**EURASIAN JOURNAL OF MATHEMATICAL AND COMPUTER APPLICATIONS**
ISSN 2306–6172
Volume 14, Issue 1 (2026) 4 – 16

# SUPERVISED PERSIAN SPEECH EMOTION RECOGNITION USING DISTILHUBERT AND OPTUNA TUNING

**Ahmadianshalchi A., Houshmand M.*, Hosseini S.**

**Abstract** Speech emotion recognition (SER) is recognized as a growing field with important applications in human computer interaction. In this study, attention is given to SER for the Persian language, which is regarded as a relatively underexplored domain. A lightweight architecture is introduced, in which a frozen DistilHuBERT model is used for feature extraction, and a bidirectional gated recurrent unit (Bi-GRU) block with a dense classification head is employed. Optuna is used for hyperparameter optimization, in which the Bi-GRUs structure, learning rate, dropout rate, L2 regularization, and optimizer choice are tuned so that a balance between simplicity and performance is achieved. The proposed model is trained and evaluated with five fold cross validation, and an acceptable accuracy is obtained. These results show that competitive performance is achieved with a compact and optimized model in comparison with larger architectures.

**Keywords:** Speech emotion recognition, Supervised learning, HuBERT, DistilHuBERT, Optuna, language, artificial intelligence, machine learning, big data, data science.

**AMS Mathematics Subject Classification:** 68T07, 68T10, 62H30.

**DOI:** 10.32523/2306-6172-2026-14-1-4-16

## 1 Introduction

Speech recognition (SR) is regarded as an important component of modern human computer interaction, in which interpretation and response to spoken language are enabled. Within SR, one of the most advanced and essential areas is speech emotion recognition (SER), in which identification of human emotions from vocal cues is performed [1]. SER has diverse applications across multiple industries. In customer service, interactive dialogue systems enhanced with SER and chatbots driven by Artificial intelligence (AI) can detect and respond to the emotional state of a customer. With emotional awareness, interactions that resemble human communication are produced. If dissatisfaction is detected, the customer is redirected to a human agent, and emotional information is provided to support more empathetic and effective engagement [2].Another application is observed in mental health monitoring, in which early signs of emotional distress are detected through analysis of speech patterns. Thus, a non invasive instrument for early intervention is provided [3]. In addition, within the context of autonomous vehicles, improvement of the in vehicle experience is achieved with SER through detection and response to driver emotions. For example, if signs of stress or frustration are detected in the driver by the SER model, adjustment of vehicle settings or provision of calming feedback is performed so that the driving experience is improved [4]. In fact, SER is applied in several areas, e.g., lie detection, security monitoring, and emergency response centers. Identification of emotions such as fear, sadness, and happiness is supported, therefore more effective and timely decisions are enabled [5].However, substantial challenges are

[1]Corresponding Author.

encountered in the development of effective SER models when non English languages such as Persian are addressed. Persian has a complex phonological structure, which includes 32 letters and 29 phonemes that are divided into vowels and consonants. In addition, distinctive spelling and syntactic characteristics are present. These properties can reduce the reliability of SER models that are originally trained with English datasets [6]. Differences between Book Persian and Conversational Persian, as well as dialectal variations, introduce additional complexity to the recognition process. These linguistic characteristics require development of specialized SER models that are designed for the phonological and syntactic structures of Persian, in which issues such as recognition of out of vocabulary words and management of dialectal variation are addressed [7].To address these challenges, AI [8]-[13], particularly deep learning (DL), is widely applied. Substantial progress is achieved with DL architectures and algorithms in areas such as computer vision and pattern recognition. As a result, analysis and interpretation of complex data are improved [14]. To further enhance the performance of DL model, particularly in tasks such as SER, especially when working with large and complex datasets, high-quality feature extraction is critical to achieving strong model performance. One widely used method is the Hidden-Unit Bidirectional Encoder Representations from Transformers (HuBERT), a self-supervised model trained on large-scale datasets such as LibriSpeech 960 hours and Libri-Light 60,000 hours. HuBERT generates rich, context-aware representations by predicting hidden units derived from clustered acoustic features. It is available in base (90M parameters), large (300M), and x-large (1B) configurations, making it a strong candidate for downstream tasks like SER [15]. However, HuBERT models are computationally demanding; the base version alone requires around 2,000 GPU hours on 32 GPUs for pre-training, making them impractical for resource-constrained environments [16]. Fig. 1 shows the HuBERT architecture. To overcome the computational demands of HuBERT, DistilHuBERT has been proposed as a lighter alternative created through knowledge distillation. The size of the original model is reduced by 75 and a 73 percent increase in inference speed is achieved, while most of the performance is preserved. In DistilHuBERT, a layer wise distillation strategy is applied, in which knowledge transfer from several internal layers of the teacher network to the student network is performed instead of only the final output layer. With this method, a practical balance between computational efficiency and representational quality is achieved. Therefore, suitability for tasks such as SER is established when resource limitations must be balanced with performance requirements [16]. Fig. 2 illustrates the DistilHuBERT architecture.Despite this efficiency, use of DistilHuBERT as a frozen feature extractor introduces the challenge of optimal hyperparameter selection for the downstream model. To address this issue, Optuna is employed as a modern hyperparameter optimization method, in which this procedure is automated with advanced optimization techniques such as tree-structured Parzen estimator (TPE), which is a form of Bayesian optimization, and covariance matrix adaptation evolution strategy (CMA-ES). In addition, efficient pruning strategies are implemented, e.g., asynchronous successive halving algorithm (ASHA), in which early termination of unpromising trials is allowed [17]. These capabilities support improvement of model performance and reduction of unnecessary computation. To complement this efficient optimization method, a bidirectional gated recurrent units (Bi-GRU) architecture is adopted as the central component of the model. GRU is recognized for an effective balance between modeling capability and computational efficiency. Fewer parameters than long short-term memory (LSTM) are required, while temporal patterns in sequential data are effectively captured [18]. Extending this structure, we employ the bidirectional variant to incorporate both past and future context within the speech signal, an essential consideration in emotion recognition, where emotional cues can occur at any point in

