Alexandria Engineering Journal 136 (2026) 89–104

Contents lists available at ScienceDirect

Alexandria Engineering Journal

journal homepage: www.elsevier.com/locate/aej

Original article

VoxVeritasNet: A new feature engineering model leveraging iterative feature selection for detecting fake or real speech

Burak Çelik [a], Burcu Zeybek [b], Mahmut Burak Karadeniz [a,*], Adem Kocyigit [a,c], Onur Arsalı [a], Ebru Efeoglu [d], Bahattin Türetken [a]

[a] Department of Electronics and Communication Engineering, Kocaeli University, Kocaeli, 58140, Türkiye
[b] Department of Public Relations and Publicity, Istanbul University, Istanbul, 34126, Türkiye
[c] Department of Electronics and Automation, Bilecik Seyh Edebali University, Bilecik, 11100, Türkiye
[d] Department of Software Engineering, Kütahya Dumlupınar University, Kütahya, 43100, Türkiye

ARTICLE INFO

Keywords:
VoxVeritasNet
Deepfake speech detection
Feature engineering
Machine learning

ABSTRACT

This study introduces VoxVeritasNet, a high-precision and computationally efficient feature engineering framework for deepfake audio detection. The methodology leverages a nine-level Multi-Level Discrete Wavelet Transform (MDWT) to capture intricate time–frequency artifacts. A key innovation is the quantum-inspired dual-path mapping algorithm, which models parallel signal dependencies and embeds features into a high-dimensional Hilbert space for enhancing geometric separability. To optimize performance, an iterative ensemble selection strategy utilizing Neighborhood Component Analysis (NCA), Chi^2, and ReliefF is employed alongside Support Vector Machines and k-Nearest Neighbors. The framework was evaluated across three public datasets with varying class distributions, achieving state-of-the-art peak accuracies of 99.96% with db4 and 99.71% with sym8 wavelets. Even using with the computationally efficient sym4 baseline, the model maintained exceptional detection rates above 98.99% and an equal error rate (EER) as low as 0.14%. VoxVeritasNet operates with a processing throughput of 6.45 segments per second on standard CPU hardware with a negligible storage footprint, offering a lightweight and explainable alternative to resource-intensive deep learning architectures.

1. Introduction

In recent years, rapid advancements in DL and neural networks have transformed generative artificial intelligence (AI), enabling systems to autonomously produce synthetic content including text, images, video, and audio files [1]. These technologies have led to innovative applications across various domains, including conversational agents (e.g., ChatGPT), film production [2], code generation [3,4], and audio synthesis [5,6]. However, the increased use of AI-powered content creation technology raises serious public safety concerns as it puts the technology at risk of being used for illicit purposes by criminals [7]. The necessity for advanced detection systems is underscored by alarming trends in synthetic media. According to the *Voice Intelligence and Security Report 2025*, the prevalence of deepfake-driven fraud is projected to increase by approximately 162% in 2025, with the volume of detected synthetic audio and video files escalating from nearly 500,000 to an estimated 8 million. Furthermore, fraud attempts have surged by 3000% globally — including a 1740% spike in North America — while overall deepfake attack rates have risen by over 1300% in the past year. These statistics reflect a rapidly intensifying wave of AI-enabled deception that further narrows the perceptual gap between authentic and artificially generated content [8]. To address this threat, researchers have developed methods to detect and mitigate the risks associated with synthetic data, commonly referred to as "deepfakes". Deepfakes are generated using deep neural networks inspired by biological nervous systems that can combine and manipulate images [9].

Unlike traditional computer-generated imagery (CGI), modern deepfakes leverage specialized neural architectures, such as autoencoders, to autonomously manipulate facial expressions, skin tone, and head orientation without requiring artistic expertise. These systems can realistically reconstruct target individuals by learning specific gestures and biometric traits. In the audio domain, these capabilities manifest as Replay, Text-to-Speech (TTS), and Voice Conversion (VC) attacks, posing severe security threats including identity theft, fraud, and the fabrication of forensic evidence [10–13]. Such high-fidelity synthetic voices particularly threaten the integrity of voice-based biometric authentication and secure communication systems used in sensitive military and political sectors [14].

* Corresponding author.
E-mail address: burak.karadeniz@kocaeli.edu.tr (M.B. Karadeniz).

https://doi.org/10.1016/j.aej.2026.01.009
Received 24 September 2025; Received in revised form 28 December 2025; Accepted 6 January 2026
1110-0168/© 2026 The Authors. Published by Elsevier B.V. on behalf of Faculty of Engineering, Alexandria University. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

The rapid evolution of Deep Learning (DL) has enabled the production of multi-modal synthetic content that is increasingly indistinguishable from reality, often exploited for social manipulation during critical global events [15–18]. Beyond audio–visual deepfakes, AI-assisted text generation facilitates the spread of disinformation and fake news, making the development of robust technical detection frameworks and legal countermeasures [19–21].

As biological perception fails to identify subtle neural vocoder artifacts, advanced Machine Learning (ML) and transfer learning techniques have become crucial for detecting micro-patterns invisible to the human ear [22,23]. For instance, methodologies focusing on rhythmic speech pauses have demonstrated that pattern-based detection can achieve significant baseline results [24]. Consequently, research has shifted towards developing specialized feature engineering models capable of distinguishing fake or real (FoR) speech by extracting granular acoustic cues from raw audio data.

Hamza et al. [25] conducted studies on detecting high-fidelity synthetic voices generated by artificial intelligence. In this approach, the "FoR" dataset, which was generated using TTS models, was divided into four subsets for analysis. Mel-frequency cepstral coefficients (MFCC) features were extracted, and various machine learning models were tested. The SVM model stood out, achieving 98.83% and 97.57% accuracy on the initial and two-second subsets, respectively. Gradient boosting achieved 92.63% accuracy on the normalized dataset, while VGG-16 performed best on the original subset.

Mcuba et al. [26] examined the impact of voice cloning technologies on digital forensics. They evaluated audio features extracted using techniques such as MFCC, spectrogram, and chromagram with different DL architectures. A specialized architecture achieved the best results with spectrogram-based features, and VGG-16 achieved the best results with MFCC. Similarly, in [27], optimized feature engineering was applied to the FoR dataset generated with TTS models, and six different ML classifiers were tested. This approach increased accuracy by 26% compared to the baseline model, with the XGBoost model achieving the best performance at 93%.

Lim et al. [28] investigated deepfake audio detection using explainable artificial intelligence (XAI) techniques adapted to the audio domain. They revealed differences in citation score patterns, formats, and speech rhythm. This approach more closely aligns with human understanding by explaining model decisions. The study suggests that security can be improved in the future with automatic annotators, eliminating the need for human recognition. Another study, [29], presented a lightweight ML framework based on traditional spectral, temporal, and color-based hand-extracted features. With 89%, 94.5%, and 94.67% accuracy on the ASVSpoof2019, FakeAVCelebV2 and In-The-Wild datasets, respectively, outperformed state-of-the-art methods. Explainability techniques also increased the transparency of model decisions.

Khochare et al. [30] focused on the threat that advanced synthetic speech generation poses to speaker verification systems and presented two approaches to audio deepfake detection. One approach is based on ML spectral audio features, while the other is an image-based approach that converts audio into Mel spectrograms and analyses them with DL models, such as Temporal Convolutional Network (TCN) and Spatial Transformer Network (STN). The comparative analysis revealed that the TCN model achieved 92% test accuracy, surpassing machine learning methods and rivaling CNN-based models of VGG16 and XceptionNet. Similarly, [31] developed and trained Deep Neural Network (DNN)-based models on various datasets including UrbanSound8K, Conversational, AMI-Corpus, FoR and integrates four key components: (1) the denoising module, which achieved 93% and 94% accuracy using Multilayer Perceptron (MLP) and CNN architectures, respectively; (2) text conversion, which was performed with 93% accuracy using Natural Language Processing methods; (3) speaker labeling, which was performed with 80% accuracy and a voicing error rate of 0.52 using a Recurrent Neural Network; and (4) the CNN-based classifier, which distinguishes real and synthetic audio with 94% accuracy. These results demonstrate the effectiveness of the proposed approach in detecting synthetic speech in complex scenarios, offering significant contributions to speech analysis.

Recent advancements in feature selection have increasingly leveraged nature-inspired and evolutionary algorithms to navigate high-dimensional search spaces effectively. For example, the Halton-enhanced Human Evolutionary Algorithm (Halton-HEA) has demonstrated superior performance in maintaining population diversity and avoiding local optima [32], while bio-inspired approaches like the Improved African Vulture Optimization Algorithm (IAVOA) have shown robustness in handling sparse data structures [33]. In addition to evolutionary methods, physics-based and chaotic optimization strategies, such as the Improved Flow Direction Algorithm (IFDA) [34] and the Chaotic RIME optimization algorithm [35], have emerged as powerful tools for enhancing convergence stability. Furthermore, multi-objective frameworks like the Non-dominated Sorting Imperialist Competitive Algorithm (NSICA) [36] have proven effective in balancing classification accuracy with feature reduction.

This study introduces VoxVeritasNet, a multi-stage framework engineered for robust deepfake audio detection through a combination of advanced signal processing and ML techniques. The methodology commences with the decomposition of 3-second audio segments using a nine-level MDWT with a sym4 wavelet, enabling the extraction of intricate audio patterns across various frequency sub-bands. The core of this framework is the novel VoxVeritasNet function, which generates a rich set of histogram-based and statistical features that capture complex audio characteristics. To enhance detection accuracy, an ensemble of feature selection methods — specifically NCA, Chi^2, and ReliefF — is applied to create a refined and highly discriminative feature set. These optimized features are then used to train and evaluate k-Nearest Neighbors (kNN) and Support Vector Machine (SVM) classifiers, with performance validated via a 10-fold cross-validation protocol. Furthermore, the framework incorporates an Information Matrix Voting (IMV) mechanism to systematically combine the outcomes from the different feature selectors, generating multiple voting-based feature vectors to boost classification performance. By employing a greedy selection process to compare the performance of both kNN and SVM classifiers across all generated feature sets, the single best-performing model is identified and its efficacy is evaluated using a confusion matrix to assess its ability to distinguish between FoR real audio classes.

The primary contributions of this work to the field of deepfake audio detection are summarized as follows. First, we introduce VoxVeritasNet, a novel feature engineering framework that integrates 9-level MDWT decomposition with a quantum-inspired dual-path mapping algorithm. This approach allows for the simultaneous capture of intricate temporal and frequency-domain artifacts that are often overlooked by conventional architectures. Second, we implement a unique mathematical mapping of audio features into a high-dimensional Hilbert space, providing a unit-vector representation that significantly enhances the geometric separability between authentic and synthetic speech classes. Third, we demonstrate that by employing an iterative ensemble feature selection strategy consisting of neighborhood component analysis, Chi^2, and ReliefF, the model achieves state-of-the-art performance with peak accuracies of 99.96% and 99.71% for db4 and sym8 wavelets, respectively. Finally, we provide a computationally efficient solution that maintains an exceptionally low EER of 0.14% and a high processing throughput on standard CPU hardware, ensuring the framework's practical applicability for real-time forensic scenarios without the need for high-end GPU resources.

2. Methodology and contributions

A central challenge in audio deepfake detection is resolving the trade-off between classification accuracy and computational efficiency.

Table 1
Methodological comparison of existing frameworks against VoxVeritasNet.

Category	Methodology	Strengths	Limitations/Gaps
Deep learning [27,31,39]	CNN/TCN	Spatial feature extraction	High GPU demand; lack of XAI
Transformers [32–36]	Hybrid ViT	Long-range dependencies	Computational overhead
Explainable AI [29,30]	XAI Systems	Model transparency	Accuracy vs. Complexity trade-off
Traditional ML [25–28]	MFCC/Acoustic	Low latency	Vulnerable to synthetic noise
VoxVeritasNet	**MDWT + Iterative FS**	**Robust artifacts capture**	**99.81% Acc.; Deterministic**

The current literature is dominated by two main approaches, both of which have significant limitations. DL models, while powerful, are they often computationally expensive "black boxes" that demand substantial resources, making them difficult to deploy in real-time systems [37]. Conversely, traditional methods using handcrafted features like MFCCs are efficient but often fail to capture the high-frequency, subtle artifacts that characterize synthetically generated audio, leading to reduced accuracy [38]. These issues highlight several critical gaps in existing research, including an over-reliance on standard acoustic features, a lack of novel feature engineering models capable of capturing intricate signal details, and underutilization of robust, multi-method feature selection strategies.

To further clarify the research gap and the contributions of our work, Table 1 provides a comparative analysis of the methodologies discussed in the literature, highlighting their strengths and inherent limitations in deepfake detection. The recent state-of-the-art (SOTA) researches, including the most recent 2024–2025 studies, are categorized on their core architectural strategies and inherent limitations. In this context, while DL models based on CNN and TCN architectures excel in spatial feature extraction, they face structural constraints such as high GPU requirements and a lack of explainable AI (XAI). Hybrid Vision Transformer (ViT) architectures possess the capability to model long-range dependencies within the data; but, they face efficiency challenges due to high computational costs and the necessity for large datasets. Although XAI systems aimed at model transparency clarify decision-making mechanisms, they typically exhibit a trade-off between accuracy and parametric complexity. Finally, traditional machine learning approaches based on MFCCs and similar acoustic features offer low-latency advantages, but they demonstrate limited performance due to their sensitivity to synthetic noise and inconsistencies in capturing the micro-frequency traces produced by modern vocoders.

