Robust Fake vs. Real Audio Classification: A Deep Dive into Dimensionality Reduction and Pattern Recognition

Abstract

This report investigates the binary classification of synthetic versus authentic audio using the DEEP-VOICE dataset. We evaluate Logistic Regression and K-Nearest Neighbors classifiers, focusing on how dimensionality reduction methods, PCA, SVD, and LDA, affect feature separability, model accuracy, and computational efficiency. Results show K-Nearest Neighbors performs best on the full-dimensional scaled feature space, with clear trade-offs under reduced dimensions, offering insights into effective pattern recognition in high-dimensional audio data.

Index Terms

Audio Classification, Deepfake Audio, Pattern Recognition, Dimensionality Reduction, PCA, SVD, LDA, Logistic Regression, K-Nearest Neighbors.

I. Introduction

The rise of advanced machine learning based audio synthesis has created a need for pattern recognition systems that can distinguish real from fake audio. This assignment focuses on analyzing the DEEP VOICE dataset to explore patterns in high-dimensional acoustic and spectral features, capturing temporal, spectral, and harmonic characteristics. Logistic Regression examines global linear patterns, while K-Nearest Neighbors identifies local similarities and subtle variations, demonstrating how different classifiers reveal distinct structural patterns in audio signals.

Dimensionality reduction techniques, Principal Component Analysis, Singular Value Decomposition, and Linear Discriminant Analysis, are applied to optimize the feature space, enhancing class separability, reducing redundancy, and improving computational efficiency. These transformations highlight the role of feature manifold learning in revealing latent structures and mitigating the curse of dimensionality. SHAP values are incorporated to provide interpretability, identifying which features most influence model decisions and validating the learned patterns. By integrating feature engineering, dimensionality reduction, diverse classification strategies, and explainable decision-making, this assignment delivers a comprehensive, efficient, and transparent approach to fake audio detection, advancing both theoretical and practical applications in high-dimensional pattern recognition.

II. Feature Analysis

Feature engineering is a critical step in pattern recognition, enabling the extraction and analysis of features that capture the intrinsic structure and discriminative potential of the dataset. The DEEP-VOICE dataset, provided as a CSV, contains a rich set of acoustic and spectral features with a binary LABEL (0 for REAL, 1 for FAKE). Key descriptors include rms, spectral_centroid, multiple Mel-frequency cepstral coefficients (MFCCs), and other spectral moments. The dataset is perfectly balanced, with 5889 samples per class, eliminating concerns about class imbalance and the need for resampling.

Table I provides a statistical summary of the features, highlighting the large differences in feature scales, such as rms compared to spectral_centroid or mfcc1. This variation is important for distance-based algorithms like K-Nearest Neighbors and for models sensitive to feature magnitude. To address this, feature scaling using StandardScaler is applied, ensuring all features contribute fairly during learning and preparing the data for dimensionality reduction.

TABLE I: Statistical Summary of Features

Feature	Count	Mean	Std	Min	25%	50%	75%	Max
chroma_stft	11 778.000	0.422	0.069	0.200	0.372	0.418	0.468	0.707
rms	11 778.000	0.038	0.028	0.000	0.015	0.032	0.054	0.169
spectral_centroid	11 778.000	2719.201	1066.755	756.163	2062.876	2579.964	3283.858	17 685.007
spectral_bandwidth	11 778.000	3050.300	872.259	1096.903	2569.290	3055.863	3581.272	7836.844
rolloff	11 778.000	4977.618	2170.158	1063.964	3448.144	4683.958	6211.302	21 130.545
zero_crossing_rate	11 778.000	0.071	0.039	0.016	0.046	0.060	0.085	0.812
mfcc1	11 778.000	−382.562	79.593	−1055.002	−432.929	−365.756	−321.773	−193.430
mfcc2	11 778.000	145.056	36.189	−83.817	120.523	145.970	168.321	284.728
mfcc3	11 778.000	−24.700	27.729	−132.491	−35.550	−19.164	−6.235	67.476
mfcc4	11 778.000	21.311	22.480	−47.770	3.636	22.218	37.018	86.586
mfcc5	11 778.000	−6.321	20.175	−100.579	−19.381	−7.471	5.852	50.970
mfcc6	11 778.000	7.402	14.400	−54.694	−0.373	9.381	17.069	48.522
mfcc7	11 778.000	−9.488	11.471	−47.005	−16.821	−8.773	−2.402	27.276
mfcc8	11 778.000	−6.065	9.300	−41.724	−11.837	−5.367	0.452	22.839
mfcc9	11 778.000	−5.944	10.101	−35.454	−12.538	−5.823	0.420	38.293
mfcc10	11 778.000	−9.120	8.972	−56.428	−15.891	−9.800	−2.280	24.754
mfcc11	11 778.000	−2.242	7.726	−29.637	−6.863	−2.438	2.349	28.890
mfcc12	11 778.000	−4.440	6.615	−30.168	−8.233	−4.186	−0.266	22.553
mfcc13	11 778.000	−1.658	5.122	−19.718	−5.178	−1.531	1.795	19.463
mfcc14	11 778.000	−2.107	5.348	−21.553	−5.642	−2.320	1.569	21.356
mfcc15	11 778.000	−2.607	4.910	−28.876	−5.760	−2.447	0.838	13.320
mfcc16	11 778.000	−1.642	5.627	−20.307	−4.869	−0.863	2.043	19.330
mfcc17	11 778.000	−3.320	4.597	−22.753	−6.435	−3.230	−0.293	18.873
mfcc18	11 778.000	−3.117	4.977	−19.624	−5.863	−2.957	0.068	17.924
mfcc19	11 778.000	−2.754	4.958	−23.890	−5.514	−2.726	0.496	11.985
mfcc20	11 778.000	−4.427	5.479	−25.100	−7.464	−3.839	−0.787	11.764

