Computers, Materials & Continua

Tech Science Press

https://doi.org/10.32604/cmc.2026.084270

ARTICLE

A Lightweight Edge Deployable Deep Learning Framework for Speech-Based Pain Classification across Heterogeneous Datasets

Nourah Fahad Janbi*

Department of Information Technology, College of Computing and Information Technology at Khulais, University of Jeddah, Jeddah, Saudi Arabia

*Corresponding Author: Nourah Fahad Janbi. Email: nfjanbi@uj.edu.sa

Received: 19 April 2026; Accepted: 11 June 2026; Published: 13 August 2026

ABSTRACT: Automatic pain assessment from speech is an emerging approach for non-invasive and objective healthcare monitoring. However, many existing methods are evaluated on single datasets or under controlled conditions, which limits their robustness under diverse real-world conditions. This paper presents a lightweight, end-to-end deep learning framework for speech-based pain level classification that explicitly targets multi-dataset robustness assessment. The proposed approach uses log-Mel spectrograms with an EfficientNetV2B0 backbone to learn discriminative acoustic features without relying on handcrafted feature engineering or multi-stage pipelines. The model is evaluated on heterogeneous datasets, including clinical recordings, controlled experimental data, and a merged dataset representing diverse conditions. Experimental results show competitive performance, achieving up to 77.8% accuracy in four-class classification, 71.7% in three-class classification, and 87.7% in binary pain detection on the merged dataset. Compared with previous works, the approach achieves competitive performance, with stronger results in clinical and reduced-class settings. The model was also converted using TensorFlow Lite to enable efficient edge deployment with minimal performance degradation and reduced computational cost, supporting future deployment in mobile and edge-based healthcare applications.

KEYWORDS: Healthcare; pain classification; pain assessment; deep learning; transfer learning; speech-based classification; audio signal processing; edge computing; healthcare AI

1 Introduction

Healthcare systems worldwide are rapidly changing due to advancements in data-driven technology and artificial intelligence (AI) [1,2]. AI has shown significant potential in improving diagnostic accuracy, enabling early disease detection, optimizing clinical workflows, and supporting decision-making in complex medical scenarios [3]. Using large-scale data and learning complex patterns, AI-based systems can provide objective, consistent, and scalable solutions that complement traditional clinical practices [4,5]. In particular, the integration of AI into healthcare has opened new opportunities for non-invasive monitoring and assessment of patient conditions using modalities such as imaging, physiological signals, and speech [6–8].

Pain is one of the fundamental human experiences and an important clinical indicator used for diagnosis, treatment planning, and patient monitoring [9]. However, accurate assessment of pain intensity is still a significant challenge in healthcare due to its inherently subjective nature and reliance on self-reporting scales. These conventional approaches are often unreliable in cases involving infants, elderly patients, individuals with communication impairments, or in emergency scenarios where immediate assessment is

required. As a result, there is a growing need for an objective, non-invasive, and automated approach to pain evaluation [10,11].

From a clinical perspective, reliable and objective pain assessment is essential for many diagnostic and care scenarios [9]. In emergency and triage settings, automated pain analysis may provide supplementary information to support patient assessment, particularly when self-reporting is limited or unreliable. In continuous monitoring scenarios, such as hospital wards or intensive care units, non-invasive assessment methods may support longitudinal observation of patient discomfort and provide complementary contextual information. Moreover, in remote healthcare and telemedicine applications, the ability to assess pain without direct physical interaction is also essential. These challenges are especially critical for vulnerable populations such as elderly individuals, infants, and patients with communication impairments, where conventional pain assessment scales are difficult to apply. Therefore, this has motivated the exploration of accessible, non-invasive modalities, such as speech, for automated pain evaluation.

The recent advances in AI technologies and algorithms have enabled the development of automatic pain recognition systems using various modalities, such as physiological signals [11–14], facial expressions [15–19], and speech [20–22]. In the literature, the multimodal approaches have demonstrated strong performance by combining complementary sources of information [23–25]. However, they typically require specialized sensors, controlled environments, and complex synchronization mechanisms, which limit their scalability and real-world applicability. On the other hand, speech-based approaches offer a practical and cost-effective alternative. This is because voice signals can be captured easily with widely available devices, without requiring intrusive equipment.

Speech carries rich paralinguistic information that reflects physiological and emotional states, including pain [26,27]. Variations in pitch, intensity, spectral distribution, and temporal dynamics provide informative indicators for pain assessment, making speech a practical, non-invasive modality for automatic pain recognition. However, despite recent progress in the literature, existing speech-based approaches face two fundamental challenges. First, many methods rely on handcrafted acoustic features [20] or multi-stage processing pipelines [24], which limit their ability to learn robust and transferable representations in complex acoustic environments. Second, most studies are evaluated on single datasets collected under controlled conditions [28], which produce models that struggle to generalize across diverse recording settings, speaker populations, and annotation schemes. For that, the development of a lightweight, end-to-end speech-based pain classification framework that maintains robustness across heterogeneous datasets remains insufficiently explored.

To overcome these limitations, this work presents a lightweight, end-to-end deep learning (DL) framework for speech-based pain classification, explicitly designed to evaluate performance under heterogeneous multi-dataset conditions. The proposed approach learns discriminative representations directly from log-Mel spectrograms using an EfficientNetV2B0 backbone, which eliminates the need for handcrafted features and complex pipelines. Unlike previous works, the model is systematically evaluated across heterogeneous datasets, including clinical recordings, controlled experimental data, and their merged combination. This aims to provide a more realistic assessment of robustness. In addition, the framework is evaluated under edge deployment conditions using TensorFlow Lite (TFLite) [29] conversion and on-device inference analysis to show the model feasibility in real-world applications.

The main contributions of this work can be summarized as follows:

1. A lightweight and deployment-oriented end-to-end DL framework for speech-based pain classification using log-Mel spectrogram representations and EfficientNetV2B0.

2. A systematic multi-dataset robustness evaluation across two heterogeneous speech-based pain datasets and their merged configuration.
3. An edge deployment evaluation demonstrating efficient on-device inference with minimal performance degradation.

Compared with many prior studies that primarily focus on improving classification accuracy using increasingly complex architectures or multimodal sensor configurations, the contribution of this work lies in the development and evaluation of a practical, lightweight, and deployment-oriented speech-based pain classification framework. The novelty of the proposed approach is not based on introducing a fundamentally new neural architecture, but rather on integrating efficient end-to-end learning, heterogeneous multi-dataset evaluation, and edge deployment analysis within a unified framework. In particular, the study emphasizes robustness under diverse recording conditions while maintaining computational efficiency suitable for resource-constrained healthcare environments.

The rest of this paper is structured as follows: Section 2 reviews related work in automatic pain recognition. Section 3 describes the data processing pipeline and feature engineering steps. Section 4 presents the proposed DL framework, detailing the model architecture, training strategy, and inference procedure. Section 5 discusses edge deployment considerations and clinical integration perspectives, outlining potential real-world applications. Section 6 presents experimental results, including both core classification performance and edge deployment evaluation under practical conditions. Finally, the paper is concluded and future work directions are outlined in Section 7.

2 Related Work

Automatic pain recognition has attracted significant attention in affective computing and healthcare AI, with research primarily focusing on multimodal sensing, physiological signals, facial expressions, and voice-based approaches. This section reviews relevant studies aligned with this work and highlights their key contributions and limitations. Table 1 provides a summary of the reviewed works.

Table 1: Summary of related work.

Ref.	Key Finding	Limitation
Oshrat et al. [20]	Speech prosody features correlate with pain and enable classification	Limited dataset size and binary classification focus
Velana et al. [30]	Multimodal dataset combining physiological, video, and audio signals for pain recognition	Requires controlled experimental setup and specialized sensors
Thiam et al. [25]	Multimodal fusion improves robustness of pain recognition	High system complexity and dependency on multiple modalities
Alhudhaif [21]	End-to-end DL using Log-Mel and GRU improves feature learning	Requires large datasets and may lack interpretability
Tsai et al. [24]	Stacked bottleneck vocal features with LSTM capture temporal and prosodic patterns for pain classification	Relies on handcrafted features and multi-stage processing pipeline, limiting end-to-end optimization and scalability

(Continued)

Table 1 (continued)

Ref.	Key Finding	Limitation
Hernández-de-la-Cruz et al. [31]	Multi-task vocal analysis framework enables simultaneous assessment of pain and related conditions using shared representations	Limited multiclass performance and strong dependency on dataset size and class balance, requiring further validation
Lu and Lu [28]	CNN-based audio classification shows strong correlation between spectral features and pain	Limited real-world generalization and reliance on controlled datasets
Aung et al. [32]	EmoPain dataset captures facial, motion, and EMG signals for chronic pain	Complex multimodal synchronization and limited scalability
Yang et al. [23]	Multimodal DL improves accuracy in clinical pain assessment	High computational cost and sensor dependency
Ricossa et al. [33]	Interpretable acoustic features (e.g., spectral entropy) correlate with pain	Limited scalability and reliance on handcrafted features

Oshrat et al. [20] explored the relationship between subjective pain reports and measurable speech prosody signals, aiming to enhance pain assessment methods for clinicians. Their work demonstrates that newly developed features improve the classification of significant vs. non-significant pain, suggesting potential for more objective pain evaluation in medical settings. The study employed a machine learning analysis using the WEKA suite for data analysis, specifically utilizing the Correlation-based Feature Selection (CFS) method to select effective attributes from the feature set.

Velana et al. [30] introduced the SenseEmotion database, a comprehensive multimodal dataset designed for automatic pain and emotion recognition that integrates physiological signals (e.g., ECG, EMG, respiration, and skin conductance), video-based facial expressions, and audio recordings recorded under controlled experimental conditions. Building on this dataset, Thiam et al. [25] further developed a multimodal pain intensity recognition framework that systematically evaluates different fusion strategies to combine audio, video, and physiological modalities, demonstrating that such fusion can significantly improve classification performance and robustness compared to single-modality approaches. Together, these works emphasize that incorporating biosignals alongside audiovisual data enables more objective and reliable pain assessment by capturing complementary physiological and behavioral responses. However, both the dataset and the associated modeling approaches rely heavily on specialized sensing equipment and controlled laboratory environments, including synchronized biosignal acquisition systems and multi-camera setups, which limit their practicality for real-world deployment.