VoxVeritasNet specifically addresses the gaps identified in these categories by combining the interpretability of feature engineering with the high-performance capabilities of ensemble selection. While DL and transformer models provide strong detection rates, their high parametric complexity and stochastic training nature make them less suitable for low-resource or real-time forensic screening. In contrast, our proposed framework utilizes MDWT for deterministic signal decomposition and an iterative ensemble selection (NCA, Chi^2, ReliefF). This methodology ensures that only the most discriminative features are retained, thereby overcoming the overfitting issues and computational bottlenecks highlighted in the recent literature. To address the research gaps identified in current literature, this work follows a comprehensive methodology based on the following three-stage framework:

- Advanced wavelet-based feature extraction: Instead of relying on conventional features, VoxVeritasNet employs MDWT. This allows for a detailed time–frequency analysis of the audio signal, effectively capturing transient, non-stationary details that are often missed by methods like MFCCs. This directly addresses the gap of underutilized, advanced signal processing techniques.
- Quantum-inspired feature engineering: The framework's key innovation is a novel algorithm that generates dual binary sequences from audio blocks, modeling dynamic signal transitions in a manner inspired by quantum state changes. This process creates rich and detailed histogram and statistical features, providing a more discriminative representation of the audio and filling the need for more innovative feature engineering models.

Table 2
Distributions of the first, second and third datasets.

Dataset	Distinct labels	Number of segments	Distribution of segments (%)	Total
First dataset	Fake sounds Real sounds	2375 2527	48.4 51.6	4902
Second dataset	Fake sounds Real sounds	2038 2892	41.3 58.7	4930
Third dataset	Fake sounds Real sounds	8769 1207	87.9 12.1	9976

- Ensemble feature selection: To ensure that only the most salient features are used for classification, the framework implements a multi-stage selection strategy. It combines the strengths of NCA, Chi^2, ReliefF, and an IMV mechanism. This ensemble approach creates a highly optimized feature set, addressing the gap of underutilized feature selection techniques and enhancing classifier performance.

By integrating these innovations, VoxVeritasNet establishes a new methodology that delivers high classification accuracy with linear time complexity, providing a powerful and efficient solution to the challenges of modern deepfake detection.

3. Audio dataset for deepfake detection with real and synthetic samples

The datasets used in this study were obtained from Kaggle and are distributed under established open-source licenses. Specifically, Dataset 1 is licensed under the CC BY-NC-SA 4.0 Licenses [40], Dataset 2 under the MIT License [41], and Dataset 3 under the CC0 Public Domain License [42].

The datasets used to evaluate the proposed VoxVeritasNet feature extraction framework are detailed in Table 2 depending on the distribution of distinct labels and number of segments. In the first dataset, 4902 sound segments was used in balanced, with 2375 fake (48%) and 2527 real (52%) audio samples. This equilibrium supports consistent model performance by preventing class dominance and ensuring fair representation during training and evaluation. The second dataset involved a slightly imbalanced of 4930 segments, including 2038 fake (41.3%) and 2892 real (58.7%) samples. Although the minor skew favors real audio, the dataset remains diverse and sufficiently representative to support meaningful classification insights. In the third dataset, a highly imbalanced distributed 9976 segments was employed, with 8769 fake (87.9%) and 1207 real (12.1%) samples. Despite the pronounced asymmetry, this dataset offers a realistic scenario for testing model robustness and generalizability under non-ideal conditions. Collectively, the three datasets provide a comprehensive framework for evaluating VoxVeritasNet across balanced, mildly imbalanced, and heavily skewed distributions, highlighting its adaptability and classification effectiveness under varied conditions.

The amplitude profile of a sound samples from the FoR sounds in the curated balanced dataset is visualized along a 3-second time axis, as depicted in the accompanying Fig. 1a, b and c for the first, second and third datasets, respectively. Third dataset holds particular significance due to its inclusion of audio samples from prominent

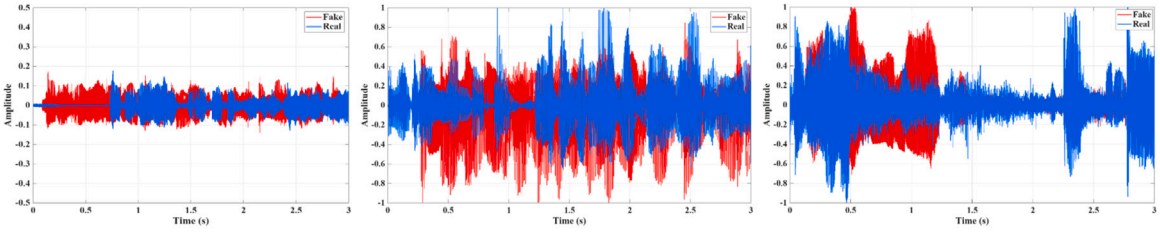


Fig. 1. Time-domain amplitude representation of sample audio signals from the constructed balanced datasets: (a) balanced, (b) slightly imbalanced and (c) heavily imbalanced datasets.

individuals, including Joe Biden, Donald Trump, Elon Musk, Barack Obama, Taylor Swift, and Margot Robbie. The presence of these high-profile voices underscores its critical role in detecting synthetic audio manipulations that could be exploited in political manipulations or perception operations. In an era where deepfake technology poses significant risks to public discourse, electoral integrity, and societal trust, the capability to identify falsified audio from such influential figures is of utmost importance [43]. The application of the VoxVeritasNet framework to this dataset could substantially contribute to countering misinformation campaigns and enhancing the reliability of audio-based communications.

4. The proposed VoxVeritasNet model

The VoxVeritasNet framework is a multi-stage methodology engineered for the robust detection of deepfake audio. The architecture systematically processes audio signals through a sequential pipeline comprising five principal stages: (i) multi level feature extraction, (ii) ensemble feature selection, (iii) classification and training, (iv) information fusion based post-processing, and (v) validation. The model is designed to first generate a rich set of discriminative features from raw audio, then refine this set by selecting the most informative attributes, and finally fuse this information to achieve high classification accuracy. A schematic representation of the entire process is illustrated in Fig. 2.

The complete pseudocode for the VoxVeritasNet methodology, presented in Algorithm 1, employs a deterministic and nested validation strategy. This ensures that the processes of feature selection and model choice are entirely isolated from the final evaluation. By processing 3-second audio segments through a nine-level multi-level discrete wavelet transform, a refined set of the 135 most discriminative features is identified within each training fold. This approach ensures that the selection of the optimal model from the 30 generated candidates is conducted using a strictly sealed validation protocol. The performance metrics reported thereafter are derived from a held-out test set that remained untouched during the validation stage. This rigorous separation eliminates the risk of circularity and confirms that the high accuracy and low equal error rate are representative of the model's genuine capability to generalize across unseen synthetic audio samples.

Algorithm 1 The VoxVeritasNet Framework

Input: Raw audio signal $\mathcal{A}$, Full Dataset $\mathcal{D}$, Class Labels Y
Output: Final classification label {Authentic, Deepfake}, Final hold-out test score P_{HO}

Main Procedure
$W \leftarrow \text{MDWT}(\mathcal{A}, \text{level} = 9)$ ▷ Stage 1.1: Signal Decomposition
$F_{total} \leftarrow \emptyset$
for each wavelet band $w_x \in W$ $(x = 1 \dots 9)$ **do**
  $f_{VVN} \leftarrow \text{VoxVeritasExtractor}(w_x)$ ▷ Stage 1.2: Dual-path patterns
  $f_{Stat} \leftarrow \text{StatisticalExtractor}(w_x)$ ▷ Stage 1.3: 16 metrics
  $f_x \leftarrow f_{VVN} \cup f_{Stat}$ ▷ Total 528 features per level
  $\mathcal{FS}_x \leftarrow \text{EnsembleFeatureSelection}(f_x, Y)$ ▷ Stage 2: Select 15 features per level
  $F_{total} \leftarrow F_{total} \cup \mathcal{FS}_x$ ▷ Accumulate to final 135 features
end for
$\{\mathcal{M}_{1..30}\} \leftarrow \text{TrainClassifiers}(F_{total}, Y)$ ▷ Stage 3: Training with 135 features
$\mathcal{M}^* \leftarrow \text{SelectBestModel}(\{\mathcal{M}_{1..30}\})$ ▷ Stage 4: Optimization
$P_{CV}, P_{HO} \leftarrow \text{ValidateModel}(\mathcal{M}^*, \mathcal{D}, F_{total})$ ▷ Stage 5: 10-fold Validation
return $\mathcal{M}^*$, P_{HO}

procedure VoxVeritasExtractor(w_x) ▷ Inter-block Comparison Logic
  Partition w_x into N non-overlapping blocks $B = \{b_1, \dots, b_N\}$
  $List_{min}, List_{max} \leftarrow \emptyset, \emptyset$
  for $i \leftarrow 1$ to $N - 8$ **do**
    $v_{min}, v_{max} \leftarrow [], []$
    for $j \leftarrow 0$ to 7 **do** ▷ Generate 8-bit patterns
      Min-Route: $bit \leftarrow 1$ **if** $\min(b_{i+j}) > \min(b_{i+j+1})$ **else** 0
      Max-Route: $bit \leftarrow 1$ **if** $\max(b_{i+j}) > \max(b_{i+j+1})$ **else** 0
      Append bit to v_{min}, v_{max}
    end for
    Append Bin2Dec(v_{min}) to $List_{min}$ and Bin2Dec(v_{max}) to $List_{max}$
  end for
  $\psi_{min}, \psi_{max} \leftarrow \text{HilbertMap}(\text{Hist}(List_{min})), \text{HilbertMap}(\text{Hist}(List_{max}))$
  return $\{\psi_{min}, \psi_{max}\}$
end procedure

procedure EnsembleFeatureSelection(f_x, Y) ▷ Iterative Subtractive Logic
  $S_1 \leftarrow \text{NCA}(f_x, Y, k = 5)$
  $f'_x \leftarrow f_x \setminus S_1$
  $S_2 \leftarrow \text{Chi}^2(f'_x, Y, k = 5)$
  $f''_x \leftarrow f'_x \setminus S_2$
  $S_3 \leftarrow \text{ReliefF}(f''_x, Y, k = 5)$
  return $\{S_1 \cup S_2 \cup S_3\}$ ▷ Returns 15 optimized features
end procedure

procedure TrainClassifiers(F_{total}, Y)
  $\mathcal{M}_{all} \leftarrow \emptyset$
  for $j \leftarrow 1$ to 15 (indices of selection configurations) **do**
    $\mathcal{M}_{j,\text{kNN}} \leftarrow \text{Train}(\text{kNN}, F_{total}[j])$
    $\mathcal{M}_{j,\text{SVM}} \leftarrow \text{Train}(\text{CubicSVM}, F_{total}[j])$
    $\mathcal{M}_{all} \leftarrow \mathcal{M}_{all} \cup \{\mathcal{M}_{j,\text{kNN}}, \mathcal{M}_{j,\text{SVM}}\}$
  end for
  return $\mathcal{M}_{all}$
end procedure

procedure SelectBestModel($\{\mathcal{M}\}$)
  $P_{\text{max}} \leftarrow 0$
  for each model $\mathcal{M}_i \in \{\mathcal{M}\}$ **do**
    $P_i \leftarrow \text{Evaluate}(\mathcal{M}_i, \text{Validation Set})$
    if $P_i > P_{\text{max}}$ **then** $P_{\text{max}} \leftarrow P_i, \mathcal{M}^* \leftarrow \mathcal{M}_i$
    end if
  end for
  return $\mathcal{M}^*$
end procedure

procedure ValidateModel($\mathcal{M}^*, \mathcal{D}, F_{total}$)
  Partition D into 10 folds $\{D_1, \dots, D_{10}\}$

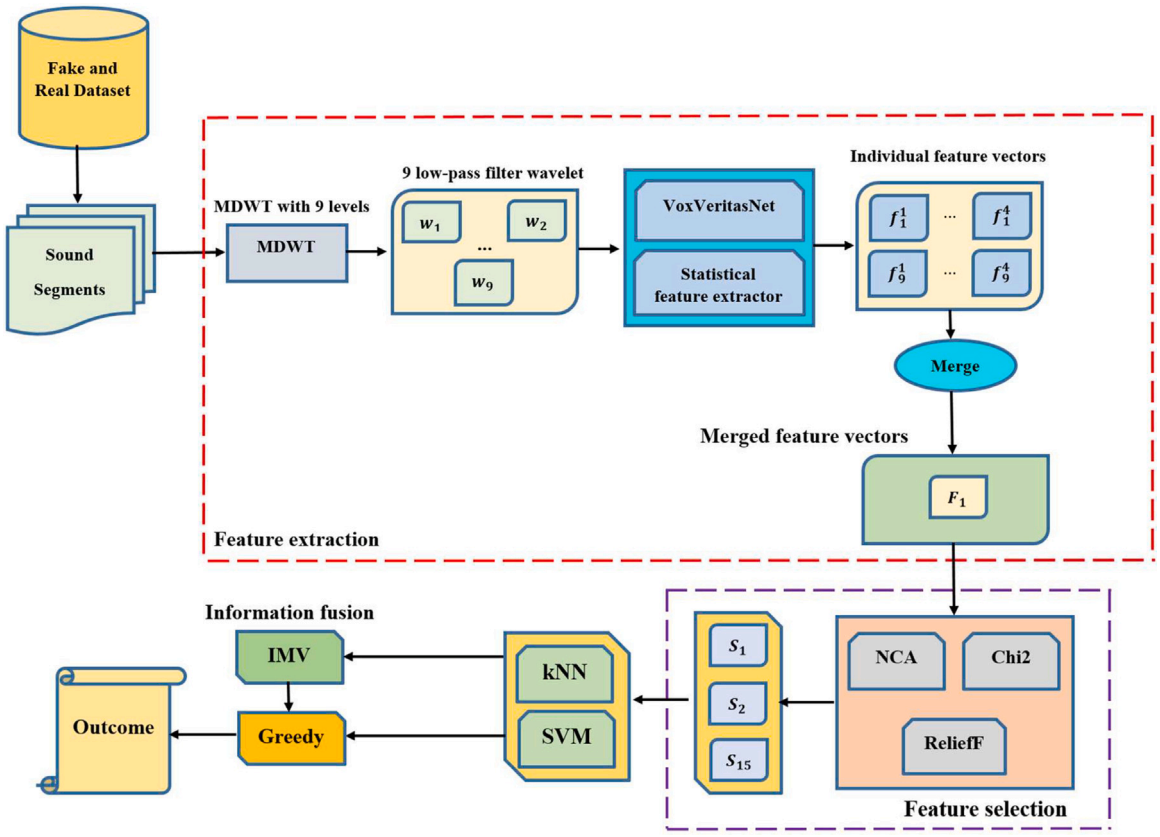


Fig. 2. Schematic representation of the proposed VoxVeritasNet method based on signum, dual-path binary mapping, and NCA-Chi^2-ReliefF for audio deepfake detections.