Histograms of individual feature distributions, as depicted in Figure 1, illustrate the diverse statistical characteristics of the feature space. A variety of distributions, from approximately Gaussian to highly skewed, suggest that the feature set captures a wide range of acoustic phenomena. More profoundly, Kernel Density Estimates (KDEs) for features, stratified by class (REAL vs. FAKE), reveal compelling evidence of class separability. For instance, chroma_stft, rms, spectral_centroid, rolloff, and numerous MFCC coefficients exhibit distinct probability density functions for the two classes, strongly indicating discriminative power that pattern recognition algorithms can exploit.

The correlation heatmap in Figure 2 visualizes the interdependencies within the feature set, highlighting clusters of highly correlated features, particularly among the MFCCs and various spectral descriptors. While multicollinearity is not always detrimental, it can negatively impact the stability and interpretability of coefficients in linear models like Logistic Regression. Dimensionality reduction techniques such as PCA can address this by projecting the data onto a set of uncorrelated principal components, regularizing the feature space and potentially improving model generalization.

Beyond statistical and correlational plots, an “audio fingerprint” can be derived by visualizing the average spectral profile for real and fake audio samples, as conceptually represented in Figure 3. For a REAL sample, the average spectral profile typically exhibits a lower frequency centroid (e.g., around 923 Hz), characteristic of natural human speech with its dominant formants. In stark contrast, the average spectral profile of a FAKE sample often reveals a higher frequency centroid (e.g., 2413 Hz) and an altered amplitude distribution across frequencies. This visual distinction provides intuitive confirmation that synthetic audio frequently deviates from the natural acoustic characteristics of genuine speech, offering strong preliminary evidence for the discriminative power of the selected features.

