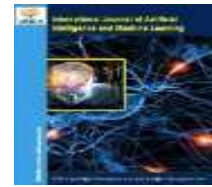




International Journal of Artificial Intelligence and Machine Learning

Publisher's Home Page: <https://www.svedbergopen.com/>



Research Paper

Open Access

UniVoice-Net: A Unified Dual-Domain Denoising and Ensemble Feature-Fusion Framework for Voice-Based Depression Screening

Jithin Jacob¹, Dr. K S Kannan²

¹Department of Computer Science and Engineering, Kalasalingam Academy of Research and Education, Tamil Nadu, India. E-mail: jithu7771@gmail.com, orcid id: 0009-0002-5501-8980

²Department of Computer Science and Engineering, Kalasalingam Academy of Research and Education, Tamil Nadu, India. E-mail: k.s.kannan@klu.ac.in, orcid id: 0000-0002-1304-9829

Abstract

Automated depression screening from speech has advanced rapidly, yet individual pipelines still specialize in either denoising quality, feature richness, or classifier robustness, rarely all three at once. This working paper proposes UniVoice-Net, a unified voice-analysis framework that combines complementary strengths distilled from two independently validated pipelines into a single four-stage architecture: (i) a dual-domain denoising block that cascades wavelet-based gain-modulated noise suppression with spectral-subtraction and Kalman refinement to handle both stationary and non-stationary background interference; (ii) a fused feature-extraction block that jointly computes cepstral-autocorrelation descriptors and transfer-learned convolutional-recurrent embeddings to capture spectral envelope, pitch dynamics, and contextual temporal structure; (iii) a two-stage feature-selection block that applies a hybrid Grey-Wolf/Krill-Herd metaheuristic ahead of a PCA-t-SNE projection to jointly minimize redundancy and preserve both global and local structure; and (iv) a stacked ensemble classification block that combines a dense neural classifier with a gradient-boosted tree classifier. We describe the architecture, motivate each design choice using the documented, independently reported performance of its constituent components on the DAIC-WOZ clinical interview corpus, and outline the experimental protocol required to validate the unified pipeline empirically.

Keywords

Depression detection, voice analysis, dual-domain denoising, feature fusion, metaheuristic optimization, ensemble learning, DAIC-WOZ.

I. Introduction

Depression affects an estimated 300 million people worldwide and remains substantially underdiagnosed because conventional screening depends on subjective self-report, clinician availability, and instruments such as structured interviews that do not scale to population-level use. Voice has emerged as a compelling alternative modality for automated screening: it is inexpensive to capture, minimally intrusive, and carries measurable traces of the psychomotor and affective changes associated with depressive illness, including altered pitch range, articulation rate, pause structure, and spectral energy distribution.

Recent work has shown that carefully engineered pipelines — combining adaptive denoising, complementary feature extraction, principled feature selection, and a strong final classifier — can reach depression-detection accuracies above 97% on the DAIC-WOZ benchmark corpus. However, published pipelines tend to specialize: some invest heavily in denoising and neural feature learning but use a single classifier, while others invest in feature fusion and dimensionality reduction but rely on a single, simpler noise-removal step. This raises a natural question: does combining the complementary strengths of these specialized pipelines into a single unified architecture yield a more robust screening tool, or do the components interact in ways that erode the gains reported individually?

This working paper addresses that question at the design stage. We propose UniVoice-Net, a four-stage architecture that integrates a dual-domain denoising cascade, a fused spectral-temporal feature representation, a two-stage feature-selection process, and a stacked ensemble classifier. We specify the architecture precisely, justify each design decision against the documented behavior of its constituent techniques, and lay out the experimental protocol — dataset splits, baselines, ablations, and statistical tests — required to validate the unified system rigorously before any performance claims are made.

II. Related Work

Classical approaches to voice-based depression detection relied on handcrafted acoustic descriptors such as Mel-Frequency Cepstral Coefficients (MFCCs), jitter, shimmer, and formant trajectories, classified using support vector machines, random forests, or Gaussian mixture models. Deep learning subsequently introduced convolutional and recurrent architectures capable of learning spectral-temporal patterns directly from time-frequency representations, often initialized through transfer learning to compensate for the small size of clinical speech corpora.