57: **for** $i \leftarrow 1$ to 10 **do**
58: $\mathcal{M}_i \leftarrow \text{Train}(\mathcal{M}^*, \{\mathcal{D} \setminus \mathcal{D}_i\}$ with $F_{total})$
59: $Score_i \leftarrow \text{Evaluate}(\mathcal{M}_i, \mathcal{D}_i$ with $F_{total})$
60: **end for**
61: **return** Average($Score_{1..10}$)
62: **end procedure**

4.1. Stage 1: Multi-level feature extraction

The initial stage focuses on generating a comprehensive feature set by capturing both statistical and structural characteristics of the input audio. This is accomplished through a multi-level framework that combines wavelet decomposition with a novel dual-channel feature extractor (Algorithm 1, lines 1–6).

4.1.1. Data preprocessing and normalization

To ensure the reproducibility of our results and provide full transparency regarding our data preparation, a standardized preprocessing pipeline was strictly followed. Each audio recording was first converted to a single-channel (mono) format and normalized based on its maximum absolute amplitude to ensure dynamic range consistency. To normalize the duration of the heterogeneous dataset, a non-overlapping segmentation strategy was adopted, partitioning each signal into fixed 3-second segments. Recordings shorter than this threshold were discarded to prevent zero-padding artifacts. Notably, no explicit resampling, noise suppression, or silence trimming was applied. Signals were processed at their native sampling rates to maintain spectral integrity, and the decision to retain background noise and silences was intentional. This approach enables the VoxVeritasNet framework to be evaluated in unconstrained, real-world scenarios, demonstrating its robustness in detecting deepfake patterns without the need for ideal laboratory-clean speech conditions.

The process begins with the segmentation of the raw audio signal. The signal ($sgnl$) is first segmented into overlapping blocks (v) of 36 samples each.

$$v(i) = sgnl(i + j - 1), \quad i \in \{1, 2, \ldots, 36\}, \quad j \in \{1, 2, 3, \ldots, len - 35\} \tag{1}$$

where len is the length of the signal. Then, $sgnl$ is decomposed using a 9-level MDWT with the sym4 wavelet. This technique decomposes the signal into various frequency sub-bands, enabling a detailed time–frequency analysis crucial for exposing subtle artifacts characteristic of synthetic speech. The decomposition is represented as:

$$[w_1, e_1] = \psi(v) \tag{2}$$

$$[w_{t+1}, e_{t+1}] = \psi(w_t), \quad t \in \{1, 2, \ldots, 8\} \tag{3}$$

where ψ represents the DWT decomposition function, w is the low-pass wavelet bands, and e is the high-pass wavelet bands. Capturing only the low-pass wavelet bands, the decomposed signals are then processed through a dual-channel extraction module:

1. The novel VoxVeritasNet feature extractor, which applies dual-path algorithm to generate detailed histogram based features.
2. A parallel statistical feature extractor, which computes conventional statistical metrics.

4.1.2. VoxVeritasNet feature extractor

VoxVeritasNet feature extractor is designed to model complex local structural patterns through a dual-path binary mapping technique. The operational steps are as follows (Algorithm 1, lines 14–28):

Dual-path generation: The graph index structure shown in Fig. 3, which is modeled on tree circuit patterns, is defined over 36 interconnected blocks. The pathways between these blocks are engineered to mirror the characteristic state transitions of tree circuit frameworks. Within

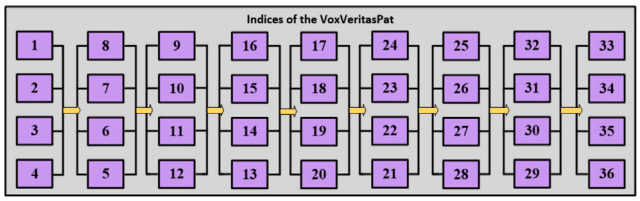


Fig. 3. Indices of the proposed VoxVeritasNet.

this structure, two distinct paths are established: one connecting the maximum values of adjacent blocks and another connecting the minimum values. This dual-path methodology emulates the principle of superposition, thereby enabling the concurrent evaluation of multiple feature states.

Binary pattern mapping: For each of the two paths, an 8-bit binary sequence is generated by sequentially comparing adjacent values as:

$$bit1_f = Y(route1_f, route1_{f+1}), \quad f \in \{0, 1, \dots, 7\} \tag{4}$$

$$bit2_f = Y(route2_f, route2_{f+1}), \tag{5}$$

where the comparison is governed by the function Y as:

$$Y(a,b) = \begin{cases} 0 & \text{if } a - b < 0 \\ 1 & \text{if } a - b \geq 0 \end{cases} \tag{6}$$

which assigns a binary output based on the relationship between consecutive values, as detailed in Eqs. (4)–(6).

The dependency between adjacent values established through this process is conceptually analogous to quantum entanglement. In this work, quantum principles are employed in a quantum-inspired manner to model parallel decision paths and symbolic dependencies within local signal blocks. Rather than simulating physical quantum circuits, the proposed framework adopts mathematical abstractions such as superposition, entanglement-inspired dependency, and Hilbert space embedding to enhance the representational capacity of binary comparison-based features.

The following equations provide the formal representation of this quantum-inspired encoding framework. Let $A = \{a_1, a_2, a_3, \dots, a_n\}$ be the input sequence and $P^{(b)} = \{(i_k, j_k)\}_{k=1}^{m}$ represent the pattern of index pairs compared in block b. Here, b denotes the index of the current local block, m is the number of binary comparisons within a single block pattern, and k is the index of an individual comparison in the set $P^{(b)}$. The resulting binary feature vector $v^{(b)}$ for block b are defined as:

$$v^{(b)} = [f(a_{i1}, a_{j1}), f(a_{i2}, a_{j2}), \dots, f(a_{im}, a_{jm})] \in \{0,1\}^m, \tag{7}$$

where the binary comparison function $f(a_i, a_j) = 1(a_i \geq a_j)$. To reflect the simultaneous evaluation of the minimum- and maximum-based routes, we form a normalized superposition of the two binary vectors as shown below. The weights α and β are set data-dependently from the relative energies of each path, ensuring full compatibility with the block-wise construction:

$$\psi^{(b)} = \alpha v_{min}^{(b)} + \beta v_{max}^{(b)}, \quad |\alpha|^2 + |\beta|^2 = 1, \tag{8}$$

where,

$$\alpha = \sqrt{\frac{\|v_{min}^{(b)}\|_2}{\|v_{min}^{(b)}\|_2 + \|v_{max}^{(b)}\|_2}}, \quad \beta = \sqrt{\frac{\|v_{max}^{(b)}\|_2}{\|v_{min}^{(b)}\|_2 + \|v_{max}^{(b)}\|_2}} \tag{9}$$

This superposed representation maintains the original bit structure while providing a single state vector for downstream feature formation. The specific comparison functions for minimum and maximum routes are defined as $f_{min}(a_i, a_j), f_{max}(a_i, a_j)$, where $1(a_i \geq a_j)$, leading to the corresponding vectors $v_{min}^{(b)}$ and $v_{max}^{(b)}$. The dependence between bit positions within a block can be quantified in order to capture coupling beyond independent paths. For bit indices $i \neq j$, the feature entanglement $Ent(i,j)$ and mutual information $I(i;j)$ quantify the dependency between symbolic transitions:

$$Ent(i,j) = E[v_i^{(b)} v_j^{(b)}] - E[v_i^{(b)}]E[v_j^{(b)}] \tag{10}$$

$$I(i;j) = \sum_{b_i, b_j \in \{0,1\}} p(b_i, b_j) \log \frac{p(b_i, b_j)}{p(b_i)p(b_j)} \tag{11}$$

This transformation encodes all comparison paths into a unique identifier $N^{(b)}$ for the block, enabling histogram-based feature extraction where $N^{(b)} \in \mathbb{N}$:

$$N^{(b)} = \sum_{k=1}^{m} v_k^{(b)} * 2^{m-k} \tag{12}$$

Finally, the histogram is mapped to a unit vector in a 2^m-dimensional Hilbert space. Let $p_k^{(l)}$ represent the empirical probabilities over codes $k \in \{0, \dots, 2^m - 1\}$ at a given MDWT level l:

$$|\psi^l\rangle = \sum_{k=0}^{2^m-1} \sqrt{p_k^{(l)}} |k\rangle, \quad \| |\psi^l\rangle \|_2 = 1 \tag{13}$$

This transformation enables inner product-based analyses to compare frequency and rhythm patterns of different sounds while preserving histogram information.

The feature extraction mechanism of VoxVeritasNet is illustrated in Fig. 4. The process operates on a given data block by establishing two parallel computational pathways, one guided by a maximum value selection criterion and the other by a minimum value criterion. Each path executes a series of sequential, pairwise comparisons between data points to generate an 8-bit binary pattern. These patterns act as concise and powerful feature descriptors, encoding the localized relational dynamics of the signal. This dual-path methodology is detailed as follows:

- Maximum value path: The process follows a sequence of comparisons to determine the path of local maxima. For example, the first comparison evaluates '19' against '18', yielding a bit '1' since 19 > 18. Subsequent comparisons, such as '18–14 >0' (Bit '1') and '14–17 < 0' (Bit '0'), continue this process. The complete sequence generates the binary pattern $(11011101)_2$, which corresponds to the decimal feature value 221.
- Minimum value path: Concurrently, a second path is traced based on local minima. This path involves comparisons like '1–4 < 0' (Bit '0') and '5–3 > 0' (Bit '1'). This sequence results in the distinct binary pattern $(00101001)_2$, yielding the decimal feature value 41.

Histogram-based feature extraction: In the final feature extraction stage, a 256-bin histogram is generated from the distribution of each of the two mapped signals as:

$$feature_z = HilbertMap(Hist(List(z))). \quad z \in \{min, max\} \tag{14}$$

These resulting histograms serve as feature vectors that encapsulate the statistical distribution of local structural patterns within the audio signal. The proposed quantum-inspired binary pattern generation significantly diverges from conventional Local Binary Pattern (LBP) techniques. While traditional 1D-LBP variants utilize a fixed, single-path comparison to encode local variations, our framework adopts a dual-route encoding mechanism inspired by the quantum superposition principle. By simultaneously evaluating 'minimum-based' and 'maximum-based' comparison paths, we represent local signal blocks as a superposed state before mapping them into a high-dimensional Hilbert space. This approach allows for the modeling of intricate symbolic dependencies and parallel decision trajectories, capturing fine-grained prosodic irregularities in synthetic speech that standard deterministic descriptors fail to resolve.

