EmotionNet: A Novel Hybrid Deep Learning Model for Arabic Speech Emotion Recognition

Mourad Belhadj
LINATI, University of Ouargla, Algeria | LESIA, University of Biskra, Algeria
belhadj.mourad@univ-ouargla.dz (corresponding author)

Mihoub Mazouz
LINATI, University of Ouargla, Algeria
mazouz.mihoub@univ-ouargla.dz

Dalal Djeridi
LINATI, University of Ouargla, Algeria
djeridi.dalal@univ-ouargla.dz

Received: 11 May 2025 | Revised: 4 June 2025 and 21 June 2025 | Accepted: 5 July 2025

 DOI: https://doi.org/10.48084/etasr.12035

ABSTRACT

This study presents EmotionNet, a novel hybrid deep learning model designed for Arabic speech emotion recognition. EmotionNet integrates a Variational Auto-Encoder (VAE) for latent representation learning with a lightweight classification branch enhanced by latent-space refinement. Evaluated on the KEDAS dataset, which includes five emotionally acted categories, the model achieved a test accuracy of 93.99% and outperformed conventional classifiers such as SVM, MLP, and Random Forest. The proposed approach employs a compound loss function and KL annealing to jointly optimize reconstruction and classification. Although the results are promising, the acted nature of KEDAS may overstate real-world performance, highlighting the need for evaluation on spontaneous, multimodal datasets, an effort currently underway in an ongoing interdisciplinary project.

Keywords-Arabic speech emotion; variational autoencoder; KEDAS dataset; lld features

I. INTRODUCTION

Understanding human emotions has long been a focal point in psychology, neuroscience, and computer science. With the rise of deep learning, automated Speech Emotion Recognition (SER) has achieved significant success, enabling applications in mental health diagnostics, intelligent virtual assistants, and human–machine interaction. Despite these advances, many SER systems suffer from two critical limitations: a lack of robustness in real-world environments and limited interpretability of model decisions.

The KEDAS dataset [1, 2], a validated, emotion-annotated Arabic speech corpus, presents a valuable opportunity to explore these challenges in a linguistically and culturally nuanced setting. With its rich phonetic diversity, Arabic poses unique challenges for emotion classification. KEDAS offers diverse emotional categories, balanced classes, and makes it a benchmark for deep learning research in underrepresented languages.

This study presents EmotionNet, an end-to-end deep learning framework designed to address the dual challenges of performance and interpretability. The architecture integrates a Variational Auto-Encoder (VAE) for compact latent representation learning with a lightweight classification branch enhanced by latent-space refinement. By leveraging a compound loss function and Kullback-Leibler (KL) divergence annealing, it gradually increases the regularization strength of the KL term during training. This approach is jointly optimized for both reconstruction and classification, leading to improved generalization.

Arabic SER has received increasing attention due to the linguistic diversity and dialectal variations inherent in the Arabic language [3]. Early studies primarily employed classical machine learning techniques such as Support Vector Machines (SVM) and k-Nearest Neighbors (k-NN) on handcrafted features, such as Mel-Frequency Cepstral Coefficients (MFCCs) [4]. However, the advent of deep learning has shifted the focus toward more robust and automated feature extraction methods. Recent advances include the development of deep neural networks tailored for Arabic SER. For instance, the SERDNN model [5] achieved a notable accuracy of 97.4% on the ANAD dataset. Similarly, in [4], Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks were used on a Saudi dialect corpus, reporting an accuracy of 95%. In [5], a comparative study of several deep learning models on Arabic emotional speech data, including

CNN, LSTM, and GRU, highlighted that hybrid models, particularly CNN-LSTM architectures, outperformed standalone networks in terms of accuracy, showing promise for complex emotion recognition tasks in low-resource languages such as Arabic. These results reinforce the growing consensus that deep learning, particularly in multilayer and hybrid forms, is well-suited to capture the nuanced emotional content in spoken Arabic.

Data augmentation techniques have been explored to address the challenges posed by limited and imbalanced datasets. In [6], noise addition and volume adjustments were applied on the Saudi Dialect and BAVED datasets, improving model performance. Autoencoders, particularly VAEs, have been employed to learn compact and informative representations of speech data. Although their application in Arabic SER is still emerging, studies in other languages have demonstrated their efficacy in capturing latent emotional features [7-9]. Integrating VAEs with classifiers can enhance the ability to generalize across different emotional expressions. Hybrid architectures, combining CNNs with Recurrent Neural Networks (RNNs) or attention mechanisms, have also shown promise. These models leverage the spatial feature extraction capabilities of CNNs and the temporal modeling strengths of RNNs to capture the dynamic nature of speech emotions [10].