Alhudhaif [21] proposed a speech-based pain classification framework using a GRU-Mixer architecture combined with Log-Mel spectrogram features. The model employs bidirectional GRU layers to capture temporal dependencies in vocal signals and applies temporal pooling to generate compact representations for classification. The approach is evaluated on the TAME dataset using speaker-independent splits, demonstrating its effectiveness for both binary and multiclass pain prediction tasks.

Tsai et al. [24] proposed a speech-based pain classification framework using stacked bottleneck acoustic features embedded within an LSTM architecture, focusing on emergency triage scenarios. Their approach

employs prosodic and spectral features extracted from patient speech and models temporal dependencies using deep recurrent networks. The study demonstrated that vocal characteristics, particularly prosodic features, contain significant information related to pain intensity and can be used for automatic classification. However, the approach relies on handcrafted acoustic features and a multi-stage processing pipeline that includes feature extraction, encoding, and classification, which may limit scalability and end-to-end optimization on more complex, diverse real-world datasets.

Hernández-de-la-Cruz et al. [31] presented a pilot study proposing a multi-task learning framework for simultaneous vocal assessment of chronic pain, presbyphonia, and age estimation using acoustic features such as MFCCs, jitter, and shimmer. The model was designed to jointly learn multiple related tasks using a shared architecture, aiming to improve computational efficiency and exploit shared representations across vocal biomarkers. While the framework demonstrated feasibility for detecting certain conditions, its performance in multiclass pain classification remained limited, highlighting challenges with generalization and a strong dependence on dataset size and class balance. The authors emphasized the need for larger datasets, hybrid architectures, and real-world validation to enhance clinical applicability

Lu and Lu [28] proposed a voice-based framework for pain level classification using acoustic features and convolutional neural networks (CNNs). Their work demonstrates that spectral features of speech, such as pitch and energy distribution, are strongly correlated with pain levels. The study presents a lightweight, cost-effective solution that can be deployed in real time on edge devices. However, existing audio-based approaches often rely on pre-recorded datasets and may struggle to generalize to real-world noisy environments.

Aung et al. [32] presented the EmoPain dataset, a multimodal dataset designed for chronic pain analysis. The dataset integrates facial videos, audio signals, full-body motion capture, and electromyographic data collected from patients performing rehabilitation exercises. Their work emphasizes the importance of capturing both communicative behaviors, such as facial expressions, and protective behaviors, such as movement patterns. However, despite its richness, the dataset introduces significant complexity due to multimodal synchronization and requires controlled experimental setups, which limit scalability and real-time deployment.

For older patients with hip fractures, Yang et al. [23] developed a multimodal pain recognition system that integrates voice and facial expressions to assess pain severity. Utilizing the ResNet-50 model for visual features and a VGGish network enhanced with BiLSTM and attention mechanisms for audio processing, followed by a weighted decision-level fusion strategy to integrate both modalities into a unified classifier. The system is trained on a self-collected clinical dataset and evaluated on the BioVid dataset. The study highlights the advantages of multimodal fusion in addressing the limitations of unimodal systems, particularly in complex, subjective scenarios such as pain assessment in older adults.

Ricossa et al. [33] investigated the automatic analysis of infant cry signals for pain assessment by extracting a set of handcrafted acoustic indicators, including cry duration, fundamental frequency, and a novel dysphonation measure based on spectral entropy. The proposed method focuses on identifying distress levels by segmenting cry signals and computing statistical features that correlate with human-based pain assessment scales. Notably, the study emphasizes interpretability and performs effectively even with a limited dataset, relying on signal-processing techniques rather than data-intensive machine learning models. The results show that certain features, particularly the spectral entropy-based dysphonation score, correlate well with perceived pain levels, highlighting the potential of acoustic analysis for objective infant pain evaluation.

Recent transformer-based and self-supervised approaches have demonstrated strong capability in affective and pain-related analysis. For example, Gkikas et al. [34] introduced a transformer-based foundation

model for multimodal automatic pain assessment, while Bonafos et al. [35] explored speech transformer architectures for extracting information from infant cries, and Yan et al. [36] proposed a multiscale masked autoencoder framework for neonatal pain speech emotion recognition. Nevertheless, such approaches generally require substantially larger datasets and higher computational resources.

Despite the significant progress in automatic pain recognition, several limitations remain across existing approaches. Multimodal methods provide comprehensive and accurate assessment by combining physiological, visual, and audio signals [23–25]; however, they rely on specialized hardware, complex data synchronization, and controlled experimental environments, which limit their scalability and practical deployment in real-world settings. Vision-based approaches [30,32,37] are sensitive to environmental conditions, such as lighting and occlusion, which reduces their robustness in uncontrolled settings. Similarly, recent transformer-based and self-supervised models [34–36], although promising, typically require large-scale datasets and substantial computational resources for effective training and generalization.

On the other hand, speech-based methods offer a promising, non-invasive, and cost-effective alternative. However, many existing works either rely on handcrafted features or perform evaluation on pre-recorded datasets, which limits their ability to generalize to diverse, noisy real-world conditions. Moreover, there are a few approaches that effectively utilize modern deep convolutional architectures to learn robust and discriminative representations from speech signals while maintaining efficiency and scalability.

Therefore, there is a clear need for a lightweight, end-to-end, and scalable audio-based framework that can reliably classify pain without relying on complex multimodal infrastructure. This gap was the motivation behind the proposed approach, which uses DL with log-Mel spectrogram representations and an EfficientNetV2B0 backbone to deliver a practical, deployable solution for real-world healthcare applications.

3 Data Processing and Feature Engineering

This section describes the data processing pipeline and feature engineering steps used to transform raw speech signals into structured representations suitable for deep learning–based pain classification. The methodology focuses on ensuring consistency across datasets, improving signal quality, and extracting discriminative time–frequency features.

3.1 Dataset and Label Harmonization

In this study, three datasets are used: the Chronic Pain Audio Dataset, the TAME Pain Dataset, and a merged dataset combining the two.

3.1.1 The Chronic Pain Audio Dataset

The Chronic Pain Audio (Chronic Pain) dataset [38] is a publicly available dataset designed to support research on speech-based pain-level classification. The dataset contains 256 audio recordings collected from Spanish-speaking patients aged 24–84 years with musculoskeletal pain.

Each audio sample is annotated according to a four-level verbal pain intensity scale: None, Low, Medium, and Strong. The recordings were obtained in real-world clinical and semi-clinical settings, including medical consultations, physiotherapy sessions, and voluntary contributions during a knee prosthesis surgery campaign. This diverse acquisition process enhances the variability in speech characteristics across different pain levels and improves the dataset's representativeness.

3.1.2 The TAME Pain Dataset

The TAME Pain dataset [39] is a large-scale and publicly available speech dataset designed to investigate the relationship between acute pain and vocal expression. The dataset contains 7039 annotated audio utterances totaling approximately 311 min of speech, collected from 51 adult participants. All participants were healthy adults aged 18 to 35 years and fluent English speakers.

Pain was experimentally induced under controlled laboratory conditions using the Cold Pressor Task, in which participants submerged one hand in cold water maintained at 0°C–4°C, with a warm water condition serving as a control. During each condition, participants read aloud phonetically balanced Harvard sentences and periodically reported their pain intensity on a numeric self-assessment scale from 1 to 10. Each spoken sentence and pain report was recorded as an individual utterance, resulting in fine-grained temporal alignment between speech signals and self-reported pain levels. The recording durations range from 0.33 to 5.88 s, with an average length of approximately 2.65 s. Each is annotated with multiple metadata fields, including raw and revised pain levels, experimental condition labels (cold or warm), and detailed quality annotations describing background noise, speech errors, audible breathing, and other non-speech vocal events.

To ensure consistency between datasets and enable a unified multi-dataset learning framework, the pain intensity annotations of the TAME Pain dataset were rescaled to match the four-level categorical scheme used in the Chronic Pain Dataset. Specifically, the original numeric pain scores in TAME Pain, provided as revised self-reported values on a scale from 1 to 10, were mapped into four discrete categories: None, Low, Medium, and Strong. Pain scores of $\{1, 2\}$ were mapped to None, $\{3, 4, 5\}$ to Low, $\{6, 7, 8\}$ to Medium, and $\{9, 10\}$ to Strong. Prior to relabeling, only high-quality utterances were retained by filtering samples with an ACTION_LABEL equal to zero, thereby excluding recordings affected by noise, speech errors, or other confounding artifacts. Following this process, the audio files were reorganized into class-specific directories based on the new categorical labels.

3.1.3 The Merged Dataset

The merged dataset is constructed by combining the Chronic Pain Dataset and the TAME Pain Dataset, yielding 4914 audio samples. Table 2 summarizes the class distribution across the three datasets, providing an overview of the number of samples in each pain level category.

Table 2: Class distribution across datasets.

Dataset	Low	Medium	Strong	None	Total
Chronic Pain	63	71	75	47	256
TAME Pain	1088	856	151	2563	4658
Merged Dataset	1151	927	226	2610	4914

The datasets used in this study differ substantially in language, recording conditions, speaker populations, and pain elicitation protocols. The Chronic Pain dataset contains Spanish clinical recordings collected under real-world conditions, whereas the TAME Pain dataset contains English speech recorded under controlled laboratory conditions using experimentally induced pain. These differences introduce considerable domain variability and may cause the model to partially learn dataset-specific or language-dependent patterns in addition to pain-related acoustic cues. The merged dataset therefore provides a more

heterogeneous evaluation setting, while also highlighting the challenges associated with robust speech-based pain classification across diverse domains.