the utterance. Therefore we use DistilHuBERT as a frozen feature extractor, to benefit from its reduced size and faster inference while retaining strong representational power. To capture the temporal dynamics of speech, we incorporate Bi-GRU, followed by a dense layer for emotion classification. The number of Bi-GRU layers and their associated hyperparameters are automatically optimized using Optuna. There is currently limited studies that integrates DistilHuBERT-based feature extraction with Bi-GRU networks, wherein the architecture's depth and other critical hyperparameters are fine-tuned using the Optuna optimization. We introduce what we believe to be the first comprehensive investigation of such a pipeline, specifically designed for the task of Persian SER (PSER). The proposed approach aims to strike a balance between simplicity and performance in low-resource emotional speech analysis by combining a lightweight yet powerful speech representation model with an automatically optimized recurrent backbone. Sec. 2 reviews related work in SER and associated technologies. Sec. 3 details the proposed methodology, including the overall method (Sec. 3.1), preprocessing (Sec. 3.2), DistilHuBERT (Sec. 3.3), Optuna (Sec 3.4), and the proposed architecture (Sec. 3.5). Sec. 4 presents the results of the proposed method, including the Sharif emotional speech database (ShEMO) (Sec. 4.1), implementation details (Sec. 4.2), training behavior and loss analysis (Sec. 4.3), and comparative evaluation and discussion (Sec. 4.4).

## 2 Related Work

Recent efforts have advanced Persian-language processing across various natural language processing (NLP) tasks, such as question answering, with systems like FarsNewsQA [19] and PersianRAG [20] addressing the unique challenges of low-resource settings. Motivated by this growing interest in Persian-language technologies, several methods have been proposed for PSER, leveraging various techniques, including statistical approaches and DL. For instance, [21] introduced a PSER method using a hidden Markov model (HMM)-based classifier with a minimal feature set. Their results demonstrate that their method achieves an average accuracy of 79.50%. Similarly, [22] presented a systematic and robust approach for implementing an emotion recognition system tailored to low-resource languages such as Persian. The method employs a one-dimensional (1D) convolutional neural network (CNN) architecture, where Mel-frequency cepstral coefficients (MFCC) are used for feature extraction and serve as the input to the neural network. Experimental results demonstrate that this approach achieves approximately 74% classification accuracy on the ShEMO.[23] have proposed for SER, leveraging both labeled and unlabeled samples. It employs a maximum mean discrepancy cost function to reduce dataset distribution differences. The method achieved a 14.13% error reduction in the INTERSPEECH 2009 challenge and improved PSER accuracy by 10% in domain adaptation. A feature extraction method for SER, based on adaptive time-frequency coefficients, was proposed by [24] to improve emotion classification. Combining fractional Fourier transform-based features using cepstral features, the method achieved high accuracy across multiple datasets, including 91.46% on the Persian drama radio emotional corpus (PDREC). A methodology using a stacked autoencoder neural network was proposed by [25] for SER, and tested on the Persian emotional speech database and Berlin emotional database (EMO-DB). The approach improved recognition accuracy significantly by leveraging both local and global features. The authors of [26] proposed a dimension reduction module (DRM) for Wav2vec 2.0 to enhance SER. Two methods- classifier combination and attention-based feature fusion- were introduced for decision-making using DRM outputs. The approach achieved competitive accuracy, using the attention-based feature fusion method, reaching 80.60% unweighted accuracy on ShEMO. In the domain of PSER, the lack of large labeled datasets has limited the widespread application of supervised learning (SL). However, recent efforts have pro-

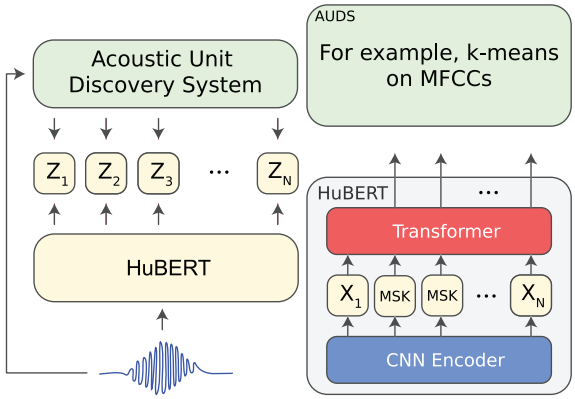


Figure 1: The HuBERT architecture. It predicts the masked frame targets based on cluster assignments obtained through one or more rounds of k-means clustering [15]

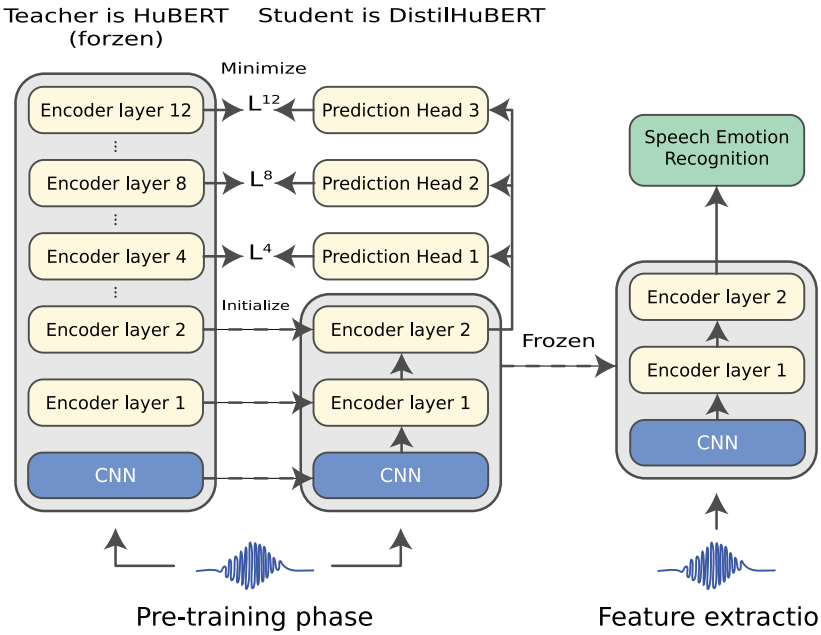


Figure 2: The DistilHuBERT architecture. First, the student model is initialized with the teacher's weights, and its prediction layers are trained to replicate the teacher's hidden representations by minimizing the combined loss $L = L_4 + L_8 + L_{12}$. Then, after pre-training, the DistilHuBERT model is frozen and used to produce speech representations for downstream tasks [16]

