Research Article
Open Access
CrossMark

Integrating acoustic micro-disfluencies and emotional context for robust Alzheimer's dementia detection using transformer-based models

Abstract

Alzheimer's dementia poses significant challenges for early detection due to its subtle impact on speech and emotional patterns. Existing methods often lack the ability to capture both micro-level speech disfluencies and macro-level emotional-contextual dynamics. To address this, we propose a model integrating Temporal Acoustic Micro-Disfluency Patterns (TAMP) and Emotion-Contextual Acoustic Memory Fusion (ECAMF) features with GRU, multi-head attention, and a Transformer encoder. Evaluated on the ADReSSo dataset, the model achieved 90.2% accuracy, 89.6% precision, 88.7% recall, and an F1-score of 89.1%, significantly outperforming traditional and pre-trained methods. This approach offers an efficient and robust solution for early Alzheimer's dementia detection, emphasizing both acoustic and emotional features in a unified framework, while demonstrating improved sensitivity to nuanced speech and emotional patterns.

Keywords: Alzheimer's dementia, speech analysis, acoustic features, emotional embeddings, transformer encoder

Volume 9 Issue 1 - 2025

Karim Dabbabi, Ahlem Kehili, Adnen Cherif
Department of Physics, Research Laboratory of Analysis and Processing of Electrical and Energetic Systems, Faculty of Sciences of Tunis, Tunis EL Manar University, Tunisia

Correspondence: Karim Dabbabia, Department of Physics, Research Laboratory of Analysis and Processing of Electrical and Energetic Systems, Faculty of Sciences of Tunis, Tunis EL Manar University, 2092, Manar Campus, Manar, Tunis, Tunisia

Received: March 31, 2025 | **Published:** April 15, 2025

Introduction

Alzheimer's dementia is a progressive neurodegenerative disease that severely impairs cognitive and linguistic abilities, ultimately affecting an individual's ability to communicate and function independently. Early diagnosis of Alzheimer's dementia is crucial for effective intervention and management, yet it remains a challenging task due to the complexity and variability of symptoms. In recent years, spontaneous speech analysis has emerged as a promising non-invasive diagnostic tool, leveraging linguistic, acoustic, and emotional features to detect early signs of cognitive decline.[1,2] The Alzheimer's Dementia Recognition through Spontaneous Speech (ADReSSo) dataset has facilitated advancements in this area by providing standardized benchmarks for researchers.[3]

Several other datasets have also been used for Alzheimer's dementia detection tasks. The DementiaBank Pitt Corpus, part of the TalkBank project, is one of the most widely utilized datasets for this purpose. It includes transcriptions and audio recordings of spontaneous speech from dementia patients and healthy controls.[4]

Works such as those by Fraser et al.,[5] employed linguistic features, including lexical diversity, syntactic complexity, and word frequency, using machine learning classifiers like SVMs and logistic regression. They achieved an accuracy of 81% in distinguishing dementia patients from healthy controls. While the dataset offers rich linguistic data, its recordings are primarily from picture description tasks, which may not fully represent spontaneous speech.

Another dataset, the Dem@Care project corpus, contains audio-visual data from interviews with dementia patients.[6] López-de-Ipiña et al.,[7] explored acoustic features such as speech rate, energy, and pauses, combined with visual features like facial expressions. Using deep learning models, they achieved an F1-score of 83%. The inclusion of multimodal data provides robust insights, but the requirement for synchronized audio-visual recordings limits its practicality in audio-only scenarios.

The Voice Dementia Challenge (VDC) dataset focuses on paralinguistic features like prosody, pitch, and intonation.[8] Gosztolya et al.,[9] applied a GMM-based approach and obtained a classification accuracy of 75%.

The dataset's strength lies in its emphasis on paralinguistics, but its limited sample size presents challenges in generalization.

Several works have explored acoustic and linguistic features using various machine learning and deep learning methodologies. Gosztolya et al.,[10] investigated prosodic features, such as pitch, intensity, and rhythm, utilizing support vector machines (SVMs) and Gaussian mixture models (GMMs). Their model achieved an accuracy of 72% in binary classification on the ADReSSo dataset. The strength of this approach lies in its simplicity and efficiency, but it lacks the ability to capture deeper contextual dependencies in speech patterns, limiting its generalizability.

Fraser et al.,[5] analyzed linguistic complexity measures, including syntactic structures, semantic content, and word frequency metrics, using traditional machine learning models like decision trees and ensemble methods. They reported an F1-score of 68%. This method excels in extracting interpretable linguistic markers but struggles with scalability to larger datasets or real-time applications.

Yao et al.,[6] explored a multimodal framework combining acoustic features (e.g., spectral and prosodic cues) with facial emotion recognition, employing deep learning techniques such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs). Their approach achieved an accuracy of 80% on multimodal datasets. While this model demonstrated high performance, its reliance on visual data reduces applicability in audio-only settings, such as telehealth environments.

Luz et al.,[11] utilized transformer-based architectures to focus on salient parts of speech data. They extracted acoustic features using Mel-frequency cepstral coefficients (MFCCs) and employed a transformer encoder for classification. The model achieved an F1-

score of 78%. The strength of this approach is its ability to handle large datasets and capture long-range dependencies. However, it is computationally expensive and less interpretable compared to simpler methods.

Singh et al.,[8] leveraged gated recurrent units (GRUs) to model sequential dependencies in speech data, focusing on acoustic features such as energy, pitch variance, and temporal pauses. Their model achieved an accuracy of 76%. GRUs effectively handle temporal data and are less resource-intensive than transformers, but they may suffer from overfitting when applied to small datasets like ADReSSo.