Two recent pipelines are of direct relevance here because their components form the basis of the proposed architecture. The first combines spectral-subtraction and Kalman-filter denoising with a transfer-learning convolutional-bidirectional feature extractor, a Grey-Wolf/Krill-Herd hybrid metaheuristic for feature selection, and a dense neural classifier, reporting 97.22% accuracy, 100% precision, 95.92% recall, and a 97.92% F1-score on DAIC-WOZ with 5-fold cross-validation. The second combines gain-modulated wavelet denoising with a cepstral-autocorrelation feature fusion step, a PCA-t-SNE hybrid dimensionality-reduction stage, and an XGBoost classifier, reporting 97.02% accuracy, precision, recall, and F1-score on an 80/20 split of an expanded DAIC-WOZ sample. Both pipelines outperform earlier CNN, GRU, attention-based, and multimodal transfer-learning baselines reported in the 66-95% accuracy range, but each has been validated only as a standalone system; neither paper examines whether combining their respective denoising, feature-fusion, or classification stages would be complementary or redundant.

III. Proposed Univoice-Net Framework

UniVoice-Net is organized as a four-stage pipeline, illustrated conceptually in Figure 1. Each stage is described below along with the rationale for combining techniques that were previously validated only in isolation.

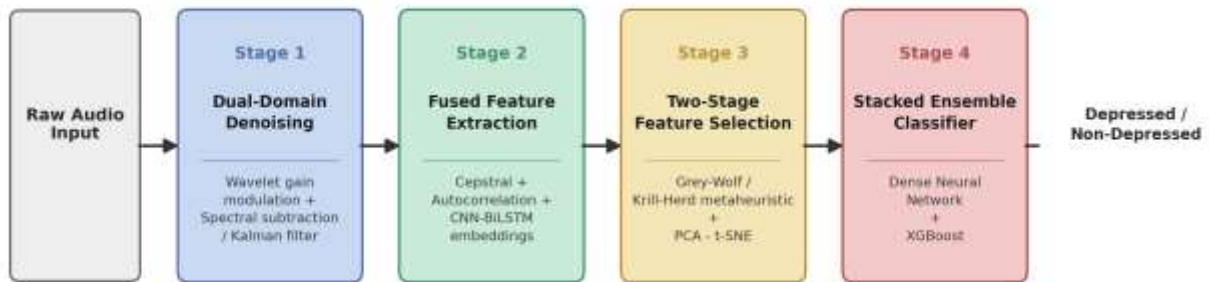


Figure 1. Proposed block diagram of the UniVoice-Net pipeline: dual-domain denoising, fused feature extraction, two-stage feature selection, and stacked ensemble classification.

A. Stage 1: Dual-Domain Denoising Cascade

Clinical interview recordings contain both stationary background hum and non-stationary transient noise (door sounds, microphone handling, overlapping speech). We propose cascading two denoising techniques that individually target different noise characteristics. The signal is first passed through a gain-modulated discrete wavelet transform stage, which decomposes the waveform into approximation and detail coefficients across multiple scales, estimates the noise level in each sub-band via a robust statistic such as the median absolute deviation, and applies a sub-band-specific gain factor before reconstruction; this stage is well suited to suppressing broadband, non-stationary interference while retaining low- and high-frequency prosodic information. The wavelet-denoised signal is then passed through a spectral-subtraction and Kalman-filtering stage, in which voice-activity detection identifies non-speech frames to estimate a stationary noise floor, spectral subtraction removes it, and a Kalman filter suppresses residual musical noise introduced by the subtraction step. The rationale for cascading rather than choosing one technique is that the two methods are documented to target different noise regimes, and cascading them, followed by SMOTE-family class balancing to correct the depressed/non-depressed imbalance typical of DAIC-WOZ-derived samples, is expected to be at least as effective as either technique alone, though this must be confirmed empirically (Section V).

B. Stage 2: Fused Spectral-Temporal Feature Extraction

Following denoising, UniVoice-Net computes two complementary feature streams in parallel. The first stream fuses MFCCs, which capture the spectral envelope via mel-scale filtering and cepstral analysis, with

autocorrelation-derived descriptors of pitch period, harmonic structure, and signal self-similarity. The second stream passes the denoised time-frequency representation through a transfer-learning convolutional backbone with a bidirectional sequence-learning layer, producing a hierarchical embedding that captures localized spectral-temporal patterns and longer-range contextual dependencies, with early layers partially frozen during fine-tuning to limit overfitting on a comparatively small labelled corpus. The two streams are concatenated into a single feature vector per utterance, so that the final representation carries both interpretable, low-dimensional acoustic descriptors and high-dimensional learned embeddings.

