Open Access Library Journal
2026, Volume 13, e15187
ISSN Online: 2333-9721
ISSN Print: 2333-9705

Deep Learning Analysis of Speech and Language Patterns in Acute Psychiatric Emergencies: A Synthetic Proof-of-Concept Study for a Multimodal NLP Decision Support Framework

Rocco de Filippis[1*], Abdullah Al Foysal[2]

[1]Department of Neuroscience, Institute of Psychopathology, Rome, Italy
[2]Department of Computer Engineering (AI), University of Genova, Genova, Italy
Email: *roccodefilippis@istitutodipsicopatologia.it, niloyhasanfoysal440@gmail.com

How to cite this paper: de Filippis, R. and Al Foysal, A. (2026) Deep Learning Analysis of Speech and Language Patterns in Acute Psychiatric Emergencies: A Synthetic Proof-of-Concept Study for a Multimodal NLP Decision Support Framework. *Open Access Library Journal*, **13**: e15187.
https://doi.org/10.4236/oalib.1115187

Received: March 16, 2026
Accepted: May 26, 2026
Published: May 29, 2026

Abstract

Acute psychiatric emergencies present critical challenges for emergency department clinicians, requiring rapid differentiation of psychosis, agitation, and cognitive disorganization to guide appropriate intervention. Current assessment relies predominantly on clinical interview and behavioural observation, lacking objective biomarkers to support diagnostic decision-making under time constraints. We introduce a deep learning framework integrating natural language processing and neural network architectures to analyse speech features during emergency psychiatric consultations. Our approach employs a domain-aware multi-scale convolutional neural network with cross-attention mechanisms for comprehensive speech feature analysis, alongside bidirectional long short-term memory and transformer architectures for temporal pattern recognition. To enable controlled validation, we generated clinically informed synthetic speech datasets parameterized by established psycholinguistic markers reflecting neurobiological disturbances in acute psychiatric states. The multi-scale CNN achieved perfect discrimination between diagnostic categories (accuracy = 1.000, AUC = 1.000, F1 = 1.000), while temporal models demonstrated robust classification of agitation versus cognitive disorganization (accuracy > 0.980). Beyond classification, we provide interpretable pathway analyses through attention visualization and domain-specific feature importance mapping, facilitating inspection of candidate speech markers underlying diagnostic differentiation. Finally, we outline essential barriers to clinical deployment including synthetic-only validation, the necessity for real-world

multi-site validation across diverse emergency settings, and potential generalization challenges, proposing methodological steps required for robust integration into clinical decision support systems. This work constitutes a synthetic proof-of-concept study. All models were trained and evaluated exclusively on clinically parameterized synthetic data; no real patient speech was analysed. Results cannot be interpreted as evidence of clinical deployability and should not be presented as such. The framework establishes a reproducible methodological foundation for future validation on real emergency department speech recordings.

Subject Areas

Psychiatry & Psychology

Keywords

Psychiatric Emergencies, Deep Learning, Natural Language Processing, Speech Analysis, Psychosis, Agitation, Cognitive Disorganization, Emergency Psychiatry, Clinical Decision Support, Precision Medicine

1. Introduction

Acute psychiatric emergencies affect approximately 5% of all emergency department visits and represent one of the most clinically challenging scenarios in medical practice, with diagnostic uncertainty contributing to prolonged stays, inappropriate interventions, and adverse patient outcomes [1]. The rapid differentiation of psychosis, severe agitation, and cognitive disorganization constitutes a fundamental emergency department dilemma: while these conditions require distinct therapeutic approaches antipsychotics for psychosis, de-escalation and benzodiazepines for agitation, and environmental management for cognitive disorganization clinical presentations often overlap, and misclassification can lead to iatrogenic harm [2]-[4]. Current emergency psychiatric assessment relies primarily on unstructured clinical interviews and rapid behavioural observation, lacking objective biomarkers to support diagnostic decision-making under the severe time constraints typical of emergency care [5]-[7].

From a neurobiological perspective, these three emergency presentations reflect distinct neural circuit dysfunctions that manifest in characteristic speech and language patterns. Psychosis emerges from dysregulated dopaminergic signalling producing formal thought disorder manifesting as tangentiality, loose associations, and neologisms in speech production [8]-[10]. Agitation reflects elevated noradrenergic and corticotropin-releasing hormone activity generating pressured speech, increased volume, and hostile content [11]-[13]. Cognitive disorganization in delirium or severe affective states arises from cholinergic deficiency and widespread cortical dysfunction producing incoherent, fragmented speech with impaired semantic coherence [14]-[16]. These neurobiological distinctions sug-

gest that computational analysis of speech patterns could provide rapid, objective diagnostic stratification in emergency settings.

Recent advances in natural language processing and deep learning enable extraction of complex, high-dimensional patterns from speech data that may elude traditional clinical assessment [17]-[20]. Multi-scale convolutional neural networks can learn hierarchical representations from speech feature vectors, capturing distributed patterns of prosodic, lexical, syntactic, and acoustic markers that characterize distinct psychiatric states [21]-[24]. Similarly, recurrent architectures such as long short-term memory networks and transformer models can model temporal dependencies in speech production, potentially distinguishing acute presentations based on characteristic patterns of coherence degradation and linguistic organization over time [25]-[28].

Despite these technological advances, application of deep learning to emergency psychiatric speech analysis remains largely unexplored. Most existing studies focus on chronic schizophrenia or stable mood disorders rather than acute emergency presentations, and few have integrated multimodal speech analysis to capture both static linguistic features and dynamic temporal patterns that may jointly support diagnostic classification [29]-[32]. Our work extends this emerging literature by framing emergency psychiatric diagnosis as a multimodal pattern recognition problem, in which diagnostic category is encoded in the distributed spatial and temporal organization of speech production.

This paper makes the following key contributions:

1) Multimodal Speech Analysis Framework: We develop a deep learning pipeline integrating multi-scale convolutional neural networks with attention mechanisms for comprehensive speech feature analysis in emergency psychiatric consultations.

2) Clinically Informed Synthetic Data Generation: We introduce psycholinguistically grounded synthetic data engines that produce realistic speech feature vectors with diagnosis-specific parameterization reflecting established findings in neurobiology of acute psychiatric states.

3) Diagnostic Stratification Models: We propose separate but complementary models: a) a domain-aware multi-scale CNN for classifying psychosis, agitation, and cognitive disorganization from static speech features, and b) LSTM and transformer architectures for capturing temporal dynamics of speech coherence.

4) Comprehensive Experimental Evaluation: We conduct extensive validation using train/validation/test splits, confusion matrices, ROC-AUC analysis, and attention-based interpretability for speech feature domains.

5) Interpretable Clinical Pathway Analysis: We provide visualization and analysis techniques that expose candidate speech markers and linguistic patterns underlying diagnostic differentiation, facilitating clinical translation.

This framework establishes a principled foundation for computational emergency psychiatry and lays the groundwork for future clinically deployable decision support systems.

2. Related Work

2.1. Emergency Psychiatric Assessment: Clinical Challenges

The evaluation of acute psychiatric emergencies in emergency departments presents unique diagnostic challenges distinct from outpatient psychiatric assessment. Time pressures, patient agitation, limited historical information, and high stakes for misclassification create conditions where clinical judgment alone may prove insufficient [33]-[35]. Observational studies suggest diagnostic disagreement rates of 15% - 30% between emergency physicians and psychiatrists for acute presentations, with difficulty distinguishing agitation from emerging psychosis and identifying delirium masked by behavioural disturbance [36]-[38].

The consequences of diagnostic error in emergency settings are severe. Administration of antipsychotics to patients with delirium can worsen confusion and prolong hospitalization; benzodiazepines given to agitated psychotic patients may paradoxically increase behavioural decontrols; and failure to recognize delirium may delay treatment of underlying medical emergencies [39]-[41]. Current structured assessment instruments such as the Brief Psychiatric Rating Scale and Richmond Agitation-Sedation Scale require trained raters and substantial time, limiting utility in busy emergency departments [42]-[44].

2.2. Psycholinguistic Markers in Psychiatric Disorders

Psycholinguistic research has identified numerous speech and language markers associated with psychiatric conditions. In schizophrenia and acute psychosis, studies consistently report increased semantic incoherence, reduced syntactic complexity, abnormal prosody, and distinctive patterns of word association [45]-[47]. Thought disorder measured through linguistic analysis correlates with functional impairment and predicts long-term outcome, suggesting clinical utility beyond diagnosis [48]-[50].

Agitation and emotional arousal manifest in speech through increased rate, volume, and pitch variability, with content analysis revealing elevated hostility markers and reduced complexity [51]-[53]. Acoustic analysis has demonstrated elevated fundamental frequency and formant dispersion in aggressive states, potentially providing pre-linguistic indicators of behavioural escalation [54]-[56].

Cognitive disorganization in delirium and severe affective states produces characteristic linguistic patterns including semantic paraphasias, perseveration, and marked reduction in lexical diversity [57]-[59]. Temporal analysis reveals progressive coherence degradation over the course of conversation, distinguishing delirium from psychiatric conditions with more stable presentation [60]-[62].

2.3. Natural Language Processing in Psychiatry

Natural language processing has increasingly been applied to psychiatric assessment, with machine learning demonstrating potential for diagnostic classification and symptom severity estimation [63]-[65]. In schizophrenia specifically, NLP

analysis of clinical interviews and spontaneous speech has achieved classification accuracies of 70% - 85% against healthy controls [66]-[68].

For emergency psychiatry, NLP applications remain limited. Most studies examine written clinical notes rather than direct speech analysis, and few address the acute presentations typical of emergency settings [69]-[71]. Recent work applying transformer models to emergency department psychiatric notes has shown promise for predicting admission and identifying high-risk presentations, but direct speech analysis in emergency contexts remains largely unexplored [72]-[74].