3.2 Audio Preprocessing and Signal Conditioning

Each audio file is loaded as a mono waveform sampled at 16 kHz. To reduce variability caused by recording gain, peak normalization is applied:

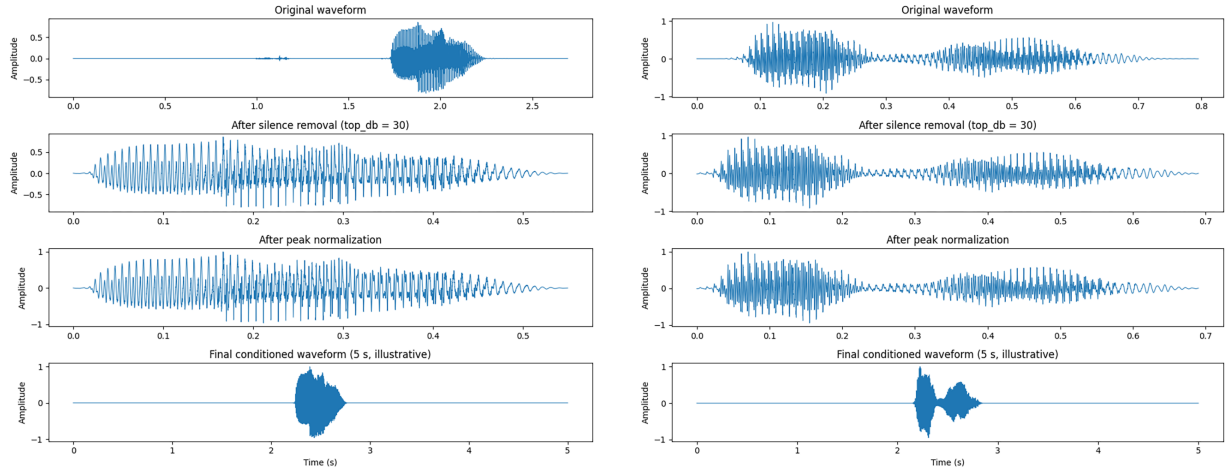
$$\tilde{x}(n) = \frac{x(n)}{\max(|x(n)|) + \varepsilon} \tag{1}$$

where $x(n)$ is the original waveform sample, and ε is a small constant for numerical stability.

To suppress non-informative low-energy regions, silence removal is applied using an energy-based segmentation approach. Non-silent intervals are detected relative to a reference level using a decibel threshold of 30 dB and concatenated while discarding very short segments. This step focuses the analysis on voiced and acoustically relevant regions and reduces unnecessary padding during subsequent processing.

During training, random temporal cropping is applied by extracting a fixed-duration segment of 5 s from each waveform with a predefined probability. If the audio signal is shorter than the target duration, zero-padding is applied to ensure a consistent input length. This strategy introduces temporal variability while maintaining a fixed-size representation required by the neural network.

Figs. 1 and 2 illustrate the effect of the proposed preprocessing pipeline on representative speech signals from the Chronic Pain and TAME Pain datasets. It can be observed that silence removal eliminates non-informative low-energy regions, while peak normalization standardizes amplitude across recordings. The resulting waveforms are more compact and emphasize voiced segments, which are critical for capturing pain-related acoustic characteristics. Although the TAME Pain dataset exhibits more uniform signal structures due to controlled recording conditions, preprocessing ensures consistent signal representation across both datasets, improving robustness for subsequent feature extraction.



(a) (c)

Figure 1: (Continued)

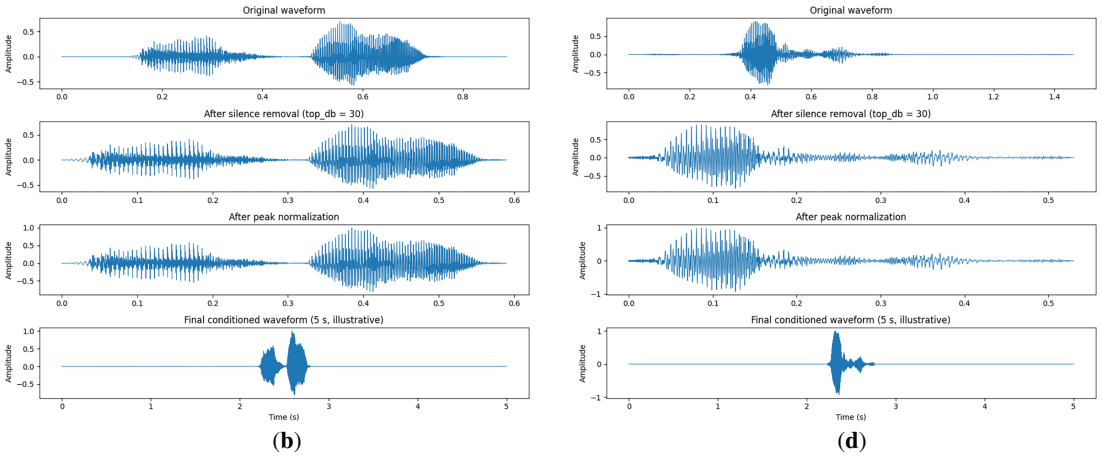


Figure 1: Representative waveform transformations for the Chronic Pain dataset across pain intensity levels: (**a**) None, (**b**) Low, (**c**) Medium, and (**d**) Strong, illustrating the effect of preprocessing, including silence removal and amplitude normalization.

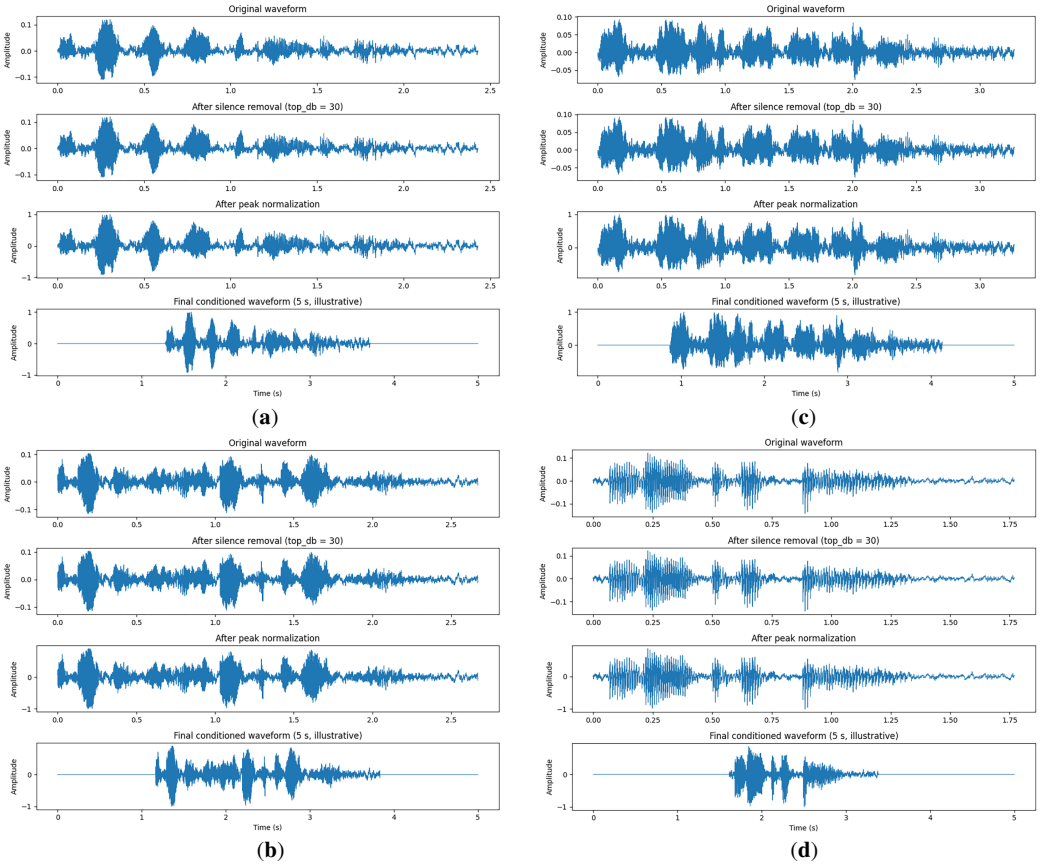


Figure 2: Representative waveform transformations for the TAME Pain Dataset across pain intensity levels: (**a**) None, (**b**) Low, (**c**) Medium, and (**d**) Strong, illustrating the effect of preprocessing, including silence removal and amplitude normalization.

3.3 Time-Frequency Representation

The normalized audio signal is mapped to a time–frequency domain using the short-time Fourier transform:

$$X(k,m) = \sum_{n=0}^{N-1} x(n+mH)w(n)e^{-j2\pi kn/N} \tag{2}$$

where N is the FFT size, H is the hop length, $w(n)$ is a Hann window, k is the frequency bin index, and m is the time frame index.

The power spectrogram is computed as:

$$P(k,m) = \mid X(k,m) \mid^2 \tag{3}$$

To better align with human auditory perception, the power spectrum is projected onto the Mel scale using a Mel filter bank:

$$M(f,m) = \sum_k P(k,m) \cdot W_f(k) \tag{4}$$

where $W_f(k)$ represents the weight of the f-th Mel filter. The Mel scale is widely used in speech processing as it approximates the nonlinear frequency perception of the human auditory system.

The Mel spectrogram is then converted to the logarithmic domain to obtain a log-Mel representation:

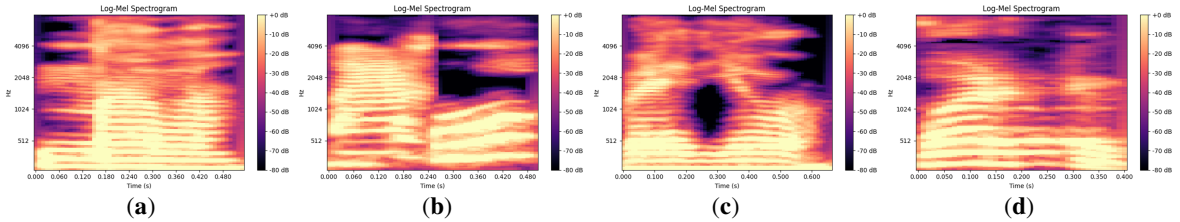
$$M_{\log}(f,m) = \log(M(f,m) + \epsilon) \tag{5}$$

To ensure consistent dynamic range across samples, the log-Mel spectrogram is clipped to a fixed range and normalized:

$$\hat{M}(f,m) = \frac{\text{clip}(M_{\log}(f,m), d_{\min}, d_{\max}) - d_{\min}}{d_{\max} - d_{\min}} \tag{6}$$

where $d_{\min} = -80$ dB and $d_{\max} = 0$ dB define the dynamic range used in this study.

Figs. 3 and 4 present the log-Mel spectrogram representations for different pain intensity levels across the Chronic Pain and TAME Pain datasets. These time–frequency representations reveal how acoustic energy is distributed across frequency bands over time. Differences between classes can be observed in spectral intensity patterns and temporal variations, indicating that pain-related speech introduces distinct acoustic signatures. The Chronic Pain dataset shows higher variability due to real-world recording conditions, whereas the TAME Pain dataset exhibits more structured patterns. This transformation enables the model to capture discriminative features that are not readily observable in the raw waveform.