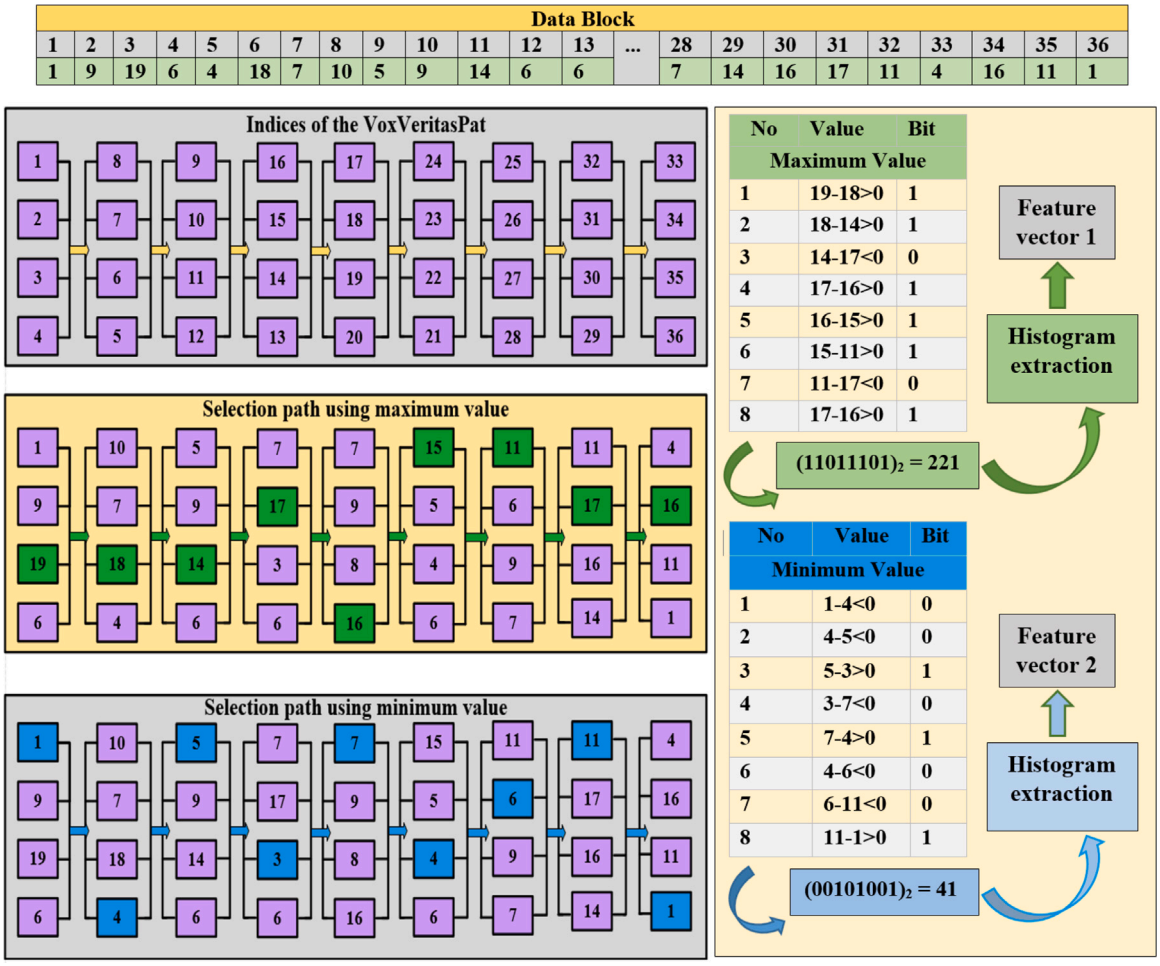


Fig. 4. Operational flow of the VoxVeritasNet feature extraction mechanism. The architecture utilizes dual-path methodology, employing parallel maximum and minimum value pathways to generate 8-bit binary descriptors. These descriptors are mapped into a high-dimensional Hilbert space to enhance geometric separability between authentic and synthetic speech.

4.1.3. Parallel statistical feature extraction

In parallel with the VoxVeritasNet extractor, a conventional statistical feature extractor is employed to compute global properties of the audio signals wavelet bands as:

$$[sts^1, sts^2] = (w_t), \quad t \in \{1, 2, 3, \ldots, 9\} \tag{15}$$

where sts^1 and sts^2 are statistical metrics and their absolute values, respectively. This extractor calculates eight statistical metrics: mean, standard deviation, skewness, kurtosis, median, minimum, maximum, and entropy. The formulations for these features are provided in Table 3. Applying these metrics to both the signal and its absolute value generates 16 statistical features.

Table 3
Formulations of statistical characteristics.

Feature	Equation
Mean (f_1)	$\frac{1}{N}\sum_{i=1}^{N} x_i$
Standard deviation (f_2)	$\sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - f_1)^2}$
Skewness (f_3)	$\frac{\frac{1}{N}\sum_{i=1}^{N}(x_i-f_1)^3}{(\frac{1}{N}\sum_{i=1}^{N}(x_i-f_1)^2)^{3/2}}$
Kurtosis (f_4)	$\frac{\frac{1}{N}\sum_{i=1}^{N}(x_i-f_1)^4}{(\frac{1}{N}\sum_{i=1}^{N}(x_i-f_1)^2)^2}$
Median (f_5)	Value separating the higher half from the lower half
Minimum (f_6)	$\min\{x_1, x_2, \ldots, x_N\}$
Maximum (f_7)	$\max\{x_1, x_2, \ldots, x_N\}$
Entropy (f_8)	$-\sum_{i=1}^{N} p(x_i)\log_2 p(x_i)$

4.1.4. Feature vector formulation

The features from the VoxVeritasNet extractor and the statistical extractor are merged to create comprehensive feature vector per wavelet level. This merging strategy ensures a rich representation by combining global statistical insights with localized structural information. Within the proposed framework, four distinct feature vectors were generated to capture diverse aspects of the audio signals. The initial two vectors, labeled as f_x^1 and f_x^2 were generated using histogram-based descriptors exclusively, extracted by the VoxVeritasNet operator across various decomposition levels. Specifically, f_x^1 and f_x^2 each comprises 256 features represent the max and min feature vectors, respectively. To enrich the feature set, two additional vectors, f_x^3 and f_x^4, were extracted by statistical feature extractor. The compositions of the feature vectors are defined as:

$$f_x^1 = feature_1, \tag{16}$$

$$f_x^2 = feature_2, \tag{17}$$

$$f_x^3 = sts_x^1, \tag{18}$$

$$f_x^4 = sts_x^2 \tag{19}$$

Table 4
Size of the individual feature vectors.

Vector	Type	Size
f_x^1	Textural	256
f_x^2	Textural	256
f_x^3	Statistical	8
f_x^4	Statistical	8

where, f_x and sts_x are the VoxVeritasNet extracted feature vectors and statistical feature vectors, respectively, at x_{th} wavelet. The sizes of the composed feature vectors are summarized in Table 4.

Finally, the features from all nine wavelets are concatenated to form final feature vector as:

$$F = \frown (f_x^1, f_x^2, f_x^3, f_x^4), \quad x \in \{1, 2, \ldots, 9\} \tag{20}$$

where $\frown$ is the concatenation operation. This operation culminates in a high-dimensional feature descriptor of length 4752 (9 wavelets × 528 combined statistical and textural features each). The strategic importance of this multi-level extraction lies in its ability to perform a multi-resolution analysis of the audio signal. Each wavelet captures information from a distinct frequency sub-band, allowing the model to simultaneously encode both coarse, low-frequency structural information and fine-grained, high-frequency textural details that are often crucial for detecting sophisticated manipulations.

4.2. Stage 2: ensemble feature selection

To reduce dimensionality and isolate the most discriminative attributes, the final feature vector undergoes a rigorous selection phase. This stage employs an ensemble of three distinct algorithms:

- NCA [44]: A distance metric learning algorithm that finds an optimal linear transformation of the data.
- Chi^2: A statistical test used to identify features with the strongest relationship to the output class.
- ReliefF: An algorithm that estimates attribute quality based on how well their values distinguish between proximal instances.

The proposed ensemble feature selection strategy is implemented as a sequential and iterative refinement process, as detailed in the EnsembleFeatureSelection procedure of Algorithm 1 (lines 29–36). To maximize the discriminative power of the final feature set while minimizing redundancy, the techniques are integrated through a subtractive logic. Initially, NCA is employed to identify the 5 most significant features based on stochastic neighbor weights. To prevent the selection of overlapping information, these features are excluded from the initial pool ($F' = F \setminus \bigcup \mathcal{FS}_{1..5}$) before applying the Chi^2 test, which evaluates the remaining attributes for their statistical independence from the class labels. In the final iteration, the ReliefF algorithm is applied to the remaining feature subset to capture localized dependencies and ensure resilience against signal noise. This sequential integration ensures that each method selectively targets complementary characteristics of the speech signal across all decomposition levels.

4.3. Stage 3: classification and training

In this stage, 15 candidate feature sets were used to train and evaluate two distinct machine learning models: a kNN classifier [45] and a Cubic SVM [46]. These particular classifiers were selected based on their demonstrated effectiveness in pattern recognition and signal processing applications (Algorithm 1, lines 37–45).

4.4. Stage 4: post-processing

Following the training of all 30 model configurations (15 candidate feature sets paired with two classifiers), a post-processing stage is implemented to identify the optimal pairing (Algorithm 1, lines 46–54). This stage employs an information fusion strategy wherein an IMV mechanism, guided by a Greedy optimization algorithm, systematically evaluates the performance of each model–feature combination. The Greedy algorithm selects the single configuration that demonstrates the highest classification accuracy. This top-performing model is then exclusively utilized for the final inference task: classifying unlabeled audio segments as either authentic or deepfake.

4.5. Stage 5: validation

To ensure the robustness and generalizability of the final model selected in Stage 4, a rigorous validation protocol is implemented using k-fold cross-validation, with k set to 10 (Algorithm 1, lines 55–62). In this procedure, the entire dataset is partitioned into ten mutually exclusive, equally-sized subsets or folds [47]. The model is then trained and validated iteratively ten times; during each iteration, nine folds serve as the training set, while the remaining fold is reserved for validation as:

$$P_{CV}, P_{HO} = \text{Test}\left(\frac{1}{k} \sum_{i=1}^{k} \text{Eval}\left(\text{Train}(\mathcal{M}, \mathcal{D} \setminus \mathcal{D}_i, \mathcal{D}_o), \mathcal{D}_i \right), \mathcal{D}_o \right), \tag{21}$$

where P_{CV} is the average cross validation accuracy of the classification model $\mathcal{M}$ on a dataset $\mathcal{D}$ from each of the k validation iterations. $\mathcal{D}_i$ and $\mathcal{D}_o$ are validation and test data, respectively. Following the identification of the optimal model configuration, a final performance evaluation, P_{HO}, is conducted using a completely independent hold-out test set. This test set remains strictly isolated and is not accessed during any stage of the ensemble feature selection or training/validation process.

5. Performance evaluation

The empirical evaluation of the proposed VoxVeritasNet model was conducted using MATLAB R2023b. The classification models, specifically Fine kNN and Cubic SVM, were implemented and trained utilizing the MATLAB Classification Learner toolbox, from which the numerical results were derived. To provide a comprehensive and rigorous assessment of the framework's performance, a suite of standard evaluation metrics was employed [48]. The model's classification effectiveness was quantified using accuracy, precision, recall, and the F1-score. For a more granular analysis of classification outcomes, confusion matrices were also generated. Additionally, the system's reliability was measured by calculating the false positive rate (FPR), the false negative rate (FNR), and the EER. The formal definitions for these metrics are provided in the following equations, consistent with established literature [49].

$$Accuracy(\%) = 100 \times \frac{TP + TN}{TP + TN + FP + FN} \tag{22}$$

$$Precision = \frac{TP}{TP + FP} \tag{23}$$

$$Recall = \frac{TP}{TP + FN} \tag{24}$$

$$F1 - Score = 2 \times \frac{Recall \times Precision}{Recall + Precision} \tag{25}$$

$$FPR = \frac{FP}{FP + TN} \tag{26}$$

$$FNR = \frac{FN}{FN + TP} \tag{27}$$

$$EER = \frac{FAR + FRR}{2} \tag{28}$$

Table 5
Detailed parameter settings and sensitivity analysis of the VoxVeritasNet framework.

Parameter	Value	Reasoning/Sensitivity
Segment duration	3 s	Chosen as an empirical trade-off between temporal resolution after testing 1–5 s intervals.
Wavelet levels	9	Determined through MDWT experimentation to capture both coarse prosodic trends and fine-grained spectral artifacts.
Histogram bins	256 × 2	Captures the full resolution of 8-bit binary mappings (Map1 and Map2) without quantization data loss.
Statistical features	16 per level	Includes statistical metrics, e.g. skewness, kurtosis which demonstrated high class-separability in ablation trials.
Feature selection count	135/inference (15 × 9 level)	Top 15 features from NCA (5), Chi^2 (5), ReliefF (5) per wavelet level to ensure diverse subset
Iterations	5 each	Ensures robustness against stochastic variability in NCA; verified to reach optimal performance stability.
kNN parameter (k)	1	Tested across $k \in \{1 \dots 7\}$; $k = 1$ provided the most precise local neighborhood boundaries for this dataset.
SVM kernel	Poly. (deg 3)	Selected over RBF and linear kernels for achieving the highest generalization on 10-fold cross-validation.

Here, TP and TN represent the number of correctly classified deepfake and authentic samples, respectively. Conversely, FP (False Positive) and FN (False Negative) denote the number of misclassified samples.

The configuration parameters for VoxVeritasNet, detailed in Table 5, are strategically selected to ensure both optimal detection performance and "intrinsic interpretability". Unlike stochastic deep learning models, the choice of MDWT decomposition levels (1–9) and histogram bin sizes is grounded in the physical properties of audio signals, where specific sub-bands are known to host the micro-spectral artifacts characteristic of neural vocoders. By documenting the deterministic reasoning behind each parameter — such as the alignment of wavelet depth with the frequency resolution required for forensic textural analysis — we provide a transparent framework where decision-making logic is traceable back to signal processing fundamentals. This methodological clarity serves as a domain-specific form of explainability, allowing for a clear understanding of how feature variations in the Hilbert space translate into classification results without the need for post-hoc surrogate models.

The sensitivity analysis conducted during these trials confirmed that the interaction between the 9-level wavelet decomposition and the iterative ensemble selection is the primary driver of the framework's stability. The composition of the 528-feature vector per wavelet level is obtained to provide a full forensic signature. Since each dual-path trajectory is encoded using an 8-bit histogram, 256 features are required per path to achieve a full representation of the local relational patterns, totaling 512 textural features. To complement these local descriptors with global context, 8 statistical feature are calculated alongside their absolute values, contributing 16 features to capture systemic signal variations. Across the nine decomposition levels, this results in a high-dimensional pool of 4752 features (9 levels × 528 features).

The computational efficiency and hardware requirements of the proposed VoxVeritasNet framework are comprehensively detailed in Table 6. Evaluated on a standard Intel-based PC with 16 GB of RAM, the system demonstrates a remarkably high inference speed, achieving an average feature extraction time of only 0.155 s per 3-second audio segment. This translates to a processing throughput of approximately 6.45 segments per second, confirming its suitability for real-time forensic applications. Furthermore, unlike deep learning-based architectures that necessitate high-end GPU clusters and substantial storage for millions of trainable parameters, VoxVeritasNet operates entirely in a CPU-based MATLAB environment with a negligible model storage footprint. Reliance on handcrafted feature engineering and classical machine learning enables the system to maintain peak detection performance while remaining highly resource-efficient and portable.

