Calvo, R. A., D'Mello, S.,"Affect detection: An interdisciplinary review of models, methods, and their applications", ***IEEE Transactions on affective computing***, 1(1), 18-37. (2010).

[14] Poria, S., Cambria, E., Bajpai, R., Hussain, A., "A review of affective computing: From unimodal analysis to multimodal fusion", ***Information fusion***, 37, 98-125, (2017).

[15] Elkobaisi, M. R., Al Machot, F.," Human emotion modeling (HEM): an interface for IoT systems", ***Journal of Ambient Intelligence and Humanized Computing***, 13(8), 4009-4017, (2022).

[16] Kumar, N., Makkar, A., "Machine learning in cognitive IoT.",***CRC Press***, (2020).

[17] Gyrard, A., Boudaoud, K., "Interdisciplinary iot and emotion knowledge graph-based recommendation system to boost mental health", ***Applied Sciences***, 12(19), 9712, (2022).

[18] Chamishka, S., Madhavi, I., Nawaratne, R., Alahakoon, D., De Silva, D., Chilamkurti, N., Nanayakkara, V.," A voice-based real-time emotion detection technique using recurrent neural network empowered feature modelling", ***Multimedia Tools and Applications***, 81(24), 35173-35194, (2022).

[19] Oruh, J., Viriri, S., Adegun, A., "Long short-term memory recurrent neural network for automatic speech recognition", ***IEEE Access***, 10, 30069-30079, (2022).

[20] Yin, T., Dong, F., Chen, C., Ouyang, C., Wang, Z., Yang, Y.," A Spiking LSTM Accelerator for Automatic Speech Recognition Application Based on FPGA", ***Electronics***, 13(5), 827, (2024).

[21] Abdelwahab, M., Busso, C., "Study of dense network approaches for speech emotion recognition", In 2018 ***IEEE international conference on acoustics, speech and signal processing (ICASSP),***5084-5088. IEEE, (2018).

[22] Latif, S., Shahid, A., Qadir, J., "Generative emotional AI for speech emotion recognition: The case for synthetic emotional speech augmentation", ***Applied Acoustics***, 210, 109425, (2023).

[23] Kaur, J., Kumar, A., "Speech emotion recognition using CNN, k-NN, MLP and random forest", ***In Computer Networks and Inventive Communication Technologies: Proceedings of Third ICCNCT 2020*** (pp. 499-509). Springer Singapore, (2021).

[24] Jabbooree, A. I., Khanli, L. M., Salehpour, P., Pourbahrami, S., "A novel facial expression recognition algorithm using geometry β–skeleton in fusion based on deep CNN", ***Image and Vision Computing***, 134, 104677, (2023).

[25] Qazi, H., Kaushik, B. N., "A hybrid technique using CNN+ LSTM for speech emotion recognition",

International Journal of Engineering and Advanced Technology (IJEAT), 9(5), 1126-1130, (2020).

[26] Chowdhury, J.H., Ramanna, S. Kotecha, K., "Speech emotion recognition with light weight deep neural ensemble model using hand crafted features", ***Sci Rep*** 15, 11824, (2025).

[27] Pichandi, S., Balasubramanian, G. Chakrapani, V.," Hybrid deep models for parallel feature extraction and enhanced emotion state classification.", ***Sci Rep*** 14, 24957, (2024).

[28] Donuk, K., Arı, A., Özdemir, M. F., Hanbay, D., "Deep feature selection for facial emotion recognition based on BPSO and SVM." ***Politeknik Dergisi***, 26(1), 131–142, (2023).

[29] Demirel, U., Çam, H., "Investigation of fluctuations in cryptocurrency transactions with sentiment analysis." ***Politeknik Dergisi***, 28(3), 773–784, (2025).

[30] Kelebekler, E., İnal, M., "Otomobil içindeki cihazların sesle kontrolüne yönelik konuşma tanıma sisteminin gerçek zamanlı laboratuar uygulaması." ***Politeknik Dergisi***, 11(2), 109-114, (2008).

[31] Turgut, Z., Üstebay, S., Ali Aydın, M., Gürkaş Aydın, G. Z., Sertbaş, A. " Performance analysis of machine learning and deep learning classification methods for indoor localization in Internet of things environment." ***Transactions on emerging telecommunications technologies***, 30(9), e3705, (2019).