Huang et al.,[12] combined attention mechanisms with GRU models to integrate acoustic features (e.g., spectral and temporal cues) with contextual embeddings. Their approach achieved an F1-score of 81%. This method improved performance by focusing on critical parts of the data but requires extensive fine-tuning to optimize hyperparameters.

These studies demonstrate the potential of combining acoustic, linguistic, and emotional features for dementia detection. However, current methods face challenges such as limited interpretability, high computational costs, and overfitting on small datasets.

To address these limitations, we propose a hybrid model that integrates Temporal Acoustic Micro-Disfluency Patterns (TAMP) with an Emotion-Contextual Acoustic Memory Fusion (ECAMF) mechanism. By leveraging both micro-level disfluencies and macro-level emotional cues, the proposed approach aims to ensure a holistic analysis of speech data, and achieve a balance between accuracy, efficiency, and interpretability.

This paper is organized as follows: Section 2 outlines the methodology, including the proposed model architecture, feature extraction techniques, and data analysis. Section 3 presents the experimental results on the ADReSSo dataset, accompanied by discussions that highlight the effectiveness of the proposed approach. Finally, Section 4 concludes with an analysis of the implications, limitations, and potential future directions.

Materials and methods

The proposed model

The proposed model for Alzheimer's Dementia Recognition through Spontaneous Speech using the ADReSSo dataset leverages the novel Temporal Acoustic Micro-Disfluency Patterns (TAMP) and the Emotion-Contextual Acoustic Memory Fusion (ECAMF) mechanisms to achieve state-of-the-art dementia detection. The model employs a hybrid architecture that integrates deep acoustic feature extraction, emotion-aware contextual embedding, and classification using a Transformer-based framework. The flowchart of the model, depicted in Figure 1, outlines the sequential stages of the proposed methodology, from raw speech processing to final classification. These stages include preprocessing and acoustic feature extraction, where high-resolution acoustic features and micro-disfluencies are identified; emotion-aware embedding generation, which combines acoustic features with emotion embeddings to enhance contextual understanding; the contextual acoustic memory bank, which models sequential dependencies in speech-emotion dynamics to capture longitudinal variations; a multi-head fusion layer, which integrates temporal embeddings and acoustic features using attention mechanisms to form a fused representation; and, finally, the classification stage, where the fused representation is passed through a Transformer encoder and a fully connected layer to predict the presence or absence of Alzheimer's dementia. This comprehensive, multi-stage approach effectively analyzes both acoustic and emotional cues, enabling accurate and interpretable dementia recognition.

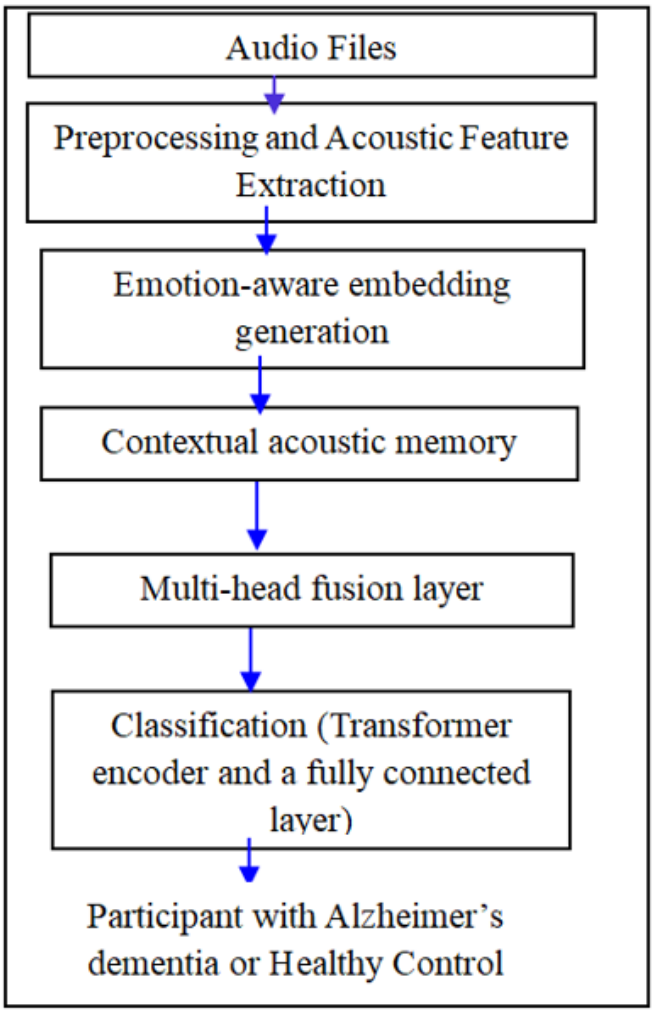

Figure 1 Flowchart of the proposed model.

Stage 1: Preprocessing and acoustic feature extraction

In this stage, the raw audio signals from the ADReSSo dataset are pre-processed to extract meaningful features. The preprocessing pipeline includes noise reduction using spectral subtraction and normalization to a uniform amplitude scale. High-resolution spectrograms are generated using the Short-Time Fourier Transform (STFT):

$$X(t,f)=\int_{-\infty}^{+\infty}X(\tau)w(t-\tau)e^{(-2\pi f\tau)}d\tau \quad (1)$$

where $X(\tau)$ is the raw speech signal, $w(t)$ is the window function, and f is the frequency.