Table 6
Computational complexity and hardware configuration of the proposed framework.

Item	Value
Dataset size	1.83 GB
Number of audio segments	4902
Segment duration	3 s
Total feature extraction time	760.25 s
Average feature extraction time per segment	0.155 s
Processing throughput	~6.45 segments/s
Hardware configuration	Intel-based PC, 16 GB RAM, 512 GB SSD
Implementation environment	MATLAB (CPU-based)
Model type	Handcrafted features + classical ML
Model storage size	Negligible (no trainable deep parameters)

6. Results and discussion

The performance of the VoxVeritasNet framework was evaluated using a systematic and hierarchical experimental design to validate its effectiveness in deepfake speech detection, computational efficiency, and environmental robustness. The core of the experimental setup was established using the sym4 wavelet as the ground reference, selected specifically for its superior trade-off between temporal resolution and computational cost-efficiency. The evaluation process began with a series of the ablation studies designed to isolate the contribution of each framework component. Initially, the performance of the multi-level feature selection strategy (incorporating NCA, Chi^2, and ReliefF) was tested under the sym4 configuration to quantify its impact on classification accuracy and its ability to minimize the FNR. In the second stage of ablation, the individual contributions of the 9-level MDWT textural and statistical features were examined using a Cubic SVM classifier to demonstrate how the hybrid integration of local structural patterns and global statistical descriptors enhances the model's discriminative power.

Following the ablation trials, the generalizability of the framework was rigorously tested by fixing the feature selection and extraction pipeline while varying the wavelet families, specifically comparing the sym4–sym8 and db4–db8 filters across multiple datasets. To ensure the reliability of the classification outcomes, a comprehensive confusion matrix and box plot analyses were conducted for both Cubic SVM and Fine kNN, providing a statistical overview of performance stability across balanced and imbalanced data distributions. To assess the framework's real-world applicability and ensure a leakage-free validation, a cross-corpus evaluation was performed by training the model on Dataset 1 and testing it in an unseen environment of Dataset 2. Furthermore, the robustness of VoxVeritasNet was challenged under extreme conditions through a stress-test analysis utilizing additive white Gaussian noise (AWGN) ranging from 5 dB to 55 dB. Finally, the proposed VoxVeritasNet model was benchmarked against the existing State-of-the-Art (SOTA) models in the literature to substantiate its leadership in terms of accuracy, model size, and processing throughput.

6.1. Ablation analysis for feature selection

The initial stage of the performance evaluation for VoxVeritasNet involves conducting detailed ablation study focusing on the impact of the iterative feature selection strategy on classification accuracy and error metrics. Using the sym4 wavelet as the ground reference, this

Table 7
Ablation study: Impact of feature selection steps and classifier choice across datasets with focus on FNR reduction and terminal EER metrics (%)(Sym4 is referenced. Best results in bold).

Dataset	Feature selection	Algorithm	Accuracy	Precision	Recall	F1-score	FPR	FNR	EER
First	Only NCA	SVM	99.33	99.33	99.32	99.33	–	0.68	–
		kNN	92.90	93.12	93.02	92.90	–	6.98	–
	NCA+Chi^2	SVM	**99.35**	**99.35**	**99.35**	**99.35**	–	0.65	–
		kNN	93.20	93.45	93.33	93.20	–	6.67	–
	NCA+Chi^2+ReliefF	SVM	**99.35**	**99.35**	**99.35**	**99.35**	**0.55**	**0.65**	**0.60**
		kNN	93.10	93.36	93.23	93.10	6.23	6.77	6.50
Second	Only NCA	SVM	99.80	99.79	99.79	99.79	–	0.21	–
		kNN	97.00	96.62	97.17	96.93	–	2.83	–
	NCA+Chi^2	SVM	99.84	99.82	99.84	99.83	–	0.16	–
		kNN	97.53	97.19	97.85	97.47	–	2.15	–
	NCA+Chi^2+ReliefF	SVM	**99.86**	**98.84**	**99.86**	**99.85**	**0.14**	**0.14**	**0.14**
		kNN	97.42	97.38	98.01	97.65	2.81	1.99	2.40
Third	Only NCA	SVM	98.90	98.29	96.62	97.43	–	3.38	–
		kNN	97.66	95.82	93.30	94.51	–	6.70	–
	NCA+Chi^2	SVM	**98.99**	**98.35**	96.95	97.65	–	3.05	–
		kNN	97.70	96.27	93.02	94.56	–	6.98	–
	NCA+Chi^2+ReliefF	SVM	**98.99**	**98.35**	**97.02**	**97.68**	**0.12**	**2.98**	**1.55**
		kNN	97.80	96.53	93.32	94.85	2.12	6.68	4.40

analysis examines the sequential integration of NCA, Chi^2, and ReliefF algorithms.

The ablation results summarized in Table 7 underscore the critical role of the iterative feature selection ensemble (NCA, Chi^2, and ReliefF) in reducing the FNR, which represents the system's failure to detect synthetic audio samples. This metric is crucial for ensuring forensic reliability. The transition from single-method selection (Only NCA) to the hybrid ensemble architecture facilitates a significant drop in FNR across all evaluated conditions. For the highly imbalanced Third Dataset, the baseline FNR of 3.38% with NCA-only was successfully reduced to 2.98% using the full ensemble selection, effectively capturing high-frequency deepfake artifacts that were previously missed. This systematic reduction confirms that by filtering redundant information through sequential statistical and distance-based tests, the model enhances its sensitivity to microscopic forgery traces. Ultimately, the synergy between the MDWT decomposition and optimized selection ensures that VoxVeritasNet achieves peak recall and minimal FNR, thereby providing a more secure framework against voice cloning and text-to-speech attacks.

6.2. Ablation analysis of feature components and wavelet levels

The second phase of the ablation analysis investigates the individual contributions of various feature categories and the stability of the model across 9 levels MDWT. By utilizing the sym4 wavelet and a cubic SVM as the fixed baseline, the performance of textural and global statistical features was compared to evaluate their discriminative potential. The comprehensive ablation analysis presented in Table 8 indicates that the textural (histogram-based) features extracted via the proposed operator are the primary drivers of performance, achieving an accuracy of 98.95% independently. In contrast, the global statistical features yield a lower accuracy of 92.99%, confirming that localized structural artifacts are more discriminative than global variations for deepfake detection. Furthermore, the evaluation of individual MDWT levels (Level 1 to Level 9) reveals a remarkably consistent and stable performance, with each level providing an accuracy between 98.08% and 98.87%. This stability demonstrates the robustness of the method in capturing microscopic forgery traces across a wide spectral range. Ultimately, the peak performance of 99.34% attained by the full feature set (4752 dimensions) validates the synergistic effect of integrating multi-level decomposition with hybrid feature types.

Table 8
Ablation study on feature subsets across all 9 MDWT levels: Performance evaluation of textural, statistical, and individual level components (Sym4 and a cubic SVM are used as the fixed baseline).

Feature subset	Dimension	Accuracy (%)
All features (9 × 528)	**4752**	**99.34**
All level textural (9 × 512)	4608	98.95
All statistical (9 × 16)	144	92.99
Only Level 1	512	98.87
Only Level 2	512	98.72
Only Level 3	512	98.62
Only Level 4	512	98.37
Only Level 5	512	98.48
Only Level 6	512	98.26
Only Level 7	512	98.38
Only Level 8	512	98.21
Only Level 9	512	98.08

6.3. Comparative analysis of wavelet families and classifiers

Following the ablation trials, the framework was subjected to a comprehensive evaluation across multiple wavelet families and classification algorithms to identify the optimal configuration for diverse dataset conditions. This stage involved fixing the ensemble feature selection and 9-level MDWT structure while varying the wavelet filters among sym4, sym8, db4, and db8. The experimental findings presented in Table 9 indicate that while all wavelet families maintain high detection rates, specific configurations excel depending on the class distribution of the data. Specifically, the db4 wavelet achieved a peak accuracy of 99.96% on the moderately imbalanced dataset. In contrast, the sym8 wavelet demonstrated superior performance on the balanced and highly imbalanced datasets with accuracies of 99.71% and 99.43%, respectively.

The peak performance observed under moderate imbalance (Second Dataset) is attributed to the richer representation of the authentic speech class. The increased volume of real audio segments allows VoxVeritasNet to establish a more robust and stable reference manifold for natural vocal variations. By providing a comprehensive baseline of authentic speech textures, the framework can more effectively isolate the microscopic anomalies introduced by neural vocoders in synthetic samples.

Despite the marginal performance gains provided by higher-order wavelets like sym8 or db8 in specific instances, the sym4 wavelet remains the preferred ground reference for the proposed framework due to its computational efficiency. While the shorter filter lengths of sym4

Table 9
Comprehensive ablation study: Evaluating the impact of wavelet families and classifier choice across three datasets (%).

Dataset	Wavelet	Algorithm	Acc.	Prec.	Rec.	F1	FPR	FNR	EER
First	sym4	Cubic SVM	99.35	99.35	99.35	99.35	0.55	0.65	0.60
		Fine kNN	93.10	93.36	93.23	93.10	6.23	6.77	6.50
	sym8	Cubic SVM	**99.71**	**99.71**	**99.71**	**99.71**	**0.27**	**0.29**	**0.28**
		Fine kNN	95.51	95.57	95.59	95.51	4.41	4.41	4.41
	db4	Cubic SVM	99.53	99.53	99.53	99.53	0.45	0.47	0.46
		Fine kNN	94.59	94.67	94.67	94.59	5.01	5.33	5.17
	db8	Cubic SVM	99.49	99.48	99.49	99.48	0.51	0.51	0.51
		Fine kNN	94.88	95.04	94.95	94.88	4.66	5.05	4.85
Second	sym4	Cubic SVM	99.86	99.84	99.86	99.85	0.14	0.14	0.14
		Fine kNN	97.42	97.38	98.01	97.65	2.81	1.99	2.40
	sym8	Cubic SVM	99.82	99.80	99.81	99.81	0.19	0.19	0.19
		Fine kNN	98.76	98.54	98.94	98.72	1.06	1.06	1.06
	db4	Cubic SVM	**99.96**	**99.95**	**99.96**	**99.96**	**0.04**	**0.04**	**0.04**
		Fine kNN	98.99	98.80	99.13	98.96	0.85	0.87	0.86
	db8	Cubic SVM	99.90	99.89	99.90	99.90	0.10	0.10	0.10
		Fine kNN	98.42	98.16	98.64	98.64	1.36	1.36	1.36
Third	sym4	Cubic SVM	98.99	98.35	97.02	97.68	0.12	2.98	1.55
		Fine kNN	97.80	96.53	93.32	94.85	2.12	6.68	4.40
	sym8	Cubic SVM	**99.43**	**99.21**	**98.16**	**98.68**	**0.06**	**1.84**	**0.95**
		Fine kNN	98.37	97.31	95.11	96.17	0.11	4.89	2.50
	db4	Cubic SVM	99.00	98.46	96.92	97.67	0.12	3.08	1.60
		Fine kNN	97.51	95.59	92.80	92.64	2.60	7.20	4.90
	db8	Cubic SVM	99.35	99.02	97.98	98.49	0.08	2.02	1.05
		Fine kNN	98.14	96.94	94.39	95.64	0.39	5.61	3.00

ensure lower latency during real-time feature extraction without compromising the system's ability to maintain an accuracy above 98.99% and an EER as low as 0.14% with the Cubic SVM, it provides an optimal trade-off between discriminative power, low computational complexity, and robust generalizability for real-world deepfake detection.

6.4. Confusion matrix, receiver operating characteristic and box plot stability analysis

The reliability and classification consistency of the VoxVeritasNet framework were further evaluated through a detailed analysis of confusion matrices and statistical box plots. The confusion matrices presented in Fig. 5 and the Receiver Operating Characteristic (ROC) curves in Fig. 6 exhibit strong diagonal dominance across all three datasets, confirming a high rate of correct predictions for both authentic and deepfake audio segments. The confusion matrices and the ROC of Cubic SVM classifier achieves near-perfect Area Under the Curve (AUC) scores of 0.9997, 0.9997, and 0.9973 for the balanced, moderately imbalanced, and highly imbalanced datasets, respectively. On the other hand, the Fine kNN classifier yields lower AUC scores comparing the Cubic SVM across all levels of data imbalance.

The empirical superiority of the Cubic SVM over the Fine kNN is primarily due to the non-linear and high-dimensional nature of the optimized 135-feature space. While the Fine kNN classifier is highly sensitive to local noise and struggles with the dimensionality due to its reliance on simple Euclidean distances, the Cubic SVM employs a third-order polynomial kernel to effectively model complex interactions between textural and statistical descriptors. This approach allows the SVM to construct a smooth and optimal separating hyperplane that maximizes the margin between authentic and synthetic classes. Unlike the rigid, proximity-based boundaries of kNN which often fail to resolve overlapping regions in high-fidelity forgeries, the Cubic SVM provides better generalization by capturing global structural patterns rather than relying on local neighborhood densities. This makes the SVM more resilient to the subtle vocoder-induced artifacts found in sophisticated deepfake samples.