duced promising results. For instance, [27] investigated the impact of gender on SER using three emotional databases: Algerian dialect (AD), EMO-DB, and ShEMO. The SER system integrates prosodic and MFCC features, using classification methods including linear discriminant analysis (LDA), deep neural networks (DNNs), and support vector machines (SVM). The findings reveal that SER systems with gender distinction achieve higher recognition rates compared to those without gender distinction. Furthermore, [28] explored the application of various DL techniques on ShEMO for SER in Persian. By leveraging signal features and employing different DNN and machine learning methods, the study achieves an unweighted accuracy of 65.20% and a weighted accuracy of 78.29%, highlighting key factors relevant to SER in the Persian language.

The emergence of HuBERT, a self-SL (SSL) model, has been a game-changer in the field of speech processing, particularly in SER. Studies such as those by [15] have demonstrated

HuBERT's ability to extract rich acoustic representations that improve the performance of emotion classifiers. Numerous SER methods now leverage HuBERT to achieve high accuracy in emotion recognition across various languages. For example, [29] applied DL techniques, including CNN and the HuBERT model, to recognize emotions from the Ryerson audio-visual database of emotional speech and song (RAVDESS) dataset. By using MFCC features and DL architectures, the HuBERT model outperformed CNN in accuracy and efficiency. The research highlights HuBERT's potential in developing more accurate and responsive SER systems, contributing valuable insights into best practices for SER. [30] presented a HuBERT-based real-time emotion detection system using acoustic and linguistic features. With transfer learning, an accuracy of 95.95 percent for English, 88.64 for Hindi, and 89.70 for Punjabi is achieved on RAVDESS. In [31], SSL feature fusion with adapter fine tuning is combined for SER, and an accuracy improvement of 5.57 percent over the HuBERT baseline is reported, which reaches 75.1 percent. With this method, improved transfer of self supervised features to downstream tasks is achieved, therefore progress in SER performance is obtained. In [32], large SSL models are investigated for SER, with emphasis on RAVDESS. Among several evaluated models, the HuBERT-large model achieves the highest accuracy of 88 percent. Superior performance with reduced training time and smaller model size in comparison with other models is also reported.Despite the increased attention to the application of HuBERT in SER, its use in Persian language processing remains limited. The authors of [33] propose a lightweight model for PSER with ShEMO, in which the HuBERT convolutional encoder is employed while its transformer layers are replaced with a conformer block. Several variations of HuBERT are also evaluated, and DistilHuBERT is included in the analysis. Its suitability for low resource conditions is demonstrated.Hyperparameter optimization has an important role in improvement of DL model performance, and Optuna is recognized as an effective method for this purpose. The capability of Optuna in exploration of hyperparameter spaces is demonstrated in multiple tasks, including SER. For example, in [34], emphasis is placed on the bootstrap of your own latent for audio (BYOL-A) model, which shows strong capability in handling audio variation and in combination of local and global features. An average accuracy of 72.4 percent in general audio tasks and 57.6 pervent on VoxCeleb1 is achieved. Optuna is applied for hyperparameter tuning, and a substantial improvement in model accuracy is reported, which demonstrates its effectiveness in optimization of SER models.In [35], a hybrid two dimensional (2D) CNN LSTM, or 2D-CNN-LSTM, model is introduced for prediction of the next day closing price of Bitcoin, and Optuna is used for hyperparameter tuning. Higher accuracy and reliability than CNN, LSTM, and GRU are achieved, therefore strong suitability for real time forecasting and informed investment decisions is established. Similarly, in [36], Bayesian optimization is applied for fine tuning of key hyperparameters, namely the initial learning rate and the number of hidden units, within an LSTM based architecture for Parkinson's disease detection and severity grading. Their model achieves notable results, i.e., 99.19 percet accuracy for detection and 92.28 percent for grading. These results demonstrate the practical advantage of Bayesian methods, such as those applied in Optuna, for efficient and effective hyperparameter search across different domains. this research [37] introduced a pruning stacked ensemble learning model combining pruned multilayer perceptron (MLP) networks and heterogeneous neural networks to enhance train delay prediction accuracy. Optimized using Optuna, the model shows significant improvements, outperforming existing benchmarks by 85.22% in prediction error and achieving up to 53.40% better accuracy. The authors of [38] introduced a horizontally cascaded LSTM model for building occupancy prediction, optimized using Optuna. The model captures both short and long-term dependencies, with the best performance achieved using 2-4 LSTMs and smaller prediction

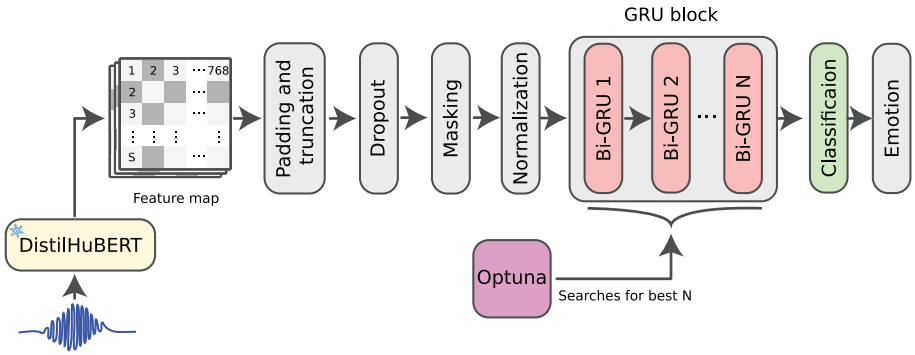


Figure 3: The overall architecture of the proposed method. The frozen DistilHuBERT model extracts a sequence of $S$ feature vectors, each with 768 dimensions. These feature sequences are then passed to the subsequent processing layers for emotion classification.