TABLE I. RECENT RELATED ARABIC SER STUDIES

Study	Model	Dataset
[5]	DNN	ANAD
[4]	CNN, LSTM	Saudi Dialect
[6]	CNN	Saudi Dialect, BAVED

II. METHODOLOGY

A. Overall Architecture

Figure 1 depicts the complete EmotionNet architecture. The input layer processes the 124-dimensional acoustic feature vector extracted from the KEDAS dataset. The VAE branch consists of an encoder network that compresses the input into a compact latent representation and a decoder that reconstructs the original features, enforcing meaningful information encoding. The latent vector is then refined through a transformation layer, concatenated with its original version, and fed to a shallow MLP classifier for final emotion prediction. This dual-branch design enables unsupervised feature extraction while directly optimizing classification, thereby improving both the expressiveness and robustness of the learned representations.

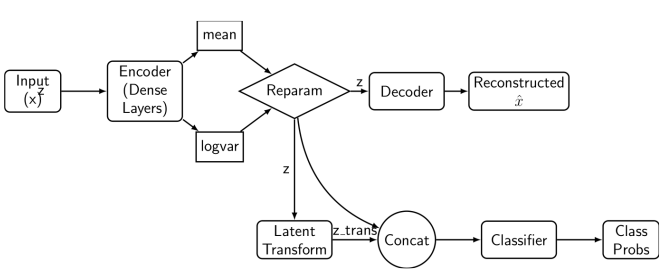

Fig. 1. Architecture diagram for EmotionNet.

EmotionNet consists of two main parts:

1. VAE Branch (unsupervised): Learns a latent representation of the high-dimensional input data (assumed 124 features in KEDAS). The 124-dimensional vectors typically encompass Low-Level Descriptor (LLD) features [11], including MFCCs and other acoustic descriptors [12, 13]. This branch has an encoder that outputs mean μ and log-variance $log(\sigma^2)$ [14], plus a reparameterization step producing latent vectors z. A decoder reconstructs the input from z.
2. Classifier Branch (supervised): Receives latent vectors from the VAE, applies a latent transform layer that refines z, and concatenates it back with z before feeding into a shallow MLP for emotion prediction (5-class softmax).

These branches train end-to-end with a compound loss that balances reconstruction fidelity, classification accuracy, and latent regularization (KL divergence) [15]. Additionally, a KL annealing schedule increases the KL contribution during early epochs, helping the model focus first on reconstruction and classification before imposing strong latent constraints.

B. Variational Autoencoder

1) Encoder

The encoder is a four-layer dense network of sizes {512, 256, 128, 2 × latent dim} with ReLU activations (except the output). For input $\mathrm{x} \in R^{124}$, a vector encompassing MFCC and other acoustic/LLD features gives:

$$(\mu,\ \log(\sigma^2)) = Encoder(\mathrm{x}) \quad (1)$$

where μ is the mean, σ^2 is the variance, x is the input feature vector, and $Encoder(\cdot)$ is the encoder neural network, $(\mu, \log(\sigma^2) \in \mathbb{R}^{latent\_dim})$. Then, reparameterization gives:

$$z = \mu + exp(0.5 \cdot \log(\sigma^2)) \odot \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, I) \quad (2)$$

where z is the latent vector, μ is the mean, σ^2 is the variance, $\exp(\cdot)$ is the exponential function, $\odot$ is element-wise multiplication, and ε is random noise sampled from a standard normal distribution with mean 0 and identity matrix I. This trick preserves differentiability with respect to the encoder parameters while injecting stochasticity into the latent vector z.

2) Decoder

The decoder receives the latent vector $(\mathrm{z} \in \mathbb{R}^{latent\_dim})$ and reconstructs a normalized version of the original input data $\hat{\mathrm{x}}$. It consists of four dense layers with sizes $[128, 256, 512, input\_dim]$. The decoder uses ReLU activations in the hidden layers. The output layer employs a sigmoid activation function to match the $[0, 1]$ normalization range. Formally, the reconstruction is defined as:

$$\hat{\mathrm{x}} = Decoder(\mathrm{z}) \quad (3)$$

Reconstruction fidelity is measured with MSE

$$\mathcal{L}rec = |x - \hat{x}|_2^2 \quad (4)$$

although other choices such as cross-entropy can also be used depending on the data type.