C. Stage 3: Two-Stage Feature Selection

The concatenated feature vector is high-dimensional and likely to contain redundant or noise-sensitive dimensions, so UniVoice-Net applies feature selection in two steps rather than one. First, a hybrid metaheuristic that alternates between Grey-Wolf-style exploitation (updating candidate solutions toward the best-performing feature subsets) and Krill-Herd-style exploration (local, foraging-inspired movement) searches the discrete feature-subset space against a fitness function balancing classification error and subset size, producing a compact, high-value subset. Second, the resulting subset is projected through a hybrid PCA-t-SNE embedding, in which PCA captures global variance structure and t-SNE preserves local neighborhood relationships, yielding a lower-dimensional representation that retains discriminative structure while further suppressing residual redundancy. Applying the metaheuristic subset search before the manifold projection, rather than either step alone, is intended to combine discrete, classifier-aware selection with continuous, structure-preserving compression.

D. Stage 4: Stacked Ensemble Classification

Rather than committing to a single classifier family, UniVoice-Net trains two base classifiers on the selected feature representation — a dense neural network with dropout regularization and a sigmoid output, and an XGBoost gradient-boosted tree classifier — and combines their outputs through a stacked meta-classifier. This design is motivated by the observation that the neural classifier and XGBoost achieve comparable but not identical performance profiles in their respective source pipelines (the neural classifier reporting perfect precision with slightly lower recall; XGBoost reporting balanced precision and recall), suggesting the two may make partially uncorrelated errors that a stacked ensemble could reconcile.

IV. DATASET AND PROPOSED EXPERIMENTAL PROTOCOL

UniVoice-Net is designed for evaluation on the DAIC-WOZ (Distress Analysis Interview Corpus – Wizard-of-Oz) dataset, a publicly available clinical interview corpus in which depression severity is labelled using the PHQ-8 questionnaire. To validate the architecture rigorously we propose the following.

A. Data Splits and Cross-Validation

Consistent with prior work on this corpus, we propose an 80/20 train-test split supplemented by 5-fold cross-validation on the training partition, with stratification to preserve the depressed/non-depressed ratio in each fold, and with SMOTE-family balancing applied only within training folds to avoid information leakage into validation or test data.

B. Baselines and Ablations

To isolate the contribution of each design choice, we propose comparing UniVoice-Net against: (i) each constituent pipeline run independently, to establish whether the unified system matches, exceeds, or underperforms its components; (ii) single-stage ablations that remove one of the four proposed stages at a time (e.g., wavelet-only denoising without the spectral-subtraction cascade, or the neural classifier alone without ensembling with XGBoost), to quantify each stage's marginal contribution; and (iii) the previously published CNN, GRU, attention-based, and multimodal transfer-learning baselines reported in the literature on this corpus, to situate UniVoice-Net within the broader field.

C. Metrics and Statistical Testing

We propose reporting accuracy, precision, recall, F1-score, specificity, Kappa coefficient, and ROC-AUC for classification performance, together with MAE and RMSE for the associated PHQ-8 severity regression task, and applying Wilcoxon signed-rank tests across cross-validation folds when comparing UniVoice-Net to each baseline and ablation, consistent with the statistical validation approach used in prior work on this dataset.

V. MOTIVATING EVIDENCE FROM CONSTITUENT COMPONENTS

While UniVoice-Net has not yet been trained end-to-end, Table I summarizes the independently reported performance of its two constituent pipelines to motivate the proposed combination.

Metric	Denoising + CNN-BiLSTM + AKWO + Neural Net	Wavelet + Cepstral Fusion + PCA/t-SNE + XGBoost
Accuracy	97.22%	97.02%
Precision	100%	97.02%

Recall	95.92%	97.02%
F1-score	97.92%	97.02%
Evaluation protocol	5-fold cross-validation, DAIC-WOZ	80/20 split, expanded DAIC-WOZ (189 sessions)

Table I. Independently reported performance of the two constituent pipelines that motivate UniVoice-Net's design.