(**a**) (**b**) (**c**) (**d**)

Figure 3: Log-Mel spectrogram representations of the Chronic Pain dataset for (**a**) None, (**b**) Low, (**c**) Medium, and (**d**) Strong classes, illustrating time–frequency characteristics associated with different pain levels.

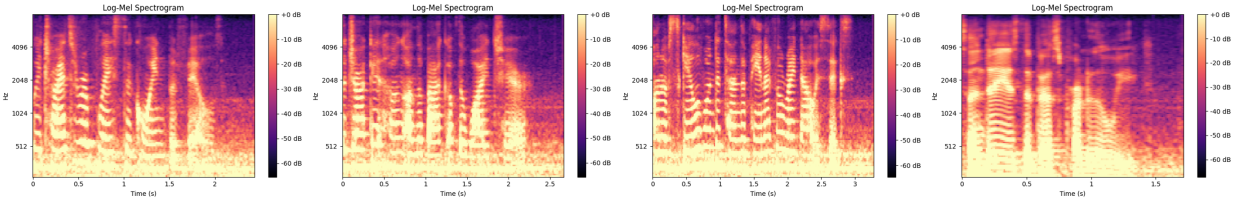


Figure 4: Log-Mel spectrogram representations of the TAME Pain dataset for (**a**) None, (**b**) Low, (**c**) Medium, and (**d**) Strong classes, showing structured spectral patterns under controlled experimental conditions.

3.4 Model Input Preparation

To construct a 3-channel input compatible with the pretrained EfficientNetV2B0 backbone, the normalized log-Mel spectrogram is replicated across three identical channels. Let $\hat{M}(f, m)$ denote the normalized log-Mel spectrogram obtained after dynamic range clipping and min-max scaling. The input tensor is defined as:

$$I(f, m, c) = \hat{M}(f, m), c \in \{1, 2, 3\} \tag{7}$$

where the same spectrogram is assigned to each channel.

This channel replication preserves the spectral structure while satisfying the 3-channel input requirement of the pretrained convolutional network. This strategy also enables direct utilization of pretrained ImageNet weights without requiring training a custom backbone from scratch. The resulting tensor is resized to 224 × 224 and normalized to the range [0, 1] to match the input expectations of the EfficientNetV2B0 model.

3.5 Data Augmentation in the Time-Frequency Domain

To further improve generalization, SpecAugment-style masking is applied during training. This involves randomly masking contiguous regions along the time axis of the log-Mel spectrogram. Time masking improves robustness to temporal variability and partial signal distortions by encouraging the model to rely on distributed temporal cues rather than localized patterns. Frequency masking was intentionally avoided to preserve fine-grained spectral structures that may contain pain-related acoustic cues.

4 DL Framework for Pain Classification

This section presents the proposed DL framework, including the system architecture, model design, training strategy, and inference procedure. The framework is implemented as an end-to-end pipeline that directly maps speech signals to pain intensity levels.

4.1 Overview of the Proposed Framework

The proposed solution consists of an end-to-end DL pipeline for automatic pain level classification from speech signals (see Fig. 5). The pipeline converts each raw audio waveform into a time-frequency representation and performs transfer learning using an EfficientNetV2B0 convolutional backbone. Each speech signal is represented using a log-Mel spectrogram, and the resulting 2D map is replicated into three identical channels to satisfy the 3-channel input requirement of the pretrained image-based model.

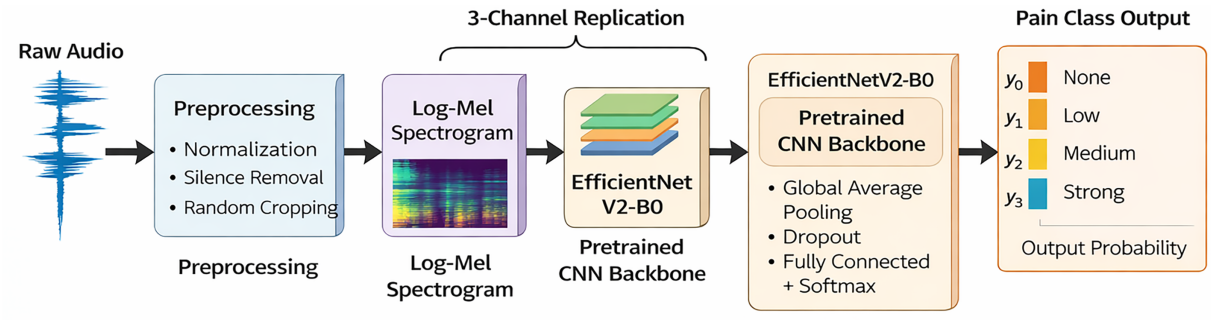


Figure 5: The proposed framework's overall architecture.

Given an input waveform x, the system learns a mapping $f(\cdot)$ that produces a probability distribution over the four pain intensity classes {None, Low, Medium, Strong}

$$\hat{y} = f(x) \tag{8}$$

where $\hat{y}$ is the predicted class distribution over the pain levels.

In addition to the primary four-class formulation, the proposed framework is also evaluated under reduced classification settings, including a three-class scenario {Low, Medium, Strong} and a binary classification scenario {Pain vs. No Pain}, to assess the robustness and generalization of the model across different task granularities. These alternative settings enable a comprehensive assessment of the model's robustness and generalization capability across varying levels of classification granularity. The results for these settings are presented in Section 6.

4.2 DL Architecture

The classifier is based on EfficientNetV2B0, a convolutional neural network optimized for fast training and strong parameter efficiency. The pretrained backbone is used as a feature extractor, followed by global average pooling, dropout, and a fully connected Softmax layer producing class probabilities:

$$\hat{y} = \text{softmax}(W\mathbf{z} + b) \tag{9}$$

where $\mathbf{z}$ is the pooled feature vector.

EfficientNetV2B0 is selected due to its favorable accuracy-to-complexity trade-off and its demonstrated effectiveness when transferring from natural images to spectrogram-based representations. In addition to its performance, the proposed architecture is designed to be lightweight. The EfficientNetV2B0 backbone contains approximately 5.9 million parameters, which is substantially smaller than larger CNN backbones such as ResNet50 and VGG-based models. This compact design significantly reduces computational cost compared to larger architectures such as ResNet50 (~25M parameters) or VGG-based models (~138M parameters), which typically contain tens of millions of parameters. As a result, the proposed framework provides a favorable balance between predictive performance and computational efficiency, supporting future deployment in resource-constrained and edge-based healthcare environments.

4.3 Training Strategy and Optimization

Training is conducted in two stages. In the first stage, the EfficientNetV2B0 backbone is frozen, and only the classification head is trained. This warmup phase stabilizes the newly initialized output layer and allows the model to adapt the final classifier to the pain classification task. In the second stage, all backbone

layers are unfrozen and fine-tuned using a lower learning rate, including the Batch Normalization layers. Updating all layers enables the pretrained network to better adapt from natural-image representations to spectrogram-based inputs and to capture dataset-specific acoustic characteristics more effectively.

To improve optimization stability, gradient clipping is applied during both training stages. In addition, an adaptive learning rate scheduling strategy is used to reduce the learning rate when validation performance plateaus, helping the model converge more reliably and avoid unstable updates during fine-tuning.

To mitigate the impact of class imbalance across datasets, balanced sampling was employed during training to increase exposure to underrepresented classes. Nevertheless, residual bias toward majority classes may still remain, particularly in the merged dataset where class distributions are highly uneven.

4.4 Inference and Test-Time Augmentation

During standard inference, each test sample is processed once using the same preprocessing pipeline without SpecAugment. In addition, a test-time augmentation (TTA) setting is evaluated by extracting five uniformly spaced 5-s crops from each audio signal. For short signals, zero-padding is applied to match the target duration. Each crop is independently processed by the model, and the final prediction is obtained by aggregating crop-level probabilities using log-probability averaging.

$$\log \bar{p}_c = \frac{1}{N} \sum_{i=1}^{N} \log p_{c,i} \tag{10}$$

where $p_{c,i}$ denotes the predicted probability for class c from the i-th crop, and N is the number of crops ($N = 5$). Unless otherwise stated, the reported tabular results correspond to the standard non-TTA evaluation setting.

4.5 Algorithmic Workflow

The complete training and inference workflow is presented in Algorithm 1, which details the full pipeline, including data loading, preprocessing, feature extraction, balanced sampling, model training, and multi-crop fusion during inference.

Algorithm 1: Log-Mel EfficientNetV2B0 training and inference

Input:

Audio dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$; sampling rate SR; Mel parameters (N_{FFT}, H, N_{MELS}); class set $\mathcal{C}$; learning rates α_1, α_2; number of epochs E_1, E_2

initialize EfficientNetV2B0 with ImageNet weights
replace classification head with task-specific softmax layer $|C|$
split dataset $\mathcal{D}$ into training, validation, and test sets
construct a balanced training sampler using uniform class sampling
freeze all backbone layers
for each training epoch $e = 1$ to E_1 **do**
    sample balanced mini-batch
    for each audio sample x_i in mini-batch **do**
        Load waveform at sampling rate SR
        Apply peak amplitude normalization
        Remove silence segments using energy-based splitting (top_db ≈ 30)

(Continued)

Algorithm 1 (continued)

Randomly crop waveform to fixed duration (5 s, with probability $p \approx 0.7$) or pad if needed
Compute STFT and power spectrum
Project spectrum onto Mel filter bank
Convert to log-Mel spectrogram and normalize
Replicate to 3 channels (for CNN input compatibility)
Apply time masking only (SpecAugment)
end for
Forward pass through frozen EfficientNetV2B0 backbone
Compute loss using categorical cross-entropy with label smoothing ($\epsilon \approx 0.05$)
Update classifier head parameters
end for
unfreeze all backbone layers (including Batch Normalization)
Set reduced learning rate α_2
for each training epoch $e = 1$ to E_2 **do**
Repeat steps 7–21 updating all model parameters
end for
for each test sample x_j **do**
Extract multiple fixed-length crops (N-crop, zero-padded if needed)
for each crop **do**
Apply same preprocessing (steps 9–16, no augmentation)
Compute prediction probabilities
end for
Fuse predictions using **log-probability averaging**
Assign final class label
end for
return trained model

This structured workflow ensures reproducibility and provides a clear mapping between the proposed methodology and its implementation.