3) KL Divergence and Annealing

To regularize the latent space, the VAE employs the KL divergence [16]:

$$\mathcal{L}_{KL} = -\frac{1}{2}\sum_{j=1}^{latent\_dim}\left[1 + log\left(\sigma_j^2\right) - \mu_j^2 - \sigma_j^2\right] \quad (5)$$

However, applying the KL divergence penalty directly from the start can sometimes destabilize training. Thus, its coefficient $\lambda_{KL}(t)$ is annealed over the first few epochs [17], increasing it from 0 to a target value, e.g., 0.01 or 0.1:

$$\lambda_{KL}(t) = \lambda_{KL}^{max} \times \min\left(1, \frac{t}{\propto T}\right) \quad (6)$$

where t is the current epoch, T is the maximum epoch, and α is typically around 0.1 (i.e., 10% of total training).

C. Latent Transform and Classifier Branch

After sampling the latent vector z from the VAE, a latent transformation is applied to refine its structure before classification. This is implemented as a fully connected dense layer with ReLU activation:

$$\mathrm{z}_{\mathrm{trans}} = ReLU(\mathrm{W}_1\mathrm{z} + b_1) \quad (7)$$

where $\mathrm{z}_{\mathrm{trans}}$ is the transformed latent vector, $ReLU(\cdot)$ is the rectified linear unit activation function, W_1 is the weight matrix $W_1 \in \mathbb{R}^{latent\ dim \times latent\ dim}$, z is the latent vector, and b_1 is the bias term. The original and transformed latent vectors are then concatenated:

$$\mathrm{z}_{\mathrm{concat}} = [\mathrm{z} \parallel \mathrm{z}_{\mathrm{trans}}] \in \mathbb{R}^{2 \times latent\ dim} \quad (8)$$

This concatenated vector is passed to a shallow MLP for emotion classification. The MLP consists of:

- A dense layer with 256 neurons, *ReLU* activation, and L2 regularization (λ=0.01).
- A dropout layer (rate = 0.3),
- A final softmax output layer of size 5 (for the five emotion classes in KEDAS):

$$p = softmax(\mathrm{W}_2 \cdot Dropout\,(ReLU(\mathrm{W}_{3\mathrm{zconcat}})) + b_2) \quad (9)$$

The classifier is optimized using cross-entropy loss $\mathcal{L}_{cls}$ (10), where C is the number of emotion classes, y_i is the true (one-hot encoded) label for class i, and $\hat{y}_i$ is the predicted probability for class i.

$$\mathcal{L}_{cls} = -\sum_{i=1}^{C} y_i \log(\hat{y}_i) \quad (10)$$

D. Compound Loss Function and Training

Let x be an input sample, z be the sampled latent vector, $\hat{\mathrm{x}}$ be the reconstruction, and p be the predicted class probabilities. The total loss $\mathcal{L}_{total}$ merges three components:

$$\mathcal{L}_{total} = \lambda_{rec}\,\mathcal{L}_{rec} + \lambda_{cls}\mathcal{L}_{cls} + \lambda_{KL}(t)\,\mathcal{L}_{KL} \quad (11)$$

where $\mathcal{L}_{rec}$ is the reconstruction loss, $\mathcal{L}_{rec}$ is the classification loss, $\mathcal{L}_{KL}$ is the KL divergence, and λ terms are weighting coefficients, with λ_{rec} and λ_{cls} typically default to 1.0 and 1.5, respectively, and $\lambda_{KL}(t)$ is the annealed KL coefficient described previously. By scheduling $\lambda_{KL}(t)$ from 0.0 to a final maximum over a fraction of the training epoch, the model can establish stable reconstruction and classification performance before strongly regularizing the latent space.

1) Optimization and Early Stopping

The Adam optimizer [18] (learning rate 10^{-3}) is used and the entire architecture is trained in mini-batches (batch size=32) for up to 100 epochs. The validation loss is tracked in each epoch, and early stopping is used with a patience of 10 epochs, restoring the best model weights (lowest validation loss). This approach balances under- and over-fitting while ensuring the final model is robust on unseen data.

In general, EmotionNet unifies a generative subnetwork (VAE) and a discriminative subnetwork (latent transform and classifier) in a single end-to-end model, guided by a compound loss function. KL annealing, L2 regularization, and early stopping further enhance stability and generalization for emotion recognition tasks on the KEDAS dataset, where each sample comprises 124-dimensional acoustic features that include MFCCs and other low-level descriptors.