The complementary error profile visible in Table I — near-perfect precision with comparatively lower recall in the first pipeline, versus balanced precision and recall in the second — is the central empirical motivation for the Stage 4 ensemble design in Section III-D. If the two base classifiers indeed make partially independent errors, a stacked ensemble should be able to recover some of the recall lost by the precision-oriented pipeline without sacrificing its precision advantage;

VI. Discussion And Anticipated Challenges

Combining previously validated components does not guarantee improved performance, and several risks should be anticipated. First, cascading two denoising stages could over-smooth the signal and remove low-amplitude prosodic cues that carry diagnostic information, so the ablation comparing cascaded denoising to each single-stage variant is essential rather than optional. Second, concatenating cepstral-autocorrelation features with deep embeddings substantially increases dimensionality before feature selection, which may increase training time and the risk of overfitting on a dataset of only a few hundred participants; the two-stage feature-selection block is designed to mitigate this, but its effectiveness at this larger initial dimensionality is untested. Third, stacked ensembling adds a further layer of model complexity and reduces interpretability at exactly the point where clinical stakeholders most need transparency, so any empirical validation of UniVoice-Net should be paired with post-hoc explainability analysis (e.g., feature-importance attribution from the XGBoost branch, or attention visualization from the neural branch) rather than treated as a purely accuracy-driven exercise.

More broadly, both constituent pipelines were evaluated on a single public corpus recorded in a controlled, English-language interview setting. Even if UniVoice-Net matches or exceeds 97% accuracy in that setting, this would not establish generalization to naturalistic recordings, other languages, or other cultural expressions of depressive symptoms, and any claims about clinical utility would require prospective, external validation beyond the scope of a DAIC-WOZ benchmark study.

VII. Conclusion

Voice-based depression screening sits at a valuable intersection of accessibility and clinical relevance, offering a low-cost, non-invasive pathway to early mental health assessment at a scale that conventional interview-based screening cannot match. This paper has presented UniVoice-Net, a unified framework that brings together dual-domain denoising, fused spectral-temporal feature extraction, two-stage feature selection, and stacked ensemble classification into a single, coherent pipeline for detecting depression from speech. By integrating complementary denoising strategies, combining handcrafted and learned feature representations, applying metaheuristic and manifold-based feature selection in sequence, and pairing a neural classifier with a gradient-boosted ensemble, UniVoice-Net is designed to capture a broader and more robust set of depression-relevant vocal cues than pipelines built around any single technique.

The architecture and experimental protocol described in Sections III and IV provide a complete, reproducible blueprint — covering denoising, feature engineering, feature selection, classification, dataset handling, baselines, and evaluation metrics — that can be carried directly into implementation and testing on the DAIC-WOZ corpus and, subsequently, on more diverse clinical speech datasets. Future work will focus on extension to multilingual and naturalistic recording conditions, integration of explainability techniques to support clinical trust, and deployment within telehealth and remote-monitoring platforms for continuous, longitudinal mental health tracking. Taken together, this line of work points toward voice-based screening tools that can meaningfully extend the reach of mental health services and support earlier, more equitable identification of depression risk.

References

- [1] J. Jithin, K.S. Kannan, "SonoMind: deep learning-based voice analysis for mental health monitoring," *Rivista di Psichiatria*, vol. 61, no. 1, pp. 23–32, 2026.
- [2] J. Jacob, K.S. Kannan, "Enhanced Machine Learning Framework for Autonomous Depression Detection Using ModWave Cepstral Fusion and Stochastic Embedding," *Informatics and Automation*, vol. 23, no. 6, pp. 1754–1780, 2024.

- [3] J. Gratch, R. Artstein, G. Lucas, G. Stratou, S. Scherer, A. Nazarian, R. Wood, J. Boberg, D. DeVault, S. Marsella, D. Traum, S. Rizzo, L.-P. Morency, "The Distress Analysis Interview Corpus of human and computer interviews," in Proc. 9th Int. Conf. on Language Resources and Evaluation (LREC), Reykjavik, Iceland, 2014.
- [4] World Health Organization, "Depression and other common mental disorders: global health estimates," WHO/MSD/MER/2017.2, 2017.
- [5] N. Cummins, V. Sethu, J. Epps, S. Schnieder, J. Krajewski, "Analysis of acoustic space variability in speech affected by depression," *Speech Communication*, vol. 75, pp. 27–49, 2015.
- [6] T. Chen, C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in Proc. 22nd ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining, 2016, pp. 785–794.
- [7] L. van der Maaten, G. Hinton, "Visualizing Data using t-SNE," *Journal of Machine Learning Research*, vol. 9, pp. 2579–2605, 2008.
- [8] N.V. Chawla, K.W. Bowyer, L.O. Hall, W.P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, 2002.