[32] Marjani, M., Nasaruddin, F., Gani, A., Karim, A., Hashem, I. A. T., Siddiqa, A., Yaqoob, I.," Big IoT data analytics: architecture, opportunities, and open research challenges." ***IEEE Access***, 5, 5247-5261, (2017).

[33] Al-Fuqaha, A., Guizani, M., Mohammadi, M., Aledhari, M., Ayyash, M.," Internet of things: A survey on enabling technologies, protocols, and applications.", ***IEEE communications surveys & tutorials,*** 17(4), 2347-2376, (2015).

[34] Atzori, L., Iera, A., Morabito, G., "The internet of things: A survey.", ***Computer networks***, 54(15), 2787-2805, (2010).

[35] Kaya, Ş. M., İşler, B., Abu-Mahfouz, A. M., Rasheed, J., AlShammari, A., "An intelligent anomaly detection approach for accurate and reliable weather forecasting at IoT edges: A case study.", ***Sensors***, 23(5), 2426, (2023).

[36] Kaya, Ş. M., Erdem, A., Güneş, A., "A smart data pre-processing approach to effective management of big health data in IoT edge.", ***Smart Homecare Technology and TeleHealth***, 9-21, (2021).

[37] O'Connor, P. J., Hill, A., Kaya, M., Martin, B., "The measurement of emotional intelligence: A critical review of the literature and recommendations for researchers and practitioners.", ***Frontiers in psychology***, 10, 429307, (2019).

[38] Song, Y., Tung, P. H., Jeon, B., "Trends in Artificial Emotional Intelligence Technology and Application." In 2022 IEEE/ACIS 7th International Conference on Big Data, Cloud Computing, and Data Science (BCD) (pp. 366-370). IEEE, (2022).

[39] Lopes, A. T., De Aguiar, E., De Souza, A. F., Oliveira-Santos, T., "Facial expression recognition with convolutional neural networks: coping with few data and the training sample order.", ***Pattern recognition***, 61, 610-628, (2017).

[40] Kong, F., "Facial expression recognition method based on deep convolutional neural network combined with improved LBP features.", ***Personal and Ubiquitous Computing***, 23, 531-539, (2019).

[41] Wen, G., Chang, T., Li, H., Jiang, L.," Dynamic objectives learning for facial expression recognition.", ***IEEE Transactions on Multimedia***, 22(11), 2914-2925, (2020).

[42] Rawat, W., Wang, Z., "Deep convolutional neural networks for image classification: A comprehensive review.", ***Neural computation***, 29(9), 2352-2449, (2017).

[43] Shrestha, G., Das, M., Dey, N., "Plant disease detection using CNN.", In 2020 IEEE applied signal processing conference (ASPCON) (pp. 109-113). IEEE, (2020).

[44] Krishnamoorthy, P., Sathiyanarayanan, M., Proença, H. P., "A novel and secured email classification and emotion detection using hybrid deep neural network.", ***International Journal of Cognitive Computing in Engineering***, 5, 44-57, (2024).

[45] Hochreiter, S., Schmidhuber, J., "Long short-term memory. Neural computation, 9(8), 1735-1780.", (1997).

[46] Pascanu, R., Mikolov, T., Bengio, Y., "On the difficulty of training recurrent neural networks.", In International conference on machine learning (pp. 1310-1318). Pmlr., (2013).

[47] Gers, F. A., Schmidhuber, J., Cummins, F., "Learning to forget: Continual prediction with LSTM", ***Neural computation***, 12(10), 2451-2471, (2000).

[48] Hinton, G., Deng, L., Yu, D., Dahl, G. E., Mohamed, A. R., Jaitly, N., ... Kingsbury, B." "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups", ***IEEE Signal processing magazine,*** 29(6), 82-97, (2012).

[49] Dahl, G. E., Sainath, T. N., Hinton, G. E., "Improving deep neural networks for LVCSR using rectified linear units and dropout", In 2013 IEEE international conference on acoustics, speech and signal processing (pp. 8609-8613). ***IEEE,*** (2013).

[50] Graves, A., Mohamed, A. R., Hinton, G., "Speech recognition with deep recurrent neural networks", In 2013 IEEE international conference on acoustics, speech and signal processing (pp. 6645-6649). ***IEEE,*** (2013).