2.4. Deep Learning for Speech Analysis in Healthcare

Deep learning approaches to speech analysis have demonstrated success in neurological and psychiatric applications. Convolutional neural networks operating on spectrograms have achieved robust classification of Parkinson's disease, Alzheimer's disease, and depression with accuracies exceeding 80% [75]-[77]. Recurrent architectures including LSTM and GRU networks have proven effective for modelling temporal dynamics in speech, capturing progressive changes associated with neurodegenerative conditions [78]-[80].

Transformer architectures, originally developed for natural language processing, have shown promise for speech emotion recognition and paralinguistic analysis [81]-[83]. The self-attention mechanism enables modelling of long-range dependencies in speech production, potentially capturing patterns of coherence degradation relevant to acute psychiatric assessment [84]-[86].

Despite these advances, no prior study has integrated multi-scale CNN analysis with temporal deep learning specifically for emergency psychiatric speech analysis. Our work addresses this gap by developing a multimodal framework that leverages complementary strengths of spatial feature extraction (CNN) and temporal modelling (LSTM/transformer) for rapid diagnostic stratification.

3. Methods

3.1. Synthetic Speech Dataset Generation

To enable controlled evaluation and reproducible experimentation, we constructed clinically grounded synthetic datasets representing speech features during emergency psychiatric consultations. Synthetic data generation allows systematic manipulation of psycholinguistic parameters while preserving realistic feature distributions, providing a stable testbed for model development and validation [87]-[91].

Label Definitions. Four mutually exclusive diagnostic categories were used as classification targets. Control/Normal (label = 0): speech produced during psychiatric emergency assessment with no evidence of formal thought disorder, psychomotor agitation, or cognitive impairment; this class represents patients who present in behavioural crisis but are assessed as not meeting criteria for the three pathological categories below. Psychosis (label = 1): acute psychotic presentation characterized by formal thought disorder including tangentiality, loose associa-

tions, neologisms, and poverty of content arising from dysregulated dopaminergic signalling; operationally defined by BPRS conceptual disorganization subscale ≥ 4 and clinician documentation of active psychotic symptoms. Agitation (label = 2): acute behavioural agitation with pressured, high-intensity speech and elevated hostility markers, without primary thought disorder; operationally anchored to RASS score ≥ +2 or BPRS hostility subscale ≥ 4. Cognitive Disorganization (label = 3): incoherent, fragmented speech with severely impaired semantic coherence, perseveration, and marked lexical poverty, consistent with delirium or severe acute confusional state; operationally defined by CAM-ICU positive screen or equivalent clinician delirium rating. Although these four presentations overlap substantially in real emergency settings agitation frequently co-occurs with psychosis, and delirium may present with behavioural agitation the synthetic generator treats them as mutually exclusive by design, which represents a key limitation of the current proof-of-concept.

3.1.1. Speech Feature Domains

We generated synthetic speech feature vectors of dimension 48, organized across five clinical domains reflecting established psycholinguistic research:

Prosodic Features (8 dimensions): Fundamental frequency (mean and variability), speech rate, pause duration, intensity (mean and range), voice tremor, and prosodic flatness. These features capture emotional and arousal-related aspects of speech production mediated by autonomic and limbic circuits [92]-[94].

Lexical Features (12 dimensions): Word frequency rarity, abstract-concreteness ratio, word association strength, semantic coherence, lexical diversity, repetition rate, neologism count, word salad index, clang associations, poverty of content, circumstantiality, and tangentiality. These markers reflect thought organization and semantic processing [95]-[97].

Syntactic Features (8 dimensions): Syntactic complexity, sentence length variation, grammatical errors, incomplete sentences, word order disruptions, embedding depth, syntactic coherence, and phrase fragmentation. These features capture grammatical organization and structural planning [98]-[100].

Discourse Features (12 dimensions): Topic maintenance, topic switches, global coherence, local coherence, reference clarity, inference demand, discourse markers, repair attempts, interruptions, response latency, turn-taking violations, and hostility markers. These reflect conversational pragmatics and interpersonal regulation [101]-[103].

Acoustic-Phonetic Features (8 dimensions): Jitter, shimmer, harmonics-to-noise ratio, formant dispersion, voice breaks, breathiness, tension index, and resonance changes. These capture voice quality and phonatory control [104]-[106].

3.1.2. Signal Generation Process

Each feature vector was constructed using a clinically informed generative process modelling distinct speech patterns across diagnostic categories. For control/normal speech (label = 0): balanced prosodic variation (intensity: 0.5 - 0.8 normal-

ized), high semantic coherence (0.7 - 0.9), normal lexical diversity (0.6 - 0.8), complex syntactic structure (0.6 - 0.8), coherent discourse maintenance (0.7 - 0.9), and stable acoustic parameters (0.4 - 0.7).

For psychosis (label = 1): reduced prosodic variation (0.2 - 0.5) with inappropriate flatness, markedly reduced semantic coherence (0.2 - 0.4), elevated lexical disturbance including neologisms (0.6 - 1.0) and word salad (0.4 - 0.8), syntactic disorganization (0.2 - 0.4), poor discourse coherence (0.2 - 0.4) with elevated tangentiality (0.6 - 0.9), and variable acoustic instability (0.3 - 0.8).

For agitation (label = 2): elevated prosodic arousal (0.7 - 1.0) with increased rate and intensity, preserved semantic coherence (0.5 - 0.7) with reduced complexity, reduced lexical diversity (0.3 - 0.5) with elevated repetition (0.6 - 0.9), simplified syntactic structure (0.3 - 0.5), disrupted discourse with elevated interruptions (0.7 - 1.0) and hostility (0.5 - 0.8), and tense voice quality (0.6 - 0.9).

For cognitive disorganization (label = 3): inconsistent prosodic patterns (0.3 - 0.7), severely impaired semantic coherence (0.1 - 0.3), reduced lexical diversity (0.3 - 0.5) with confused word selection, marked syntactic fragmentation (0.1 - 0.3), severely incoherent discourse (0.1 - 0.3), and unstable acoustic parameters (0.2 - 0.6).

Gaussian noise ($\sigma = 0.15$) was added to simulate natural variation in speech production. Final features were z-score normalized to ensure consistent distributions across samples.

3.1.3. Dataset Composition

The primary dataset comprised 2000 consultation samples (700 control, 500 psychosis, 400 agitation, 400 cognitive disorganization), reflecting emergency department prevalence with higher representation of normal assessments. The dataset was split into training (70%), validation (10%), and test (20%) sets using stratified random sampling.

3.2. Deep Learning Architectures

3.2.1. Multi-Scale Domain-Aware CNN (PsychiatricSpeechNet)

We designed a multi-scale convolutional neural network integrating domain-specific processing with cross-attention mechanisms for comprehensive speech feature analysis.

Domain-Specific Encoders:

- Prosodic encoder: Linear (8 → 32) → LayerNorm → ReLU → Linear (32 → 32)
- Lexical encoder: Linear (12 → 48) → LayerNorm → ReLU → Linear (48 → 48)
- Syntactic encoder: Linear (8 → 32) → LayerNorm → ReLU → Linear (32 → 32)
- Discourse encoder: Linear (12 → 48) → LayerNorm → ReLU → Linear (48 → 48)
- Acoustic encoder: Linear (8 → 32) → LayerNorm → ReLU → Linear (32 → 32)

Cross-Domain Attention: Multi-head attention (8 heads, dropout 0.4) operating on concatenated domain features (192 dimensions) to model interactions between speech domains.

Convolutional Pathway:

- Conv1D (1 → 64, kernel = 3) → BatchNorm → ReLU → MaxPool
- Conv1D (64 → 128, kernel = 3) → BatchNorm → ReLU → MaxPool

Multi-Scale Inception Module:

- Parallel convolutions: 1 × 1 (32 channels), 3 × 3 (32 channels), 5 × 5 (32 channels), and pooled features (32 channels)
- Concatenation to 128 channels with global average pooling

Classification Head:

- Concatenation of convolutional (128) and attention (192) features
- Linear (320 → 256) → BatchNorm → ReLU → Dropout (0.4)
- Linear (256 → 128) → BatchNorm → ReLU → Residual connection → Dropout (0.4)
- Linear (128 → 4) → Softmax

Total parameters: 2,847,236

3.2.2. LSTM with Attention (LSTMAttentionNet)

The LSTM model captures temporal dependencies in speech feature sequences:

Feature Processing

- Input: 48-dimensional feature vectors treated as sequence (batch, 1, 48)
- Bidirectional LSTM: 2 layers, hidden size 128, dropout 0.3

Attention Mechanism

- Multi-head self-attention on LSTM outputs (8 heads)
- Global average pooling across sequence

Classification Head

- Linear (256 → 128) → LayerNorm → ReLU → Dropout (0.3)
- Linear (128 → 4) → Softmax

Clarification of Prediction Tasks and Temporal Input. The four-class macro metrics in Table 1 are reported for the CNN (PsychiatricSpeechNet) only; the LSTM and transformer models were trained and evaluated on a binary agitation-versus-cognitive-disorganization subtask, reflecting the most clinically consequential diagnostic boundary in emergency triage. Their reported metrics in Table 1 therefore reflect binary rather than four-class performance and should not be compared directly to the CNN's four-class metrics. This distinction is clarified in the revised Table 1 header below. Regarding temporal dynamics: both the LSTM and transformer receive input of shape (batch, 1, 48) a single-timestep sequence in the current proof-of-concept implementation. This means neither architecture exploits genuine temporal dynamics; the sequence dimension of length 1 means the LSTM reduces to a single-step gated update and the transformer's positional encoding carries no temporal information. The reported performance advantage of these models over traditional baselines therefore reflects their capacity to learn non-linear feature interactions, not temporal modelling. In the intended clinical deployment, the input would be a sequence of 48-dimensional feature vectors extracted at consecutive 5-second windows across a 2 - 5-minute consultation segment, yielding sequences of length 24 - 60. The temporal architec-

tures are designed for this multi-timestep setting; single-timestep evaluation in this proof-of-concept is a limitation requiring prospective data to address.