windows. the other research [39] explored using bidirectional LSTM (Bi-LSTM) networks to estimate the remaining useful life (RUL) of Lithium-ion batteries. By capturing both forward and backward long-term dependencies, Bi-LSTMs offer improved RUL predictions compared to traditional recurrent neural network (RNN). The model is optimized using Optuna for hyperparameter tuning. [40] combined ResNet50 for image classification and BERT for text analysis to address radiologist shortages. The late fusion model, integrating these methods, achieved 94.2% accuracy, outperforming other approaches. Optuna-optimized hyperparameters further enhanced performance, demonstrating the power of multi-modal AI in diagnostics. A significant strides have been made in applying DL techniques to PSER, thus, the use of DistilHuBERT and hyperparameter optimization methods such as Optuna remains underexplored in the Persian language context. Future work should focus on further adapting these advanced methods to improve performance across various Persian speech tasks.

# 3 The Proposed Methodology

## 3.1 Overall Method

The overall workflow begins with preprocessing the raw audio data. We then extract features using a frozen DistilHuBERT model, which provides robust, pre-trained speech representations. These feature sequences are padded or truncated to a uniform length. To improve model stability, we apply dropout, zero masking, and layer normalization. The normalized sequences are then passed into a Bi-GRU block. The number of Bi-GRU layers and some hyperparameters, for instance, dropout rate and learning rate, are automatically optimized using the Optuna framework. Finally, the output from the Bi-GRU layers is passed through a fully connected classification layer to predict the emotion class of each input audio sample. Fig. 3 illustrates the overall architecture of the proposed method.

## 3.2 Preprocessing

As part of the preprocessing pipeline, all audio recordings are resampled to a uniform sampling rate of 16,000 Hz to match the input requirements of the DistilHuBERT model. To prevent speaker-dependent bias and ensure a fair evaluation, we split the dataset so that there is no speaker overlap between the training, validation, and test sets. That is, each speaker appears in only one of the three subsets. Furthermore, to maintain consistency and prevent class imbalance from affecting model performance, the distribution of emotion labels is preserved across all subsets. This means each set contains approximately the same proportion of samples from each emotion class, allowing for a more stable training process and reliable comparison during evaluation. Fig. 4 shows the proportion of each emotion class within each set.

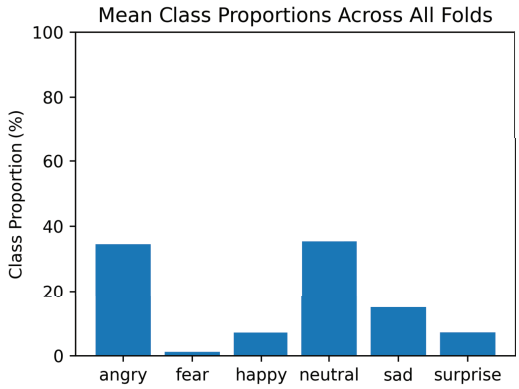


Figure 4: The proportion of each emotion class within the training, validation, and test sets.

### 3.3 DistilHuBERT

DistilHuBERT is a lightweight self-supervised speech representation model developed through knowledge distillation from the original HuBERT architecture. In this process, HuBERT is the teacher model, and the smaller student model learns to approximate its internal representations. DistilHuBERT achieves almost a 75% reduction in model size and a 73% speedup in inference time, preserving most of the HuBERT's performance and making it an efficient choice for resource-constrained tasks like PSER. The architecture of DistilHuBERT consists of two main components: a CNN feature extractor, which encodes raw audio signals, and a compact transformer encoder, which processes the extracted features. Unlike conventional distillation, which focuses only on final outputs, DistilHuBERT employs a layer-wise multi-task distillation strategy. Each layer in the student is trained to approximate the corresponding layer in the teacher, allowing deeper and more structured knowledge transfer [16]. Let $\hat{\boldsymbol{h}}_t^{(l)}$ and $\boldsymbol{h}_t^{(l)}$ represent the $D$-dimensional hidden representations at time step $t$ from the $l^{\text{th}}$ layer of the teacher and student models, respectively [16]. The layer-wise loss function in [16] is defined as:

$$L^{(l)} = L_{l1}^{(l)} + \lambda L_{\cos}^{(l)} = \sum_{t=1}^{T} \left[ \frac{1}{D} \left\| \boldsymbol{h}_t^{(l)} - \hat{\boldsymbol{h}}_t^{(l)} \right\|_1 - \lambda \log \sigma \left( \cos \left( \boldsymbol{h}_t^{(l)}, \hat{\boldsymbol{h}}_t^{(l)} \right) \right) \right]. \tag{1}$$

In this formulation, the loss $L^{(l)}$ for the $l^{th}$ layer is computed over all $T$ time steps. It consists of two components: a distance-based term scaled by the inverse of the feature dimension $D$ and a cosine similarity term weighted by a positive scalar $\lambda$. The two terms are combined additively. The cosine similarity term uses a sigmoid activation followed by a logarithm to stabilize training. This formulation encourages the student's representation $\boldsymbol{h}_t^{(l)}$ to align with the teacher's target $\hat{\boldsymbol{h}}_t^{(l)}$ both in magnitude and direction. The hyperparameter $\lambda$ modulates the influence of the similarity constraint relative to the distance component [16].