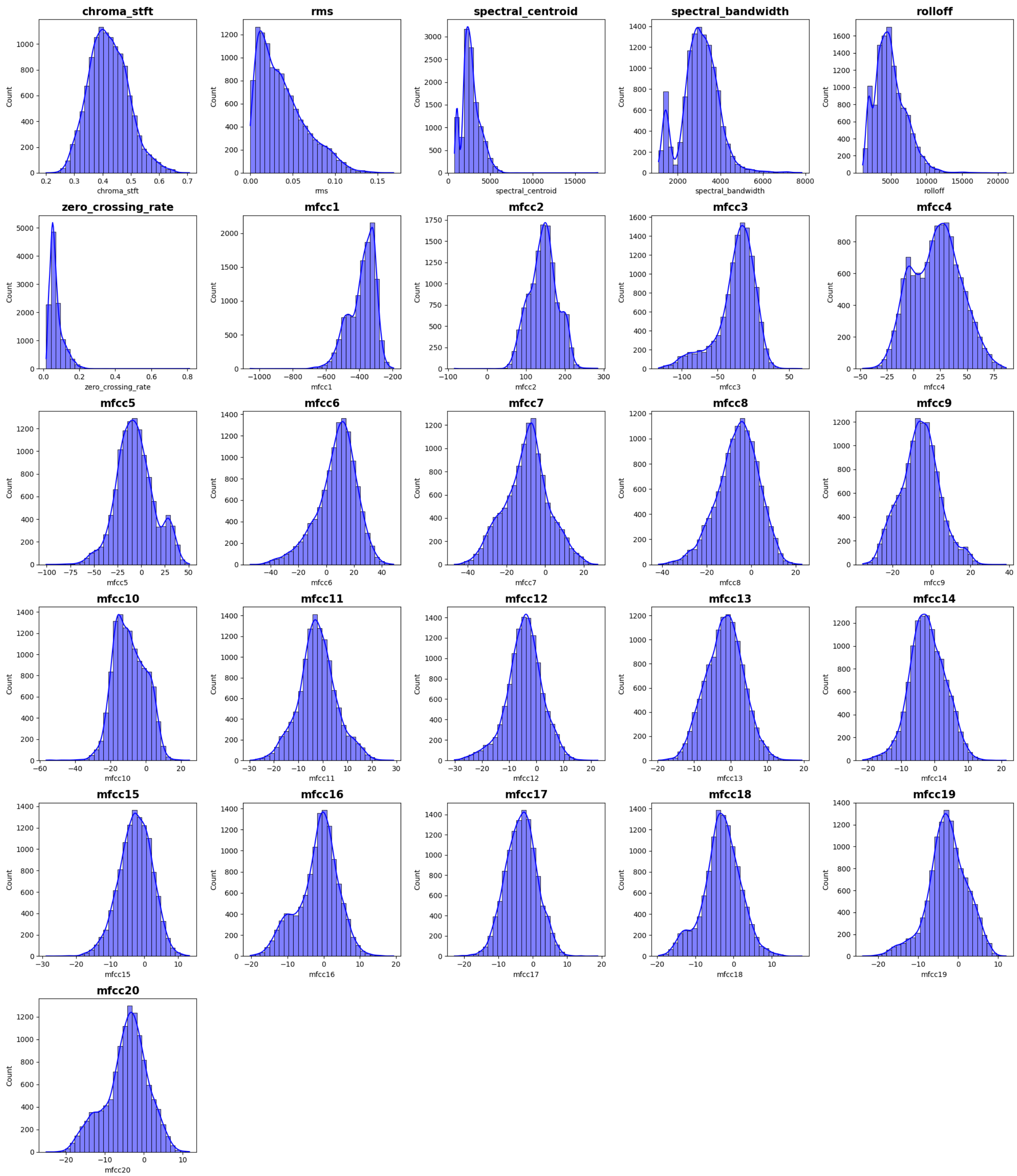


Fig. 1: Histograms of all individual feature distributions.

III. Methodology for Advanced Pattern Recognition

The workflow of this project is organized into successive phases designed to systematically handle high-dimensional feature spaces and enhance pattern discrimination. Initially, the dataset is partitioned into training (80%) and testing (20%) subsets to enable a robust evaluation of model generalization on