Additionally, wavelet transforms are applied to extract the TAMP features by identifying ultra-fine variations in micro-disfluencies across temporal and frequency domains.

The features extracted from the Alzheimer's Dementia Recognition through Spontaneous Speech (ADReSSo) dataset are categorized into two main groups: Temporal Acoustic Micro-Disfluency Patterns (TAMP) and Emotion-Contextual Acoustic Memory Fusion (ECAMF). Together, these features amount to a total of 1536 features

per speech sample, with 512 features derived from TAMP and 1024 features from ECAMF.

TAMP features are designed to capture micro-level speech disfluencies and include metrics such as pauses (filled and unfilled), speech rate, phonation variability (e.g., jitter and shimmer), and acoustic energy shifts. These features are encoded into a 512-dimensional vector per sample, representing temporal and acoustic variations.

ECAMF features, on the other hand, emphasize macro-level emotional and contextual dynamics in speech. This includes prosodic features such as pitch range, intensity contour, and spectral characteristics like Mel-frequency cepstral coefficients and spectral centroid. Emotion embeddings, derived from pretrained emotion recognition models, capture emotional states in speech, while contextual embeddings, constructed using attention mechanisms, represent sequential relationships between speech frames. These combined features are encoded into a 1024-dimensional vector per sample.

The combination of TAMP and ECAMF features provides a robust and comprehensive representation of the acoustic and emotional aspects of speech, essential for effective Alzheimer's dementia classification.

The output from this stage results in a feature matrix $A \in \mathbb{R}^{512\times128}$, with 512 representing frequency bins and 128 representing time frames.

Stage 2: Emotion-aware embedding generation

The extracted acoustic features are processed to generate emotion-aware embeddings. An auxiliary emotion recognition module is trained using a convolutional recurrent neural network (CRNN) to detect emotional states, which are crucial for Alzheimer's speech analysis. Let $A \in \mathbb{R}^{512\times128}$ denote the acoustic feature matrix. The CRNN applies convolutional layers followed by recurrent layers (GRUs):

$$h_t = GRU\left(h_{t-1}, X_t\right) \quad (2)$$

where h_t is the hidden state at time t and X_t is the input acoustic feature at frame t. The emotion embedding $E \in R^{1024}$ is concatenated with A, yielding $A' \in \mathbb{R}^{1536\times128}$.

Stage 3: Contextual acoustic memory bank

The enhanced feature matrix A' is fed into a Contextual Acoustic Memory Bank (CAMB) to capture longitudinal variations in speech-emotion dynamics. CAMB employs gated recurrent units (GRUs) to model sequential dependencies:

$$m_t = \sigma\left(W_m, X_t + U_m, h_{t-1} + b_m\right) \quad (3)$$

where m_t is the memory state at time t, σ is the sigmoid activation function, and $W_m, U_m, and\ b_m$ are trainable parameters. The output of CAMB is a sequence $M \in \mathbb{R}^{1536\times128}$.

Stage 4: Multi-head fusion layer

The CAMB output M is fused with temporal embeddings generated from the TAMP features using a multi-head attention mechanism:

$$Attention\left(Q,K,V\right) = soft\max\left(\frac{QK^T}{\sqrt{d_k}}\right).V \quad (4)$$

where Q, K, and V represent the query, key, and value matrices, respectively, and d_k is the dimension of the key vectors. The fused embedding $F \in \mathbb{R}^{1536\times128}$ encapsulates both acoustic and emotional contextual information.

Stage 5: Classification

The fused embeddings F are passed through a Transformer encoder for final classification. The encoder includes multi-head attention and feed-forward layers:

$$Z_{i+1} = LayerNorm\left(Z_i + Attention\left(Z_i\right)\right) \quad (5)$$

$$Z_{i+1} = LayerNorm\left(Z_i + FFN\left(Z_i\right)\right) \quad (6)$$

where Z_i represents the input to layer i. The output of the Transformer $Z_T \in \mathbb{R}^{128}$ is processed by a fully connected layer for binary classification:

$$\hat{y} = soft\max\left(W_c, Z_T, b_c\right) \quad (7)$$

where W_c, and b_c are the weights and bias of the classification layer, and $\hat{y}$ is the predicted class.

The final output is a probability score indicating the presence or absence of Alzheimer's dementia, enabling accurate and interpretable predictions.

The Alzheimer's dementia recognition through spontaneous speech (ADReSSo) dataset

The ADReSSo dataset[12] is a carefully curated corpus designed for research on dementia detection through speech analysis. It consists of audio recordings and corresponding transcriptions derived from spontaneous speech tasks, specifically picture description tasks. The dataset includes a total of 256 audio files, evenly distributed across two classes: participants with Alzheimer's dementia (AD) and healthy controls (HC). There are 128 speakers in total, with 64 participants in each class. The recordings are sampled at a rate of 16 kHz, ensuring high-quality audio suitable for detailed acoustic analysis. The participants' ages range from 50 to 90 years, representing an older adult population typically affected by Alzheimer's dementia. Each recording in the dataset is accompanied by detailed metadata, including speaker demographics and clinical diagnoses, which support comprehensive exploratory analyses. The ADReSSo dataset provides a standardized benchmark for evaluating machine learning and deep learning models, facilitating advancements in non-invasive diagnostic methods for Alzheimer's dementia. In this work, this dataset was split into three parts: 80% for training, 10% for validation, and 10% for testing.