### 3.4 Optuna

Optuna is a next-generation hyperparameter optimization framework designed to tackle the challenges of tuning complex machine learning models, particularly in DL applications. It introduces a define-by-run API, enabling users to dynamically construct the search space during optimization, a significant improvement over static define-and-run frameworks. This flexibility allows for loops and conditionals, facilitating the creation of diverse and complex parameter spaces. For instance, it can optimize neural network architectures with varying layers and units [17]. Optuna employs advanced optimization algorithms, including the TPE and CMA-ES, to efficiently search for optimal hyperparameters. These methods effectively

balance exploration and exploitation, often outperforming alternative methods regarding objective value or computational efficiency. Optuna begins with independent methods such as TPE, which are known to perform well even without modeling parameter correlations. As the optimization progresses, Optuna can infer latent relationships among parameters and transition to relational sampling algorithms like CMA-ES to exploit these interdependencies [17]. To improve resource usage during optimization, Optuna implements a robust pruning mechanism that operates in two phases: it (1) periodically evaluates intermediate objective values and (2) terminates trials that fail to satisfy a predefined performance threshold. This early stopping strategy prevents unnecessary computation on unpromising configurations [17]. Optuna adopts a variant of ASHA, a state-of-the-art method specifically designed for scalability in distributed environments. ASHA extends the original Successive Halving algorithm by enabling asynchronous evaluation, allowing each worker to independently decide whether to continue or halt a trial based on interim performance metrics. Unlike synchronous approaches, asynchronous pruning avoids synchronization barriers; workers do not wait for others to complete their evaluations, significantly improving parallel efficiency. This design makes ASHA particularly effective in large-scale and distributed hyperparameter optimization workloads, where minimizing idle time is critical [17].

### 3.5 Architecture

The proposed architecture starts with raw audio waveforms, which are resampled to 16,000 Hz and then provided to a pre trained and frozen DistilHuBERT model. DistilHuBERT is used in a frozen state so that its learned speech representations are fully utilized without weight updates, therefore advantages from prior training on large scale speech corpora are obtained. The model outputs a sequence of high level feature vectors for each input audio sample, in which a 768 dimensional embedding represents each time step. However, the number of time steps, i.e., sequence length, differs across audio samples because of different durations.For batch processing and computational efficiency, all sequences are either zero padded or truncated to a fixed length threshold. If the number of time steps in a sample is below this threshold, zero padding is applied. Otherwise, truncation of the sequence is performed so that consistent dimensionality across samples is ensured. This preprocessing step is required because hardware limitations must be considered and efficient GPU utilization during training must be ensured.After this stage, a dropout layer is applied to the feature sequences so that overfitting is reduced. The optimal dropout rate is not manually selected, but it is determined through the Optuna hyperparameter optimization method. Zero masking is then applied before the subsequent layer so that learning from padded zeros is prevented. In this way, padded time steps do not influence the learning process. After masking, layer normalization is applied to the input sequences. With this operation, normalization of activations across the feature dimension for each time step is achieved, the learning process is stabilized, and convergence is accelerated. The output of the normalization layer is then passed to a Bi-GRU block. This block captures temporal dependencies in the data in both forward and backward directions. The Bi-GRU structure is not fixed. Instead, the number of GRU layers, the number of hidden units in each layer, and the learning rate are determined through automated hyperparameter tuning with Optuna. In addition, L2 regularization is applied within the GRU layers so that weight sparsity is encouraged and generalization is improved. For simplicity and computational efficiency, the same L2 coefficient is applied to all GRU layers.Moreover, Optuna is also used for selection of the most suitable optimizer among three candidates, namely Adam, SGD, and RMSprop. With this procedure, exploration of the optimization strategy that provides the most stable and accurate training behavior for the task is achieved.After temporal feature processing through the Bi-GRU layers, the final hidden states are provided to a fully con-

nected dense layer. This output layer applies a softmax activation function so that probability distribution across the target emotion classes is produced. In addition, early stopping based on validation and training loss is employed so that overfitting is prevented and unnecessary training time is reduced. The loss function is categorical cross entropy, which is defined as follows:

$$L_{CE} = -\sum_{i=1}^{C} y_i \log(\hat{y}_i), \tag{2}$$

where $C$ is the number of emotion classes, $y_i$ is the ground truth label for class $i$, and $\hat{y}_i$ is the predicted probability for that class.

# 4 Results

## 4.1 ShEMO

ShEMO is a comprehensive collection of Persian emotional speech data meticulously curated for emotion recognition tasks. It operates at a 44,100 Hz frequency, ensuring high-quality audio suitable for detailed analysis and processing. The dataset comprises 3,000 speeches, totaling approximately 3 hours and 25 minutes of recorded material. This diverse dataset features contributions from 87 native Persian speakers, including 1,263 speeches by female speakers and the remainder by male speakers. The emotional classes covered by ShEMO, Angry, Fear, Happy, Neutral, Sad, and Surprise, are identical to those used in the proposed method, providing a robust foundation for training and evaluating the proposed model [41].

## 4.2 Implementation Details

To ensure reproducibility, we fixed the random seed to 40. The model was trained using 5-fold cross-validation. In each fold, approximately 70% of the data was used for training, 10% for validation, and 20% for testing. We trained the model for a maximum of 150 epochs with a batch size 16. To manage variable-length sequences from DistilHuBERT, we applied a fixed truncation limit of 300. Sequences shorter than this limit were padded with zeros, while longer sequences were truncated. Early stopping was employed with a patience of 3 epochs, monitoring the validation loss.

Specifically, three key values were optimized by Optuna to define the GRU architecture: a base unit count selected from the range 128 to 256 (step size 32), a final unit count from 32 to 128 (step size 32), and the number of GRU layers, constrained to either 1 or 2 due to hardware limitations. The unit counts for each GRU layer was generated using a linear spacing from the base to the final value, based on the selected number of layers. In addition, Optuna searched for the optimal values of several training hyperparameters:

- Learning rate: between 0.001 and 0.01 (step size 0.001)
- Dropout rate: between 0.0 and 0.5 (step size 0.05)
- L2 regularization coefficient: between 0.01 and 0.03 (step size 0.003)
- Optimizer: selected from Adam, SGD, RMSprop

Optuna was configured to use the TPE sampler with multivariate sampling and parameter grouping both enabled (multivariate=True, group=True) to improve search efficiency. The number of startup trials was set to 5. We also enabled pruning to accelerate tuning by terminating underperforming trials early. The pruner was set to start evaluating after 5 epochs, with a reduction factor of 4, preserving the top 25% of trials. The total number of optimization trials was 10, and the maximum runtime was limited to 32,400 seconds (9 hours).

