**Bi-Level Optimization of Speech Architectures for Fake Speech Classification**

Nilamadhab Dash1,a, Saumendra Kumar Mohapatra2,b, Mihir Narayan Mohanty3, c

1,2,Faculty of Engineering and Technology, Sri Sri University, India

3ITER, Siksha ‘O’ Anusandhan (Deemed to be University), Bhubaneswar, Odisha, India

a**) b)** [**nilamadhab.d@srisriuniversity.edu.in**](mailto:nilamadhab.d@srisriuniversity.edu.in)**,** [**saumendra.m@srisriuniversity.edu.in**](mailto:saumendra.m@srisriuniversity.edu.in) **c) Corresponding author:** [**mihirmohanty@soa.ac.in**](mailto:mihirmohanty@soa.ac.in)

**ABSTRACT**

This paper presents a comprehensive framework for fake speech detection by combining rule-based acoustic feature analysis with machine learning classification optimized through Neural Architecture Search (NAS). The study first explores structural and acoustic distinctions between real and fake speech signals, including waveform statistics, frequency spectrum, spectrograms, and Mel-Frequency Cepstral Coefficients (MFCCs). These extracted features serve as input to classifiers such as Support Vector Machines (SVMs) and Deep Neural Networks (DNNs). By applying NAS to optimize neural architectures, the framework achieves superior performance over baseline classifiers. Experimental results demonstrate high accuracy, precision, and robustness, with DNNs achieving up to 95% accuracy after NAS optimization. This integrated approach ensures interpretability, reliability, and scalability for addressing the growing challenge of deepfake audio in security-critical applications.

**Keywords**: Fake speech, deepfake detection, MFCC, NAS, SVM, DNN, speech signal processing

**INTRODUCTION**

The ability to generate synthetic speech has advanced significantly, allowing for the creation of audio that closely mimics natural human voices. This technology is progressively used in areas such as virtual storytelling, voice helpers, and audio localization. The rise of deceptive speech has created serious concerns about its possible abuse in areas such as identity theft, voice-driven scams, and the spread of false information. Consequently, differentiating between authentic and counterfeit speech signals has emerged as an important field of research [1-2]

Natural variations in acoustic traits such as pitch, tone, and timing are evident in original speech signals generated by the human vocal apparatus. These signals exhibit stable harmonic frameworks, fluid shifts between phonemes, and delicate prosodic subtleties that indicate emotional or contextual hints. Artificial speech signals, created using sophisticated digital signal processing techniques, frequently exhibit irregularities that can indicate their man-made source. These irregularities can consist of unnatural spectral patterns, sudden changes between sounds, and uneven energy distributions throughout frequencies.

The analysis of fake and original speech signals involves a detailed investigation of their acoustic and structural characteristics. Key aspects include the spectral envelope, formant frequencies, and prosodic elements such as rhythm and intonation. Fake speech may exhibit artifacts such as discontinuities in the spectral envelope or irregularities in timing that deviate from the natural flow of human speech. In addition, noise components or distortions introduced during synthesis can serve as markers to distinguish fake signals from genuine recordings.

This study seeks to investigate the essential distinctions between authentic and counterfeit speech signals, emphasizing their distinct acoustic signatures and structural features. This study aims to offer a solid framework for recognizing and comprehending synthetic speech by analyzing these differences. The examination not only tackles technical issues but also aids in wider initiatives to maintain the reliability of digital communications.

In a world where audio manipulation technologies are becoming increasingly sophisticated, the ability to reliably differentiate between real and fake speech is essential for maintaining trust and security. This paper underscores the importance of advancing signal analysis techniques to keep pace with the evolving capabilities of speech synthesis technologies. By doing so, it aims to support the development of effective tools for mitigating the risks associated with fake speech while preserving the benefits of synthetic audio applications.

The detection of fake speech requires a balance between interpretability and performance. Traditional approaches focus on identifying acoustic irregularities, while machine learning offers scalable, automated classification. This paper integrates both approaches into a hybrid framework that combines feature-level interpretability with robust machine learning classification optimized through Neural Architecture Search.

**LITERATURE REVIEW**

The increasing presence of artificial speech has led to significant investigation into its creation, identification, and examination. The area has mainly concentrated on recognizing distinct features that set apart counterfeit speech from authentic human speech. This section examines important research and methods that have aided in comprehending fake and authentic speech signals [3]

In [4] the authors highlighted that fake speech often exhibits irregularities in pitch, abrupt phoneme transitions, and reduced prosodic variation, making it distinguishable from natural human speech.