To gain a deeper understanding of model stability, box plot analyses were utilized to examine the variance in accuracy across different experimental setups. The statistical distribution of classification accuracies, as shown in Fig. 7, quantitatively validates the rigidity of the VoxVeritasNet framework. Across all experimental setups, the Cubic SVM demonstrably outperforms the Fine kNN classifier, maintaining median accuracies consistently above 98.5% in balanced and moderately imbalanced scenarios. Specifically, the SVM achieves peak performance scores of 99.34%, 99.81%, and 98.99% for the respective datasets, whereas the kNN exhibits a significantly higher variance and wider interquartile ranges (IQR), particularly in the first dataset where its accuracy distribution spans from 95.5% to 98.5%.

Regarding the feature selection phase, the influence of the ensemble components varies according to the data distribution. In the balanced dataset (a), NCA emerges as the most robust selector with a median accuracy exceeding 99%, while Chi^2 and ReliefF show more distributed performances. However, in the moderately imbalanced dataset (b), Chi^2 reaches a superior median of approximately 99.5%. Even in the most challenging highly imbalanced dataset (c), all methods maintain median accuracies above 97.5%, with NCA providing the most compact distribution. These findings confirm that the synergy of MDWT-based feature extraction and iterative selection effectively mitigates the impact of class skewness while maintaining high detection precision.

A detailed inspection of the confusion matrices reveals that misclassifications are primarily concentrated in high-fidelity samples where vocoder-induced artifacts are at their threshold of detection. Specifically, the slight increase in FNR in certain wavelet configurations suggests that while the dual-path mapping captures most synthetic echoes, some extremely subtle phase inconsistencies may occasionally evade detection if the wavelet's frequency resolution is not optimally aligned. However, the consistent stability shown in Fig. 7 indicates that the ensemble feature selection effectively mitigates these failure cases, ensuring that the FPR remains remarkably low even under highly imbalanced conditions. This balance between sensitivity and specificity confirms that the Cubic SVM successfully identifies the global structural signatures of deepfakes that local distance-based classifiers fail to resolve.

6.5. Cross-corpus robustness and SNR stress test analysis

The real-world applicability and generalizability of the VoxVeritasNet framework were evaluated through a rigorous cross-corpus

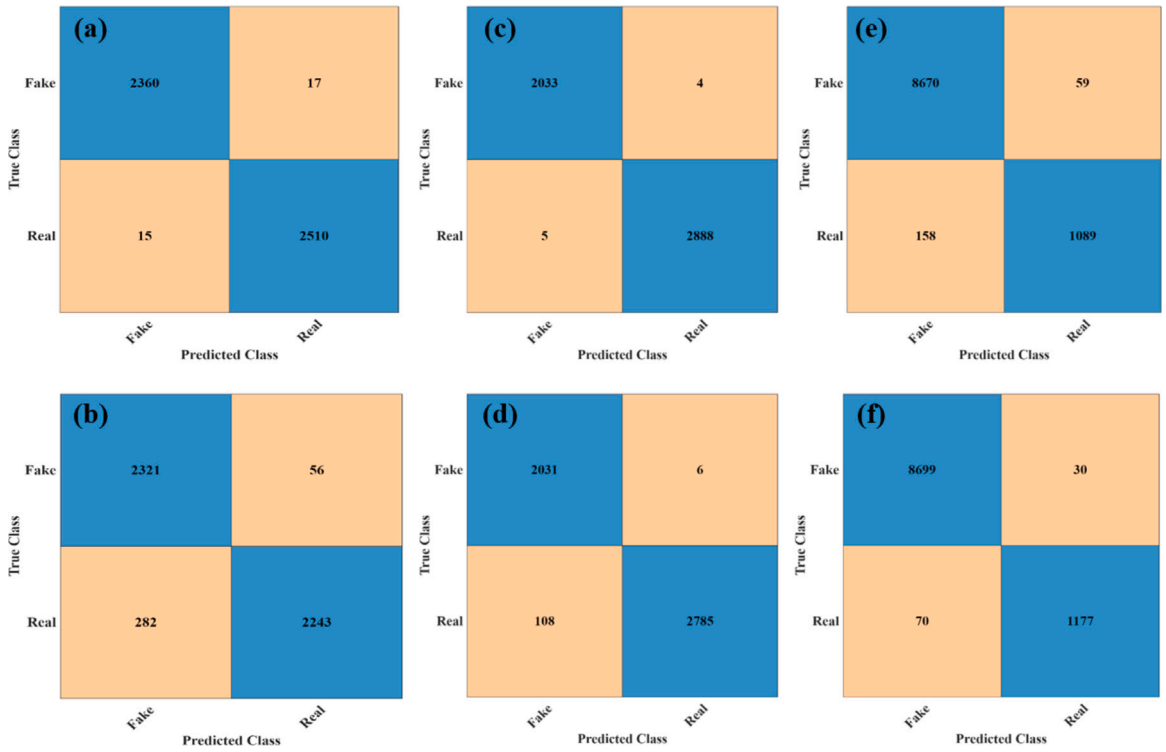


Fig. 5. Confusion matrices of the (a) Cubic SVM and (b) Fine kNN algorithm results with VoxVeritasNet, using a balanced dataset (c) Cubic SVM and (d) Fine kNN algorithm results with using a moderating imbalanced dataset, (e) Fine kNN and (f) Cubic SVM using a highly imbalanced dataset.

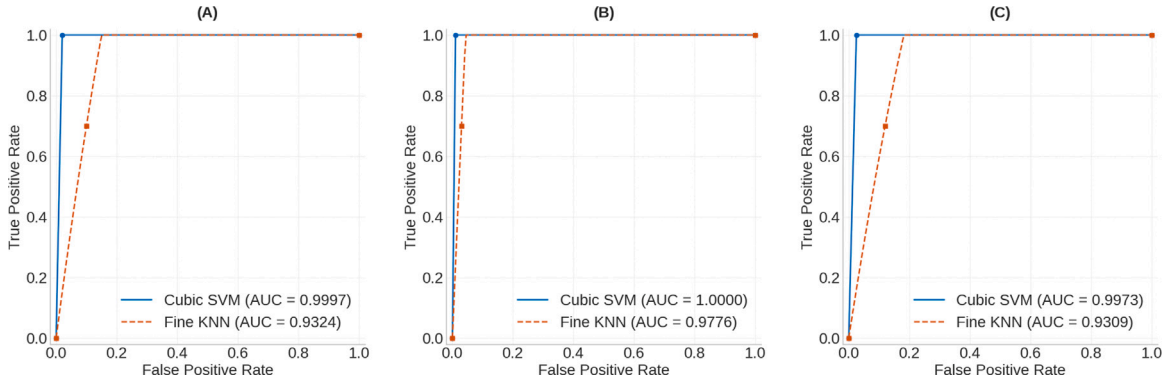


Fig. 6. Performance evaluation of VoxVeritasNet model in the case of validation ROC curve for Cubic SVM and Fine kNN on (a) balanced dataset, (b) moderately imbalanced dataset and (c) highly imbalanced dataset.

validation and environmental stress testing. To ensure that the model does not suffer from over-fitting or data leakage, it was trained in a balanced environment of using the first dataset and subsequently tested on using an entirely unseen moderately imbalanced second dataset. As illustrated in Fig. 8(a), the model utilizes a nested 10-Fold Cross-Validation strategy to ensure a leakage-free assessment. The statistical distribution shows that the gap between the median inner validation accuracy and the outer test accuracy is less than 1%, maintaining an exceptional performance level above 98.5%.

This marginal difference confirms that the model does not suffer from over-fitting or data leakage, proving its ability to generalize across different recording conditions and class distributions. Furthermore, Fig. 8(b) presents the fold-specific performance, where the Outer Test accuracies (unseen data) closely track the best inner validation scores across all 10 iterations. The fact that this consistency is maintained across folds, despite changes in the characteristics of the dataset, serves as empirical evidence of the model's robustness against channel mismatch and varying acoustic environments. This validates its reliability for real-world forensic applications.

In addition to that, the framework was subjected to a comprehensive stress-test analysis to determine its resilience against environmental and transmission distortions. We specifically focused on additive white Gaussian noise (AWGN) with SNR levels ranging from 5 dB to 55 dB, as illustrated in Fig. 9. The framework demonstrates remarkable stability. Even under the extreme stress condition of a 5 dB SNR, the model utilizing the all feature subset maintains a high accuracy of approximately 95%, with degradation remaining below 4.4% compared to the high-SNR baseline. This resilience is further evident at a 15 dB SNR, where performance remains above 95.5%, showcasing the proposed method's ability to isolate discriminative features from heavily corrupted signals. Since the datasets utilized in this work consist of raw, un-preprocessed audio from diverse real-world sources, they inherently encompass variations in compression (e.g., MP3 artifacts), bandwidth limitations, and room reverberation. The consistent performance across three independent datasets, which utilize different synthesis engines and unseen

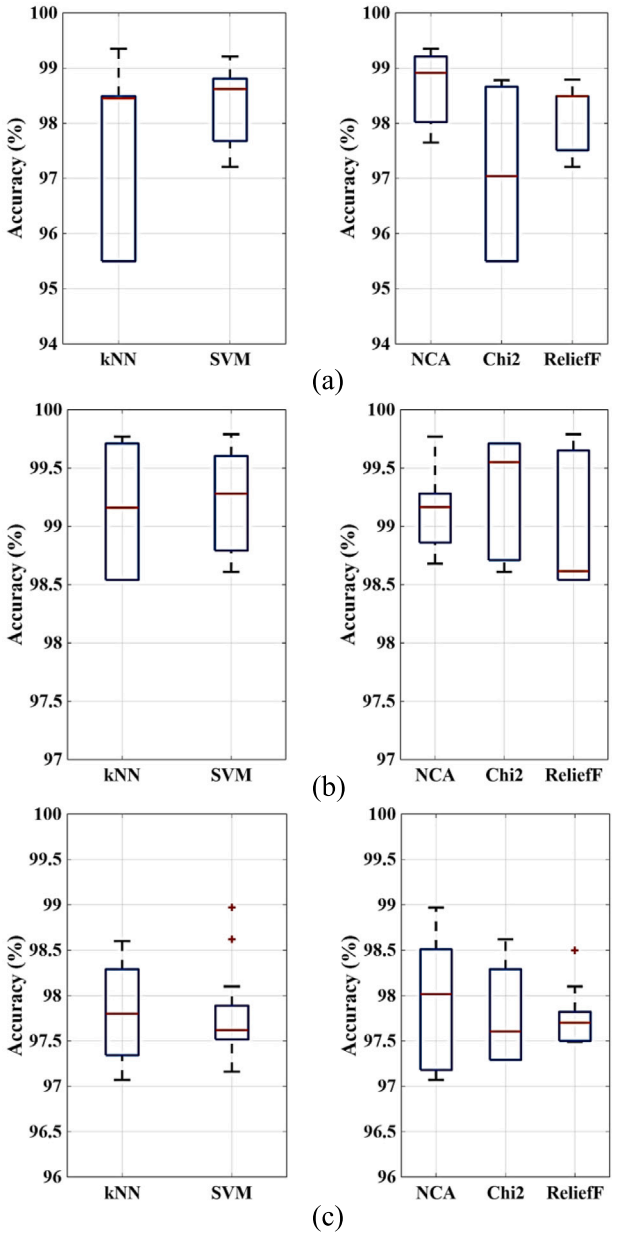


Fig. 7. Comparative distribution of classification accuracies across balanced, moderately imbalanced, and highly imbalanced datasets, evaluating (a–c) the performance stability of kNN vs. SVM classifiers and the efficacy of NCA, Chi^2, and ReliefF feature selection algorithms.

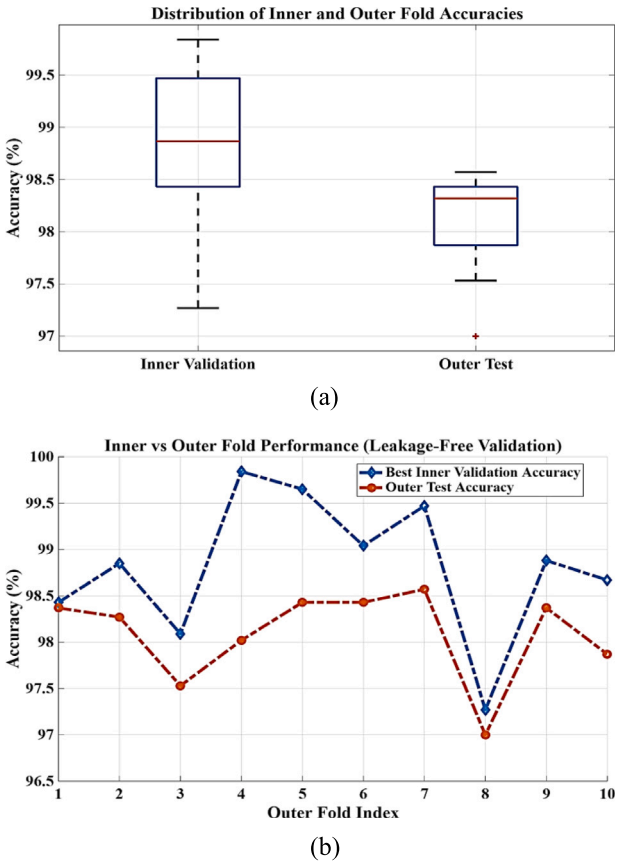


Fig. 8. Robustness and generalizability analysis of VoxVeritasNet: (a) distribution of accuracies across inner validation (training on dataset 1) and outer test (testing on dataset 2) folds. (b) fold-by-fold comparison of inner validation vs. outer test accuracies across the 10-fold CV process.