5 Edge Deployment and Clinical Integration Perspectives

The proposed framework is designed with practical deployment considerations in mind, particularly in terms of computational efficiency, scalability, and compatibility with edge devices. While the primary focus of this study is on model development and evaluation, this section outlines potential pathways for translating the proposed system into real-world healthcare environments.

Unlike multimodal approaches that rely on specialized sensors and controlled acquisition setups, the proposed method operates solely on speech signals, which can be captured using widely available devices such as smartphones and embedded microphones. This characteristic supports potential integration into existing healthcare workflows and mobile platforms. However, it is important to emphasize that the current system represents an early-stage prototype, and further validation using larger, more diverse, and clinically representative datasets is required before real-world clinical deployment.

5.1 Edge Deployment and Mobile Implementation

To assess deployment feasibility, the trained model was exported and integrated into a mobile-compatible inference pipeline. The model, originally developed using TensorFlow/Keras, was saved in standard formats and subsequently converted into TensorFlow Lite (TFLite), which is optimized for efficient execution on resource-constrained devices.

Two deployment variants were generated to balance accuracy and efficiency. The full-precision (FP32) TFLite model has a size of approximately 22.3 MB, while a dynamically quantized version reduces the model size to approximately 6.3 MB, corresponding to nearly a 3.5-fold reduction in storage requirements. These compact representations are consistent with the relatively small number of model parameters (approximately 5.9 million), supporting the suitability of the proposed architecture for edge deployment scenarios.

An Android-based prototype application was developed to enable real-time inference (see Fig. 6). The prototype application captures audio input directly from the device microphone and applies the same preprocessing pipeline described in Section 3, including waveform normalization, silence handling, and log-Mel spectrogram generation. The resulting time-frequency representation is then passed to the embedded TFLite model for prediction.

(**a**) (**b**)

Figure 6: Android-based prototype interface for on-device pain classification using the deployed TensorFlow Lite model, showing (**a**) a No Pain prediction example and (**b**) a Pain prediction example with corresponding confidence scores and class probabilities.

Inference is performed entirely on-device using the TensorFlow Lite interpreter, without reliance on external servers. This design reduces latency by eliminating network communication overhead and enhances data privacy, as raw audio signals remain on the device. It also enables operation in environments with limited or unstable connectivity, which is particularly relevant for remote healthcare applications.

From a systems perspective, the deployment pipeline preserves consistency with the training configuration, ensuring that the same feature representation and input format are used during inference. Additional evaluation confirms that the converted TFLite models produce predictions consistent with the original Keras model under identical test conditions.

It is important to note that the current implementation serves as a proof-of-concept rather than a finalized system. Further optimization is required to reduce preprocessing overhead, improve robustness under real-world acoustic conditions, and evaluate performance across diverse mobile hardware platforms.

5.2 Clinical Integration Scenarios

The proposed framework can support a range of healthcare applications where non-invasive and objective pain assessment is beneficial. These scenarios are presented as prospective use cases intended to guide future system development and validation (see Fig. 7).

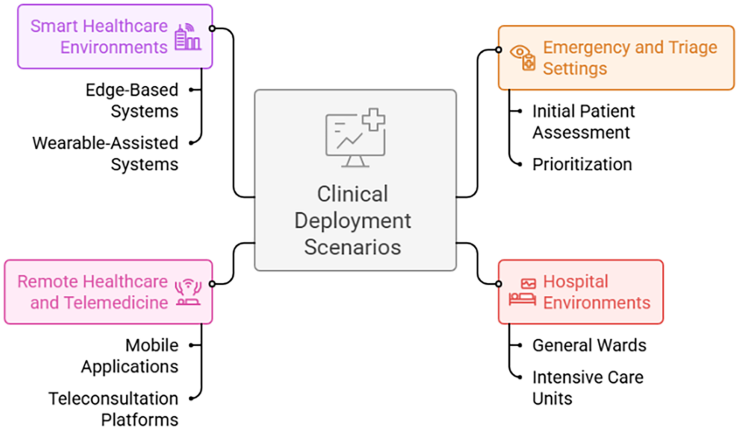


Figure 7: Clinical deployment scenarios of the proposed framework.

In emergency and triage settings, speech-based pain analysis may provide supplementary information to support rapid initial assessment, particularly when patients are unable to communicate effectively. The system could assist clinicians by offering an additional signal for prioritization, although it is not intended to replace clinical judgment.

In hospital environments, including general wards and intensive care units, the framework may serve as an experimental supportive monitoring tool. By analyzing patient speech during routine interactions, it could support the tracking of discomfort over time and provide additional context for clinical observation, especially for patients with limited ability to self-report pain.

In remote healthcare and telemedicine applications, the framework could be explored as a supportive speech-analysis component within mobile or teleconsultation platforms. This is particularly relevant for chronic pain management, where continuous monitoring and frequent evaluation are often necessary.

More broadly, the lightweight design and on-device inference capability suggest potential integration into edge-based healthcare systems and mobile health platforms. However, achieving reliable real-world deployment will require further improvements in model generalization, robustness to diverse acoustic environments, and validation across different populations, languages, and clinical conditions.

Overall, these scenarios highlight the potential applicability of the proposed framework while acknowledging that additional data collection, clinical validation, and system refinement are necessary before operational deployment in healthcare settings.

6 Results and Discussion

This section presents the experimental evaluation of the proposed framework under heterogeneous datasets and deployment-oriented conditions, including core classification performance, backbone ablation analysis, cross-dataset generalization, comparison with previous work, and edge deployment evaluation.

6.1 Core Classification Results

The experimental results demonstrate the effectiveness of the proposed framework for speech-based pain classification under multiple datasets and evaluation settings. Table 3 summarizes the overall performance on the Chronic Pain dataset, the TAME Pain dataset, and the merged dataset under four-class (None, Low, Medium, Strong), three-class (Low, Medium, Strong), and binary (Pain/No Pain) classification tasks. Results are reported as mean ± standard deviation over three independent runs under the standard non-TTA evaluation setting.

Table 3: Performance of the proposed method under different datasets and classification settings.

Dataset	Classification Task	Acc. (%)	Pre. (%)	Rec. (%)	F1 (%)
Chronic Pain	4-Class	78.7 ± 9.2	79.5 ± 10.9	78.7 ± 9.2	78.0 ± 9.7
TAME Pain	4-Class	78.3 ± 1.5	77.9 ± 1.7	78.3 ± 1.5	77.8 ± 2.0
Merged	4-Class	77.8 ± 0.7	77.8 ± 1.2	77.8 ± 0.7	77.1 ± 0.8
Chronic Pain	3-Class	76.4 ± 4.8	78.5 ± 3.7	76.4 ± 4.8	76.4 ± 5.2
TAME Pain	3-Class	71.0 ± 0.7	70.8 ± 0.8	71.0 ± 0.7	70.6 ± 0.8
Merged	3-Class	71.7 ± 1.2	72.9 ± 0.7	71.7 ± 1.2	71.2 ± 1.6
Chronic Pain	Binary	82.7 ± 4.6	83.8 ± 4.4	82.7 ± 4.6	83.1 ± 4.5
TAME Pain	Binary	90.3 ± 1.6	90.5 ± 1.4	90.3 ± 1.6	90.3 ± 1.6
Merged	Binary	87.7 ± 0.1	87.8 ± 0.1	87.7 ± 0.1	87.7 ± 0.1

For the four-class classification task, the model achieves 78.7% ± 9.2 on the Chronic Pain dataset, 78.3% ± 1.5 on the TAME Pain dataset, and 77.8% ± 0.7 on the merged dataset. Unlike typical expectations of performance degradation under domain heterogeneity, the merged dataset achieves performance comparable to the individual datasets, suggesting that the model can maintain stable performance under heterogeneous multi-source training conditions. This suggests that increased data diversity may improve robustness, although some dataset-specific or language-dependent patterns may still influence model behavior. However, the relatively high standard deviation observed for the Chronic Pain dataset (±9.2) indicates that performance remains less stable under this setting. This variability is likely related to the limited dataset size and the higher heterogeneity of the real-world clinical recordings; therefore, the corresponding results should be interpreted as preliminary and exploratory, requiring further validation on larger clinical datasets.

For the three-class classification task, the model achieves 76.4% ± 4.8 on the Chronic Pain dataset, 71.0% ± 0.7 on the TAME Pain dataset, and 71.7% ± 1.2 on the merged dataset. In this case, the Chronic dataset achieves the highest performance, suggesting that removing the "None" class reduces ambiguity and allows the model to better distinguish between pain intensity levels in real-world clinical recordings.

For the binary classification task (Pain vs. No Pain), the model achieves the highest overall performance, reaching 82.7% ± 4.6 on the Chronic dataset, 90.3% ± 1.6 on the TAME Pain dataset, and 87.7% ± 0.1 on

the merged dataset. These results suggest that coarse-grained pain detection is significantly easier than fine-grained intensity classification, as it relies on broader acoustic distinctions rather than subtle variations between adjacent pain levels.

Across all datasets, a consistent pattern emerges where performance improves as the classification task becomes less granular. This behavior indicates that the model captures strong global pain-related features, while finer distinctions between adjacent classes such as Low, Medium, and Strong remain more challenging due to overlapping acoustic characteristics.

To provide deeper insight into class-level performance, Fig. 8 presents the confusion matrices for all datasets under both standard evaluation and TTA settings. For the Chronic Pain dataset, the confusion matrix indicates moderate performance across classes, with relatively balanced predictions but noticeable confusion between adjacent intensity levels. In particular, misclassifications occur between Low and Medium, and between Strong and Medium, reflecting the inherent variability of real-world clinical speech. In contrast, the TAME Pain dataset shows strong performance, particularly for the None class, which exhibits clear separability and high accuracy. However, confusion persists between Low and Medium classes, indicating that even in controlled conditions, subtle differences in pain intensity remain difficult to distinguish.