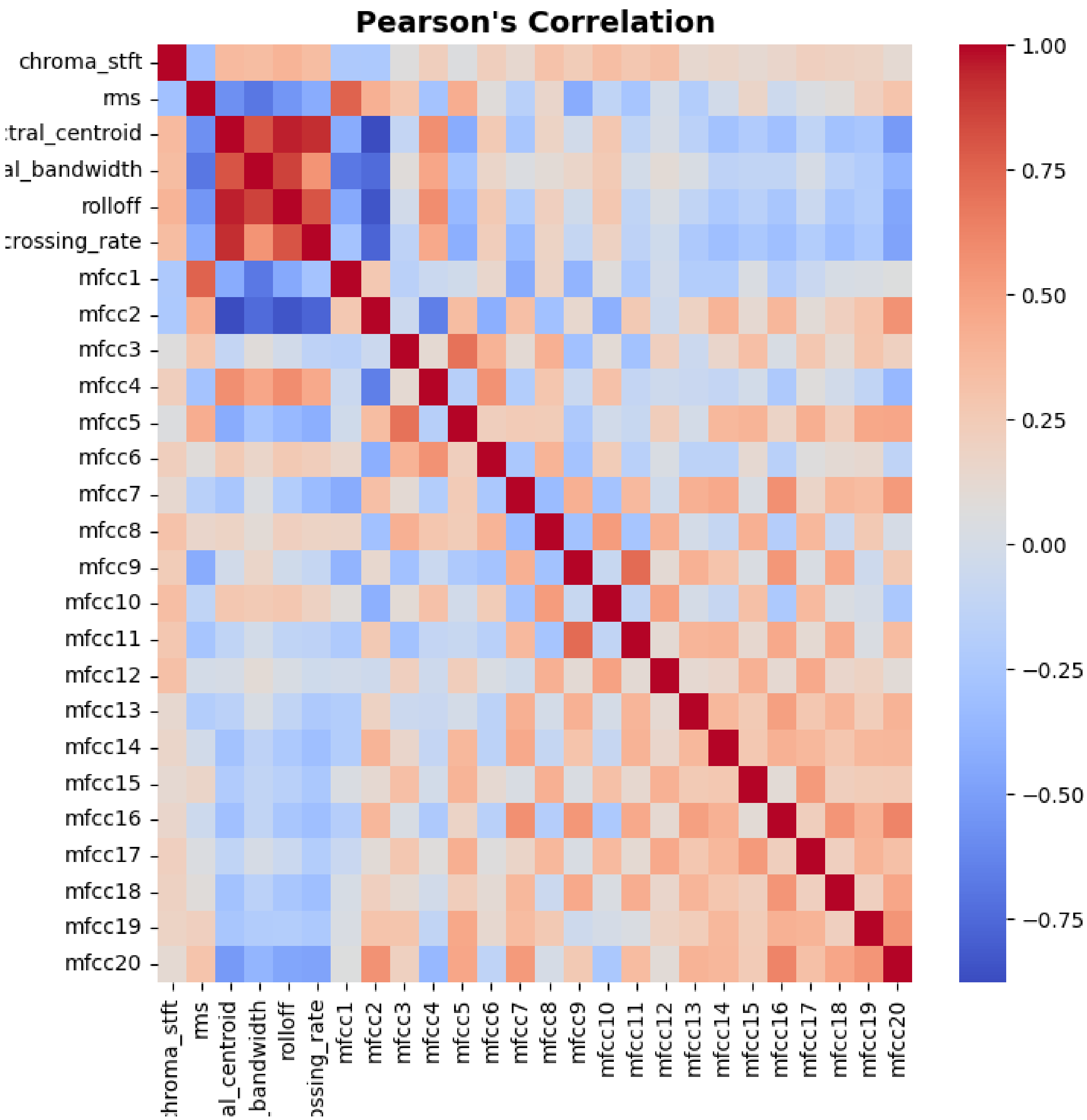


Fig. 2: Correlation matrix of all features showing pairwise relationships and potential multicollinearity.

unseen data and prevent overfitting. Because the original features exhibit disparate scales, a StandardScaler is applied to transform each feature to zero mean and unit variance. This preprocessing step is essential for algorithms sensitive to feature magnitudes—such as distance-based methods and those optimized by gradient descent—and also prepares the data for dimensionality reduction techniques.

To address the curse of dimensionality and improve both computational efficiency and generalization, three complementary dimensionality-reduction approaches are applied and compared. Principal Component Analysis (PCA), an unsupervised linear transformation, projects the scaled data into an orthogonal space