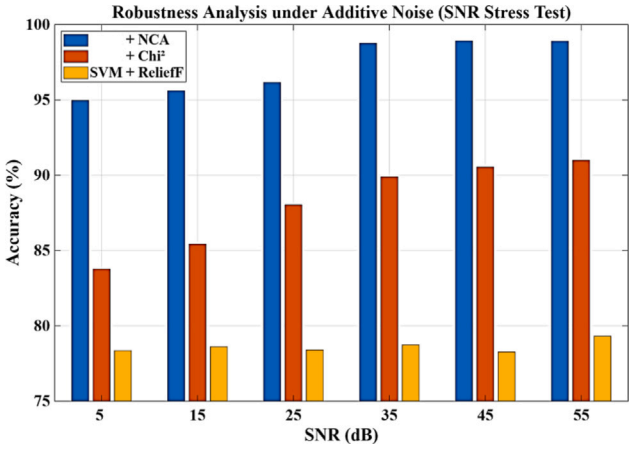


Fig. 9. Robustness analysis of VoxVeritasNet under varying SNR conditions. The performance of ensemble feature selection subsets (SVM+NCA, +Chi^2, +ReliefF) is evaluated under AWGN ranging from 5 dB to 55 dB.

VC/TTS families during specific training cycles, confirms the generalizability and stress-resilience of the proposed feature engineering approach. This indicates that VoxVeritasNet effectively mitigates the performance degradation typically associated with real-world acoustic mismatches and unseen attack families.

The noise resilience of the VoxVeritasNet framework is fundamentally tied to the integration of NCA within its feature selection ensemble. As illustrated in the SNR stress test analysis in Fig. 9, models relying solely on ReliefF or the ReliefF plus Chi^2 combination exhibit significant vulnerability to environmental distortions, particularly at lower SNR. Under extreme stress conditions such as a 5 dB SNR, configurations without NCA support suffer a substantial performance drop, with accuracies plunging towards the 80% to 85% range. In contrast, the inclusion of NCA as a primary feature selector allows the model to maintain a robust accuracy of approximately 95% even under these harsh conditions, restricting the total performance degradation to less than 4.4% compared to the high-SNR baseline. This quantitative gap proves that NCA acts as a critical discriminative anchor, effectively isolating salient forgery traces from additive white Gaussian noise where distance-based and purely statistical selectors fail. Consequently, while ReliefF and Chi^2 contribute to statistical refinement, NCA provides

Table 10
Comprehensive benchmarking of VoxVeritasNet against SOTA models: A multidimensional evaluation of detection accuracy, hardware requirements, and computational efficiency metrics (best results in bold).

Study	Method	Dataset	Hardware	Model size	Latency	Year	Accuracy
VoxVeritasNet (This work)	**MDWT + Ensemble FS**	3 datasets [40–42]	**CPU**	**Negligible* (No deep params)**	155 ms **(6.45 seg/s Throughput)**	**2025**	**db4: 99.96** [41] sym8: 99.71 [40] sym8: 99.43 [42]
Gujjar et al. [50]	MFCC-GNB	Deep-voice dataset [41]	CPU	–	–	2024	99.93
Chakravarty et al. [51]	GTCC, MFCC ResNet50	DECRO, ASVspoof2019	GPU	>100 MB	Medium	2024	99.70
Valente et al. [52]	CNN	FoR dataset	–	–	–	2024	99.00
Ilyas et al. [53]	AVFakeNet	Celeb-DF, ASV	GPU	>150 MB	Medium	2023	98.73
Ballesteros et al. [54]	Deep4SNet	4 datasets	GPU	~40 MB	Low	2021	98.5
Shaaban et al. [55]	CNN	ASVspoof 2019	GPU	>100 MB	Medium	2025	98.00
Muruganandham et al. [56]	LSTM-AE-DRDE	5 datasets	GPU	6.5 GB	120 ms	2025	97.50
Hamza et al. [25]	MFCC + VGG-16	FoR dataset	–	–	–	2022	93.00
Pham et al. [57]	CNN	ASVspoof 2019	–	–	–	2024	90.00
Altalahin et al. [58]	CNN-LSTM	ASVspoof 2019	–	~20 MB	Low	2023	88.00
Mcuba et al. [26]	VGG-16	Fake voice classif.	GPU	~512 MB	High	2023	86.90
Kulangareth et al. [24]	Adaboost	Speech recogn.	–	–	–	2024	81.00

the essential computational strength required for forensic reliability in adverse real-world acoustic environments.

6.6. Benchmarking against state-of-the-art models

The comprehensive benchmarking analysis presented in Table 10 evaluates VoxVeritasNet against established SOTA methodologies, focusing on the critical trade-off between detection precision and hardware accessibility. VoxVeritasNet establishes a new performance frontier by achieving a peak accuracy of 99.96% utilizing the db4 wavelet on Dataset 2, effectively outperforming nearly all deep learning-based architectures while maintaining an exceptionally lightweight and efficient operational profile. A granular comparison reveals the definitive superiority of our framework in terms of computational overhead and data handling compared to high-accuracy baselines. While [50] reports a competitive accuracy of 99.93% on a similar dataset, their approach relies on extensive preprocessing techniques to balance the data, which inherently increases computational complexity and limits its deployment as an end-to-end efficient solution for real-world scenarios. In contrast, VoxVeritasNet manages raw, un-preprocessed audio segments with minimal latency, ensuring that the integrity of the acoustic artifacts is preserved without the need for resource-heavy augmentation.

The framework's most striking strength lies in its ability to achieve high-speed processing on restricted hardware, effectively bridging the gap between GPU-dependent models and portable forensic tools. Large-scale models such as the LSTM-AE-DRDE framework by [56] require massive hardware resources, including 6.5 GB of GPU memory, to achieve a throughput latency of 120 ms. Remarkably, VoxVeritasNet achieves a comparable throughput of 6.45 segments per second with 155 ms latency using only a standard CPU, demonstrating its exceptional suitability for forensic environments where high-end GPU acceleration is unavailable. Unlike ResNet50 or VGG-16 based architectures [26,51] which demand substantial memory footprints exceeding 100 MB to 512 MB, our framework possesses a negligible model size because it eliminates the necessity of storing millions of trainable deep neural weights. By surpassing the detection precision of advanced networks such as AVFakeNet (98.73%), Deep4SNet (98.5%) and CNN-based approaches (98.00%) [53–55] while maintaining a state-of-the-art EER of 0.14%, VoxVeritasNet proves that the synergy of MDWT signal decomposition and iterative ensemble selection provides a more secure and scalable framework for the practical detection of voice cloning and TTS attacks.

Unlike CNN-based models that tend to overfit to specific training generators, VoxVeritasNet employs a multi-resolution MDWT structure that captures universal micro-anomalies across 9 sub-bands, ensuring superior generalization against novel synthesis techniques. Furthermore, the ensemble feature selection mitigates dependence on single-feature families, providing robust resilience against the "unseen attack" scenarios.

7. Conclusion and future work

This study introduces VoxVeritasNet, a high-precision and computationally efficient framework for deepfake audio detection that effectively bridges the gap between advanced signal processing and forensic applicability. By integrating a 9-level MDWT with a quantum-inspired dual-path mapping algorithm, the framework captures intricate temporal and spectral artifacts often missed by conventional deep learning architectures. The key contribution of this work is the embedding of extracted features into a high-dimensional Hilbert space, which provides a unit-vector representation that significantly enhances the geometric separability between authentic and synthetic speech classes.

The critical results obtained through rigorous testing on three public datasets demonstrate the framework's leadership in the field, with peak detection accuracies reaching 99.96% using db4 and 99.71% percent using sym8 wavelets. Even under the most efficient operational setup using the sym4 baseline, the model maintained an accuracy above 98.99% and established a state-of-the-art equal error rate of 0.14% . Furthermore, the model exhibits remarkable robustness under environmental stress, maintaining over 95% accuracy even in extreme 5 dB signal-to-noise ratio conditions.

A major practical achievement of VoxVeritasNet is its hardware accessibility, as it delivers high processing throughput on standard CPU-based systems without requiring high-end GPU acceleration. The deterministic nature of the feature engineering process, coupled with an iterative ensemble selection strategy consisting of NCA, Chi^2, and ReliefF, ensures both explainability and reliability in forensic analysis. These developments position our framework as a lightweight yet robust alternative to resource-heavy deep learning models.

Despite the high discriminative performance of VoxVeritasNet, certain limitations remain that provide a roadmap for future research. While the framework is exceptionally robust against current neural vocoders and GAN-based synthesis methods, its sensitivity could potentially be challenged by emerging diffusion-based generative models, which produce audio with even fewer spectral artifacts. Additionally,

the presence of adversarial attacks — specifically crafted to mask the micro-spectral inconsistencies targeted by our wavelet-based mapping — presents a sophisticated threat to forensic reliability. Furthermore, external factors such as extreme environmental noise or heavy lossy compression typical of social media pipelines may obscure the fine-grained textural signatures captured by the dual-path binary mapping.

Future work will focus on enhancing cross-domain robustness and integrating adaptive noise-cancellation layers to maintain detection precision under these adverse conditions. We plan to expand the framework's generalizability by evaluating its performance across a wider array of unseen voice conversion and text-to-speech attack families, specifically targeting multi-lingual datasets. Additionally, the integration of explainable artificial intelligence techniques will be explored to provide deeper insights into the specific acoustic cues used for detection. To address increasingly complex scenarios, future efforts will also investigate the framework's efficacy in detecting overlapping or multi-speaker synthetic audio and its adaptability to real-time mobile and edge-computing environments.

CRediT authorship contribution statement

Burak Çelik: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Resources, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Burcu Zeybek:** Writing – review & editing, Writing – original draft, Formal analysis, Data curation, Conceptualization. **Mahmut Burak Karadeniz:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Methodology, Investigation, Formal analysis, Conceptualization. **Adem Kocyigit:** Writing – review & editing, Writing – original draft, Supervision, Investigation, Data curation, Conceptualization. **Onur Arsalı:** Writing – review & editing, Investigation, Data curation, Conceptualization. **Ebru Efeoglu:** Writing – review & editing, Investigation, Formal analysis, Data curation, Conceptualization. **Bahattin Türetken:** Writing – review & editing, Visualization, Validation, Supervision, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The authors wish to acknowledge Nokta Muhendislik A.S. R&D Center for the financial support.

References

[1] Hina Fatima Shahzad, Furqan Rustam, Emmanuel Soriano Flores, Juan Luís Vidal Mazón, Isabel de la Torre Diez, Imran Ashraf, A review of image processing techniques for deepfakes, Sens. (ISSN: 1424-8220) 22 (12) (2022) http://dx.doi.org/10.3390/s22124556, URL https://www.mdpi.com/1424-8220/22/12/4556.

[2] J. Wu, W. Gan, Z. Chen, S. Wan, H. Lin, AI-generated content (AIGC): A survey, 2023.

[3] B. Yetiştiren, I. Özsoy, M. Ayerdem, E. Tüzün, Evaluating the code quality of AI-assisted code generation tools: An empirical study on GitHub copilot, amazon CodeWhisperer, and ChatGPT, 2023.

[4] S. Yeo, Y.S. Ma, S.C. Kim, H. Jun, T. Kim, Framework for evaluating code generation ability of large language models, ETRI J. 46 (1) (2024) 106–117, http://dx.doi.org/10.4218/etrij.2023-0357.

[5] B. Sisman, J. Yamagishi, S. King, H. Li, An overview of voice conversion and its challenges: From statistical modeling to deep learning, IEEE/ACM Trans. Audio Speech Lang. Process. 29 (2021) 132–157, http://dx.doi.org/10.1109/TASLP.2020.3038524.

[6] A. Natsiou, S. O'Leary, Audio representations for deep learning in sound synthesis: A review, in: Proceedings of the IEEE/ACS International Conference on Computer Systems and Applications, AICCSA, 2021, pp. 1–8, http://dx.doi.org/10.1109/AICCSA53542.2021.9686838.

[7] David Ghiurău, Daniela Elena Popescu, Distinguishing reality from AI: Approaches for detecting synthetic content, Comput. (ISSN: 2073-431X) 14 (1) (2025) http://dx.doi.org/10.3390/computers14010001, URL https://www.mdpi.com/2073-431X/14/1/1.

[8] Pindrop, Voice intelligence and security report 2025, 2025, URL https://www.pindrop.com/voice-intelligence-report. (Accessed 15 December 2025).

[9] J. Kietzmann, L.W. Lee, I.P. McCarthy, T.C. Kietzmann, Deepfakes: Trick or treat? Bus. Horiz. 63 (2) (2020) 135–146, http://dx.doi.org/10.1016/j.bushor.2019.11.006.

[10] A. Pianese, D. Cozzolino, G. Poggi, L. Verdoliva, Training-free deepfake voice recognition by leveraging large-scale pre-trained models, in: Proceedings of the 2024 ACM Workshop on Information Hiding and Multimedia Security, 2024, pp. 289–294, http://dx.doi.org/10.1145/3658664.3659662.

[11] T.C. Helmus, Artificial intelligence, deepfakes, and disinformation: A primer, 2022.

[12] Á.F. Gambín, A. Yazidi, A. Vasilakos, H. Haugerud, Y. Djenouri, Deepfakes: current and future trends, Artif. Intell. Rev. 57 (3) (2024) http://dx.doi.org/10.1007/s10462-023-10679-x.

[13] Samer Al-khazraji, Hassan Saleh, Adil KHALID, Israa Mishkhal, Impact of deepfake technology on social media: Detection, misinformation and societal implications, Eurasia Proc. Sci. Technol. Eng. Math. 23 (2023) 429–441, http://dx.doi.org/10.55549/epstem.1371792.