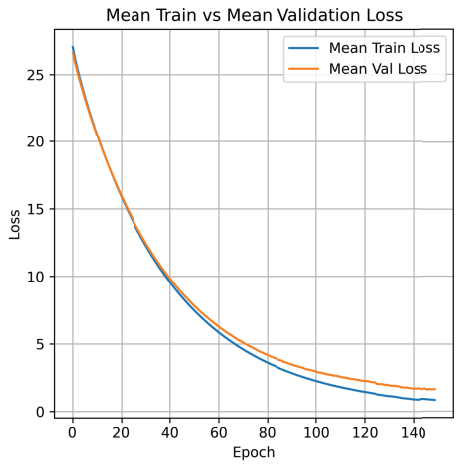


Figure 5: Mean training and validation loss across five folds for 150 epochs with the best-performing model configuration.

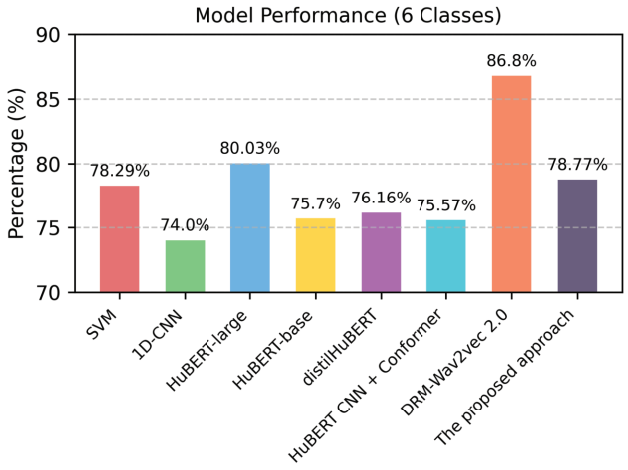


Figure 6: Accuracy comparison of different models for speech emotion recognition. The accuracy of the proposed method represents the average across five-fold cross-validation.

### 4.3 Training Behavior and Loss Analysis

Using the best hyperparameters selected by Optuna, the final model configuration consists of two bidirectional GRU layers. The first layer contains 192 units, and the second contains 32 units. The learning rate was set to 0.003, and a dropout rate of 0.2 was applied after the feature extraction stage. L2 regularization was used with a rate of 0.016, and the best-performing optimizer selected by Optuna was SGD. With these hyperparameters, the model achieved a peak accuracy of 78.77%. Fig. 5 illustrates the training dynamics of the model in terms of mean training loss and mean validation loss across five folds for 150 epochs. The plot shows a consistent downward trend for both curves, indicating effective convergence. Although the validation loss remains slightly higher than the training loss, the gap is narrow and stable, which suggests that overfitting is minimal. This behavior confirms that the model generalizes well to unseen validation samples under the selected configuration.

### 4.4 Comparative Evaluation and Discussion

As previously discussed, the objective is formulation of a method in which simplicity and performance are balanced. The results that are presented in Fig. 6 show that an accuracy of 78.77 percent is achieved by the proposed model, which exceeds the performance of several other approaches that are illustrated in the chart. Notably, higher performance than simple architectures such as 1D-CNN [22] and HuBERT CNN + Conformer [33] is obtained, and superior results compared to the DistilHuBERT approach that is proposed in [33] are also observed. These observations demonstrate the effectiveness of the proposed approach, even in comparison with models that rely on the same feature extractor.Furthermore, higher performance than the HuBERT-base model that is used in [33] is achieved by the proposed approach. Although a more complex and deeper feature extractor is used in HuBERT-base, lower performance is obtained in comparison. This observation supports the view that higher model complexity does not always result in superior performance. Therefore, the importance of careful architectural design and optimization is emphasized. Moreover, although HuBERT-large [33] and DRM-Wav2Vec 2.0 employ substantially larger and more complex configurations, results that are reasonably close to those of HuBERT-large are achieved by the proposed approach. This observation indicates that Optuna is an effective strategy for achievement of strong SER performance with a compact model, while performance that ap-

proaches more complex approaches is obtained. Although DRM-Wav2Vec 2.0 still achieves the highest accuracy, the difference in performance suggests that this gap can be reduced with further refinement of the proposed approach. The comparisons therefore confirm that the proposed approach provides a competitive and efficient solution, which distinguishes it among both lightweight and more complex alternatives.

## Conclusions

In this study, a lightweight and practical approach to PSER is presented through use of the pre trained DistilHuBERT model for feature extraction, followed by a Bi-GRU block for temporal modeling and a Dense layer for classification. The most suitable combination of hyperparameters and Bi-GRU layers is identified with Optuna, while a balance between model simplicity and performance is maintained. An accuracy of 78.77 percent is achieved on the ShEMO dataset with five fold cross validation. These results indicate the effectiveness of compact architectures combined with efficient hyperparameter optimization methods.For future work, integration of multiple Persian emotional speech datasets is planned so that model generalization is improved. In addition, application of lightweight augmentation techniques is intended, and investigation of methods for reduction of redundant features is also planned. With this approach, reduction of model complexity is expected while the most informative representations are preserved.

## References

[1] Vashishtha S., Susan S., *Unsupervised Fuzzy Inference System for Speech Emotion Recognition Using Audio and Text Cues*. Proc. 2020 IEEE Sixth International Conference on Multimedia Big Data (BigMM), (2020), 394-403. https://doi.org/10.1109/BigMM50055.2020.00067

[2] Bertero D., Siddique F., Wu C.-S., Wan Y., Chan R., Fung P., *Real-Time Speech Emotion and Sentiment Recognition for Interactive Dialogue Systems*. Proc. 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP), (2016), 1042-1047. https://doi.org/10.18653/v1/D16-1110

[3] Pourrostami H., et al., *Deep Learning for Electroencephalography Emotion Recognition*. AIMS Public Health, 12(3), (2025), 812.