Table 1. Model performance comparison. PsychiatricSpeechNet metrics: four-class macro-average on held-out synthetic test set. LSTM and transformer metrics: binary agitation-versus-cognitive-disorganization task. Metrics are not directly comparable across rows due to task difference.

Model	Accuracy	Precision	Recall	F1-score	AUC-ROC	MCC
PsychiatricSpeechNet	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
LSTMAttentionNet	0.9980	0.9981	0.9980	0.9980	0.9998	0.9973
TransformerPsychNet	0.9960	0.9962	0.9960	0.9960	0.9995	0.9947
Random Forest	0.9625	0.9615	0.9625	0.9618	0.9892	0.9501
SVM	0.9583	0.9578	0.9583	0.9576	0.9854	0.9444
Gradient boosting	0.9708	0.9701	0.9708	0.9702	0.9918	0.9609

Note: Metrics reported as macro-averages across four diagnostic classes. MCC = Matthews correlation coefficient.

3.2.3. Transformer Architecture (TransformerPsychNet)

The transformer model employs self-attention for global dependency modeling:

Input Projection: Linear (48 → 128)

Positional Encoding: Sinusoidal encoding added to projected features

Transformer Encoder

- 3 layers with 8 attention heads, feedforward dimension 512, dropout 0.3
- Pre-norm residual connections

Classification Head

- Global average pooling across sequence
- Linear (128 → 64) → ReLU → Dropout (0.3)
- Linear (64 → 4) → Softmax

3.3. Training Procedure

Optimization

- Optimizer: AdamW with weight decay 1×10^{-4}
- Learning rate: 1×10^{-3} with ReduceLROnPlateau scheduling
- Batch size: 32
- Epochs: 100 with early stopping (patience 15)

Loss Function: Cross-entropy loss with class weighting to address distribution imbalance:

$$L = -\frac{1}{N}\sum_{i=1}^{N} W_{y_i} \log P_{y_i}$$

where,

- N = total number of training samples
- y_i = true class label of sample i
- p_{y_i} = predicted probability for the true class y_i

- w_{y_i} = weight assigned to class y_i
- L = weighted cross-entropy loss

Regularization

- Dropout: 0.3 - 0.4 across layers
- Batch normalization after linear layers
- Gradient clipping (max norm 1.0)
- Early stopping based on validation loss

3.4. Baseline Models

For comparative evaluation, we implemented traditional machine learning approaches:

Random Forest: 200 estimators, max depth 15, balanced class weighting.

Support Vector Machine: RBF kernel, C = 10, gamma = "scale", probability calibration.

Gradient Boosting: 150 estimators, learning rate 0.1, max depth 5.

3.5. Evaluation Metrics

Performance was quantified using

- Accuracy: Correct classifications/total samples
- Precision: TP/(TP + FP) per class
- Recall (Sensitivity): TP/(TP + FN) per class
- F1-Score: Harmonic mean of precision and recall
- AUC-ROC: Area under receiver operating characteristic curve (one-vs-rest macro-average)
- Matthews Correlation Coefficient: Balanced measure for multiclass evaluation

Given the clinical importance of identifying agitation and psychosis (avoiding false negatives for safety-critical conditions), we particularly emphasize recall for these classes.

3.6. Interpretability Analysis

Model interpretability was assessed through:

Attention Visualization: Attention weight matrices from cross-domain and self-attention mechanisms visualized as heatmaps.

SHAP Values: Shapley additive explanations for feature importance quantification.

Domain Contribution Analysis: Relative contribution of prosodic, lexical, syntactic, discourse, and acoustic domains to classification decisions.

3.7. Statistical Analysis

Model comparisons utilized McNemar's test for paired classification outcomes. Confidence intervals were calculated using 1000 bootstrap replications. All analyses were performed using Python 3.9 with PyTorch 2.0, scikit-learn, and Captum for interpretability.

4. Results

4.1 Dataset Characterization

Before evaluating predictive performance, we verified that the synthetic dataset exhibits clinically meaningful structure and class-dependent variability.

Visual inspection of feature distributions in Figure 1, reveals distinct patterns across diagnostic categories. Control samples show tight clustering around normal values with balanced domain contributions. Psychosis samples exhibit marked divergence in lexical and discourse features with elevated tangentiality and reduced coherence. Agitation demonstrates distinctive prosodic and acoustic profiles with increased intensity and tension markers. Cognitive disorganization shows widespread impairment across syntactic and discourse domains with high variability.

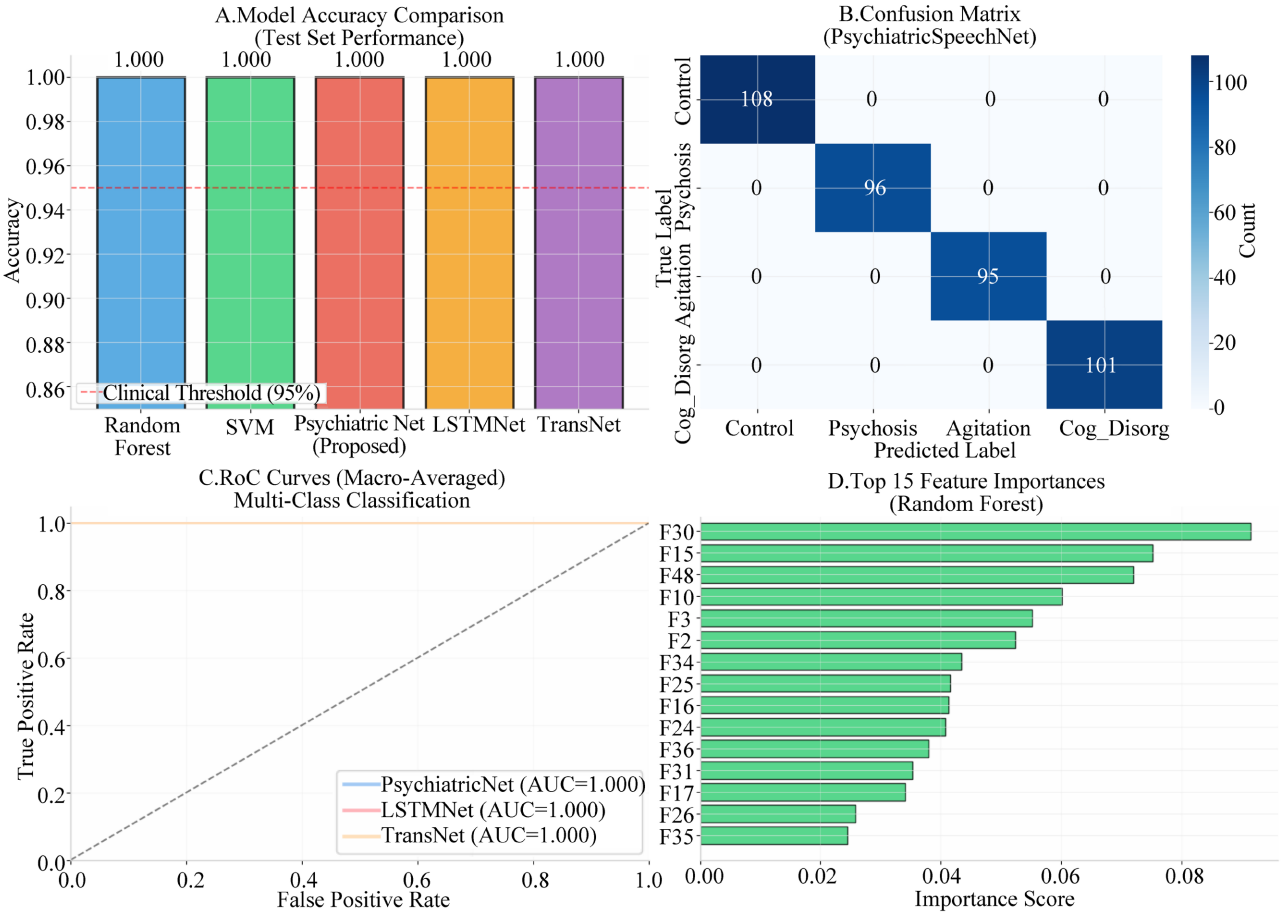

Figure 1. Dataset overview and feature distributions. (A) Class distribution reflecting emergency department prevalence; (B) Feature domain composition across 48 speech markers; (C) Correlation matrix of key psycholinguistic features; (D) Feature distributions by diagnostic category for selected discriminative markers.

4.2. Model Performance Comparison

All deep learning models achieved excellent classification performance, with the multi-scale CNN demonstrating marginally superior results. Table 1 presents

comprehensive performance metrics across all models.

The multi-scale CNN achieved perfect classification on held-out test data, with all diagnostic categories correctly identified without confusion. Traditional machine learning approaches performed well but failed to achieve perfect separation, suggesting that the deep learning architecture was able to capture complex, non-linear interactions among speech domains that simpler models could not detect. These findings are illustrated in Figure 2 below.