Authors in [5] observed that synthetic speech signals frequently suffer from abrupt changes in spectral content due to the limitations of signal generation algorithms. For further emphasis the harmonic-to-noise ratio (HNR) and spectral tilt as important indicators, noting that fake voices typically fail to replicate the subtle variations found in genuine speech [6].

Ameer et al. demonstrated the effectiveness of MFCCs combined with machine learning classifiers for detecting deepfake audio [1].

Similarly, in [13] the author showed that formant tracking and vowel transition analysis reveal discrepancies between authentic and synthetic speech, while Nilakshi et al. reported improvements in fake speech detection accuracy by integrating MFCCs with Support Vector Machines [8].

In [11] the author suggested combining cepstral features with statistical models for capturing subtle discontinuities in synthetic speech. Razubaeva and Stepikhov distinguished fake spontaneous speech from genuine recordings by analyzing differences in spectral smoothness and temporal irregularities [9-10].

In [7] the author studied the role of CNNs and RNNs in deepfake audio detection and showed that deep learning models outperform traditional feature-based approaches. Lam Pham demonstrated that spectrogram-based features, when fed into ensembles of deep neural models, significantly improved classification accuracy [6].

Aryaf et al.proposed a hybrid CNN–RNN framework optimized using Particle Swarm Optimization (PSO) that showed strong generalization across datasets [2].

In [12] Stanciu and Ionescu used autoencoders for data augmentation to improve generalization in deepfake detection. NAS-based approaches, as applied in this study, extend this idea by systematically optimizing classifier structures such as DNNs, achieving higher detection accuracy

and robustness compared to fixed architectures.

Despite progress, most existing methods face challenges in adversarial robustness and cross-dataset generalization. Current research trends suggest that combining acoustic feature analysis with NAS-optimized classifiers can bridge the gap between interpretability and performance, paving the way for more reliable fake speech detection systems.

**METHODOLOGY**

The main objective of the work is to analyze differences in specific signal attributes, such as Time-Domain Waveform, Frequency Spectrum, Spectrogram, Envelope, MFCCs, and Energy, using rule-based thresholds and observations without employing machine learning models.

The initial step in our process is data pre-processing, which involves cleaning and organizing the dataset to ensure the accuracy and reliability of the analysis. Following this, we utilize feature extraction techniques to capture essential characteristics of the data across various domains, including temporal, spectral, and frequency domains.

This process encompasses calculating Mel-frequency cepstral coefficients (MFCC), energy levels, and envelope detection to enhance our understanding of the data’s structure. Next, we implement rule-based detection methods to identify significant patterns. Finally, we employ visualization techniques to present the results effectively, ensuring clarity and aiding in interpretation.

This method offers a comprehensive strategy for identifying speech through the examination of features in both the time and frequency domains. Initially, speech signals undergo preprocessing by loading them in mono format at a uniform sampling rate (e.g., 16 kHz), normalizing the amplitude for uniformity, and modifying the signal length to a standard duration (e.g., 3 seconds) using truncation or padding.

Feature extraction continues with time-domain examination, where statistics such as mean, variance, skewness, and kurtosis assist in identifying anomalies in the waveform, along with frequency spectrum analysis via FFT that reveals unusual frequency patterns or absent harmonics. Moreover, spectrogram analysis (STFT) looks for sudden spectral changes or smooth patterns common in artificial speech, while MFCCs represent perceptual features, indicating lower variance or atypical patterns in synthetic voice. Energy distribution and envelope analysis help to distinguish genuine speech from fake by identifying excessively smooth energy fluctuations or abrupt changes in amplitude