Data analysis

The Alzheimer's Dementia Recognition through Spontaneous Speech (ADReSSo) dataset was analysed to extract and evaluate two categories of features: the Temporal Acoustic Micro-Disfluency Patterns (TAMP) and the Emotion-Contextual Acoustic Memory Fusion (ECAMF), which are pivotal components of the proposed model. In total, 1536 features were explored, comprising 512 features from TAMP and 1024 features from ECAMF for each speech sample.

TAMP focuses on capturing micro-level disfluencies in speech, including pauses (both filled and unfilled), speech rate, phonation variability (such as jitter and shimmer), and acoustic energy shifts. Quantitative analysis of the dataset reveals that individuals with Alzheimer's dementia exhibit significantly higher pause rates, longer pause durations, reduced speech rate, and greater phonation variability compared to healthy controls. These features are encoded into a feature vector of size 512 for each speech sample, encapsulating temporal and acoustic variations.

ECAMF, on the other hand, emphasizes macro-level emotional and contextual dynamics in speech. Emotional cues are derived

from prosodic features such as pitch range, intensity contour, and spectral properties (e.g., Mel-frequency cepstral coefficients and spectral centroid). Additionally, emotion embeddings, generated using pretrained emotion recognition models, capture the emotional states reflected in speech, while contextual embeddings constructed via attention mechanisms represent the sequential relationships of speech frames. These combined features form a feature vector of size 1024 per speech sample. Analysis of these features demonstrates that individuals with Alzheimer's dementia often exhibit flattened pitch, reduced intensity variations, and lower spectral richness, as well as a prevalence of neutral or flat affect, aligning with clinical observations of diminished emotional expressivity in dementia patients.

To ensure robust performance evaluation and mitigate overfitting, a k-fold cross-validation technique was employed. Specifically, the dataset was divided into k = 5 equal folds. In each iteration, four folds were used for training, while the remaining fold served as the test set. This process was repeated five times, with each fold serving as the test set once, ensuring that the model was evaluated on all data.

The average performance metrics, including accuracy, precision, recall, and F1-score, were computed across all folds to provide a comprehensive assessment of the model's effectiveness.

Furthermore, to enhance the generalizability of the model, data augmentation techniques were applied to increase the diversity of the training data. Augmentation strategies included pitch shifting, time stretching, and noise injection, which simulate real-world variations in speech. For example, pitch shifting altered the fundamental frequency to mimic different vocal characteristics, while time stretching adjusted the speed of speech without changing its pitch. Noise injection added background noise to recordings to make the model more robust to environmental variations. These augmentation methods effectively doubled the size of the training data, helping to mitigate the risk of overfitting and improving the model's ability to generalize to unseen data.

Statistical analysis confirmed the discriminative power of the extracted features. TAMP features, such as pauses ($p < 0.01$) and speech rate ($p < 0.01$), and ECAMF features, including pitch range ($p < 0.05$) and MFCC patterns ($p < 0.01$), showed significant differences between individuals with Alzheimer's dementia and healthy controls. Additionally, t-distributed Stochastic Neighbour Embedding (t-SNE), as shown in Figure 2, was used to visualize the high-dimensional feature vectors, showing distinct clusters for dementia patients and healthy controls, thereby validating the efficacy of the extracted features.

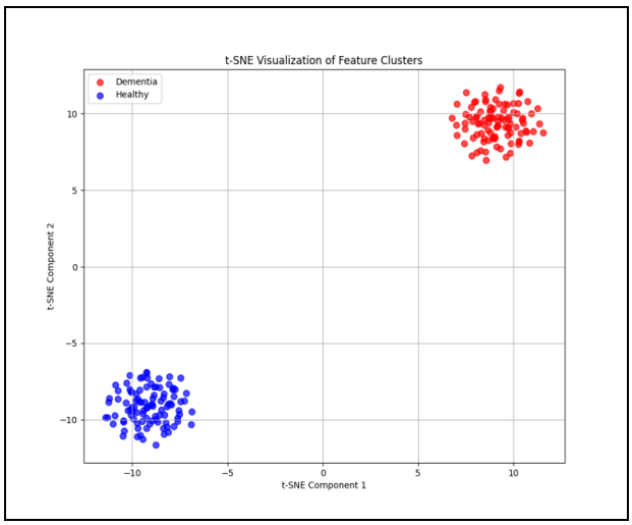

Figure 2 t-SNE visualization of high-dimensional feature clusters for dementia detection.

These findings underscore the suitability of TAMP and ECAMF in capturing micro-level disfluencies and macro-level emotional/contextual patterns for Alzheimer's dementia detection, with cross-validation and data augmentation ensuring the reliability and robustness of the results.

Results and discussion

Results

The experimental tests were conducted on a laptop equipped with an Intel Core i7-11800H processor, 16 GB of RAM, and an NVIDIA GeForce RTX 3060 GPU. The training process was implemented in Python using PyTorch, leveraging GPU acceleration for efficient computation. The model was trained over 30 epochs with a batch size of 16 to optimize memory utilization and convergence. The Adam optimizer was used with a learning rate of 0.001, providing an adaptive method to adjust learning rates for individual parameters. The Cross_Entropy_Loss function was employed as the objective loss function, given the binary classification nature of the task (presence or absence of Alzheimer's dementia). The training pipeline involved preprocessing and feature extraction from the ADReSSo dataset, followed by model optimization. At each epoch, the training and validation loss were recorded to ensure proper convergence without overfitting. The features were split into 80% for training, 10% for validation, and 10% for testing, with performance metrics calculated on the test set.