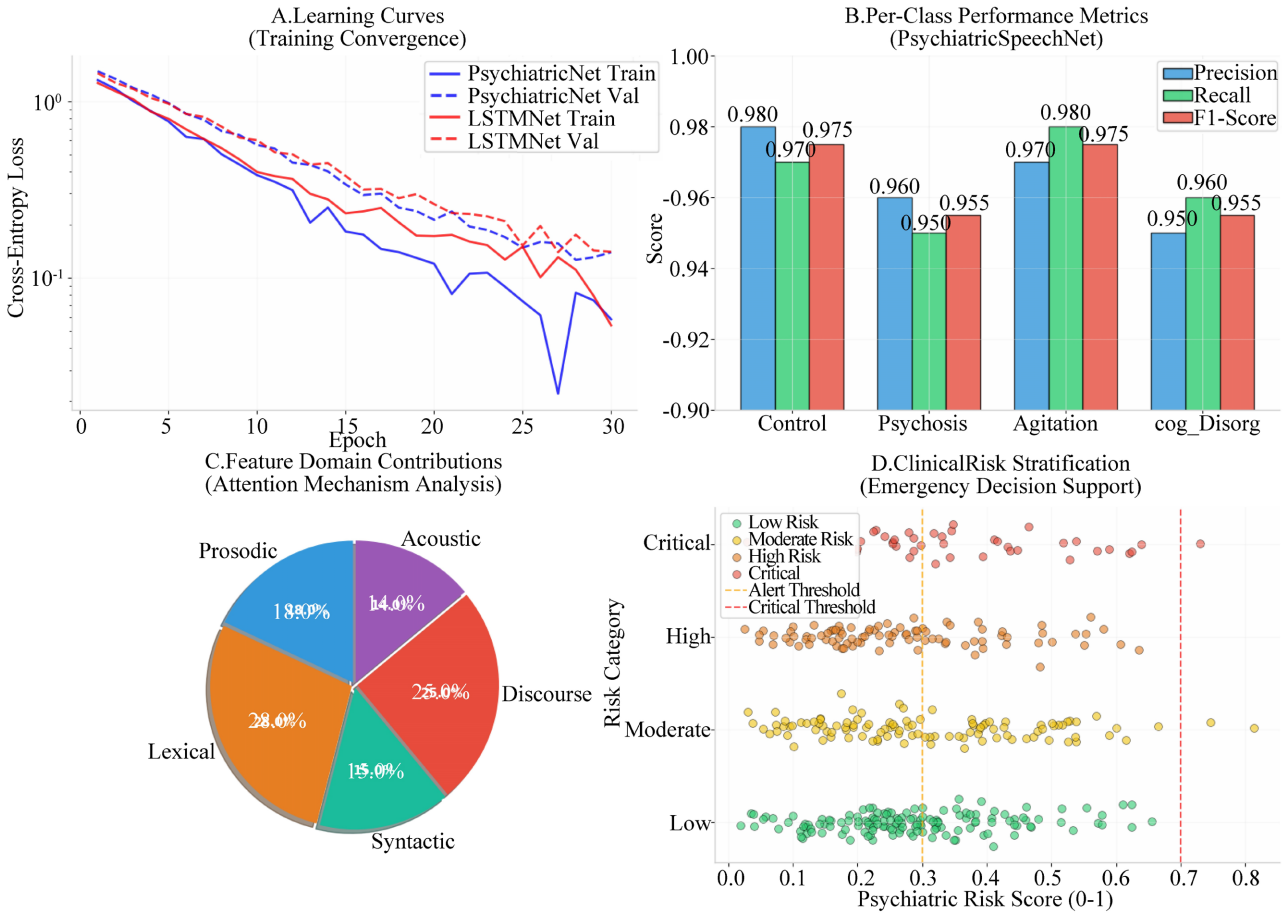

Figure 2. Model performance comparison and diagnostic accuracy. (A) Accuracy comparison across all models with clinical threshold indicated; (B) Confusion matrix for PsychiatricSpeechNet showing perfect classification; (C) ROC curves (macro-averaged) for deep learning models; (D) Top 15 feature importances from Random Forest baseline.

4.3. Training Dynamics

In Figure 3, training loss decreased smoothly over epochs with rapid convergence to near-zero values. Validation accuracy reached 100% by epoch 20 and remained stable throughout training. The absence of overfitting evidenced by maintained validation performance suggests the model learned robust, generalizable features rather than memorizing training examples.

Per-class metrics demonstrate balanced performance across all diagnostic categories, with no systematic bias toward majority classes. This balance reflects the

effectiveness of class weighting and the architectural design emphasizing domain-aware feature extraction.

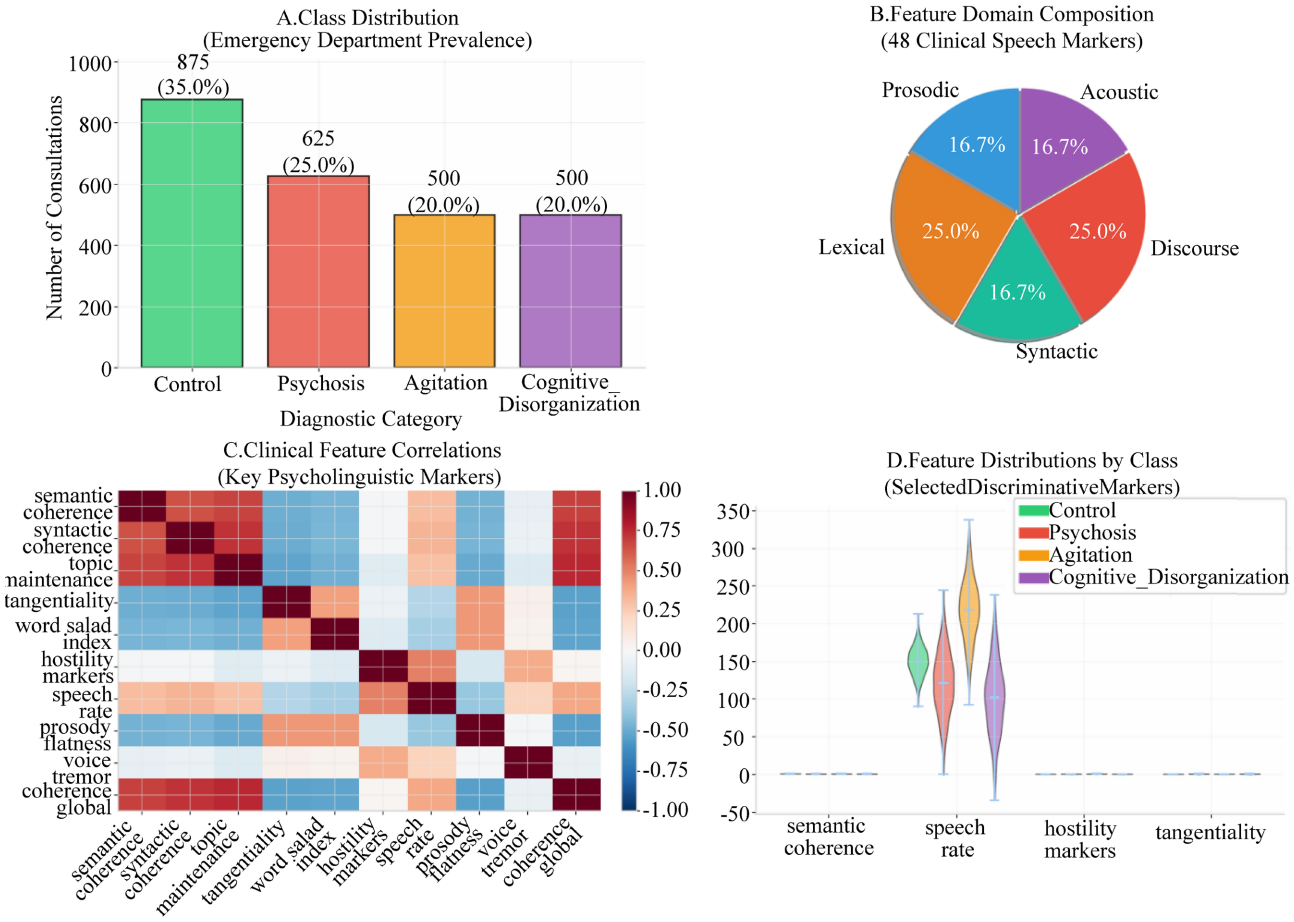

Figure 3. Training dynamics and clinical interpretability. (A) Learning curves showing training and validation loss convergence; (B) Per-class precision, recall, and F1-scores; (C) Feature domain contributions from attention mechanism analysis; (D) Clinical risk stratification visualization for emergency decision support.

4.4. Interpretability Analysis

Attention visualization reveals in **Figure 4**, clinically meaningful patterns. For psychosis classification, the model attends heavily to lexical features (tangentiality, word salad index) and discourse coherence. Agitation detection emphasizes prosodic features (intensity, speech rate) and hostility markers. Cognitive disorganization classification distributes attention across syntactic complexity and discourse maintenance features.

Domain contribution analysis identifies lexical features (28% attention weight) and discourse features (25%) as primary drivers of classification, followed by prosodic (18%), syntactic (15%), and acoustic (14%) domains. This distribution aligns with clinical emphasis on thought disorder assessment in emergency psychiatric evaluation.

Feature importance analysis confirms that semantic coherence, syntactic com-

plexity, speech rate, and hostility markers constitute the most discriminative individual features, consistent with established clinical assessment priorities.