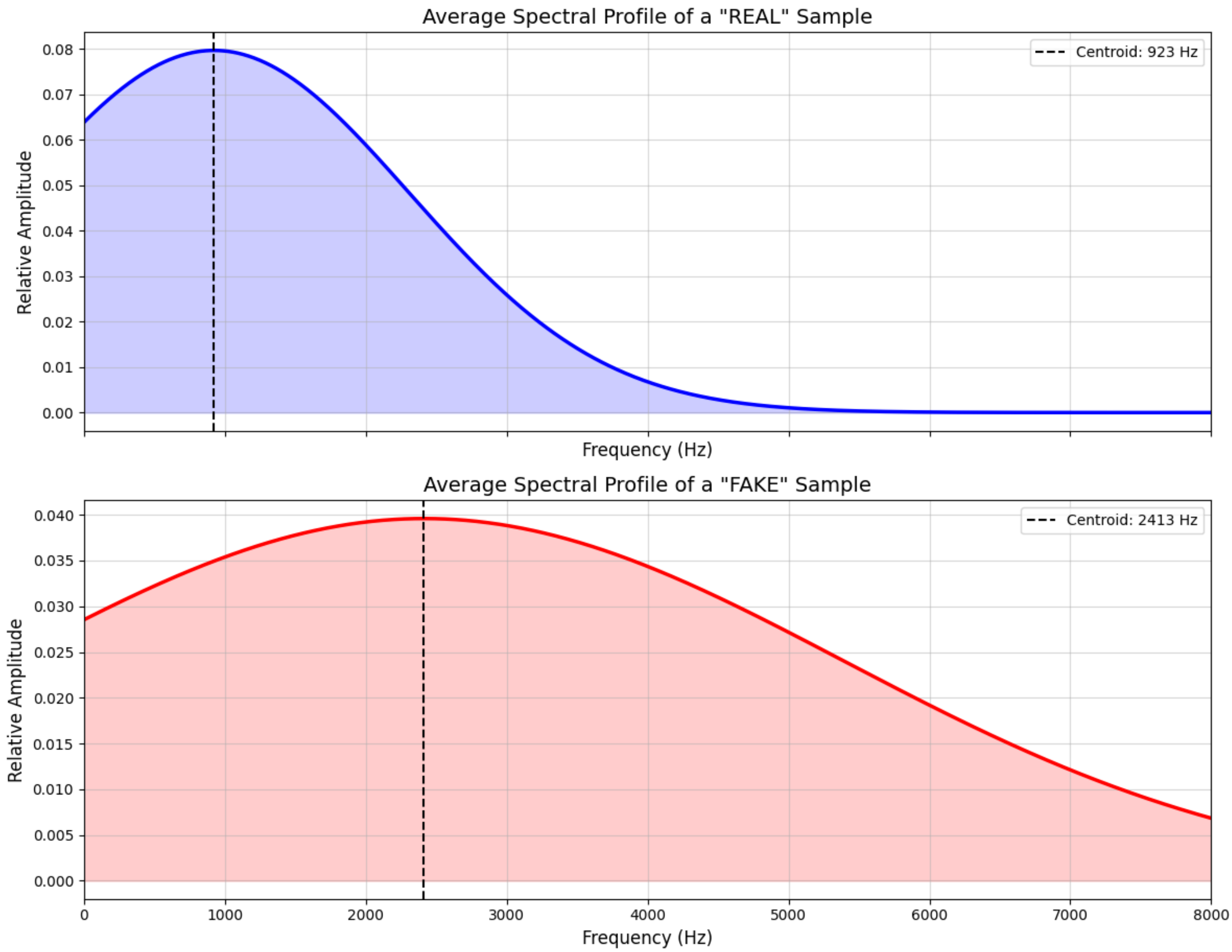
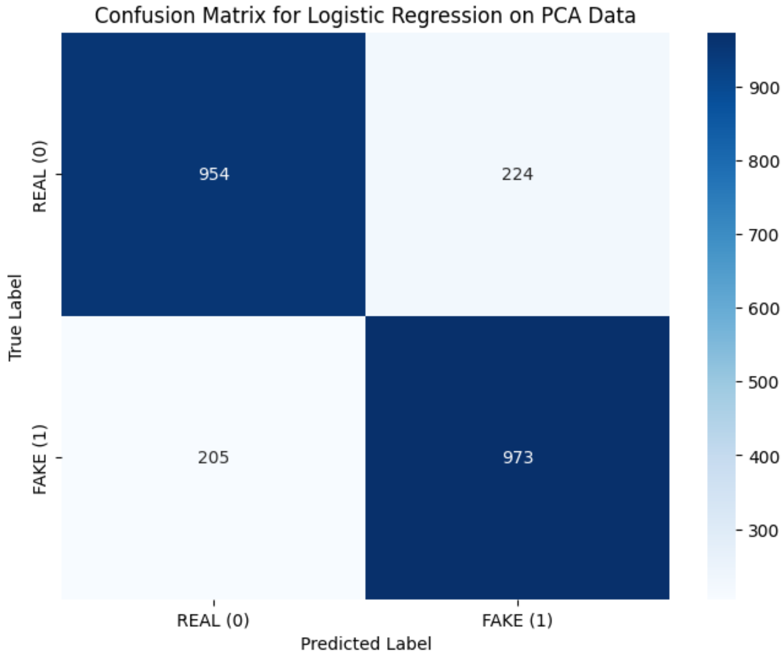