III. EXPERIMENTAL SETTINGS

A. Dataset

The KEDAS dataset is a validated Arabic emotional speech corpus annotated for five emotion classes: anger, happiness, sadness, neutral, and fear [1, 2]. Each instance consists of 124 LLD features, including MFCCs, pitch, intensity, and prosodic attributes. The dataset was loaded from a CSV file and checked for missing or infinite values. Rows containing NaNs or infinite values were excluded before training.

B. Preprocessing

- Normalization: All numeric features were normalized to the [0, 1] range using min-max scaling.
- Label encoding: Emotion labels were factorized into integer values and converted into one-hot encoding for use in neural models.
- Data splitting: 70 % for training, 15% for validation, and 15% for testing. The split was stratified to preserve class distribution.

C. Model Configuration

The proposed EmotionNet is a hybrid deep neural network integrating a VAE for unsupervised latent representation and a refined classifier branch for emotion prediction.

- Encoder:
  - 4 dense layers with sizes [512, 256, 128, 64]
  - Outputs both mean and log-variance for latent sampling
- Latent Dimensionality: 32
- Decoder:
  - 4 dense layers [64, 128, 256, 512] and an output layer
- Latent transformation layer: 1 dense layer of 32 neurons (ReLU)

- Classifier Branch:
  - Dense layer (128 neurons, L2 regularization)
  - Dropout layer (rate = 0.5)
  - Output: 5-class softmax

D. Training Strategy

- Optimizer: Adam (learning rate = 0.001)
- Loss function: Compound loss:
  - Reconstruction loss (MSE).
  - Classification loss (Categorical Crossentropy)
  - KL divergence with annealed weight over epochs
- Batch size: 32
- Epochs: Up to 100 with early stopping (patience = 10)
- Class imbalance handling: Class weights are inversely proportional to frequency.

KL divergence was annealed during the initial 10% training epochs to gradually introduce latent space regularization. The best-performing model was saved based on the lowest validation loss.

E. Evaluation Metrics

The following metrics were used for model evaluation:

- Accuracy
- Precision, Recall, F1-score (macro-averaged)
- ROC curves and AUC per class
- Confusion matrix
- Classification reports

To benchmark EmotionNet, the following models were also implemented using the same feature set and data splits:

- Logistic Regression (LR)
- Support Vector Machine (SVM)
- Random Forest (RF)
- k-Nearest Neighbors (k-NN)
- Baseline Multi-Layer Perceptron (MLP)

All comparative models were trained and evaluated with their default parameters, except otherwise specified, such as $max\_iter = 500$ for LR.

All experiments were conducted in a Google Colab environment using Python 3 and TensorFlow 2.x, accelerated by a NVIDIA Tesla T4 or P100 GPU. The validation set was used for hyperparameter tuning and to monitor early stopping during training. Each input sample was represented by a 124-dimensional feature vector composed of MFCCs and other LLDs, which were normalized to the [0, 1] range using min-max scaling. Emotion labels were one-hot encoded to represent five distinct emotion classes. The proposed EmotionNet model combines a VAE for unsupervised latent feature extraction with a supervised classification branch. Table II provides key hyperparameters, including the latent dimension size, learning rate, batch size, and regularization weights.

TABLE II. HYPERPARAMETERS FOR EMOTIONNET EXPERIMENTS

Parameter	Value
Input dimension	124
Latent dimension	32
Batch size	32
Max epochs	100
Learning rate	0.001
KL annealing	0.0→0.1 (first 10% epochs)
L2 regularization	0.01 (classifier dense layer)
Dropout rate	0.3 (classifier)
Optimizer	Adam

IV. RESULTS AND DISCUSSION

A. Results

Table III shows that EmotionNet achieved a test accuracy of 0.9399 (≈94%) on 749 held-out samples, with a cross-entropy loss of 0.2085, indicating well-calibrated probability estimates. On a per-class basis, it scored a precision of 0.92, recall of 0.93 and F1-score of 0.93 for Angry (150 samples), 0.95, 0.96 and 0.95 for Sadness (150 samples), 0.92, 0.95 and 0.93 for Fear (150 samples), 0.95, 0.95 and 0.95 for Happy (150 samples), and 0.96, 0.90 and 0.93 for Neutral (149 samples). Averaging across all classes, both unweighted (macro-average) and weighted by support, the model maintains a precision of 0.94, a recall of 0.94, and an F1-score of 0.94.