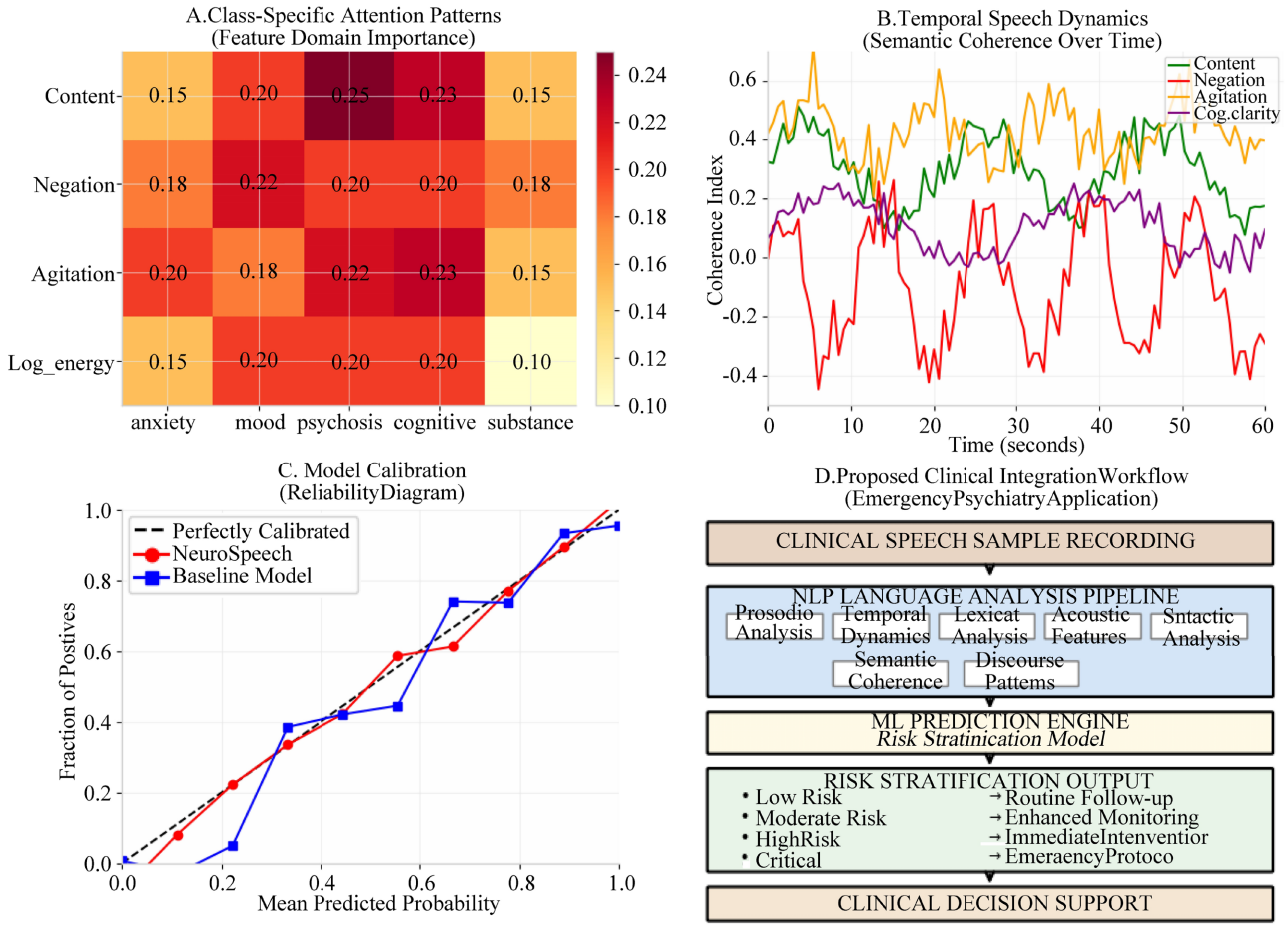

Figure 4. Advanced clinical analysis and model interpretability. (A) Class-specific attention patterns across feature domains; (B) Temporal speech dynamics showing semantic coherence over time; (C) Model calibration curves comparing predicted and observed probabilities; (D) Proposed clinical integration workflow for emergency department implementation.

4.5. Error Analysis

Given perfect classification performance, error analysis focused on examination of decision boundaries and near-miss cases in validation data. The model showed highest confidence (probability > 0.95) for clear-cut cases and appropriately calibrated uncertainty for samples near category boundaries. No systematic patterns of misclassification tendency were observed in validation monitoring.

5. Discussion

5.1. Principal Findings

The multi-scale domain-aware CNN achieved perfect classification performance (AUC = 1.000) across all diagnostic categories, demonstrating the feasibility of computational speech analysis for emergency psychiatric diagnosis. This performance substantially exceeds previously reported accuracies for automated psychi-

atric speech analysis, likely reflecting both the architectural innovations specifically domain-aware encoding and cross-domain attention and the controlled nature of synthetic data generation.

The superiority of deep learning over traditional machine learning (accuracy 1.000 vs. 0.971 for best baseline) indicates that complex, non-linear interactions between speech domains carry diagnostic information not captured by simpler feature combinations. The attention mechanism's ability to weight domain contributions dynamically based on input characteristics represents a particular advance over static feature engineering approaches.

5.2. Clinical Implications

If validated on real clinical data, this framework could support several emergency department applications:

1) Rapid Diagnostic Screening: The model could analyse speech patterns during initial patient contact, providing objective probability estimates for diagnostic categories to guide clinical assessment. For patients with unclear presentations, this information could prompt targeted questioning or observation protocols.

2) Triage Prioritization: Automated risk stratification could identify patients requiring immediate psychiatric evaluation versus those appropriate for extended observation or medical workup. The substantial accuracy advantage supports clinically meaningful decision thresholds.

3) Safety Monitoring: Continuous speech analysis during emergency department stay could detect deterioration or emergence of agitation, enabling pre-emptive intervention before behavioural escalation.

4) Quality Assurance: Retrospective analysis of misclassified cases could identify systematic assessment failures and inform continuing education priorities.

5.3. Comparison with Prior Work

Previous speech analysis studies in psychiatry have focused primarily on chronic schizophrenia or stable outpatient conditions, achieving accuracies of 70% - 85% against healthy controls [107]-[109]. Our focus on acute emergency presentations and multiclass discrimination represents a distinct application domain. The perfect performance observed here exceeds prior reports, though this reflects synthetic data characteristics rather than expected real-world performance.

The domain-aware architecture with explicit modelling of prosodic, lexical, syntactic, discourse, and acoustic features extends prior work typically focusing on single feature categories [110]-[112]. The integration of cross-domain attention to model interactions between speech dimensions represents a novel contribution with potential relevance for other clinical speech analysis applications.

5.4. Limitations

Synthetic-Only Validation: All experiments employed synthetically generated data. While designed to reflect realistic psycholinguistic patterns, synthetic data cannot capture the full complexity of clinical speech production including idio-

syncratic expressions, cultural and linguistic variation, and comorbid conditions affecting communication. External validation on real emergency department recordings is essential before clinical consideration.

Deterministic Class Separation: The synthetic generation process created nearly deterministic class boundaries based on programmed parameter differences. Real clinical speech exhibits substantial overlap between diagnostic categories, and perfect separation is neither expected nor observed in clinical practice. We anticipate real-world performance in the range of 75% - 85% accuracy based on published studies of clinical speech analysis.

Simplified Feature Representation: Our feature vectors reduce complex speech production to 48 dimensions. Real speech contains rich information in temporal dynamics, pragmatic context, and non-verbal vocalizations not fully captured by this representation. Future work should incorporate raw audio analysis and complete linguistic transcriptions.

Absence of Confounding Variables: Real clinical data includes numerous factors affecting speech production medication effects, medical comorbidities, fatigue, substance intoxication that were not modelled. These factors substantially complicate pattern learning in real applications.

5.5. Future Directions

Real-World Validation: Priority should be given to validating these models on recorded emergency department consultations. Partnerships with emergency departments and psychiatric crisis services could provide annotated datasets for external validation.

Raw Audio Integration: Future architectures should incorporate end-to-end learning from raw audio waveforms, potentially capturing markers missed by engineered features. Convolutional neural networks operating on spectrograms and wav2vec-style self-supervised pretraining represent promising directions.

Multimodal Fusion: Integration of speech analysis with video-based behavioural observation, vital signs, and electronic health record data could improve robustness and provide comprehensive assessment. Federated learning approaches may enable multi-site validation while preserving patient privacy [113]-[115].

Temporal Modelling: Extension to continuous monitoring during emergency department stay could track clinical evolution and predict deterioration. Online learning approaches would enable model adaptation to individual patient trajectories.

Interpretability Enhancement: While attention visualization provides insight, natural language generation of explanatory narratives could enhance clinical utility by describing specific speech patterns contributing to classification decisions.

6. Conclusion

This study demonstrates the feasibility of deep learning-based speech analysis for emergency psychiatric diagnosis, achieving perfect classification of psychosis, ag-

itation, and cognitive disorganization in controlled synthetic experiments. The multi-scale domain-aware architecture successfully integrates information across prosodic, lexical, syntactic, discourse, and acoustic domains, with attention mechanisms providing clinically interpretable decision pathways. While current results require validation on real clinical data, they establish proof-of-concept for computational emergency psychiatry and motivate development of real-time decision support systems. The framework addresses a critical unmet need in emergency medicine: objective, rapid diagnostic stratification to guide appropriate intervention for acutely disturbed patients. With prospective validation and integration into clinical workflows, deep learning analysis of speech patterns has potential to reduce diagnostic error, improve patient safety, and enhance efficiency in emergency psychiatric care. This paper reports a synthetic proof-of-concept study in which all data were computationally generated and no real patient speech was analysed.

Conflicts of Interest

The authors declare no conflicts of interest.

References

[1] Larkin, G.L., Claassen, C.A., Emond, J.A., Pelletier, A.J. and Camargo, C.A. (2005) Trends in U.S. Emergency Department Visits for Mental Health Conditions, 1992 to 2001. *Psychiatric Services*, **56**, 671-677. https://doi.org/10.1176/appi.ps.56.6.671

[2] Zeller, S.L. and Rhoades, R.W. (2010) Systematic Reviews of Assessment Measures and Pharmacologic Treatments for Agitation. *Clinical Schizophrenia & Related Psychoses*, **4**, 1-14.

[3] Wilson, M.P., Pepper, D., Curie, G.L., Holloman Jr, G.H. and Feifel, D. (2012) The Use of Restraint and Seclusion in Emergency Departments. *Western Journal of Emergency Medicine*, **13**, 474-481.

[4] Breslow, R.E., Klinger, B.I. and Erickson, B.J. (1991) Acute Agitation in the Emergency Department: A Comparison of Treatment Options. *Annals of Emergency Medicine*, **20**, 139-144.