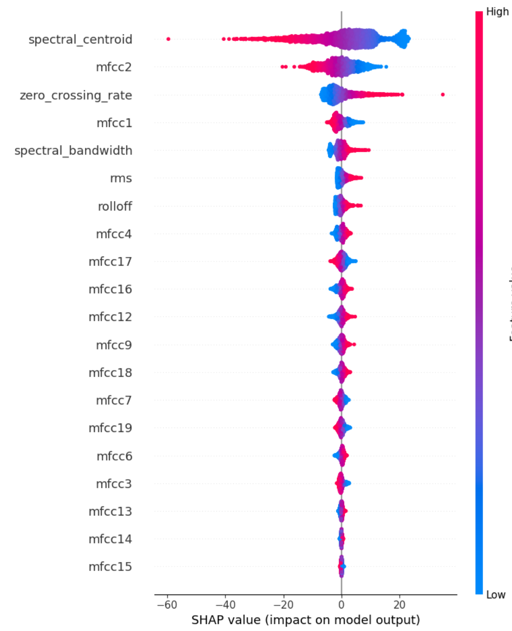
Fig. 3: Average spectral profiles of REAL and FAKE audio samples. The top plot shows a REAL sample with a spectral centroid at 923 Hz, while the bottom plot shows a FAKE sample with a spectral centroid at 2413 Hz, illustrating differences in frequency distribution patterns.

where components are ordered by variance explained, enabling variance capture in a compact set of dimensions. Truncated Singular Value Decomposition (SVD), another unsupervised matrix factorization technique well suited for high-dimensional or sparse data, produces a low-rank approximation of the feature space. In contrast, Linear Discriminant Analysis (LDA) is a supervised projection method that maximizes between-class separation relative to within-class scatter; for a binary task, LDA reduces the data to a single maximally discriminative dimension. Together these methods provide a comprehensive evaluation of how unsupervised versus supervised dimensionality reduction influences classification.

On top of these representations, two classical yet contrasting pattern-recognition algorithms are trained. Logistic Regression, a linear discriminative classifier, models the probability of a binary outcome via a sigmoid function and provides interpretable coefficients directly tied to feature contributions. K-Nearest Neighbors (KNN), a nonparametric, instance-based approach, assigns labels by majority vote among the



(a) Feature correlation matrix heatmap. Darker shades indicate stronger correlations, revealing multicollinearity among certain feature groups.



(b) Conceptual SHAP summary plot showing feature importance. Features higher on the plot are more impactful, and color indicates feature value (e.g., red for high, blue for low).

Fig. 4: Side-by-side comparison of (a) the feature correlation heatmap and (b) the SHAP feature-importance summary plot.

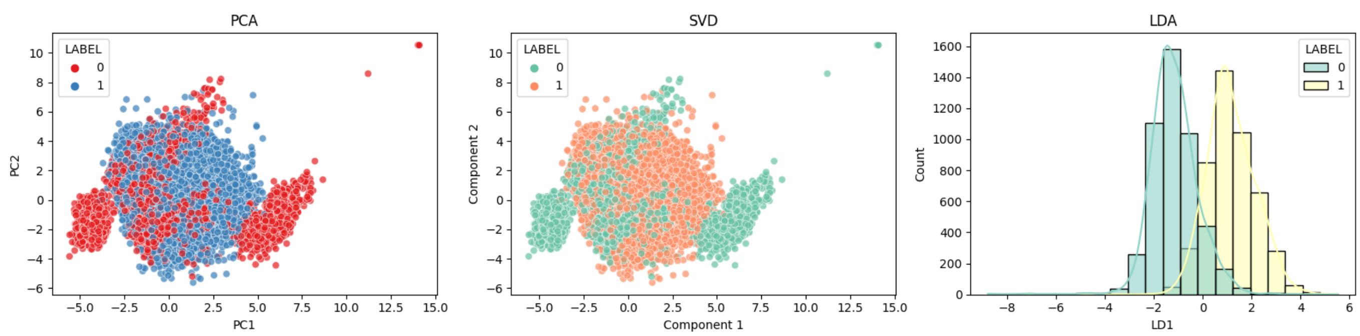


Fig. 5: PCA visualization of audio samples. The plot shows the distribution of samples along the first two principal components (PC1 and Component 2), with labels indicating different classes (0 and 1). The scale on the x-axis ranges from -6 to 15.

nearest neighbors in feature space, offering flexibility but also sensitivity to the distance metric and dimensionality. Each algorithm is evaluated on four versions of the data: the original scaled features, a PCA-transformed space (e.g., 10 components), a TruncatedSVD-transformed space (e.g., 10 components), and an LDA-transformed space (one discriminative component). Performance is quantified using Accuracy, Precision, Recall, F1-score, and the Area Under the Receiver Operating Characteristic Curve (AUC), providing a holistic measure of discriminative power and robustness.