TABLE III. EMOTIONNET RESULTS

Emotion	Precision	Recall	F1-score	Support
Angry	0.92	0.93	0.93	150
Sadness	0.95	0.96	0.95	150
Fear	0.92	0.95	0.93	150
Happy	0.95	0.95	0.95	150
Neutral	0.96	0.90	0.93	149
Accuracy			0.94	749
Macro avg	0.94	0.94	0.94	749
Weighted avg	0.94	0.94	0.94	749
Test accuracy	0.9399			
Test loss	0.2085			

Among the classic machine-learning baselines (Table IV), SVM leads with a test accuracy of 89.29% and correspondingly high F1-scores (macro- and weighted-averages of 0.89), while LR, RF, and k-NN all cluster around the low-to-mid 84% accuracy range with macro and weighted F1-scores of approximately 0.84-0.85. In contrast, the SERDNN model strikes a balance between accuracy and class-balance, achieving an overall test accuracy of 87.00%, a macro F1 of 0.88, and a weighted F1 of 0.87. Moreover, its cross-entropy test loss of 0.4280 suggests that its probability estimates remain reasonably well calibrated, even though its accuracy does not quite match the SVM. Overall, while SVM still delivers the highest raw accuracy, SERDNN offers stronger F1-score consistency across classes and provides more informative probabilistic outputs.

TABLE IV. PERFORMANCE COMPARISON.

Model	Test Acc	Macro F1	Weighted F1	Test loss
LR	0.8418	0.84	0.84	-
SVM	0.8929	0.89	0.89	-
RF	0.8468	0.85	0.85	-
k-NN	0.8408	0.84	0.84	-
SERDNN	0.8700	0.88	0.87	0.4280

EmotionNet clearly outperforms both the classical classifiers and the SERDNN approach in raw accuracy and calibration, whereas SERDNN offers more balanced class-wise performance than most traditional methods.

B. Discussion

The exceptionally high accuracy (≈94%) and uniformly strong F1-scores achieved by EmotionNet on the KEDAS test set demonstrate the model's ability to learn discriminative acoustic features from the dataset's utterances. In particular, the hybrid VAE-classification architecture appears to capture both spectral (MFCC) and temporal patterns that distinguish prototypical expressions of anger, sadness, fear, happiness, and neutral speech. Moreover, the low test loss (0.2085) indicates that EmotionNet's probability estimates are well calibrated, and the balanced macro- and weighted-average F1 of 0.94 underscores its robustness across all five emotion classes.

However, these performance results must be interpreted with caution. KEDAS is an acted dataset in which speakers intentionally portray distinct emotional states under controlled recording conditions. Such acted utterances tend to exhibit exaggerated prosodic contours, clear pitch modulations, amplitude patterns, and speech rate changes, which make class boundaries more separable compared to spontaneous, naturalistic speech. As a result, classification models often demonstrate inflated performance on acted corpora, not because they generalize equally well in real-world scenarios, but because the training and test examples contain highly prototypical, low-variability signals.

However, an accuracy of 97.4% was reported in [5] for the SERDNN model on the ANAD dataset, whereas the reimplementation of the same model achieved 87.0% on the KEDAS dataset. Although both ANAD and KEDAS are acted emotional speech corpora, differences in dataset size, speaker diversity, emotional class distribution, and recording conditions likely contribute to the observed performance gap. To ensure a fair comparison, the SVM model was also evaluated on KEDAS using the same feature extraction and data splits, achieving 89.29% accuracy. While SVM shows higher raw accuracy, SERDNN offers more consistent F1 scores and calibrated probability estimates, highlighting its advantage in class-balanced performance. This limitation is reflected in two key observations. First, the confusion matrix in Figure 2 shows minimal overlap between classes, an outcome rarely seen when applying models to spontaneous recordings, where emotional expression is subtler and more context-dependent. Second, the neutral class, while achieving the highest precision (0.96), suffers from lower recall (0.90), suggesting that even in acted data, speakers sometimes underplay or inconsistently portray neutrality, with the model mislabelling these less emphatic utterances as low-energy emotions such as sadness.