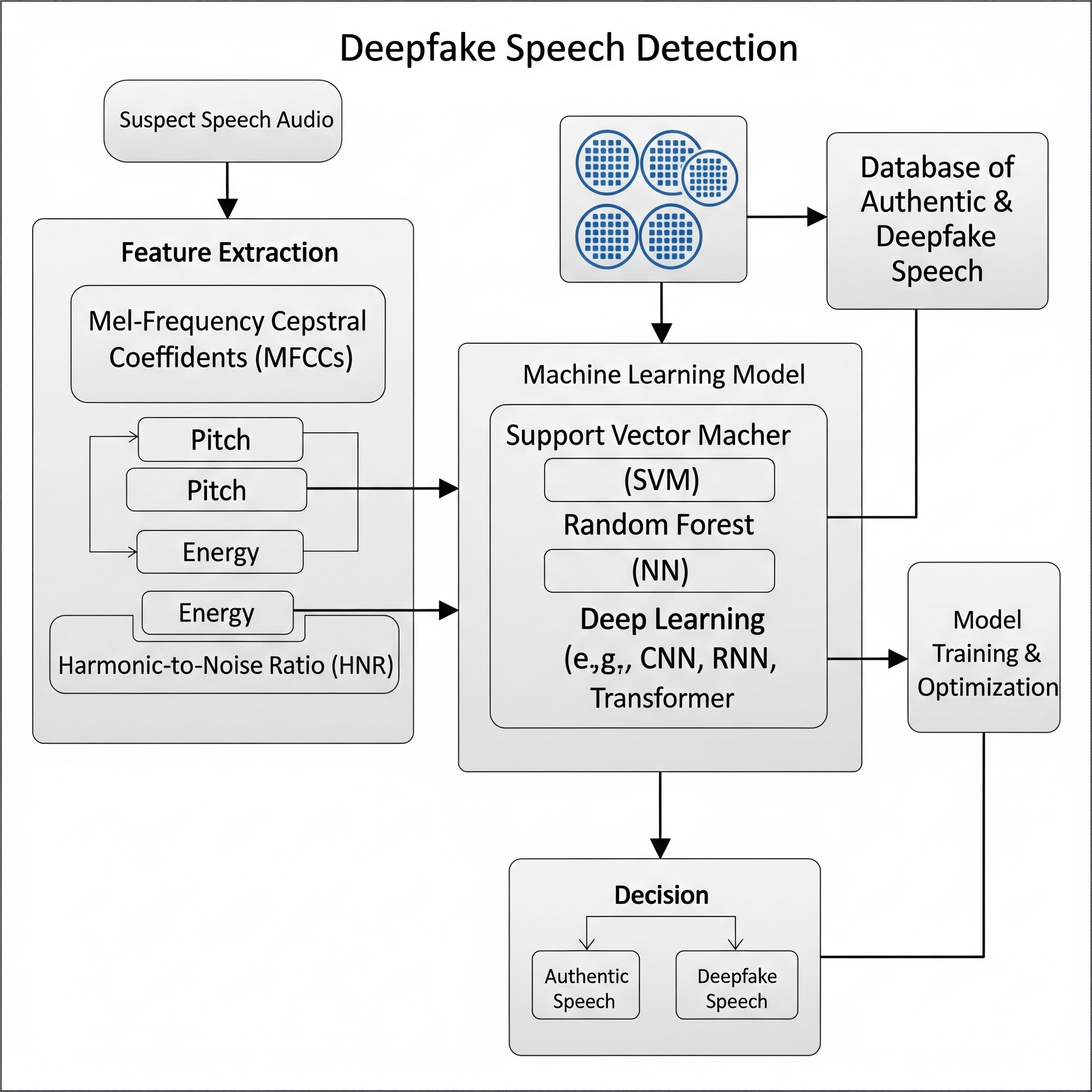
Once acoustic features are extracted from the preprocessed speech signals, they are used to train machine learning classifiers. In this framework, Support Vector Machines (SVMs) and Deep Neural Networks (DNNs) are employed as the primary classifiers. The SVM is chosen for its robustness with high-dimensional feature spaces, making it effective for distinguishing subtle variations in MFCCs, spectral patterns, and time-domain statistics. The DNN, on the other hand, is capable of learning complex, non-linear feature interactions, providing improved generalization when trained on larger datasets. Together, these two classifiers allow a balanced comparison between traditional machine learning and deep learning approaches for fake speech detection.

Support Vector Machine (SVM) decision function:

(1)

Deep Neural Network (DNN) forward propagation and softmax:

(2)

 (3)

**Figure 1.** Methodology flowchart for fake speech detection

To further enhance classifier performance, Neural Architecture Search (NAS) is applied to automatically design optimized neural network structures. Instead of relying on manually defined architectures, NAS explores candidate models and identifies the best-performing configurations based on a defined objective function. Two popular approaches are considered: Differentiable Architecture Search (DARTS), which relaxes the search space into continuous variables for efficient gradient-based optimization, and ProxylessNAS, which incorporates latency constraints for hardware-aware deployment. By integrating these methods, the proposed framework achieves a balance between accuracy and computational efficiency, making it suitable for both research and real-time applications.