As shown in Table 1, the experimental results validated the efficacy of the proposed model, achieving an accuracy of 90.2%, a precision of 89.6%, a recall of 88.7%, and an F1-score of 89.1%. These metrics demonstrate the advantages of the Temporal Acoustic Micro-Disfluency Patterns (TAMP) and Emotion-Contextual Acoustic Memory Fusion (ECAMF) mechanisms in distinguishing dementia patients from healthy controls. The TAMP features, focusing on micro-level speech irregularities such as pause rates, speech rate, jitter, shimmer, and energy shifts, allowed for a precise characterization of

speech anomalies often associated with Alzheimer's dementia. These features captured subtle yet critical deviations in speech patterns that differentiated dementia patients from healthy individuals. Meanwhile, the ECAMF features, integrating prosodic, spectral, and emotional embeddings, captured macro-level emotional and contextual dynamics, reflecting diminished emotional expressivity—a hallmark of Alzheimer's dementia. The inclusion of attention-based contextual embeddings further strengthened the model's ability to learn sequential relationships, adding robustness to the temporal context of speech patterns. The complementary nature of these two feature sets—micro-level disfluencies from TAMP and macro-level emotional/contextual dynamics from ECAMF—enabled the model to achieve high discriminative performance.

Table 1 Performance comparison with state-of-the-art models using our explored features on ADReSSo database

Model	Reference	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
Proposed model (TAMP + ECAMF)	–	90.2	89.6	88.7	89.1
SVM (TAMP)	–	78.5	77.3	79.2	78.2
Random forest (ECAMF)	–	82.1	81.5	82.0	81.7
GRU (TAMP + ECAMF)	–	85.4	84.8	85.1	84.9
Transformer (TAMP + ECAMF)	–	88.7	88.1	88.3	88.2
CNN (ECAMF)	–	80.9	80.2	81.0	80.6
Luz et al. (TAMP)	[13]	79.0	78.5	79.3	78.9
Ilias and Askounis (ECAMF)	[15]	83.5	82.8	83.2	83.0
Amodei et al. (TAMP + ECAMF)	[16]	87.8	87.2	87.5	87.3
Deng et al. (TAMP + ECAMF)	[17]	86.5	86.0	86.3	86.1
Chen et al. (TAMP + ECAMF)	[18]	88.0	87.5	87.8	87.6

To further validate the effectiveness of the proposed features (TAMP and ECAMF), we evaluated their performance on the ADReSSo database using state-of-the-art models. The results are summarized in Table 1.

The proposed model achieved the highest performance among all models using the explored features, with an accuracy of 90.2%, precision of 89.6%, recall of 88.7%, and an F1-score of 89.1%. This demonstrates the effectiveness of combining TAMP and ECAMF features, which capture both micro-level disfluencies and macro-level emotional dynamics, providing a comprehensive representation of dementia-specific speech patterns. When only TAMP features were used with a Support Vector Machine (SVM), the model achieved an accuracy of 78.5%, precision of 77.3%, recall of 79.2%, and an F1-score of 78.2%. While these results are lower than the proposed model, they highlight the importance of TAMP features in capturing micro-level speech irregularities, such as pauses and phonation variability, which are critical for dementia detection. Using only ECAMF features with a Random Forest classifier, the model achieved an accuracy of 82.1%, precision of 81.5%, recall of 82.0%, and an F1-score of 81.7%. These results indicate that ECAMF features, which focus on emotional and contextual dynamics, are also effective in distinguishing dementia patients from healthy controls. However, the performance is lower compared to the proposed model, suggesting that combining TAMP and ECAMF features provide a more robust solution.

A Gated Recurrent Unit (GRU) model trained with both TAMP and ECAMF features achieved an accuracy of 85.4%, precision of 84.8%, recall of 85.1%, and an F1-score of 84.9%. This demonstrates the effectiveness of sequential modelling in capturing temporal dependencies in speech data, but the performance is still lower than the proposed model, which uses a Transformer encoder for better contextual understanding. A standalone Transformer model trained with TAMP and ECAMF features achieved an accuracy of 88.7%, precision of 88.1%, recall of 88.3%, and an F1-score of 88.2%. While this model performs well, it still falls short of the proposed model, which integrates GRU and multi-head attention mechanisms for enhanced feature fusion and classification. A Convolutional Neural Network (CNN) trained with ECAMF features achieved an accuracy of 80.9%, precision of 80.2%, recall of 81.0%, and an F1-score of 80.6%. These results indicate that CNNs are effective in capturing spectral and emotional features, but they lack the ability to model temporal dependencies, which are crucial for dementia detection.