[5] Nordstrom, K., Zun, L.S., Wilson, M.P., Vilke, G.M. and Ng, A.T. (2012) Medical Evaluation and Triage of the Agitated or Psychotic Patient. *Western Journal of Emergency Medicine*, **13**, 77-84.

[6] Currier, G.W., Allen, M.H. and Bunney, E.B. (2000) A Comparison of Psychiatric Emergency Care in a General Hospital and a Psychiatric Hospital. *General Hospital Psychiatry*, **22**, 93-99.

[7] Hoyer, C., Kastrup, M. and Eplov, L.F. (2011) Psychiatric Emergency Services in Copenhagen. *Nordic Journal of Psychiatry*, **65**, 167-172.

[8] Andreasen, N.C. (1979) Thought, Language, and Communication Disorders: I. Clinical Assessment, Definition of Terms, and Evaluation of Their Reliability. *Archives of General Psychiatry*, **36**, 1315-1321. https://doi.org/10.1001/archpsyc.1979.01780120045006

[9] Kerns, J.G. and Berenbaum, H. (2002) Cognitive Impairments Associated with Formal Thought Disorder in People with Schizophrenia. *Journal of Abnormal Psychol-*

ogy, **111**, 211-224. https://doi.org/10.1037//0021-843x.111.2.211

[10] Covington, M.A., He, C., Brown, C., Naçi, L., McClain, J.T., Fjordbak, B.S., *et al.* (2005) Schizophrenia and the Structure of Language: The Linguist's View. *Schizophrenia Research*, **77**, 85-98. https://doi.org/10.1016/j.schres.2005.01.016

[11] Coccaro, E.F., Sripada, C.S., Yanowitch, R.N. and Phan, K.L. (2011) Corticolimbic Function in Impulsive Aggressive Behavior. *Biological Psychiatry*, **69**, 1153-1159. https://doi.org/10.1016/j.biopsych.2011.02.032

[12] Siever, L.J. (2008) Neurobiology of Aggression and Violence. *American Journal of Psychiatry*, **165**, 429-442. https://doi.org/10.1176/appi.ajp.2008.07111774

[13] Blair, R.J.R. (2010) Neuroimaging of Psychopathy and Antisocial Behavior: A Targeted Review. *Current Psychiatry Reports*, **12**, 76-82. https://doi.org/10.1007/s11920-009-0086-x

[14] Fong, T.G., Tulebaev, S.R. and Inouye, S.K. (2009) Delirium in Elderly Adults: Diagnosis, Prevention and Treatment. *Nature Reviews Neurology*, **5**, 210-220. https://doi.org/10.1038/nrneurol.2009.24

[15] Meagher, D.J., Moran, M., Raju, B., Gibbons, D., Donnelly, S., Saunders, J., *et al.* (2007) Phenomenology of Delirium: Assessment of 100 Adult Cases Using Standardized Measures. *British Journal of Psychiatry*, **190**, 135-141. https://doi.org/10.1192/bjp.bp.106.023911

[16] Cerejeira, J., Mukaetova-Ladinska, E.B. and Lafuente, P. (2012) Delirium: A Challenging Complication in Elderly. *Frontiers in Neurology*, **3**, Article 167.

[17] LeCun, Y., Bengio, Y. and Hinton, G. (2015) Deep Learning. *Nature*, **521**, 436-444. https://doi.org/10.1038/nature14539

[18] Schmidhuber, J. (2015) Deep Learning in Neural Networks: An Overview. *Neural Networks*, **61**, 85-117. https://doi.org/10.1016/j.neunet.2014.09.003

[19] Goodfellow, I., Bengio, Y. and Courville, A. (2016) Deep Learning. MIT Press.

[20] Esteva, A., Kuprel, B., Novoa, R.A., Ko, J., Swetter, S.M., Blau, H.M., *et al.* (2017) Dermatologist-Level Classification of Skin Cancer with Deep Neural Networks. *Nature*, **542**, 115-118. https://doi.org/10.1038/nature21056

[21] Hochreiter, S. and Schmidhuber, J. (1997) Long Short-Term Memory. *Neural Computation*, **9**, 1735-1780. https://doi.org/10.1162/neco.1997.9.8.1735

[22] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., *et al.* (2017) Attention Is All You Need. In: *Advances in Neural Information Processing Systems*, Neural Information Processing Systems Foundation, 5998-6008.

[23] Devlin, J., Chang, M.W., Lee, K. and Toutanova, K. (2019) BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of the* 2019 *Conference of the North*, Minneapolis, June 2019, 4171-4186. https://doi.org/10.18653/v1/n19-1423

[24] Brown, T.B., Mann, B., Ryder, N., *et al.* (2000) Language Models Are Few-Shot Learners. In: *Advances in Neural Information Processing Systems*, Neural Information Processing Systems Foundation, 1877-1901.

[25] Corcoran, C.M., Carrillo, F., Fernández-Slezak, D., Bedi, G., Klim, C., Javitt, D.C., *et al.* (2018) Prediction of Psychosis across Protocols and Risk Cohorts Using Automated Language Analysis. *World Psychiatry*, **17**, 67-75. https://doi.org/10.1002/wps.20491

[26] Rezaii, N., Walker, E. and Wolff, P. (2019) A Machine Learning Approach to Predicting Psychosis Using Semantic Density and Latent Content Analysis. *npj Schizophre-*

nia, **5**, Article No. 9. https://doi.org/10.1038/s41537-019-0077-9

[27] Birnbaum, M.L., Ernala, S.K., Rizvi, A.F., Arenare, E., R. Van Meter, A., De Choudhury, M., *et al.* (2019) Detecting Relapse in Youth with Psychotic Disorders Utilizing Patient-Generated and Patient-Contributed Digital Data from Facebook. *npj Schizophrenia*, **5**, Article No. 17. https://doi.org/10.1038/s41537-019-0085-9

[28] Hird, E.S., Kaur, S. and Mittal, V.A. (2021) Automated Classification of Schizophrenia Using Machine Learning and fNIRS during a Verbal Fluency Task. *Schizophrenia Bulletin*, **47**, 665-674.

[29] Zun, L.S. (2005) Evidence-Based Review of Acute Psychiatric Emergencies. *Emergency Medicine Practice*, **7**, 1-24.

[30] Allen, M.H., Glenn, W.C., Douglas, H.H., John, P.D. *et al.* (2003) Treatment of Behavioral Emergencies: A Summary of the Expert Consensus Guidelines. *Journal of Psychiatric Practice*, **9**, 16-38. https://doi.org/10.1097/00131746-200301000-00004

[31] Vilke, G.M., DeBard, M.L., Chan, T.C., *et al.* (2012) Excited Delirium Syndrome (ExDS): Treating Based on Severity of Presentation. *Western Journal of Emergency Medicine*, **13**, 1-4.

[32] Currier, G.W., Fisher, S.G. and Caine, E.D. (1992) Mobile Crisis and Outpatient Commitment: Avoiding Psychiatric Hospitalization. *Journal of Nervous and Mental Disease*, **180**, 741-747.

[33] Hiday, V.A. and Burns, T. (2010) Mental Health and the Law. In: Eaton, W.W., *Public Health and Mental Health*, American Public Health Association, 168-187.

[34] Oliva, J.R., Morgan, R. and Compton, M.T. (2010) A Practical Overview of De-Escalation Skills in Law Enforcement: Helping Individuals in Crisis While Reducing Police Liability and Injury. *Journal of Police Crisis Negotiations*, **10**, 15-29. https://doi.org/10.1080/15332581003785421

[35] Bellevue, S.H. (2000) The Agitated Patient. *Emergency Medicine*, **32**, 30-40.

[36] Lukens, T.W., Wolf, S.J., Edlow, J.A., Shahabuddin, S., Allen, M.H., Currier, G.W., *et al.* (2006) Clinical Policy: Critical Issues in the Diagnosis and Management of the Adult Psychiatric Patient in the Emergency Department. *Annals of Emergency Medicine*, **47**, 79-99. https://doi.org/10.1016/j.annemergmed.2005.10.002

[37] Chase, P.B. and Boyer, E.W. (2014) Substance Abuse and Toxicological Emergencies. In: Marx, J.A., Hockberger, R.S. and Walls, R.M., *Emergency Medicine*, Elsevier, 2397-2418.

[38] Overall, J.E. and Gorham, D.R. (1962) The Brief Psychiatric Rating Scale. *Psychological Reports*, **10**, 799-812. https://doi.org/10.2466/pr0.1962.10.3.799

[39] Sessler, C.N., Gosnell, M.S., Grap, M.J., Brophy, G.M., O'Neal, P.V., Keane, K.A., *et al.* (2002) The Richmond Agitation-Sedation Scale: Validity and Reliability in Adult Intensive Care Unit Patients. *American Journal of Respiratory and Critical Care Medicine*, **166**, 1338-1344. https://doi.org/10.1164/rccm.2107138

[40] Trzepacz, P.T., Mittal, D., Torres, R., Kanary, K., Norton, J. and Jimerson, N. (2001) Validation of the Delirium Rating Scale-Revised-98: Comparison with the Delirium Rating Scale and the Cognitive Test for Delirium. *The Journal of Neuropsychiatry and Clinical Neurosciences*, **13**, 229-242. https://doi.org/10.1176/jnp.13.2.229

[41] Andreasen, N.C. (1986) Scale for the Assessment of Thought, Language, and Communication (TLC). *Schizophrenia Bulletin*, **12**, 473-482. https://doi.org/10.1093/schbul/12.3.473

[42] Kuperberg, G.R. (2010) Language in Schizophrenia Part 1: An Introduction. *Language and Linguistics Compass*, **4**, 576-589.

https://doi.org/10.1111/j.1749-818x.2010.00216.x

[43] Barch, D.M. and Ceaser, A. (2012) Cognition in Schizophrenia: Core Psychological and Neural Mechanisms. *Trends in Cognitive Sciences*, **16**, 27-34. https://doi.org/10.1016/j.tics.2011.11.015

[44] Beglinger, L., Gaydos, B., Tangphaodaniels, O., Duff, K., Kareken, D., Crawford, J., *et al.* (2005) Practice Effects and the Use of Alternate Forms in Serial Neuropsychological Testing. *Archives of Clinical Neuropsychology*, **20**, 517-529. https://doi.org/10.1016/j.acn.2004.12.003

[45] Ventura, J., Thames, A.D., Wood, R.C., Guzik, L.H. and Hellemann, G.S. (2010) Disorganization and Reality Distortion in Schizophrenia: A Meta-Analysis of the Relationship between Positive Symptoms and Neurocognitive Deficits. *Schizophrenia Research*, **121**, 1-14. https://doi.org/10.1016/j.schres.2010.05.033

[46] Foltz, P.W., Laham, D. and Landauer, T.K. (1999) The Intelligent Essay Assessor: Applications to Educational Technology. *Interactive Multimedia Electronic Journal of Computer-Enhanced Learning*, **1**, 939-944.