[14] Abdulazeez Alali, George Theodorakopoulos, Partial fake speech attacks in the real world using deepfake audio, J. Cybersecur. Priv. (ISSN: 2624-800X) 5 (1) (2025) http://dx.doi.org/10.3390/jcp5010006, URL https://www.mdpi.com/2624-800X/5/1/6.

[15] C. Vaccari, A. Chadwick, Deepfakes and disinformation: Exploring the impact of synthetic political video on deception, uncertainty, and trust in news, Soc. Media + Soc. 6 (1) (2020) http://dx.doi.org/10.1177/2056305120903408.

[16] M. Pawelec, Deepfakes and democracy (theory): How synthetic audio-visual media for disinformation and hate speech threaten core democratic functions, Philos. Technol. 1 (2) (2022) http://dx.doi.org/10.1007/s44206-022-00010-6.

[17] P.B. Hirsch, At the crossroads: generative AI and corporate risk management, J. Bus. Strategy 44 (6) (2023) 426–429, http://dx.doi.org/10.1108/JBS-09-2023-0196.

[18] J. Twomey, D. Ching, M.P. Aylett, M. Quayle, C. Linehan, G. Murphy, Do deepfake videos undermine our epistemic trust? A thematic analysis of tweets that discuss deepfakes in the Russian invasion of Ukraine, PLoS One 18 (10) (2023) 1–22, http://dx.doi.org/10.1371/journal.pone.0291668.

[19] W.M. Lim, Fact or fake? The search for truth in an infodemic of disinformation, misinformation, and malinformation with deepfake and fake news, J. Strateg. Mark. (2023) 1–37, http://dx.doi.org/10.1080/0965254X.2023.2253805.

[20] A. Naitali, M. Ridouani, F. Salahdine, N. Kaabouch, Deepfake attacks: Generation, detection, datasets, challenges, and research directions, Comput. 12 (10) (2023) 1–26, http://dx.doi.org/10.3390/computers12100216.

[21] E.I. Galyashina, V.D. Nikishin, The protection of megascience projects from deepfake technologies threats: Information law aspects, in: Journal of Physics: Conference Series, Vol. 2210, 2022, http://dx.doi.org/10.1088/1742-6596/2210/1/012007.

[22] Kimberly T. Mai, Sergi Bray, Toby Davies, Lewis D. Griffin, Warning: Humans cannot reliably detect speech deepfakes, in: Yogan Jaya Kumar (Ed.), PLoS One (ISSN: 1932-6203) 18 (8) (2023) e0285333, http://dx.doi.org/10.1371/journal.pone.0285333.

[23] Kevin Warren, Tyler Tucker, Anna Crowder, Daniel Olszewski, Allison Lu, Caroline Fedele, Magdalena Pasternak, Seth Layton, Kevin Butler, Carrie Gates, Patrick Traynor, 'Better be computer or i'm dumb": A large-scale evaluation of humans as audio deepfake detectors, 2024, pp. 2696–2710, http://dx.doi.org/10.1145/3658644.3670325.

[24] N.V. Kulangareth, J. Kaufman, J. Oreskovic, Y. Fossat, Investigation of deepfake voice detection using speech pause patterns: Algorithm development and validation, JMIR Biomed. Eng. 9 (2024) 1–11, http://dx.doi.org/10.2196/56245.

[25] A. Hamza, et al., Deepfake audio detection via MFCC features using machine learning, IEEE Access 10 (2022) 134018–134028, http://dx.doi.org/10.1109/ACCESS.2022.3231480.

[26] M. McUba, A. Singh, R.A. Ikuesan, H. Venter, The effect of deep learning methods on deepfake audio detection for digital investigation, Procedia Comput. Sci. 219 (2023) 211–219, http://dx.doi.org/10.1016/j.procs.2023.01.283.

[27] F. Iqbal, A. Abbasi, A. R. Javed, Z. Jalil, J. Al-Karaki, Deepfake audio detection via feature engineering and machine learning, in: CEUR Workshop Proceedings, Vol. 3318, 2022.

[28] S.Y. Lim, D.K. Chae, S.C. Lee, Detecting deepfake voice using explainable deep learning techniques, Appl. Sci. 12 (8) (2022) http://dx.doi.org/10.3390/app12083926.

[29] C. Bisogni, V. Loia, M. Nappi, C. Pero, Acoustic features analysis for explainable machine learning-based audio spoofing detection, Comput. Vis. Image Underst. 249 (2024) 104145, http://dx.doi.org/10.1016/j.cviu.2024.104145.

[30] J. Khochare, C. Joshi, B. Yenarkar, S. Suratkar, F. Kazi, A deep learning framework for audio deepfake detection, Arab. J. Sci. Eng. 47 (3) (2022) 3447–3458, http://dx.doi.org/10.1007/s13369-021-06297-w.

[31] R.C. Wijethunga, D.M.K. Matheesha, A. Al Noman, K.H.V.T.A. De Silva, M. Tissera, L. Rupasinghe, Deepfake audio detection: A deep learning based solution for group conversations, in: 2nd International Conference on Advanced Computing, ICAC, 2020, pp. 192–197, http://dx.doi.org/10.1109/ICAC51239.2020.9357161.

[32] Mahmoud Abdel-Salam, Amit Chhabra, Malik Braik, Farhad Soleimanian Gharehchopogh, Nebojsa Bacanin, A halton enhanced solution-based Human Evolutionary Algorithm for complex optimization and advanced feature selection problems, Knowl.-Based Syst. (ISSN: 0950-7051) 311 (2025) 113062, http://dx.doi.org/10.1016/j.knosys.2025.113062, URL https://www.sciencedirect.com/science/article/pii/S0950705125001091.

[33] Aitak Shaddeli, Farhad Soleimanian Gharehchopogh, Mohammad Masdari, Vahid Solouk, An improved African vulture optimization algorithm for feature selection problems and its application of sentiment analysis on movie reviews, Big Data Cogn. Comput. (ISSN: 2504-2289) 6 (4) (2022) http://dx.doi.org/10.3390/bdcc6040104, URL https://www.mdpi.com/2504-2289/6/4/104.

[34] Yuqi Fan, Sheng Zhang, Yaping Wang, Di Xu, Qisong Zhang, An improved flow direction algorithm for engineering optimization problems, Math. (ISSN: 2227-7390) 11 (9) (2023) http://dx.doi.org/10.3390/math11092217, URL https://www.mdpi.com/2227-7390/11/9/2217.

[35] Mahmoud Abdel-Salam, Gang Hu, Emre Çelik, Farhad Soleimanian Gharehchopogh, Ibrahim M. EL-Hasnony, Chaotic RIME optimization algorithm with adaptive mutualism for feature selection problems, Comput. Biol. Med. (ISSN: 0010-4825) 179 (2024) 108803, http://dx.doi.org/10.1016/j.compbiomed.2024.108803, URL https://www.sciencedirect.com/science/article/pii/S0010482524008886.

[36] Mehdi Ayar, Ayaz Isazadeh, Farhad Soleimanian Gharehchopogh, MirHojjat Seyedi, NSICA: Multi-objective imperialist competitive algorithm for feature selection in arrhythmia diagnosis, Comput. Biol. Med. (ISSN: 0010-4825) 161 (2023) 107025, http://dx.doi.org/10.1016/j.compbiomed.2023.107025, URL https://www.sciencedirect.com/science/article/pii/S0010482523004900.

[37] Zahra Moussavi Ali Mohammad Alqudah, A review of deep learning for biomedical signals: Current applications, advancements, future prospects, interpretation, and challenges, Comput. Mater. Continua (ISSN: 1546-2226) 83 (3) (2025) 3753–3841, http://dx.doi.org/10.32604/cmc.2025.063643, URL http://www.techscience.com/cmc/v83n3/61052.

[38] Yunjie Ge, Qian Wang, Jingfeng Zhang, Juntao Zhou, Yunzhu Zhang, Chao Shen, WaveFuzz: A clean-label poisoning attack to protect your voice, 2022, URL https://arxiv.org/abs/2203.13497.

[39] T. Tuncer, S. Dogan, A novel octopus based Parkinson's disease and gender recognition method using vowels, Appl. Acoust. 155 (2019) 75–83, http://dx.doi.org/10.1016/j.apacoust.2019.05.019.

[40] Huzain Azis, Nurul Rismayanti, Munaisyah Binti Abdullah, Suriana Binti Ismail, Speech dataset of human and AI-generated voices, 2025, URL https://www.kaggle.com/dsv/10957363.

[41] Jordan J. Bird, Ahmad Lotfi, DEEP-VOICE: Real-time Detection of AI-Generated Speech for DeepFake Voice Conversion, Licensed under MIT License, 2023.

[42] Shivam Kapoor, Deepfakee_data, 2023, Licensed under CC0 Public Domain License.

[43] Reshma Sunil, Parita Mer, Anjali Diwan, Rajesh Mahadeva, Anuj Sharma, Exploring autonomous methods for deepfake detection: A detailed survey on techniques and evaluation, Heliyon (ISSN: 2405-8440) 11 (3) (2025) e42273, http://dx.doi.org/10.1016/j.heliyon.2025.e42273, URL https://www.sciencedirect.com/science/article/pii/S240584402500653X.

[44] Burak Celik, Ayhan Akbal, A trio-based feature extraction framework for bird sounds classification, Appl. Acoust. (ISSN: 0003-682X) 242 (2026) 111064, http://dx.doi.org/10.1016/j.apacoust.2025.111064, URL https://www.sciencedirect.com/science/article/pii/S0003682X25005365.

[45] M.K. Singh, Feature extraction and classification efficiency analysis using machine learning approach for speech signal, Multimedia Tools Appl. 83 (16) (2024) 47069–47084, http://dx.doi.org/10.1007/s11042-023-17368-5.

[46] Y. El Belghiti, K. Daoudi, K. Minaoui, A.D. MacIntyre, Automatic detection of speech inspiration using SVM and decision trees, in: 12th IEEE International Symposium on Signal, Image, Video, and Communication, ISIVC, 2024, pp. 1–5, http://dx.doi.org/10.1109/ISIVC61350.2024.10577864.

[47] M.B. Karadeniz, Ebru Efeoğlu, Burak Çelik, Adem Kocyigit, Bahattin Türetken, Clipper: An efficient cluster-based data pruning technique for biomedical data to increase the accuracy of machine learning model prediction, Egypt. Inform. J. (ISSN: 1110-8665) 30 (2025) 100641, http://dx.doi.org/10.1016/j.eij.2025.100641, URL https://www.sciencedirect.com/science/article/pii/S1110866525000349.

[48] Karthick Mookkandi, Malaya Kumar Nath, Sanghamitra Subhadarsini Dash, Madhusudhan Mishra, Radak Blange, A robust lightweight vision transformer for classification of crop diseases, AgriEngineering (ISSN: 2624-7402) 7 (8) (2025) http://dx.doi.org/10.3390/agriengineering7080268, URL https://www.mdpi.com/2624-7402/7/8/268.

[49] A. Akbal, StellarisPat: A novel signum and LU-ternary feature extraction approach for speech-based Parkinson's disease detection, 2025.

[50] M. Usama Tanveer Gujjar, K. Munir, M. Amjad, A.U. Rehman, A. Bermak, Unmasking the fake: Machine learning approach for deepfake voice detection, IEEE Access 12 (2024) 197442–197453, http://dx.doi.org/10.1109/ACCESS.2024.3521026.

[51] N. Chakravarty, M. Dua, A lightweight feature extraction technique for deepfake audio detection, Multimedia Tools Appl. 83 (26) (2024) 67443–67467, http://dx.doi.org/10.1007/s11042-024-18217-9.

[52] L.P. Valente, M.M.S. De Souza, A.M.D. Rocha, Speech audio deepfake detection via convolutional neural networks, in: 2024 IEEE Conference on Evolving and Adaptive Intelligent Systems, EAIS, 2024, pp. 1–6, http://dx.doi.org/10.1109/EAIS58494.2024.10569111.

[53] H. Ilyas, A. Javed, K.M. Malik, AVFakeNet: A unified end-to-end dense swin transformer deep learning model for audio-visual deepfakes detection, Appl. Soft Comput. 136 (2023) 110124, http://dx.doi.org/10.1016/j.asoc.2023.110124.

[54] D.M. Ballesteros, Y. Rodriguez-Ortega, D. Renza, G. Arce, Deep4SNet: deep learning for fake speech classification, Expert Syst. Appl. 184 (2021) 115465, http://dx.doi.org/10.1016/j.eswa.2021.115465.

[55] O.A. Shaaban, R. Yildirim, Audio deepfake detection using deep learning, Eng. Rep. 7 (3) (2025) 1–28, http://dx.doi.org/10.1002/eng2.70087.

[56] P. Muruganandham, G.R. Thangasamy, S. Jayaraman, R. Dharmarajan, LSTM autoencoder based parallel architecture for deepfake audio detection with dynamic residual encoding and feature fusion, Sci. Rep. 15 (1) (2025) 1–18, http://dx.doi.org/10.1038/s41598-025-08198-6.

[57] L. Pham, P. Lam, T. Nguyen, H. Nguyen, A. Schindler, Deepfake audio detection using spectrogram-based feature and ensemble of deep learning models, in: 2024 IEEE 5th International Symposium on Internet of Sounds, IS2, 2024, pp. 1–5, http://dx.doi.org/10.1109/IS262782.2024.10704095.

[58] I. Altalahin, S. Alzu'Bi, A. Alqudah, A. Mughaid, Unmasking the truth: A deep learning approach to detecting deepfake audio through MFCC features, in: 2023 International Conference on Information Technology and Cybersecurity Challenges in Smart Cities, ICIT, 2023, pp. 511–518, http://dx.doi.org/10.1109/ICIT58056.2023.10226172.