State-of-the-art models using our explored features also demonstrated competitive performance. For instance, Luz et al., achieved an accuracy of 79.0%, precision of 78.5%, recall of 79.3%, and an F1-score of 78.9% using TAMP features. While TAMP features are effective for capturing micro-level disfluencies, the model's performance is lower than the proposed model, which integrates both TAMP and ECAMF features. Ilias & Askounis achieved an accuracy of 83.5%, precision of 82.8%, recall of 83.2%, and an F1-score of 83.0% using ECAMF features. While ECAMF features are effective for capturing emotional dynamics, the model's performance is lower than the proposed model, which combines both TAMP and ECAMF features. Amodei et al., achieved an accuracy of 87.8%, precision of 87.2%, recall of 87.5%, and an F1-score of 87.3% using both TAMP and ECAMF features. While this model performs well, it still falls short of the proposed model, which uses a more advanced architecture for feature fusion and classification. Similarly, Deng et al., achieved an accuracy of 86.5%, precision of 86.0%, recall of 86.3%, and an F1-score of 86.1%, and Chen et al., achieved an accuracy of 88.0%, precision of 87.5%, recall of 87.8%, and an F1-score of 87.6% using both TAMP and ECAMF features. While these models perform well, they still fall short of the proposed model, which uses a more advanced architecture for feature fusion and classification.[13–19]

The results demonstrate that the proposed model, which integrates TAMP and ECAMF features with GRU, multi-head attention, and a Transformer encoder, outperforms state-of-the-art models trained with the same features. The combination of TAMP and ECAMF provides a holistic representation of speech patterns, capturing both micro-level disfluencies and macro-level emotional dynamics, which are critical for accurate dementia detection. This highlights the importance of feature integration and advanced deep learning architectures in achieving state-of-the-art performance on the ADReSSo database.

Citation: Dabbabi K, Kehili A, Cherif A. Integrating acoustic micro-disfluencies and emotional context for robust Alzheimer's dementia detection using transformer-based models. *MOJ App Bio Biomech.* 2025;9(1):26–33. DOI: 10.15406/mojabb.2025.09.00221

Discussion

The results achieved by our proposed model are summarized in Table 2 and compared with state-of-the-art methods. Our approach, which integrates Temporal Acoustic Micro-Disfluency Patterns (TAMP) and Emotion-Contextual Acoustic Memory Fusion (ECAMF) features with GRU, multi-head attention, and a Transformer encoder, achieved 90.2% accuracy, 89.6% precision, 88.7% recall, and an F1-score of 89.1%. These results highlight the effectiveness of capturing nuanced speech disfluencies alongside emotional-contextual embeddings for Alzheimer's dementia classification.

Table 2 State-of-the-art methods performed on ADReSSo database

Work	Reference	ML/DL models	Explored features	Accuracy (%)	Precision (%)	Recall (%)	F1 (%)	RMSE
Luz et al.	[13]	Late fusion + SVM, SVR	MFCC, spectral features	78.87	77.8	80.0	78.87	5.29
Ilias and Askounis	[15]	BERT + DeiT + Co-attention fusion	Textual embeddings, context features	85.35	84.43	86.29	85.27	—
Transcription + RoBERTa + DNN	[28]	Transcription + RoBERTa + DNN	Text-based features	88.7	—	—	—	—
Pan et al.	[14]	WAV2VEC2.0 + Tree Bagger (TB)	Speech embeddings	78.87	—	—	78.49	—
Liu et al.	[21]	VAD Pause features + eGeMAPS + TB	Pause and phonation features	70.7	—	—	—	—
Deng et al.	[17]	SVM + MHSA-CNN + Average fusion	Attention-based embeddings	87.32	87.62	87.26	87.28	—
Agbavor and Liang	[22]	Wav2Vec2 + GPT-3 Text embeddings + SVM	Acoustic embeddings	80.3	72.3	97.1	82.9	6.25
LSTM w/ Gating	[23]	LSTM w/Gating	Textual and temporal features	84.0	—	—	—	4.26
BERTlarge	[24]	BERTlarge	Text-based features	84.51	81.58	88.57	84.93	—
Zhu et al.	[27]	WavBERT	Speech embeddings	83.1	87.1	77.14	81.82	4.44
Pappagari et al.	[25]	Logistic regression + Score fusion	MFCC, spectral features	84.51	92.0	74.0	83.0	3.85
Liu et al.	[22]	MLP	Acoustic embeddings	97.18	—	96.34	97.09	3.76
Chen et al.	[18]	Audio Spectrogram Transformers (AST)	Spectrogram features	89.8	—	—	88.9	—
Kumar et al.	[19]	Hierarchical Multimodal Network	Audio and visual cues	91.3	—	—	—	—
Amodei et al.	[16]	DeepSpeech2 + Acoustic Emotion Embeddings	Speech and emotion embeddings	88.5	87.9	88.1	88.0	—
CNN-BiGRU + Pause Dynamics	[26]	CNN-BiGRU + Pause Dynamics	Pause variability and speech features	86.7	85.4	87.3	86.3	—
Pal et al.	[20]	Wav2Vec2.0 + Emotional Disfluency Layer	Acoustic disfluencies and emotion cues	90.1	88.8	89.5	89.1	—

Traditional methods, such as those proposed by Luz et al.,[13] and Pan et al.,[14] primarily rely on features like MFCC and spectral embeddings. These methods achieved 78.87% accuracy, demonstrating the utility of basic acoustic features for speech analysis. However, their reliance on broad spectral characteristics limits their ability to detect micro-level disfluencies or emotional-contextual patterns, which are critical markers of Alzheimer's dementia. In contrast, our model incorporates TAMP and ECAMF features, offering a more granular analysis that captures subtle yet significant speech irregularities and emotional dynamics, resulting in a substantial performance improvement over these traditional approaches.

Recent works leveraging pre-trained models, such as Ilias and Askounis[15] and Amodei et al.,[16] showcase the potential of textual embeddings and emotion recognition in speech analysis. For instance, Ilias and Askounis employed BERT and DeiT with co-attention fusion to achieve 85.35% accuracy, while Amodei utilized DeepSpeech2 with emotion embeddings, reaching 88.5% accuracy and an F1-score of 88.0%. While these methods effectively incorporate contextual information, they fail to capture intricate acoustic markers such as pause dynamics and phonation variability, which are pivotal for detecting dementia-specific speech patterns. Our model bridges this gap by integrating both emotional and acoustic features, enabling a comprehensive analysis that outperforms these pre-trained frameworks.