[4] Liang T., Yu K., Lin L., Cheng X., Srivastava G., Lin J., Wei W., *Speech Emotion Recognition Enhanced Traffic Efficiency Solution for Autonomous Vehicles in a 5G-Enabled Space-Air-Ground Integrated Intelligent Transportation System*. IEEE Transactions on Intelligent Transportation Systems, (2021), 1-13. https://doi.org/10.1109/TITS.2021.3119921

[5] Zubair Asghar M. et al., *Performance Evaluation of Supervised Machine Learning Techniques for Efficient Detection of Emotions from Online Content*. Computer Modeling in Engineering and Sciences, 63(3), (2020), 1093-1118.

[6] Vafaie M., Dehdari J., *4 Speech Recognition for Persian*. In: K. MarszaE,ek-Kowalewska, Ed., Persian Computational Linguistics and NLP. De Gruyter Mouton, Berlin, Boston, (2023), 85-104. https://doi.org/10.1515/9783110619225-004

[7] Rahimi Pour M. H., Rastin N., Kermani M. M., *Persian Automatic Speech Recognition by the use of Whisper Model*. Proc. CSI Int. Symp. Artificial Intelligence Signal Processing (AISP), (2024), 1-7. https://doi.org/10.1109/AISP61396.2024.10475300

[8] Modaresnia Y., Torghabeh F. A., Hosseini S. A., *Enhancing Multi-Class Diabetic Retinopathy Detection Using Tuned Hyper-Parameters and Modified Deep Transfer Learning*. Multimedia Tools and Applications, 83.34 (2024), 81455-81476. https://doi.org/10.1007/s11042-024-18506-3

[9] Moghadam E. A., Torghabeh F. A., Hosseini S. A., Moattar M. H., *Improved ADHD Diagnosis Using EEG Connectivity and Deep Learning through Combining Pearson Correlation Coefficient and Phase-Locking Value*. Neuroinformatics, 22.4 (2024), 521-537. https://doi.org/10.1007/s12021-024-09685-3

[10] Ahmad H., et al., *A Hybrid Deep Learning Technique for Personality Trait Classification from Text*. IEEE Access, 9, (2021), 146214-146232.

[11] Modaresnia Y., Torghabeh F. A., Hosseini S. A., *EfficientNetB0's Hybrid Approach for Brain Tumor Classification from MRI Images Using Deep Learning and Bagging Trees*. Proc., (2023), 234-239. https://doi.org/10.1109/ICCKE60553.2023.10326290

[12] Soltani M., Shakeri H., *A DICOM Encryption Algorithm to Increase Security and Privacy in Health Data Management Based on Patient Biometrics Data, Artificial Intelligence, and RNA Encryption Algorithm*. Eurasian Journal of Mathematical and Computer Applications, 13 (2025), 137-153. https://doi.org/10.32523/2306-6172-2025-13-1-137-153

[13] Jahandoost A., Torghabeh F. A., Hosseini S. A., Houshmand M., *Crude Oil Price Forecasting Using K-means Clustering and LSTM Model Enhanced by Dense-Sparse-Dense Strategy*. Journal of Big Data, 11.1 (2024), 117. https://doi.org/10.1186/s40537-024-00977-8

[14] Young T., Hazarika D., Poria S., Cambria E., *Recent Trends in Deep Learning Based Natural Language Processing*. IEEE Computational Intelligence Magazine, 13 (2018), 55-75. https://doi.org/10.1109/MCI.2018.2840738

[15] Hsu W.-N., Bolte B., Tsai Y.-H., Lakhotia K., Salakhutdinov R., Mohamed A., *HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units*. IEEE/ACM Transactions on Audio, Speech, and Language Processing, (2021), 1. https://doi.org/10.1109/TASLP.2021.3122291

[16] Chang H.-J., Yang S.-W., Lee H., *DistilHuBERT: Speech Representation Learning by Layer-Wise Distillation of Hidden-Unit BERT*. Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP), (2022), 7087-7091. https://doi.org/10.1109/ICASSP43922.2022.9747490

[17] Akiba T., Sano S., Yanase T., Ohta T., Koyama M., *Optuna: A Next-Generation Hyperparameter Optimization Framework*. Proc. 25th ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining (KDD), (2019), 2623-2631. https://doi.org/10.1145/3292500.3330701

[18] Safavi S., Jalali M., *Toward Point-of-Interest Recommendation Systems: A Critical Review on Deep-Learning Approaches*. Proc. Electronics, 11 (2022). https://doi.org/10.3390/electronics11131998

[19] Kazemi A., Zojaji Z., Malverdi M., Mozafari J., Ebrahimi F., Abadani N., Varasteh M. R., Nematbakhsh M. A., *FarsNewsQA: a deep learning-based question answering system for the Persian news articles*. Information Retrieval Journal, 26.1 (2023), 3. doi: https://doi.org/10.1007/s10791-023-09417-2

[20] Hosseini H., Zare M. S., Mohammadi A. H., Kazemi A., Zojaji Z., Nematbakhsh M. A., *PersianRAG: A Retrieval-Augmented Generation System for Persian Language*. Proc. 2024 15th Int. Conf. on Information and Knowledge Technology (IKT), (2024), 272-278. doi: https://doi.org/10.1109/IKT65497.2024.10892726

[21] Savargiv M., Bastanfard A., *Persian Speech Emotion Recognition*. Proc. IKT, (2015), 1-5. https://doi.org/10.1109/IKT.2015.7288756

[22] Siadat S., Voronkov I., Kharlamov A., *Emotion Recognition from Persian Speech with 1D Convolution Neural Network*. Proc. CNN, (2022), 152-157. https://doi.org/10.1109/CNN56452.2022.9912532

[23] Pourebrahim Y., Razzazi F., Sameti H., *Semi-Supervised Parallel Shared Encoders for Speech Emotion Recognition*. Digital Signal Processing, 118 (2021), 103205. https://doi.org/10.1016/j.dsp.2021.103205

[24] Langari S., Marvi H., Zahedi M., *Efficient Speech Emotion Recognition Using Modified Feature Extraction*. Informatics in Medicine Unlocked, 20 (2020), 100424. https://doi.org/10.1016/j.imu.2020.100424

[25] Bastanfard A., Abbasian A., *Speech Emotion Recognition in Persian Based on Stacked Autoencoder by Comparing Local and Global Features*. Multimedia Tools and Applications, 82.1 (2023), 1-18. https://doi.org/10.1007/s11042-023-15132-3