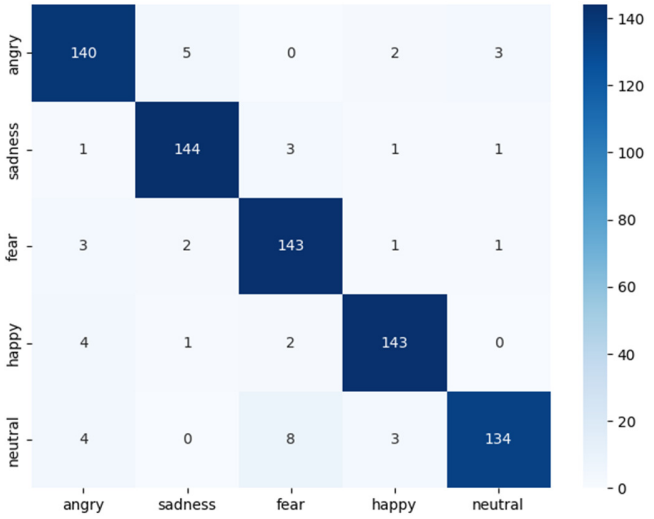

Fig. 2. Confusion matrix.

Another factor contributing to the high accuracy is the balanced class distribution and relatively clean audio in KEDAS, which minimizes issues of class imbalance and background noise. In real-world applications, datasets are often skewed toward neutral or mild affective states, and recordings contain environmental distortions, thereby degrading classification accuracy.

To comprehensively assess EmotionNet's real-world applicability, future work should include cross-corpus evaluation on spontaneous emotional speech and in-the-wild recordings, as well as domain-adaptation techniques [19] to bridge the gap between acted and naturalistic data. Incorporating multimodal cues, such as facial expressions or physiological signals, could further improve robustness in less controlled environments. An ablation study that removes the VAE pre-training branch would also quantify the specific gains provided by unsupervised feature learning versus the model's reliance on exaggerated, acted examples.

Figure 3 shows the ROC curves for each emotion class. All curves are clustered near the top-left corner, indicating that the model achieves high sensitivity and a low false alarm rate across all classes. The Area Under the Curve (AUC) values are very high, ranging from 0.99 to 1.00, demonstrating excellent discrimination between emotions. This result confirms that the model reliably distinguishes each emotion with minimal confusion, further validating its strong classification performance. The learning curves (Figure 4) align with the quantitative results for the same model, offering a richer, more dynamic view of its training behavior. During the first five epochs, EmotionNet rapidly learns to distinguish the five prototypical emotional categories: training accuracy rises rapidly from about 63% to over 90%, and training loss plummets from roughly 1.0 to below 0.3. Validation accuracy follows a very similar trajectory, after an initial dip (likely due to the model finding its footing on unseen samples), it too crosses the 90% mark by epoch 6 and remains remarkably stable thereafter. Likewise, validation loss quickly converges to the 0.20-0.25 range and exhibits only minor fluctuations.

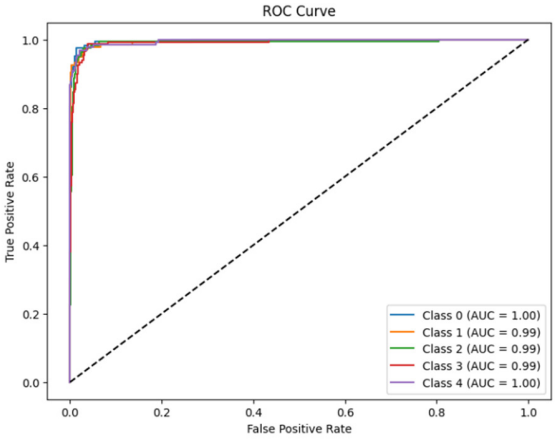

Fig. 3. ROC curve for each emotion class.