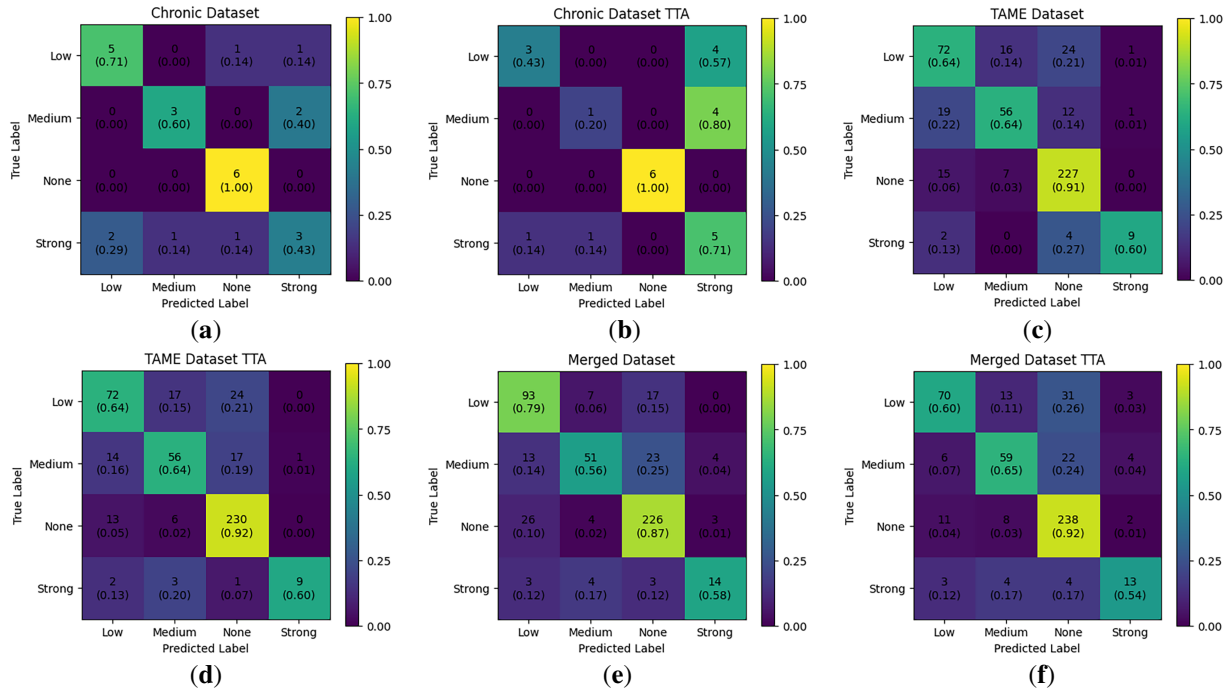


Figure 8: Confusion matrices across datasets: (**a**) Chronic, (**b**) Chronic with TTA, (**c**) TAME, (**d**) TAME with TTA, (**e**) Merged dataset, and (**f**) Merged dataset with TTA.

For the merged dataset, the confusion matrix demonstrates improved overall structure compared to the Chronic dataset, but with increased dispersion relative to TAME. Misclassifications are primarily concentrated between adjacent classes, especially Low versus None and Medium versus Low, reflecting the impact of combining heterogeneous data sources. This behavior highlights the challenge of maintaining consistent decision boundaries across different domains.

Fig. 9 illustrates the per-class accuracy across datasets, providing a complementary view to the confusion matrix analysis. The figure shows that the None class consistently achieves the highest accuracy across all datasets, reflecting strong separability of neutral speech patterns. The Medium class demonstrates relatively stable performance, acting as an intermediate category with moderate discriminability. In contrast, the Low and Strong classes exhibit lower accuracy and greater variability across datasets, highlighting the difficulty of distinguishing subtle and extreme pain expressions.

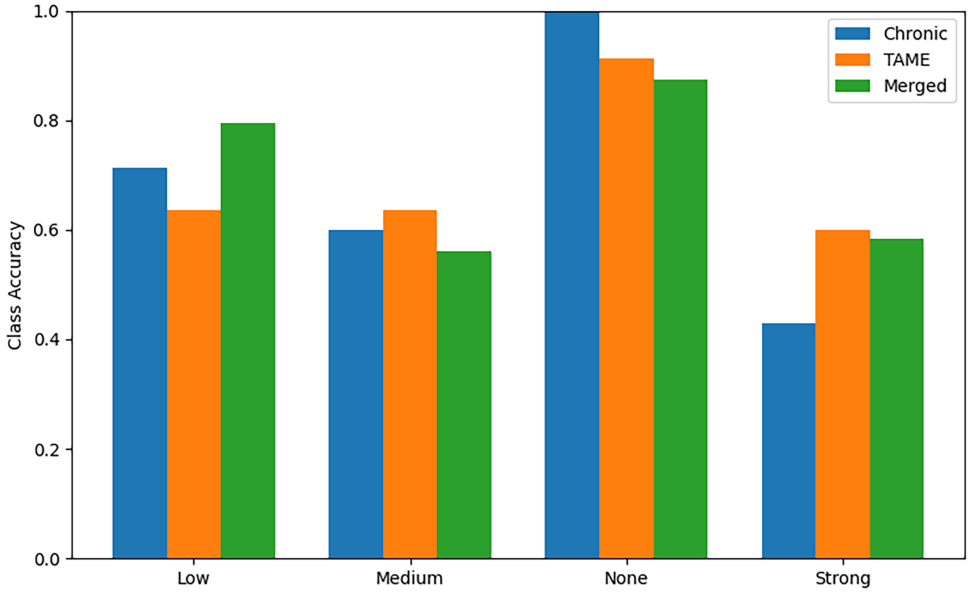


Figure 9: Per-Class accuracy comparison across experiments.

6.2 Backbone Ablation Study

To evaluate the effectiveness of the selected EfficientNetV2B0 backbone, additional backbone comparison experiments were conducted using MobileNetV2, EfficientNetB0, and ResNet50 under the same preprocessing pipeline, training configuration, and four-class classification setting. Table 4 summarizes the obtained results under the standard non-TTA evaluation setting. Additional internal ablation experiments were also conducted during development to evaluate preprocessing, augmentation, sampling, fine-tuning, and inference configurations; however, these experiments are not reported in detail due to manuscript space limitations.

The results indicate that EfficientNetV2B0 provides the most favorable balance between classification performance and computational efficiency across the evaluated datasets. Although ResNet50 achieves slightly higher accuracy on the TAME Pain and merged datasets, EfficientNetV2B0 consistently attains stronger overall F1 scores while requiring substantially fewer parameters (≈5.9M vs. ≈25.6M). In contrast, MobileNetV2 (≈3.5M parameters) demonstrates lower robustness, particularly on the Chronic Pain dataset, suggesting limited representational capacity under highly heterogeneous clinical conditions. EfficientNetB0 (≈5.3M parameters) achieves moderate performance but remains less stable than EfficientNetV2B0 across datasets. These findings support the selection of EfficientNetV2B0 as an effective accuracy-efficiency trade-off for lightweight and deployment-oriented speech-based pain classification.

Table 4: Backbone ablation study under the four-class classification setting.

Dataset	Model	Acc. (%)	Pre. (%)	Rec. (%)	F1 (%)
Chronic Pain	EfficientNetV2B0	78.7 ± 9.2	79.5 ± 10.9	78.7 ± 9.2	78.0 ± 9.7
	MobileNetV2	33.3 ± 4.6	24.5 ± 10.9	32.9 ± 3.8	23.9 ± 7.1
	EfficientNetB0	54.7 ± 15.1	62.4 ± 24.9	54.4 ± 15.0	51.5 ± 19.4
	ResNet50	73.3 ± 8.3	71.5 ± 15.7	71.2 ± 9.6	67.2 ± 10.5
TAME Pain	EfficientNetV2B0	78.3 ± 1.5	77.9 ± 1.7	78.3 ± 1.5	77.8 ± 2.0
	MobileNetV2	71.8 ± 2.0	68.4 ± 7.4	57.4 ± 5.3	60.2 ± 4.8
	EfficientNetB0	72.7 ± 2.1	64.6 ± 2.3	59.4 ± 2.8	61.0 ± 2.2
	ResNet50	78.5 ± 1.5	76.2 ± 8.4	69.0 ± 5.1	70.9 ± 2.4
Merged	EfficientNetV2B0	77.8 ± 0.7	77.8 ± 1.2	77.8 ± 0.7	77.1 ± 0.8
	MobileNetV2	71.5 ± 1.2	68.6 ± 4.9	59.8 ± 2.1	62.4 ± 2.5
	EfficientNetB0	72.6 ± 2.9	67.0 ± 3.9	66.7 ± 1.9	66.6 ± 3.0
	ResNet50	79.2 ± 1.5	75.9 ± 6.7	69.6 ± 2.4	71.1 ± 1.3

6.3 Cross-Dataset Generalization

To further evaluate the robustness of the proposed framework under heterogeneous conditions, additional cross-dataset experiments were conducted using train-on-one, test-on-another evaluation settings. Unlike the merged-dataset experiments presented previously, these experiments assess the ability of the model to generalize across substantially different domains without exposure to the target dataset during training. Table 5 summarizes the obtained results for binary, three-class, and four-class classification tasks under the standard non-TTA evaluation setting.

Table 5: Cross-dataset evaluation.

Classification Task	Direction	Acc. (%)	Pre. (%)	Rec. (%)	F1 (%)
2-class	TAME → Chronic	69.4 ± 3.3	57.5 ± 0.2	60.1 ± 1.3	57.5 ± 0.7
2-class	Chronic → TAME	45.0 ± 0.0	22.5 ± 0.0	50.0 ± 0.0	31.0 ± 0.0
3-class	TAME → Chronic	34.8 ± 2.4	28.9 ± 3.3	36.7 ± 2.3	29.5 ± 1.9
3-class	Chronic → TAME	36.5 ± 6.6	33.3 ± 1.4	32.1 ± 1.5	27.4 ± 1.9
4-class	TAME → Chronic	26.1 ± 0.8	21.8 ± 2.9	27.9 ± 2.5	19.0 ± 4.0
4-class	Chronic → TAME	20.7 ± 1.6	24.8 ± 0.9	24.8 ± 1.8	16.8 ± 2.4

The results indicate that cross-dataset generalization is substantially more challenging than within-dataset or merged-dataset evaluation due to the considerable domain mismatch between the datasets, including differences in language, recording conditions, speaker populations, and pain elicitation protocols. Nevertheless, the binary classification setting demonstrates relatively stronger transferability, particularly for the TAME → Chronic direction, suggesting that coarse-grained pain-related acoustic patterns are more transferable across domains than fine-grained pain intensity distinctions. In contrast, the three-class and four-class settings exhibit substantially lower performance, highlighting the difficulty of learning domain-invariant representations for subtle pain intensity differences.