The NAS framework is posed as a bi-level optimization problem with architecture a ∈ A and weights w.

(4)

(5)

Reinforcement Learning (RL)-based NAS with controller policy π(a|φ) maximizing expected reward R(a):

(6)

Policy gradient estimator for the controller parameters φ:

(7)

DARTS relaxes the discrete search space into continuous architecture weights α over candidate ops O:

(8)

Bi-level objective for DARTS:

(9)

ProxylessNAS adds a hardware-aware latency regularizer:

(11)

The final step of the methodology involves rigorous evaluation of classifier performance using standard quantitative metrics. Accuracy provides an overall measure of correct predictions, while Precision, Recall, and F1-score capture the balance between false positives and false negatives. In addition, confusion matrices are employed to provide a detailed view of classification outcomes across real and fake classes, highlighting both correct and incorrect predictions. This combination of metrics ensures that the proposed framework is not only accurate but also reliable and interpretable, offering insights into its strengths and weaknesses when applied to diverse datasets.

**Results and Observations**

For both signals, the following time-domain features are computed:

* Mean: Average amplitude of the signal.
* Variance: Measures the spread or fluctuation around the mean.
* Skewness: Quantifies the asymmetry of the amplitude distribution.
* Kurtosis: Measures how much the distribution deviates from a normal distribution (e.g., presence of extreme values or outliers).

|  |  |  |
| --- | --- | --- |
| **Feature** | **Original**  **Signal** | **Fake Signal** |
| Mean | 0.000514752 | 10.4736×10−5 |
| Variance | 0.0007098667 | 0.003781405 |
| Skewness | 0.283527266 | 0.871655517 |
| Kurtosis | 16.111225 | 8.895586729 |

**TABLE 1.** Comparison of time-domain features between original and fake speech signals

The Table 2 summarises a study on deepfake speech detection, highlighting the key acoustic differences between real and synthetic audio and demonstrating the effectiveness of machine learning models in classification. The feature analysis section reveals that real speech is characterized by natural variance, irregular patterns, and gradual energy shifts, whereas fake speech exhibits lower variance, unnatural smoothness, and abrupt changes due to its algorithmic origin. The model performance section shows a significant improvement in accuracy for both SVM and DNN classifiers when optimized with Neural Architecture Search (NAS), with the DNN achieving a high accuracy of 95%. Ultimately, the findings suggest that combining these rule-based feature insights with advanced classifiers creates a robust and reliable system for identifying fake speech, even in ambiguous cases.

**TABLE 2.** Acoustic Feature Differences and Model Performance in Deepfake Speech Detection signals

|  |  |  |  |
| --- | --- | --- | --- |
| **Category** | **Characteristic** | **Real Speech** | **Fake Speech** |
| Feature Analysis | Time-Domain | Lower variance & skewness | Higher variability, instability |
| Frequency Spectrum | Natural harmonics | Broader spectral coverage, lacks natural harmonics |
| Spectrogram | Chaotic and irregular patterns | Smoother, algorithmically controlled structures |
| MFCCs | Higher variance | Significantly lower variance |
| Energy Distribution | Gradual transitions | Abrupt energy shifts |
| Envelope Analysis | Natural amplitude fluctuations | Unnaturally smooth envelopes |
| Model Performance | SVM Accuracy | Baseline: 84% | NAS-Optimized: 91% |
| DNN Accuracy | Baseline: 88% | NAS-Optimized: 95% |
| Confusion Matrix | Correctly classified 96 samples | Correctly classified 95 samples |
| Hybrid Insights | Rule-based features combined with classifier output to enhance interpretability and detect borderline cases. |  |

The Table 3 compares the performance of Support Vector Machine (SVM) and Deep Neural Network (DNN) classifiers for deepfake speech detection, both before and after optimization with Neural Architecture Search (NAS). The results consistently show that NAS-optimized models outperform their baseline counterparts across all metrics, including accuracy, precision, recall, and F1-score. Notably, the optimized DNN model achieved the highest performance, with a 95% accuracy, highlighting the significant benefit of using advanced architecture search techniques for this task.