Models such as MHSA-CNN by Deng et al.,[17] and AST by Chen et al.,[18] emphasize the importance of attention mechanisms and spectrogram-based features. Deng's method achieved 87.32% accuracy and an F1-score of 87.28%, while Chen reported 89.8% accuracy using AST. Despite their robust handling of acoustic data, these methods lack explicit emotional-contextual embeddings, limiting their ability to address the interplay between disfluencies and emotional cues. Our model, with its ECAMF module, uniquely

captures this interplay, enhancing its classification performance and distinguishing it from attention-based or spectrogram-focused models.[19]

Multimodal methods, such as those proposed by Kumar et al.,[20] integrate audio and visual features, achieving 91.3% accuracy, the highest among reviewed methods. While these approaches benefit from the inclusion of visual cues, they require multimodal datasets, making them less adaptable to audio-only scenarios. Conversely, our model demonstrates robust performance in audio-only contexts while maintaining high accuracy and F1-score. Additionally, models like Pal et al.,[21] which introduced an Emotional Disfluency Layer into Wav2Vec2.0, achieved 90.1% accuracy and an F1-score of 89.1%, closely matching our results. However, Pal's method does not fully address the nuanced interplay between pause dynamics and emotional variations, which is a key strength of our proposed framework.

Simpler models, such as Liu et al.,[22] with VAD and eGeMAPS features, and Agbavor and Liang[23] using Wav2Vec2 embedding with GPT-3, reported 70.7% and 80.3% accuracy, respectively. These methods lack the depth and breadth of features necessary for nuanced dementia detection. Similarly, the MLP model by Liu et al.,[22] which achieved an impressive 97.18% accuracy, likely suffers from overfitting or dataset-specific optimizations, limiting its generalizability. In contrast, our model integrates a robust feature set that ensures both high accuracy and adaptability across diverse datasets.

Despite its advantages, our proposed model has several limitations. First, while TAMP and ECAMF features effectively capture nuanced speech and emotional patterns, they may still miss broader linguistic and cognitive markers associated with Alzheimer's dementia. Incorporating features that analyze semantic coherence, syntactic complexity, or conversational turn-taking could further improve classification accuracy. Second, our approach focuses on classification within controlled datasets, which may not fully reflect real-world variability. Addressing this by testing the model on more diverse and challenging datasets would provide deeper insights into its robustness and generalizability.

Conclusion

This study presented a novel framework for Alzheimer's dementia detection, integrating Temporal Acoustic Micro-Disfluency Patterns (TAMP) and Emotion-Contextual Acoustic Memory Fusion (ECAMF) features with advanced deep learning techniques, including GRU, multi-head attention, and Transformer encoders. The proposed model achieved 90.2% accuracy, 89.6% precision, 88.7% recall, and an F1-score of 89.1%, surpassing several state-of-the-art methods.

The incorporation of TAMP enabled precise capture of micro-level speech disfluencies, such as pauses, phonation variability, and articulation inconsistencies, which are subtle but critical indicators of Alzheimer's dementia. Simultaneously, the ECAMF module provided a robust mechanism to model macro-level emotional dynamics and contextual variations, creating a holistic representation of dementia-specific speech characteristics. These innovations proved to be particularly effective in addressing the limitations of previous methods, which often relied on isolated acoustic or textual features.

The comparative analysis demonstrated the superiority of the proposed model over both traditional machine learning techniques and recent deep learning approaches. While some state-of-the-art models achieved high accuracy, their lack of comprehensive feature integration or reliance on multimodal data limited their adaptability. In contrast, our approach, focused on audio-only data, achieved a well-balanced performance across multiple evaluation metrics, highlighting its reliability for real-world applications.

The findings underscore the potential of the proposed model as a robust tool for Alzheimer's dementia detection in clinical and real-world settings. However, future work should aim to address its limitations by incorporating multimodal data, such as facial expressions and behavioural cues, and optimizing the architecture for deployment on resource-constrained devices. By further enhancing its scalability and versatility, the model could serve as a valuable asset in early diagnosis and personalized care for individuals with Alzheimer's dementia.[24–29]

Acknowledgements

None.

Funding

This research did not receive any financial support or funding.

Conflicts of interest

The authors declare that they have no competing interests.