6.4 Comparison with Previous Work

Table 6 compares the proposed framework with representative prior studies in speech-based pain classification. Although direct comparison across studies should be interpreted cautiously due to differences in datasets, recording conditions, label definitions, preprocessing strategies, and evaluation protocols, the proposed approach demonstrates competitive performance across multiple classification settings. In particular, the framework achieves strong performance on the Chronic Pain dataset, which contains real-world clinical recordings with higher acoustic variability than many controlled experimental datasets.

Table 6: Comparison with previous work on speech-based pain classification.

Ref.	Dataset	Task	Accuracy (%) (non-TTA)
Oshrat et al. [20]	Own Dataset	Binary (Pain/No Pain)	77.25
Alhudhaif [21]	TAME Pain	Binary (Pain/No Pain)	83.86
Alhudhaif [21]	TAME Pain	3-Class (Mild, Moderate, Severe)	75.36
Lu and Lu [28]	TAME + VIVAE (non-vocal)	3-Class (Mild, Moderate, Severe)	72.74
Tsai et al. [24]	Own Dataset	Binary (Pain/No Pain)	72.3
Tsai et al. [24]	Own Dataset	3-Class (Mild, Moderate, Severe)	54.2
Hernández-de-la-Cruz et al. [31]	Chronic Pain	4-Class (None, Low, Medium, Strong)	23.1

Compared with prior studies that focus on single datasets or controlled laboratory conditions, the proposed framework is evaluated under more heterogeneous conditions using both clinical and experimentally induced pain recordings. The results suggest that the combination of log-Mel spectrogram representations, transfer learning, and lightweight end-to-end optimization provides an effective balance between predictive performance and computational efficiency. In addition, the deployment-oriented design and TensorFlow Lite evaluation support the practical feasibility of the proposed framework for future edge-based healthcare applications.

For the TAME three-class classification task, the proposed framework achieves 71.0% accuracy, which is lower than the 75.36% reported by Alhudhaif [21]. This performance gap may be attributed to differences in model design and optimization specifically tailored to the TAME dataset. In contrast, the proposed framework was designed to maintain consistent performance across heterogeneous datasets and multiple classification settings while preserving lightweight deployment capability. Despite the lower TAME three-class result, the framework provides a more comprehensive evaluation across heterogeneous datasets and classification settings, offering a broader assessment of performance under diverse real-world conditions.

From a clinical perspective, the results indicate that the model appears more effective in coarse-grained pain assessment settings, especially in binary pain detection where performance reaches up to 90.3% on TAME and 87.7% on the merged dataset. This level of performance suggests potential feasibility for future supportive healthcare applications, although further clinical validation is required before practical deployment. However, fine-grained classification remains more challenging, reflecting the inherent subjectivity of pain perception and the overlap between adjacent intensity levels, which is also observed in traditional clinical scales such as the Visual Analog Scale (VAS) and Numeric Rating Scale (NRS) [40,41]. In practice, even human assessments exhibit variability across observers and contexts. Importantly, most misclassifications occur between adjacent classes, such as Low and Medium, rather than between extreme categories like None and Strong, indicating that the model largely preserves clinically meaningful distinctions. Accordingly, the proposed system is best viewed as a supportive tool that complements, rather than replaces, clinical judgment.

In summary, these results emphasize that evaluating models under heterogeneous conditions is essential for developing reliable and deployable speech-based pain assessment systems, and demonstrate that the proposed framework provides a practical balance between accuracy, generalization, and computational efficiency.

6.5 Edge Deployment Evaluation

To evaluate the practical deployment feasibility of the proposed framework, additional experiments were conducted to assess model conversion, inference efficiency, and on-device execution performance. The evaluation was performed using the binary classification model (Pain vs. No Pain) trained on the merged dataset, as this setting represents the primary deployment scenario considered in this work. The trained Keras model was converted into TensorFlow Lite (TFLite) format using both full-precision (FP32) and dynamic range quantization (Dynamic) to investigate the trade-off between model efficiency and predictive performance. The analysis includes both TensorFlow Lite conversion experiments and device-level benchmarking on Android smartphones.

As shown in Table 7, the FP32 model preserves the original classification accuracy (87.6%), while the Dynamic model achieves 87.4%, corresponding to only a 0.2 percentage-point reduction. At the same time, model size decreases from 68.4 to 22.3 MB for FP32 and 6.3 MB for the Dynamic version. Inference time is also substantially reduced, with the FP32 and Dynamic models achieving approximately 3× and 4× speedups, respectively, compared with the original Keras implementation. These results indicate that TensorFlow Lite conversion can significantly improve computational efficiency while maintaining predictive performance.

Table 7: Edge deployment evaluation results (binary classification, merged dataset).

Model Type	Model Size	Test Accuracy (%)	Avg. Inference Time/Sample (s)	Agreement with Keras
Original	68.4 MB	87.6	0.0747	—
FP32	22.3 MB	87.6	0.0239	1.000
Dynamic	6.3 MB	87.4	0.0170	0.976

To further evaluate deployment performance on real mobile hardware, both TFLite models were benchmarked on four Android smartphones with different hardware configurations. As reported in Table 8, dynamic quantization consistently reduces both latency and memory consumption across all tested devices. For the Dynamic model, end-to-end latency ranges from 0.075 to 0.202 s, while peak memory consumption ranges from 7.5 to 31.5 MB. Although execution speed varies across hardware platforms, all devices are capable of real-time inference, demonstrating the practicality of the proposed framework for edge deployment.

The deployment results demonstrate that the proposed framework can maintain predictive performance while achieving substantial reductions in model size, latency, and memory usage. These findings support the feasibility of real-time on-device pain classification using commercially available smartphones. Future work will further investigate device-level characteristics, including CPU utilization, energy consumption, and thermal behavior during prolonged operation.

Table 8: Device-level android benchmark results on different smartphone platforms.

Device	SoC (Processor)	RAM (GB)	Model Type	Avg. Inference Time/Sample (s)	End-to-End Latency (s)	Peak Memory (MB)
Motorola Razr 50	MediaTek Dimensity 7300X	7.2	FP32	0.172 ± 0.075	0.311 ± 0.148	46.1
			Dynamic	0.084 ± 0.002	0.202 ± 0.003	27.5
HONOR VER-N49	Qualcomm Snapdragon 8 Gen 3	14.9	FP32	0.058 ± 0.013	0.139 ± 0.019	72.6
			Dynamic	0.034 ± 0.001	0.105 ± 0.001	13.2
Samsung SM-F956B	Qualcomm Snapdragon 8 Gen 3 for Galaxy	10.9	FP32	0.073 ± 0.012	0.166 ± 0.028	92.2
			Dynamic	0.052 ± 0.003	0.155 ± 0.005	31.5
Huawei VOG-L29	HiSilicon Kirin 980	7.4	FP32	0.072 ± 0.002	0.103 ± 0.003	81.6
			Dynamic	0.044 ± 0.001	0.075 ± 0.002	7.5

7 Conclusion and Future Work

This paper presented a lightweight, end-to-end DL framework for automatic pain level classification from speech signals. The proposed approach employs log-Mel spectrogram representations and a pretrained EfficientNetV2B0 backbone to learn discriminative time–frequency features without relying on hand-crafted feature engineering or multi-stage processing pipelines. In addition, the study explicitly addresses multi-dataset robustness assessment and extends the analysis to edge deployment scenarios, providing a comprehensive assessment of both model effectiveness and practical feasibility.

Experimental results demonstrate competitive performance across heterogeneous datasets, with the strongest results observed in reduced-class settings. For the four-class classification task, the model reaches 78.7 ± 9.2% accuracy on the Chronic dataset, 78.3 ± 1.5% on the TAME dataset, and 77.8 ± 0.7% on the merged dataset. Performance further improves in reduced-classification settings, achieving up to 87.7 ± 0.1% accuracy for binary pain detection on the merged dataset. These results indicate that the model can capture pain-related acoustic patterns while highlighting the increased difficulty of fine-grained pain-intensity classification. Furthermore, the cross-dataset experiments demonstrate that generalization across different languages, recording conditions, and pain elicitation protocols remains challenging, particularly for multiclass classification tasks, emphasizing the importance of heterogeneous evaluation settings when developing speech-based pain assessment systems.

Beyond standard evaluation, the edge deployment analysis confirms that the proposed model can be effectively deployed on resource-constrained devices. The TensorFlow Lite models maintain predictive consistency with the original model while significantly reducing model size and inference time, achieving up to a fourfold speedup with minimal accuracy loss. This experiment supports the feasibility of future on-device, privacy-preserving healthcare applications and provides a foundation for future integration into edge-based healthcare systems. However, additional real-world and clinical validation is required before practical deployment.

From a clinical perspective, the proposed system shows promise as a supportive tool for pain assessment. However, the current study remains research-oriented and does not constitute clinical validation. Its strongest performance in binary classification suggests potential feasibility for future supportive applications in screening, triage-assistance, and continuous monitoring scenarios, particularly in situations where self-reporting may be limited or unreliable. At the same time, the difficulty of distinguishing adjacent pain levels

reflects the inherent subjectivity of pain and the variability of pain expression across individuals. Therefore, the system should be viewed as a complementary tool rather than a replacement for clinical judgment.

Despite these promising results, several directions remain for future work. Further improvements are needed to enhance generalization across diverse acoustic conditions and to optimize preprocessing efficiency for real-time deployment on a wider range of hardware platforms. In particular, future research will focus on collecting larger, more diverse, and clinically representative speech datasets, incorporating variability in speakers, languages, and recording environments. Additional directions include investigating domain adaptation techniques to improve cross-dataset generalization, leveraging self-supervised learning to better utilize unlabeled data, and exploring multimodal extensions when complementary signals are available.

Acknowledgement: Not applicable.

Funding Statement: The author received no specific funding for this study.

Availability of Data and Materials: The datasets used in this study are publicly available. The Chronic Pain Audio Dataset is available at https://doi.org/10.17632/TDKB5RPSCH.1, and the TAME Pain dataset is available at https://doi.org/10.1038/s41597-025-04733-2.

Ethics Approval: Not applicable.

Conflicts of Interest: The author declares no conflicts of interest.