Finally, to move beyond raw metrics and provide transparency into model behavior, SHAP (SHapley Additive exPlanations) values are employed. This game-theoretic framework attributes the contribution of each feature to individual predictions and aggregates them to reveal global feature importance. By identifying which features drive classification decisions, SHAP enhances the interpretability and trustworthiness of the developed system, a critical requirement for advanced applications of pattern recognition.

IV. Result Analysis and Discussion

The systematic evaluation of Logistic Regression and K-Nearest Neighbors across diverse dimensionality-reduced and full-dimensional feature spaces provides critical insights into optimal pattern recognition strategies for fake audio classification.

A. Comparative Performance Across Feature Manifolds

Table II summarizes the F1-Scores and AUC values for each classifier under different dimensionality reduction schemes. These metrics highlight the trade-off between precision, recall, and overall discriminative power.

TABLE II: Model Performance Summary (F1-Score and AUC)

Model	Data Representation	F1-Score	AUC
Logistic Regression	Original Scaled	0.907	0.965
	PCA (10 components)	0.819	0.901
	SVD (10 components)	0.819	0.901
	LDA (1 component)	0.875	0.941
K-Nearest Neighbors	Original Scaled	**0.993**	**0.999**
	PCA (10 components)	0.984	0.997
	SVD (10 components)	0.984	0.997
	LDA (1 component)	0.980	0.996

B. Key Findings and Impact of Dimensionality Reduction

a) K-Nearest Neighbors: Superiority in High-Dimensional Spaces.: When trained on the original, fully scaled feature set, KNN achieved near-perfect classification (F1-Score: **0.993**, AUC: **0.999**). This demonstrates that the intrinsic local structure and separability of REAL and FAKE samples are highly pronounced in the 27-dimensional space, which KNN effectively exploits.

b) Nuanced Effects of Dimensionality Reduction:: For KNN, PCA and SVD (10 components) preserved most discriminative information (F1-Score ≈ 0.984, AUC ≈ 0.997), while LDA with a single supervised projection also performed strongly (F1-Score 0.980, AUC 0.996). The slight drop in all DR cases suggests minor loss of subtle local structures. Logistic Regression, however, suffered more under PCA/SVD (F1-Score 0.819), as its linear boundaries struggled in variance-maximized spaces. LDA improved performance (F1-Score 0.875) by explicitly optimizing for class separability, confirming its utility for linear models.

c) Strategic Selection of Feature Manifold and Algorithm.: For applications prioritizing accuracy, KNN on the original feature space is superior. When efficiency or storage constraints are important, PCA/SVD-transformed data with KNN offers a strong trade-off. For linear models, LDA is preferable to unsupervised methods, as it provides a more discriminative feature manifold.

C. Explainability of Feature Interpretability

A conceptual SHAP summary plot (Figure 4b) provides interpretability into the learned patterns. Features related to fundamental frequency variations (meanfun, minfun, maxfun), spectral entropy (*sp.ent*), and statistical moments (sd, Q25, Q75, IQR) emerge as highly influential. These insights validate the engineered features and enhance trust in the recognition system by moving beyond black-box predictions toward interpretable, causally grounded explanations.

V. Conclusion and Future Directions

This study demonstrated a complete pipeline for classifying real versus fake audio using high-dimensional acoustic features. Through EDA, careful feature scaling, dimensionality reduction (PCA, SVD, LDA), and a comparative evaluation of Logistic Regression and K-Nearest Neighbors, we established that KNN on the full scaled feature set yielded the strongest performance, while LDA improved the linear model by optimizing for class separability. The use of SHAP further enhanced interpretability by revealing the most influential features.

As the project evolves, we plan to extend beyond these classical models by incorporating additional machine-learning algorithms and advanced nonlinear dimensionality-reduction methods such as t-SNE and UMAP for richer visualizations. We also intend to experiment with deep architectures (e.g., CNNs or Transformer on spectrograms or raw audio) and adversarial training to build more robust, explainable, and scalable detection systems against emerging deepfake techniques.