[51] Yu, D., Deng, L., "Automatic speech recognition (Vol. 1)" Berlin: Springer, (2016).

[52] Davis, S., Mermelstein, P., "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences", ***IEEE transactions on acoustics, speech, and signal processing***, 28(4), 357-366, (1980).

[53] Bakır, H., Çayır, A. N., Navruz, T. S., "A comprehensive experimental study for analyzing the effects of data augmentation techniques on voice classification" ***Multimedia Tools and Applications***, 83(6), 17601-17628, (2024).

[54] Liu, Z. T., Wu, M., Cao, W. H., Mao, J. W., Xu, J. P., Tan, G. Z., "Speech emotion recognition based on feature selection and extreme learning machine decision tree" ***Neurocomputing***, 273, 271-280, (2018).

[55] Li, F., Luo, J., Wang, L., Liu, W., Sang, X., "GCF2-Net: Global-aware cross-modal feature fusion network for speech emotion recognition", ***Frontiers in Neuroscience***, 17, 1183132, (2023).

[56] Wang, J., Xue, M., Culhane, R., Diao, E., Ding, J., Tarokh, V., "Speech emotion recognition with dual-sequence LSTM architecture", In ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 6474-6478). IEEE, (2020).

[57] Zhao, R., Jiang, X., Yu, F. R., Leung, V., Wang, T., Zhang, S., "Leveraging Cross-Attention Transformer and Multi-Feature Fusion for Cross-Linguistic Speech Emotion Recognition", arXiv preprint arXiv:2501.10408, (2025).

[58] Wang, C., Shen, X., "Feature-Enhanced Multi-Task Learning for Speech Emotion Recognition Using Decision Trees and LSTM", ***Electronics***, 13(14), 2689, (2024).

[59] Oyucu, S., Polat, H., "A language model optimization method for Turkish automatic speech recognition system", ***Politeknik Dergisi***, 26(3), 1167-1178, (2023).

[60] Mishra, S., Bhatnagar, N., P, P. et al., "Speech emotion recognition and classification using hybrid deep CNN and BiLSTM model", ***Multimed Tools Appl*** 83, 37603–37620, (2024).

[61] Costa, Y. M., Oliveira, L. S., Silla Jr, C. N.,"An evaluation of convolutional neural networks for music classification using spectrograms", ***Applied soft computing***, 52, 28-38, (2017).

[62] Pandiammal, K. S., Karishma, S., Sakthe, K. H., Manimaran, V., Kalaiselvi, S., Anitha, V., "Emotion Recognition from Speech–an LSTM approach with the Tess Dataset", In 2024 5th International Conference on Innovative Trends in Information Technology (ICITIIT) (pp. 1-6). IEEE, (2024).

[63] Chaugule, Irfan.,"Enhancing Generalizatıon In Speech Emotıon Recognıtıon: A Comprehensıve Review Of Domaın Adaptatıon And Transfer Learnıng Technıques", ***International Journal of Research in Management & Social Science Volume 13***, Issue 2 April- June 2025, (2025).

[64] Zhao, Y., Wang, J., Zong, Y., Zheng, W., Lian, H., Zhao, L., "Deep implicit distribution alignment networks for cross-corpus speech emotion recognition", In ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 1-5). IEEE, 2023).

[65] Wang, X., Jia, W., "Optimizing edge AI: a comprehensive survey on data, model, and system strategies", arXiv preprint arXiv:2501.03265, (2025).

[66] Pandelea, V., Hulea, D., Suciu, G., "Real-Time Emotion Recognition System on Jetson Nano Edge Device", **Sensors**, 21(13), 4496, (2021).

[67] Lin, S., Li, B., Cook, C., Berisha, V., Liss, J.," EdgeSpeechNets: Highly efficient deep neural networks for speech recognition on the edge", arXiv preprint arXiv:1810.08559, (2018).

[68] Wong, A., Famuori, M., Shafiee, M. J., Chwyl, B., "TinySpeech: Attention Condensers for Deep Speech Recognition on Edge Devices", arXiv preprint arXiv:2008.04245, (2020).

[69] Ozdemır, M. F., Hanbay, D.," Real-time scalable system for face tracking in multi-camera", ***Politeknik Dergisi***, 1-1, (2024).