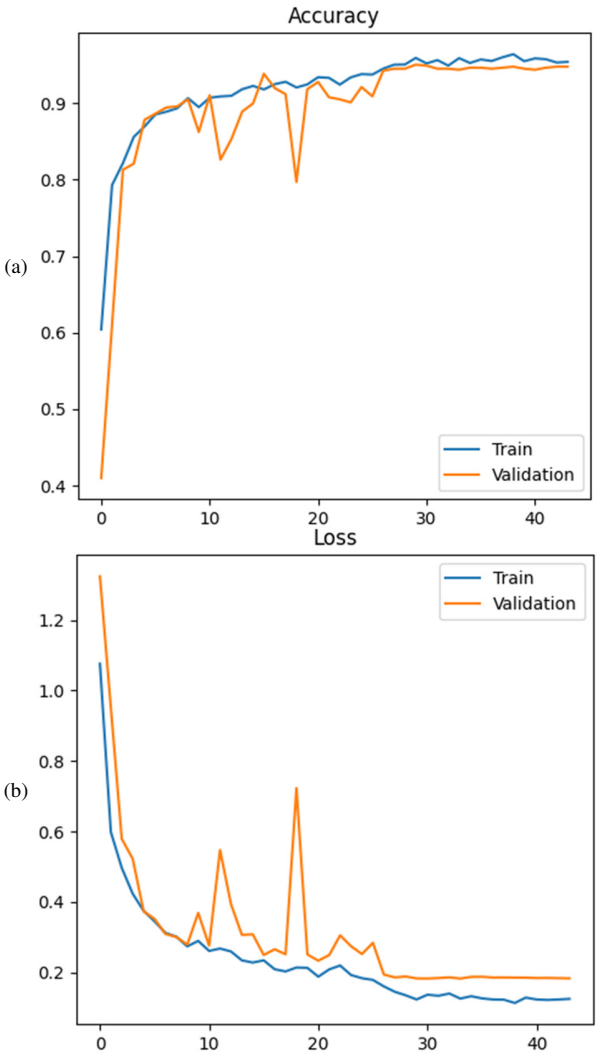

Fig. 4. Learning curves for accuracy (a) and loss (b).

From the two figures, the following points can be extracted:

- Fast Convergence and Minimal Overfitting: Both training and validation curves plateau by epoch 10-12, with only a slight gap (≤ 3 %) between them. This suggests that, under the controlled conditions of KEDAS, EmotionNet does not overfit dramatically, even with its relatively large capacity, because the acted utterances share highly consistent prosodic patterns.
- Transient Validation Drop: The sharp dip in validation accuracy and corresponding spikes in validation loss around epoch 6 hint at a temporary mismatch between the model's learned parameters and the validation split. In practice, this could arise from slight class imbalance or from particularly challenging neutral samples being overrepresented in that fold.
- Plateau and Diminishing Returns: Beyond epoch 12, accuracy gains are marginal (≤1%), and loss improvements are negligible. This plateau underscores that the model has already captured most of the signal in the actual data, and further training primarily refines low-variance features rather than discovering new patterns.
- Effect of Acted Data. The exceptionally clean separation between train and validation performance, and their joint rapid convergence, reflects the very prototypical, exaggerated expressions present in KEDAS. In spontaneous speech corpora, one would expect both slower convergence and a larger train/validation gap, as natural emotions are more diffuse and have acoustic overlap.

The learning curves confirm that EmotionNet is highly effective at modeling the clear, low-variability prosodic cues in an acted dataset such as KEDAS. However, the speed and consistency of its convergence also serve as a reminder that real-world emotional speech, with its subtler nuance and background noise, will pose a greater challenge and will likely require more epochs, stronger regularization, or multimodal inputs to achieve comparable robustness.

V. CONCLUSION AND FUTURE WORK

This study introduced EmotionNet, a deep learning model for recognizing emotions in Arabic speech. EmotionNet is unique because it combines a VAE for unsupervised feature learning directly from the speech data with a classifier that uses these features to identify emotions. This hybrid approach allows the model to understand the complex patterns of Arabic speech more effectively. This combined design helps the model learn useful information even from limited or unbalanced data, such as the KEDAS dataset. In addition, a training method carefully balanced how the model rebuilds input data, classifies emotions, and organizes its internal features. This process leads to smooth learning and high accuracy.

REFERENCES

[1] M. Belhadj, I. Bendellali, and E. Lakhdari, "KEDAS: A validated Arabic Speech Emotion Dataset," in *2022 International Symposium on iNnovative Informatics of Biskra (ISNIB)*, Biskra, Algeria, Dec. 2022, pp. 1–6, https://doi.org/10.1109/isnib57382.2022.10075694.

[2] M. Belhadj, I. Bendellali, and E. Lakhdari, "Kasdi-Merbah (University) Emotional Database in Arabic Speech." Linguistic Data Consortium, 2023, https://doi.org/10.35111/QQER-QZ15.

[3] L. Iben Nasr, A. Masmoudi, and L. Hadrich Belguith, "Survey on Arabic speech emotion recognition," *International Journal of Speech*

Technology, vol. 27, no. 1, pp. 53–68, Mar. 2024, https://doi.org/10.1007/s10772-024-10088-7.

[4] H. Alamri and H. S. Alshanbari, "Emotion Recognition in Arabic Speech from Saudi Dialect Corpus Using Machine Learning and Deep Learning Algorithms," *International Journal of Computer Science and Network Security*, vol. 23, no. 8, pp. 9–16, Aug. 2023, https://doi.org/10.22937/IJCSNS.2023.23.8.2.