References

1. Freitas AT. Data-driven approaches in healthcare: challenges and emerging trends. Law Gov Technol Ser. 2024;58:65–80. doi:10.1007/978-3-031-41264-6_4.
2. Ahmadi A, RabieNezhad Ganji N. AI-driven medical innovations: transforming healthcare through data intelligence. Int J BioLife Sci. 2023;2(2):132–42. doi:10.22034/IJBLS.2023.185475.
3. Desai A, Panda P, Dash AP, Panchal SD. AI-driven diagnostics: bridging medicine, data science, and clinical decision support. Int J Adv Signal Image Sci. 2025;11(6s):297–308. doi:10.29284/rsd1j059.
4. Samuel OJ. AI-powered decision support systems in traditional medicine. Joster. 2025;2(2):66–86. doi:10.64206/dncatz20.
5. Alowais SA, Alghamdi SS, Alsuhebany N, Alqahtani T, Alshaya AI, Almohareb SN, et al. Revolutionizing healthcare: the role of artificial intelligence in clinical practice. BMC Med Educ. 2023;23(1):689. doi:10.1186/s12909-023-04698-z.
6. Abbasi N, Fnu N, Zeb S. AI in healthcare: integrating advanced technologies with traditional practices for enhanced patient care. J Multidisiplin Ilmu. 2023;2(3):546–56.
7. Latif S, Qadir J, Qayyum A, Usama M, Younis S. Speech technology for healthcare: opportunities, challenges, and state of the art. IEEE Rev Biomed Eng. 2021;14:342–56. doi:10.1109/rbme.2020.3006860.
8. Xu X, Li J, Zhu Z, Zhao L, Wang H, Song C, et al. A comprehensive review on synergy of multi-modal data and AI technologies in medical diagnosis. Bioengineering. 2024;11(3):219. doi:10.3390/bioengineering11030219.
9. Jibb L, Stinson J. Pain assessment. In: managing pain in children and young people: a clinical guide. Treasure Island, FL, USA: StatPearls Publishing; 2024. p. 73–93.
10. Ben Aoun N. A review of automatic pain assessment from facial information using machine learning. Technologies. 2024;12(6):92. doi:10.3390/technologies12060092.
11. Li J, Luo J, Wang Y, Jiang Y, Chen X, Quan Y. Automatic pain assessment based on physiological signals: application of multi-scale networks and cross-attention cross-attention. In: Proceedings of the 2024 13th International Conference on Bioinformatics and Biomedical Science; 2024 Oct 18–20; Hong Kong, China. p. 113–22. doi:10.1145/3704198.3704212.

12. Posada-Quintero HF, Kong Y, Chon KH. Objective pain stimulation intensity and pain sensation assessment using machine learning classification and regression based on electrodermal activity. Am J Physiol Regul Integr Comp Physiol. 2021;321(2):R186–96. doi:10.1152/ajpregu.00094.2021.
13. Gkikas S, Chatzaki C, Tsiknakis M. Multi-task neural networks for pain intensity estimation using electrocardiogram and Demographic factors. Commun Comput Inf Sci. 2023;1856:324–37. doi:10.1007/978-3-031-37496-8_17.
14. Phan KN, Iyortsuun NK, Pant S, Yang HJ, Kim SH. Pain recognition with physiological signals using multi-level context information. IEEE Access. 2023;11(1):20114–27. doi:10.1109/access.2023.3248654.
15. Tan CW, Du T, Teo JC, Chan DXH, Kong WM, Sng BL. Automated pain detection using facial expression in adult patients with a customized spatial temporal attention long short-term memory (STA-LSTM) network. Sci Rep. 2025;15(1):13429. doi:10.1038/s41598-025-97885-5.
16. Semwal A, Londhe ND. ECCNet: an ensemble of compact convolution neural network for pain severity assessment from face images. In: Proceedings of the 2021 11th International Conference on Cloud Computing, Data Science & Engineering (Confluence); 2021 Jan 28–29; Noida, India. p. 761–6. doi:10.1109/confluence51648.2021.9377197.
17. Alghamdi T, Alaghband G. Facial expressions based automatic pain assessment system. Appl Sci. 2022;12(13):6423. doi:10.3390/app12136423.
18. Semwal A, Londhe ND. A shallow convolutional neural network for pain severity assessment in uncontrolled environment. In: Proceedings of the 2021 IEEE 11th Annual Computing and Communication Workshop and Conference (CCWC); 2021 Jan 27–30. Virtual. p. 800–6. doi:10.1109/ccwc51732.2021.9376052.
19. Karamitsos I, Seladji I, Modak S. A modified CNN network for automatic pain identification using facial expressions. J Softw Eng Appl. 2021;14(8):400–17. doi:10.4236/jsea.2021.148024.
20. Oshrat Y, Bloch A, Lerner A, Cohen A, Avigal M, Zeilig G. Speech prosody as a biosignal for physical pain detection. Proc Int Conf Speech Prosody. 2016;2016:420–4. doi:10.21437/speechprosody.2016-86.
21. Alhudhaif A. Pain level classification from speech using GRU-mixer architecture with log-mel spectrogram features. Diagnostics. 2025;15(18):2362. doi:10.3390/diagnostics15182362.
22. Szczapa B, Daoudi M, Berretti S, Pala P, Del Bimbo A, Hammal Z. Automatic estimation of self-reported pain by trajectory analysis in the manifold of fixed rank positive semi-definite matrices. IEEE Trans Affect Comput. 2022;13(4):1813–26. doi:10.1109/taffc.2022.3207001.
23. Yang S, Luo W, Yang T, Chen X, Shen S, Wang L, et al. Automatic pain classification in older patients with hip fracture based on multimodal information fusion. Sci Rep. 2025;15(1):21562. doi:10.1038/s41598-025-09046-3.
24. Tsai FS, Weng YM, Ng CJ, Lee CC. Embedding stacked bottleneck vocal features in a LSTM architecture for automatic pain level classification during emergency triage. In: Proceedings of the 2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII); 2017 Oct 23–26; San Antonio, TX, USA. p. 313–8. doi:10.1109/acii.2017.8273618.
25. Thiam P, Kessler V, Amirian M, Bellmann P, Layher G, Zhang Y, et al. Multi-modal pain intensity recognition based on the *SenseEmotion* database. IEEE Trans Affect Comput. 2021;12(3):743–60. doi:10.1109/taffc.2019.2892090.
26. Jha T, Kavya R, Christopher J, Arunachalam V. Machine learning techniques for speech emotion recognition using paralinguistic acoustic features. Int J Speech Technol. 2022;25(3):707–25. doi:10.1007/s10772-022-09985-6.
27. Anagnostopoulos CN, Iliou T, Giannoukos I. Features and classifiers for emotion recognition from speech: a survey from 2000 to 2011. Artif Intell Rev. 2015;43(2):155–77. doi:10.1007/s10462-012-9368-5.
28. Lu AY, Lu W. Voice-based pain level classification for sensor-assisted intelligent care. Sensors. 2026;26(3):892. doi:10.3390/s26030892.
29. TensorFlow Lite|ML for Mobile and Edge Devices. [cited 2020 Oct 12]. Available from: https://www.tensorflow.org/lite.
30. Velana M, Gruss S, Layher G, Thiam P, Zhang Y, Schork D, et al. The SenseEmotion database: A multimodal database for the development and systematic validation of an automatic pain- and emotion-recognition system. In: Multimodal pattern recognition of social signals in human-computer-interaction. Cham, Switzerland: Springer; 2017. p. 127–39. doi:10.1007/978-3-319-59259-6_11.

31. Hernández-de-la-Cruz M, Hernández-Hernández JL, Zavala-Hurtado M, Evangelista-Alcocer Y, Barrera SRZ, Peña MV. Pilot study of a multitasking framework for the simultaneous vocal assessment of chronic pain, presbyphonia, and gender: a feasibility analysis of 256 clinical samples. In: Technologies and innovation. Cham, Switzerland: Springer; 2026. p. 208–21. doi:10.1007/978-3-032-11494-5_14.
32. Aung MSH, Kaltwang S, Romera-Paredes B, Martinez B, Singh A, Cella M, et al. The automatic detection of chronic pain-related expression: requirements, challenges and the multimodal EmoPain dataset. IEEE Trans Affect Comput. 2016;7(4):435–51. doi:10.1109/taffc.2015.2462830.
33. Ricossa D, Baccaglini E, Di Nardo E, Parodi E, Scopigno R. On the automatic audio analysis and classification of cry for infant pain assessment. Int J Speech Technol. 2019;22(1):259–69. doi:10.1007/s10772-019-09601-0.
34. Gkikas S, Rojas RF, Tsiknakis M. PainFormer: a vision foundation model for automatic pain assessment. IEEE Trans Affect Comput. 2025;16(4):3369–86. doi:10.1109/taffc.2025.3605475.
35. Bonafos G, Rouch J, Lego L, Reby D, Patural H, Mathevon N, et al. Speech transformer models for extracting information from baby cries. In: Proceedings of the Workshop on Child Computer Interaction—WOCCI 2025; 2025 Aug 22–24; Nijmegen, The Netherlands. p. 11–5. doi:10.21437/wocci.2025-3.
36. Yan J, Sun W, Sun B, Zhou X, Lu G, Li X, et al. Neonatal pain speech emotion recognition based on multiscale multilevel MAE network. IEEE Trans Comput Soc Syst. 2026;13(1):37–50. doi:10.1109/tcss.2025.3625592.
37. Zarghami Y, Rad MG, Mafeld S, Taati B, Conway A. Computer vision for pain detection during procedural sedation. Sci Rep. 2026;16(1):16776. doi:10.1038/s41598-026-45130-y.
38. de la Cruz MH, Ramirez Arcos MI, Morales Morales C, Evangelista Alcocer Y, Zavala Hurtado M, Valencia Díaz EF. Chronic pain audio dataset. Mendeley Data. 2024. doi:10.17632/TDKB5RPSCH.1.
39. Dao TQ, Schneiders E, Williams J, Bautista JR, Seabrooke T, Vigneswaran G, et al. TAME Pain data release: using audio signals to characterize pain. Sci Data. 2025;12(1):595. doi:10.1038/s41597-025-04733-2.
40. Raja SN, Carr DB, Cohen M, Finnerup NB, Flor H, Gibson S, et al. The revised international association for the study of pain definition of pain: concepts, challenges, and compromises. Pain. 2020;161(9):1976–82. doi:10.1097/j.pain.0000000000001939.
41. Hjermstad MJ, Fayers PM, Haugen DF, Caraceni A, Hanks GW, Loge JH, et al. Studies comparing numerical rating scales, verbal rating scales, and visual analogue scales for assessment of pain intensity in adults: a systematic literature review. J Pain Symptom Manag. 2011;41(6):1073–93. doi:10.1016/j.jpainsymman.2010.08.016.