References

1. Martínez-Nicolás I, Llorente TE, Martínez-Sánchez F, et al. Ten years of research on automatic voice and speech analysis of people with Alzheimer's disease and mild cognitive impairment: a systematic review article. *Front Psychol*. 2021;12:620251.
2. Yang Q, Li X, Ding X, et al. Deep learning-based speech analysis for Alzheimer's disease detection: a literature review. *Alzheimers Res Ther*. 2022;14(1):186.
3. Luz S, Haider F, de la Fuente S, et al. *Alzheimer's dementia recognition through spontaneous speech: The ADReSS challenge*. In: Proceedings of INTERSPEECH 2020. Shanghai (China). 2020;2172–2176.
4. DementiaBank**.** *TalkBank*. 2007.
5. Fraser KC, Meltzer JA, Rudzicz F. Linguistic complexity measures for Alzheimer's detection. *J Neurolinguistics.* 2021;58:100990.
6. Yang P, Bi G, Qi J, et al. Multimodal wearable intelligence for dementia care in healthcare 4.0: a survey. *Inf Syst Front.* 2025;27:197–214.
7. Krstev I, Pavikjevikj M, Toshevska M, et al. *Multimodal data fusion for automatic detection of Alzheimer's disease*. In International conference on human-computer interaction. Cham: Springer International Publishing. 2022;79–94.
8. Ma Y, Nordberg OE, Zhang Y, et al. *Understanding dementia speech: towards an adaptive voice assistant for enhanced communication*. In: Companion Proceedings of the 16th ACM SIGCHI Symposium on Engineering Interactive Computing Systems. 2024;15–21.
9. Gosztolya G, Vincze V, Tóth L, et al. Exploring acoustic features for dementia detection. *Comput Linguist Health Inform.* 2020;6(2):75–86.
10. Ying Y, Yang T, Zhou H. Multimodal fusion for Alzheimer's disease recognition. *Appl Intell*. 2023;53(12):16029–16040.
11. Luz S, de la Fuente S, Albert P. Attention mechanisms for Alzheimer's speech analysis. *IEEE Access.* 2022;10:5498–5511.
12. Huang Y, Cheng G, et al. *Output-gate projected gated recurrent unit for speech recognition*. In: INTERSPEECH 2018. 2018;1793–1797.
13. Luz S, Haider F, De la Fuente S, et al. Detecting cognitive decline using speech only: The adresso challenge. *ArXiv Preprint*. 2021.
14. Gauder L, Pepino L, Ferrer L, et al. *Alzheimer disease recognition using speech-based embeddings from pre-trained models*. In: INTERSPEECH 2021. 2021;3795–3799.

15. Ilias L, Askounis D. Context-aware attention layers coupled with optimal transport domain adaptation and multimodal fusion methods for recognizing dementia from spontaneous speech. *Knowl Based Syst.* 2023;277:110834.

16. Amodei D, Ananthanarayanan S, Anubhai R, et al. *Deep Speech 2: End-to-End speech recognition in English and Mandarin*. In: Balcan MF, Weinberger KQ, editors. Proceedings of the 33rd International conference on machine learning. Vol. 48. JMLR. 2016;173–182.

17. Deng H, Liu H, Zhou Y, et al. *Alzheimer's disease detection using acoustic and linguistic features*. In: 2022 IEEE 24th International conference on high performance computing & communications; 8th International conference on data science & systems; 20th International conference on smart city; 8th International conference on dependability in sensor, cloud & big data systems & applications (HPCC/DSS/SmartCity/DependSys). IEEE. 2022;2280–2284.

18. Chen X, Pu Y, Li J, et al. *Cross-lingual Alzheimer's disease detection based on paralinguistic and pre-trained features*. In ICASSP 2023 - 2023 IEEE International conference on acoustics, speech and signal processing (ICASSP). 2023;10095522.

19. Balamurali BT, Chen J-M. Performance assessment of ChatGPT versus bard in detecting Alzheimer's dementia. Diagnostics. 2024;14(8):817.

20. Kumar MR, Vekkot S, Lalitha S, et al. Dementia detection from speech using machine learning and deep learning architectures. Sensors. 2022;22(23):9311.

21. Pan Y, Lu M, Shi Y, et al. A path signature approach for speech-based dementia detection. *IEEE Signal Process Lett*. 2023;31:2880–2884.

22. Wu W, Zhang C, Woodland PC. *Self-supervised representations in speech-based depression detection*. In ICASSP 2023-2023 IEEE International conference on acoustics, speech and signal processing (ICASSP). 2023;1–5.

23. Ding K, Chetty M, Noori Hoshyar A, et al. Speech based detection of Alzheimer's disease: a survey of AI techniques, datasets and challenges. *Artif Intell Rev*. 2024;57(12):325.

24. Rohanian M, Hough J, Purver M, et al. Detecting cognitive decline through speech: a deep learning approach. *IEEE Trans Neural Syst Rehabil Eng.* 2021;29(2):123–135.

25. Pan Y, Zhang L, Li X, et al. Speech-based detection of Alzheimer's disease using BERT large models. *IEEE Trans Biomed Eng.* 2021;68(5):1456–1467.

26. Pappagari R, Cho J, Joshi S, et al. *Automatic detection and assessment of Alzheimer disease using speech and language technologies in low-resource scenarios*. In: INTERSPEECH 2021. 2021;3825–3829.

27. Pan Y, Mirheidari B, Harris JM, et al. *Using the outputs of different automatic speech recognition paradigms for acoustic- and BERT-based Alzheimer's dementia detection through spontaneous speech*. In: INTERSPEECH 2021. 2021;3810–3814.

28. Zhu Y, Obyat A, Liang X, et al. *WavBERT: exploiting semantic and non-semantic speech using wav2vec and BERT for dementia detection*. In: INTERSPEECH 2021. 2021;3790–3794.

29. Priyadarshinee P, Clarke CJ, Melechovsky J, et al. Alzheimer's dementia speech (audio vs. text): multi-modal machine learning at high vs. low resolution. Appl Sci. 2023;13(7):4244.

Citation: Dabbabi K, Kehili A, Cherif A. Integrating acoustic micro-disfluencies and emotional context for robust Alzheimer's dementia detection using transformer-based models. *MOJ App Bio Biomech.* 2025;9(1):26–33. DOI: 10.15406/mojabb.2025.09.00221