[5] W. Ismaiel, A. Alhalangy, A. O. Y. Mohamed, and A. I. A. Musa, "Deep Learning, Ensemble and Supervised Machine Learning for Arabic Speech Emotion Recognition," *Engineering, Technology & Applied Science Research*, vol. 14, no. 2, pp. 13757–13764, Apr. 2024, https://doi.org/10.48084/etasr.7134.

[6] W. Bouchelligua, R. Al-Dayil, and A. Algaith, "Effective Data Augmentation Techniques for Arabic Speech Emotion Recognition Using Convolutional Neural Networks," *Applied Sciences*, vol. 15, no. 4, Jan. 2025, Art. no. 2114, https://doi.org/10.3390/app15042114.

[7] Y. Xiao, Y. Bo, and Z. Zheng, "Speech Emotion Recognition based on Semi-Supervised Adversarial Variational Autoencoder," in 2023 IEEE 10th International Conference on Cyber Security and Cloud Computing (CSCloud)/2023 IEEE 9th International Conference on Edge Computing and Scalable Cloud (EdgeCom), Xiangtan, Hunan, China, Jul. 2023, pp. 275–280, https://doi.org/10.1109/cscloud-edgecom58631.2023.00054.

[8] A. V. Porco and D. Kang, "Enhancing Emotion Classification Through Speech and Correlated Emotional Sounds via a Variational Auto-Encoder Model with Prosodic Regularization," in *2023 IEEE International Conference on Computer Vision and Machine Intelligence (CVMI)*, Gwalior, India, Dec. 2023, pp. 1–6, https://doi.org/10.1109/cvmi59935.2023.10464855.

[9] S. Sadok, S. Leglaive, L. Girin, X. Alameda-Pineda, and R. Séguier, "A multimodal dynamical variational autoencoder for audiovisual speech representation learning," *Neural Networks*, vol. 172, Apr. 2024, Art. no. 106120, https://doi.org/10.1016/j.neunet.2024.106120.

[10] S. Latif, R. Rana, S. Khalifa, R. Jurdak, J. Qadir, and B. Schuller, "Survey of Deep Representation Learning for Speech Emotion Recognition," *IEEE Transactions on Affective Computing*, vol. 14, no. 2, pp. 1634–1654, Apr. 2023, https://doi.org/10.1109/taffc.2021.3114365.

[11] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," in *Proceedings of the 18th ACM international conference on Multimedia*, Firenze, Italy, Oct. 2010, pp. 1459–1462, https://doi.org/10.1145/1873951.1874246.

[12] M. S. Likitha, S. R. R. Gupta, K. Hasitha, and A. U. Raju, "Speech based human emotion recognition using MFCC," in *2017 International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET)*, Chennai, India, Mar. 2017, pp. 2257–2260, https://doi.org/10.1109/wispnet.2017.8300161.

[13] S. Patnaik, "Speech emotion recognition by using complex MFCC and deep sequential model," *Multimedia Tools and Applications*, vol. 82, no. 8, pp. 11897–11922, Mar. 2023, https://doi.org/10.1007/s11042-022-13725-y.

[14] Y. Chen, J. Liu, L. Peng, Y. Wu, Y. Xu, and Z. Zhang, "Auto-Encoding Variational Bayes," *Cambridge Explorations in Arts and Sciences*, vol. 2, no. 1, Feb. 2024, https://doi.org/10.61603/ceas.v2i1.33.

[15] J. M. Wu and P. H. Hsu, "Annealed Kullback–Leibler divergence minimization for generalized TSP, spot identification and gene sorting," *Neurocomputing*, vol. 74, no. 12–13, pp. 2228–2240, Jun. 2011, https://doi.org/10.1016/j.neucom.2011.03.002.

[16] S. Wu, T. H. Falk, and W. Y. Chan, "Automatic speech emotion recognition using modulation spectral features," *Speech Communication*, vol. 53, no. 5, pp. 768–785, May 2011, https://doi.org/10.1016/j.specom.2010.08.013.

[17] S. R. Bowman, L. Vilnis, O. Vinyals, A. Dai, R. Jozefowicz, and S. Bengio, "Generating Sentences from a Continuous Space," in *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, Berlin, Germany, 2016, https://doi.org/10.18653/v1/k16-1002.

[18] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization." arXiv, Jan. 30, 2017, https://doi.org/10.48550/arXiv.1412.6980.

[19] X. Wang, "Toward Domain Adaptive Learning-Based Variation Autoencoder Emotional Analysis in English Teaching," *International Journal on Artificial Intelligence Tools*, vol. 33, no. 07, Nov. 2024, https://doi.org/10.1142/s0218213024400062.