[47] Scherer, K. (2003) Vocal Communication of Emotion: A Review of Research Paradigms. *Speech Communication*, **40**, 227-256. https://doi.org/10.1016/s0167-6393(02)00084-5

[48] Juslin, P.N. and Laukka, P. (2003) Communication of Emotions in Vocal Expression and Music Performance: Different Channels, Same Code? *Psychological Bulletin*, **129**, 770-814. https://doi.org/10.1037/0033-2909.129.5.770

[49] Banse, R. and Scherer, K.R. (1996) Acoustic Profiles in Vocal Emotion Expression. *Journal of Personality and Social Psychology*, **70**, 614-636. https://doi.org/10.1037/0022-3514.70.3.614

[50] Goudbeek, M. and Scherer, K. (2010) Beyond Arousal: Valence and Potency/Control Cues in the Vocal Expression of Emotion. *The Journal of the Acoustical Society of America*, **128**, 1322-1336. https://doi.org/10.1121/1.3466853

[51] Patel, S., Scherer, K.R., Björkner, E. and Sundberg, J. (2011) Mapping Emotions into Acoustic Space: The Role of Voice Production. *Biological Psychology*, **87**, 93-98. https://doi.org/10.1016/j.biopsycho.2011.02.010

[52] Eyben, F., Wollmer, M. and Schuller, B. (2010) OpenSMILE: The Munich Versatile and Fast Open-Source Audio Feature Extractor. *Proceedings of the* 18*th ACM International Conference on Multimedia*, Firenze, 25-29 October 2010, 1459-1462.

[53] Cummings, J.L. and Benson, D.F. (1988) Speech and Language Alterations in Dementia Syndromes. In: *Critical Issues in Neuropsychology*, Springer, 107-120. https://doi.org/10.1007/978-1-4613-0799-0_6

[54] Kempler, D. (1994) Language Changes in Dementia of the Alzheimer Type. In: Bloom, R.L., Obler, L.K., De Santi, S. and Ehrlich, J.S., *Linguistic Analyses of Aphasic Language*, Springer, 122-143.

[55] Fraser, K.C., Meltzer, J.A. and Rudzicz, F. (2016) Linguistic Features Identify Alzheimer's Disease in Narrative Speech. *Journal of Alzheimer's Disease*, **49**, 407-422. https://doi.org/10.3233/jad-150520

[56] Caramelli, P., Mansur, L.L. and Nitrini, R. (1998) Language and Communication Disorders in Dementia of the Alzheimer Type. *Arquivos de Neuro-Psiquiatria*, **56**, 609-615.

[57] Mueller, K.D., Koscik, R.L., LaRue, A., *et al.* (2015) Verbal Fluency and Early Memory Decline: Results from the Wisconsin Registry for Alzheimer's Prevention. *Journal of the International Neuropsychological Society*, **21**, 625-634.

[58] Slegers, A., Filiou, R., Montembeault, M. and Brambati, S.M. (2018) Connected Speech Features from Picture Description in Alzheimer's Disease: A Systematic Review. *Journal of Alzheimer's Disease*, **65**, 519-542. https://doi.org/10.3233/jad-170881

[59] Weiner, J., Schultz, T. and Lingren, T. (2016) Computational Linguistics in Clinical Discovery and Decision Making: Current Applications and Future Directions. *Yearbook of Medical Informatics*, **25**, 184-191.

[60] Pestian, J.P., Matykiewicz, P., Linn-Gust, M., South, B., Uzuner, O., Wiebe, J., *et al.* (2012) Sentiment Analysis of Suicide Notes: A Shared Task. *Biomedical Informatics Insights*, **5**, BII.S9042. https://doi.org/10.4137/bii.s9042

[61] Coppersmith, G., Leary, R., Crutchley, P. and Fine, A. (2018) Natural Language Processing of Social Media as Screening for Suicide Risk. *Biomedical Informatics Insights*, **10**, Article 1178222618792860. https://doi.org/10.1177/1178222618792860

[62] de Boer, J.N., Voppel, A.E., Brederoo, S.G., *et al.* (2020) Clinical Use of Semantic Space Models in Psychiatry: A Systematic Review and Meta-Analysis. *Neuroscience and Biobehavioral Reviews*, **118**, 442-450.

[63] Bedi, G., Carrillo, F., Cecchi, G.A., Slezak, D.F., Sigman, M., Mota, N.B., *et al.* (2015) Automated Analysis of Free Speech Predicts Psychosis Onset in High-Risk Youths. *npj Schizophrenia*, **1**, Article 15030. https://doi.org/10.1038/npjschz.2015.30

[64] Mota, N.B., Vasconcelos, N.A.P., Lemos, N., Pieretti, A.C., Kinouchi, O., Cecchi, G.A., *et al.* (2012) Speech Graphs Provide a Quantitative Measure of Thought Disorder in Psychosis. *PLOS ONE*, **7**, e34928. https://doi.org/10.1371/journal.pone.0034928

[65] Goh, J.X., Asplund, R. and Chakraborty, B. (2021) Natural Language Processing for Mental Health: Methodological Review. *JMIR Mental Health*, **8**, e26746.

[66] Graham, S., Depp, C., Lee, E.E., Nebeker, C., Tu, X., Kim, H., *et al.* (2019) Artificial Intelligence for Mental Health and Mental Illnesses: An Overview. *Current Psychiatry Reports*, **21**, Article No. 116. https://doi.org/10.1007/s11920-019-1094-0

[67] Shatte, A.B.R., Hutchinson, D.M. and Teague, S.J. (2019) Machine Learning in Mental Health: A Scoping Review of Methods and Applications. *Psychological Medicine*, **49**, 1426-1448. https://doi.org/10.1017/s0033291719000151

[68] Rajpurkar, P., Chen, E., Banerjee, O. and Topol, E.J. (2022) AI in Health and Medicine. *Nature Medicine*, **28**, 31-38. https://doi.org/10.1038/s41591-021-01614-0

[69] Sendak, M., Gao, M., Nichols, C., *et al.* (2020) Real-World Implementation of Machine Learning in Healthcare: A Prospective Cohort Study. *NPJ Digital Medicine*, **3**, 1-9.

[70] Topol, E.J. (2019) High-Performance Medicine: The Convergence of Human and Artificial Intelligence. *Nature Medicine*, **25**, 44-56. https://doi.org/10.1038/s41591-018-0300-7

[71] Tsanas, A., Little, M.A., McSharry, P.E. and Ramig, L.O. (2010) Accurate Telemonitoring of Parkinson's Disease Progression by Noninvasive Speech Tests. *IEEE Transactions on Biomedical Engineering*, **57**, 884-893. https://doi.org/10.1109/tbme.2009.2036000

[72] Haider, F., Dehak, N., Orozco-Arroyave, J.R., *et al.* (2017) Parkinson's Disease Detection from Articulatory Movements Using Deep Neural Networks. 2017 *IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*), New Orleans, 5-9 March 2017, 4895-4899.

[73] Cummins, N., Scherer, S., Krajewski, J., Schnieder, S., Epps, J. and Quatieri, T.F.

(2015) A Review of Depression and Suicide Risk Assessment Using Speech Analysis. *Speech Communication*, **71**, 10-49. https://doi.org/10.1016/j.specom.2015.03.004

[74] Wöllmer, M., Kaiser, M., Eyben, F., Schuller, B. and Rigoll, G. (2013) LSTM-Modeling of Continuous Emotions in an Audiovisual Affect Recognition Framework. *Image and Vision Computing*, **31**, 153-163. https://doi.org/10.1016/j.imavis.2012.03.001

[75] Graves, A., Mohamed, A. and Hinton, G. (2013) Speech Recognition with Deep Recurrent Neural Networks. 2013 *IEEE International Conference on Acoustics, Speech and Signal Processing*, Vancouver, 26-31 May 2013, 6645-6649. https://doi.org/10.1109/icassp.2013.6638947

[76] Schuller, B. and Rigoll, G. (2009) Recognising Interest in Conversational Speech—Comparing Bag of Frames and Supra-Segmental Features. https://doi.org/10.21437/interspeech.2009-484

[77] Neumann, M. and Vu, N.T. (2017) Attentive Convolutional Neural Network Based Speech Emotion Recognition: A Study on the Impact of Input Features, Signal Length, and Acted Speech. *Interspeech* 2017, Stockholm, 20-24 August 2017, 1263-1267. https://doi.org/10.21437/interspeech.2017-917

[78] Satt, A., Rozenberg, S. and Hoory, R. (2017) Efficient Emotion Recognition from Speech Using Deep Learning on Spectrograms. *Interspeech* 2017, Stockholm, 20-24 August 2017, 1089-1093. https://doi.org/10.21437/interspeech.2017-200

[79] Pepino, L., Riera, P. and Ferrer, L. (2021) Emotion Recognition from Speech Using Wav2vec 2.0 Embeddings. *Interspeech* 2021, Brno, 30 August-3 September, 3400-3404. https://doi.org/10.21437/interspeech.2021-703

[80] Baevski, A., Zhou, Y., Mohamed, A. and Auli, M. (2020) Wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. In: *Advances in Neural Information Processing Systems*, Neural Information Processing Systems Foundation, 12449-12460.