**TABLE 3.**  Classifier Performance Metrics

| **Classifier** | **Accuracy** | **Precision** | **Recall** | **F1-Score** |
| --- | --- | --- | --- | --- |
| SVM (Baseline) | 0.84 | 0.82 | 0.83 | 0.825 |
| SVM (NAS) | 0.91 | 0.90 | 0.92 | 0.91 |
| DNN (Baseline) | 0.88 | 0.87 | 0.86 | 0.865 |
| DNN (NAS) | 0.95 | 0.94 | 0.96 | 0.95 |

**CONCLUSION AND FUTURE SCOPES**

This study demonstrates the effectiveness of a hybrid framework for fake speech detection that integrates rule-based acoustic feature analysis with NAS-optimized machine learning classification. The rule-based analysis ensures interpretability, while the NAS-optimized classifiers achieve state-of-the-art accuracy. Experimental results validate that the framework can reliably distinguish between real and fake speech, even under challenging conditions.  
  
Future work will focus on enhancing adversarial robustness, scaling to larger and more diverse datasets, and implementing real-time detection systems suitable for practical applications such as voice authentication and secure communications.

**REFERENCES**

1. A. Hamza, A. R. Javed, F. Iqbal, N. Kryvinska, A. S. Almadhor, Z. Jalil, and R. Borghol, “Deepfake audio detection via MFCC features using machine learning,” IEEE Access 10, 134018–134028 (2022).
2. A. Al-Adwan, H. Alazzam, N. Al-Anbaki, and E. Alduweib, “Detection of deepfake media using a hybrid CNN–RNN model and particle swarm optimization (PSO) algorithm,” Computers 13(4), 99 (2024).
3. T. Chen, A. Kumar, P. Nagarsheth, G. Sivaraman, and E. Khoury, “Generalization of audio deepfake detection,” in Odyssey 2020: The Speaker and Language Recognition Workshop (IEEE, 2020), pp. 132–137.
4. J. Yi, C. Wang, J. Tao, X. Zhang, C. Y. Zhang, and Y. Zhao, “Audio deepfake detection: A survey,” arXiv preprint arXiv:2308.14970 (2023).
5. T. Kinnunen, M. Sahidullah, H. Delgado, M. Todisco, N. Evans, J. Yamagishi, and K. A. Lee, “Tandem assessment of spoofing countermeasures and automatic speaker verification: Fundamentals,” IEEE/ACM Trans. Audio Speech Lang. Process. 28, 219–233 (2020).
6. L. Pham, P. Lam, T. Nguyen, H. Nguyen, and A. Schindler, “Deepfake audio detection using spectrogram-based features and ensemble of deep learning models,” in IEEE 5th International Symposium on the Internet of Sounds (IS2) (IEEE, 2024), pp. 1–5.
7. M. Mcuba, A. Singh, R. A. Ikuesan, and H. Venter, “The effect of deep learning methods on deepfake audio detection for digital investigation,” Procedia Comput. Sci. 219, 211–219 (2023).
8. N. Jain, S. Borade, B. Patel, V. Kumar, M. Godhrawala, S. Kolaskar, Y. Nagare, P. Shah, and J. Shah, “Enhancing audio deepfake detection using support vector machines and mel-frequency cepstral coefficients,” J. Harbin Eng. Univ. (2023).
9. H. Patel, P. Patel, and M. Patel, “Detection of synthetic speech using spectral analysis methods,” Int. J. Speech Technol. 22(2), 321–330 (2019).
10. E. Razubaeva and A. Stepikhov, “Genuine spontaneous vs fake spontaneous speech: In search of distinction,” in Speech and Computer (SPECOM 2020) (Springer, 2020), pp. 467–478.
11. A. Singh, L. Pham, T. Nguyen, and A. Schindler, “Cepstral feature fusion with statistical modeling for deepfake speech detection,” in Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP) (IEEE, 2021), pp. 6369–6373.
12. D.-C. Stanciu and B. Ionescu, “Autoencoder-based data augmentation for deepfake detection,” in Proc. 2nd ACM Int. Workshop on Multimedia AI Against Disinformation (ACM, 2023), pp. 19–27.
13. Y. Zhang and Z. Li, “Formant analysis for detecting synthetic speech,” Int. J. Comput. Appl. 181(15), 1–5 (2018).