[26] Nasersharif B., Namvarpour M., *Exploring the Potential of Wav2vec 2.0 for Speech Emotion Recognition Using Classifier Combination and Attention-Based Feature Fusion*. The Journal of Supercomputing, 80 (2024), 1-22. https://doi.org/10.1007/s11227-024-06158-x

[27] Horkous H., *Study the Effect of Gender on Speech Emotion Recognition Using Algerian, German, and Persian Databases*. Proc. 2024 IEEE International Conference on Electrical, Electronics, and Computer Engineering (ICEEAC), (2024), 1-6. https://doi.org/10.1109/ICEEAC61226.2024.10576414

[28] Yazdani A., Simchi H., Shekofteh Y., *Emotion Recognition In Persian Speech Using Deep Neural Networks*. Proc. 2021 IEEE International Conference on Knowledge and Engineering (ICCKE), (2021), 374-378. https://doi.org/10.1109/ICCKE54056.2021.9721504

[29] Gismelbari M., Vixnin I., Kovalev G., Gogolev E., *Speech Emotion Recognition Using Deep Learning*. Proc. 2024 IEEE International Conference on Smart Computing (SCM), (2024), 380-384. https://doi.org/10.1109/SCM62608.2024.10554271

[30] Singh K., Sehgal L., Aggarwal N., *Multilingual Emotion Recognition from Continuous Speech Using Transfer Learning*. In Emergent Converging Technologies and Biomedical Systems, Springer Nature Singapore, (2024), 197-211. https://doi.org/10.1007/978-981-99-8646-0_17

[31] Li T., Hou J., *Utilizing Self-Supervised Learning Features and Adapter Fine-Tuning for Enhancing Speech Emotion Recognition*. In 2023 5th International Conference on Machine Learning, Big Data and Business Intelligence (MLBDBI), (2023), 79-84. https://doi.org/10.1109/MLBDBI60823.2023.10482145

[32] Gavali M. P., Verma A., *Automatic recognition of emotions in speech with large self-supervised learning transformer models*. 2023 IEEE International Conference on Artificial Intelligence, Blockchain, and Internet of Things (AIBThings), (2023), 1-7. https://doi.org/10.1109/AIBThings58340.2023.10292462

[33] Nasersharif B., Azad M., *Speech emotion recognition using transfer learning and self-supervised speech representation learning*. 2023 31st International Conference on Electrical Engineering (ICEE), (2023), 684-689. https://doi.org/10.1109/ICEE59167.2023.10334799

[34] Niizumi D., Takeuchi D., Ohishi Y., Harada N., Kashino K., *BYOL for audio: Exploring pre-trained general-purpose audio representations*. IEEE/ACM Trans. Audio, Speech and Lang. Proc., 31 (2023), 137-151. https://doi.org/10.1109/TASLP.2022.3221007

[35] Kazeminia S., Sajedi H., Arjmand M., *Real-time bitcoin price prediction using hybrid 2D-CNN LSTM model*. Proc. 9th Int. Conf. Web Research (ICWR), (2023), 173-178. https://doi.org/10.1109/ICWR57742.2023.10139275

[36] Abedinzadeh Torghabeh F., Modaresnia Y., Hosseini S. A., *An efficient tool for Parkinson's disease detection and severity grading based on time-frequency and fuzzy features of cumulative gait signals through improved LSTM networks*. Medicine in Novel Technology and Devices, 22 (2024), 100297. https://doi.org/10.1016/j.medntd.2024.100297

[37] Boateng V. A., Yang B., *A global modeling pruning ensemble stacking with deep learning and neural network meta-learner for passenger train delay prediction*. IEEE Access, 11 (2023), 62605-62615. https://doi.org/10.1109/ACCESS.2023.3287975

[38] Kanthila C., Boodi A., Beddiar K., Amirat Y., Benbouzid M., *Occupancy prediction in buildings using cascaded LSTM model*. IECON 2023 - 49th Annual Conference of the IEEE Industrial Electronics Society, (2023), 1-6. https://doi.org/10.1109/IECON51785.2023.10311629

[39] Sahay R., Pugalenthi K., Raghavan N., *Remaining useful life estimation of lithium-ion batteries via hyperparameter optimized bi-long short-term memory recurrent neural networks*. 2023 Global Reliability and Prognostics and Health Management Conference (PHM-Hangzhou), (2023), 1-8. https://doi.org/10.1109/PHM-Hangzhou58797.2023.10482631

[40] Upadhya J., Poudel K., Ranganathan J., *Advancing medical image diagnostics through multi-modal fusion: Insights from MIMIC chest X-ray dataset analysis*. 2024 IEEE 3rd International Conference on Computing and Machine Intelligence (ICMI), (2024), 1-8. https://doi.org/10.1109/ICMI60790.2024.10586129

[41] Mohamad Nezami O., Lou P., Karami M., *ShEMO: A large-scale validated database for Persian speech emotion detection*. Language Resources and Evaluation, 53 (2019). https://doi.org/10.1007/s10579-018-9427-x

[42] Agarla M., Bianco S., Celona L., Napoletano P., Petrovsky A., Piccoli F., Shanin I., *Semi-supervised cross-lingual speech emotion recognition*. Expert Systems with Applications, 237 (2023), 121368. https://doi.org/10.1016/j.eswa.2023.121368

Amirmohammad Ahmadianshalchi,
Department of Computer Engineering, Ma.C.,
Islamic Azad University, Mashhad, Iran,
Email: am.ahmadianshalchi@iau.ir,

Mahboobeh Houshmand *,
Department of Computer Engineering, Ma.C.,
Islamic Azad University, Mashhad, Iran,
Email:ma.houshmand@iau.ac.ir,

Seyyedabed Hosseini,
Department of Electrical Engineering, Ma.C.,
Islamic Azad University, Mashhad, Iran,
Email:sa.hosseini@iau.ac.ir

Received 10.04.2025, Revised 15.12.2025, Accepted 08.01.2026, Available online 31.03.2026.