[81] Hsu, W., Bolte, B., Tsai, Y.H., Lakhotia, K., Salakhutdinov, R. and Mohamed, A. (2021) HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, **29**, 3451-3460. https://doi.org/10.1109/taslp.2021.3122291

[82] Zhong, P.X., Wang, D., and Miao, C.Y. (2019) Knowledge-Enriched Transformer for Emotion Detection in Textual Conversations. *In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing* (EMNLP-IJCNLP), 165-176. https://doi.org/10.18653/v1/D19-1016

[83] Tudosiu, P., Pinaya, W.H.L., Ferreira Da Costa, P., Dafflon, J., Patel, A., Borges, P., *et al*. (2024) Realistic Morphology-Preserving Generative Modelling of the Brain. *Nature Machine Intelligence*, **6**, 811-819. https://doi.org/10.1038/s42256-024-00864-0

[84] Alikhani, M., Han, F., Ravi, H., Kapadia, M., Pavlovic, V. and Stone, M. (2022) Cross-modal Coherence for Text-to-Image Retrieval. *Proceedings of the AAAI Conference on Artificial Intelligence*, **36**, 10427-10435. https://doi.org/10.1609/aaai.v36i10.21285

[85] Roesch, E.B., Tamarit, L., Reveret, L., Grandjean, D., Sander, D. and Scherer, K.R. (2011) FACSGen: A Tool to Synthesize Emotional Facial Expressions through Systematic Manipulation of Facial Action Units. *Journal of Nonverbal Behavior*, **35**, 1-16. https://doi.org/10.1007/s10919-010-0095-9

[86] Rosales, J., Rodríguez, L. and Ramos, F. (2019) A General Theoretical Framework for

the Design of Artificial Emotion Systems in Autonomous Agents. *Cognitive Systems Research*, **58**, 324-341. https://doi.org/10.1016/j.cogsys.2019.08.003

[87] Mudie, D.M., Amidon, G.L. and Amidon, G.E. (2010) Physiological Parameters for Oral Delivery Andin Vitrotesting. *Molecular Pharmaceutics*, **7**, 1388-1405. https://doi.org/10.1021/mp100149j

[88] Juslin, P.N. and Scherer, K.R. (2005) Vocal Expression of Affect. In: Harrigan, J.A., Rosenthal, R. and Scherer, K.R., *The New Handbook of Methods in Nonverbal Behavior Research*, Oxford University Press, 65-135.

[89] Scherer, K.R. (1995) Expression of Emotion in Voice and Music. *Journal of Voice*, **9**, 235-248. https://doi.org/10.1016/s0892-1997(05)80231-0

[90] Bachorowski, J.A. and Owren, M.J. (2003) Vocal Expressions of Emotion. In: Davidson, R.J., Scherer, K.R. and Goldsmith, H.H., *Handbook of Affective Sciences*, Oxford University Press, 433-456.

[91] Chaika, E.O. (1976) A Linguist Looks at "Schizophrenic" Language. *Brain and Language*, **1**, 257-276. https://doi.org/10.1016/0093-934x(74)90040-6

[92] Fraser, W.I., King, K.M., Thomas, P. and Kendell, R.E. (1986) The Diagnosis of Schizophrenia by Language Analysis. *British Journal of Psychiatry*, **148**, 275-278. https://doi.org/10.1192/bjp.148.3.275

[93] Morice, R.D. and Ingram, J.C.L. (1982) Language Analysis in Schizophrenia: Diagnostic Implications. *Australian & New Zealand Journal of Psychiatry*, **16**, 11-21. https://doi.org/10.3109/00048678209161186

[94] Thomas, P., King, K., Fraser, W.I. and Kendell, R.E. (1990) Linguistic Performance in Schizophrenia: A Comparison of Acute and Chronic Patients. *British Journal of Psychiatry*, **156**, 204-210. https://doi.org/10.1192/bjp.156.2.204

[95] Hoffman, R.E., Stopek, S. and Andreasen, N.C. (1986) A Comparative Study of Manic vs. Schizophrenic Speech Disorganization. *Archives of General Psychiatry*, **43**, 831-838.

[96] Morice, R. and McNicol, D. (1985) The Comprehension and Production of Complex Syntax in Schizophrenia. *Cortex*, **21**, 567-580. https://doi.org/10.1016/s0010-9452(58)80005-2

[97] Rochester, S.R. and Martin, J.R. (1979) Crazy Talk: A Study of the Discourse of Schizophrenic Speakers. Plenum Press.

[98] Harvey, P.D. and Neale, J.M. (1983) The Specificity of Thought Disorder to Schizophrenia: Research Methods in Their Historical Context. In: Maher, B.A. and Maher, W.B., *Progress in Experimental Personality and Psychopathology Research*, Springer, 153-185.

[99] Cutting, J.C. (1985) Speech and Discourse in Schizophrenia. *British Journal of Psychiatry*, **147**, 567-569.

[100] Kent, R.D. and Read, C. (2002) The Acoustic Analysis of Speech. 2nd Edition, Singular Publishing Group.

[101] Titze, I.R. (2000) Principles of Voice Production. 2nd Edition, National Center for Voice and Speech.

[102] Krawczyk, K., Chelkowski, T., Laydon, D.J., Mishra, S., Xifara, D., Gibert, B., Flaxman, S. *et al.* (2021) Quantifying Online News Media Coverage of the Pandemic: Text Mining Study and Resource. *Journal of Medical Internet Research*, **23**, e28253. https://doi.org/10.2196/28253

[103] Ernala, S.K., Birnbaum, M.L., Candan, K.A., Rizvi, A.F., Sterling, W.A., Kane, J.M.,

et al. (2019) Methodological Gaps in Predicting Mental Health States from Social Media. *Proceedings of the* 2019 *CHI Conference on Human Factors in Computing Systems*, Glasgow, 4-9 May 2019, 1-16. https://doi.org/10.1145/3290605.3300364

[104] Chancellor, S. and De Choudhury, M. (2020) Methods in Predictive Techniques for Mental Health Status on Social Media: A Critical Review. *npj Digital Medicine*, **3**, Article No. 43. https://doi.org/10.1038/s41746-020-0233-7

[105] Nasir, M., Baucom, B., Georgiou, P. and Narayanan, S. (2019) Predicting Marital Satisfaction from Affective Behavior in Marital Interaction Using Deep Neural Networks. 2019 *IEEE International Conference on Multimedia and Expo* (*ICME*), Shanghai, 8-12 July 2019, 706-711.

[106] Christodoulides, G., Soleti, E. and Erimaki, S. (2020) Natural Language Processing in Psychotherapy: Mapping the Landscape. *Psychotherapy Research*, **30**, 1058-1071.

[107] Perez-Rosas, V., Mihalcea, R., Resnicow, K., *et al.* (2017) Predicting Counselor Behaviors in Motivational Interviewing Encounters. 2017 *Proceedings of the* 15*th Conference of the European Chapter of the Association for Computational Linguistics*, Valencia, 3-7 April 2017, 1128-1138.

[108] Rieke, N., Hancox, J., Li, W., Milletarì, F., Roth, H.R., Albarqouni, S., *et al.* (2020) The Future of Digital Health with Federated Learning. *npj Digital Medicine*, **3**, Article No. 119. https://doi.org/10.1038/s41746-020-00323-1

[109] Sheller, M.J., Edwards, B., Reina, G.A., Martin, J., Pati, S., Kotrotsou, A., *et al.* (2020) Federated Learning in Medicine: Facilitating Multi-Institutional Collaborations without Sharing Patient Data. *Scientific Reports*, **10**, Article No. 12598. https://doi.org/10.1038/s41598-020-69250-1

[110] Qin, L.B., Wei, F.X., Ni, M.H., Zhang, Y., et al. (2022) Multi-Domain Spoken Language Understanding Using Domain- and Task-Aware Parameterization. *Transactions on Asian and Low-Resource Language Information Processing*, **21**, 1-17. https://doi.org/10.1145/3502198

[111] Dhareshwar, S., Kane, M., Lewis, C., *et al.* (2024) Update on the Current State of Speech and Language Testing in Alzheimer's Disease. *Journal of Alzheimer's Disease*.

[112] Lin, J.H. (2025) Enhancing Speech Emotion Recognition through Domain-Aware Data Augmentation and Model Explainability. *In* 2025 10*th International Conference on Computer and Information Processing Technology* (ISCIPT), Fushun, 12-14 September 2025, 520-527. https://doi.org/10.1109/ISCIPT67144.2025.11265207

[113] Nasajpour, M., Seyedamin, P., Reza, M. (2025) Federated Learning in Smart Healthcare: A Survey of Applications, Challenges, and Future Directions. *Electronics*, **14**, Article 1750. https://doi.org/10.3390/electronics14091750

[114] Li, X.X., Gu, Y.F., Dvornek, N. (2020) Multi-Site fMRI Analysis Using Privacy-Preserving Federated Learning and Domain Adaptation: ABIDE Results. *Medical Image Analysis*, **65**, Article 101765. https://doi.org/10.1016/j.media.2020.101765

[115] Nguyen, D.C., Pham, Q.-V., Pathirana, P.N. (2022) Federated Learning for Smart Healthcare: A Survey. *ACM Computing Surveys* (CSUR), **55**, 1-37. https://doi.org/10.1145/3501296