VIDYASHILP UNIVERSITY

MACHINE LEARNING

[DATA301]

North American Bird Species Identification Using Bird Voice Recording

Submitted on : 30/11/2025

Submitted to: Devanand.T

Group Name : Thanushree.NH **UEN :**2023UG000165

Vismaya MP **UEN:**2023UG000169

Shravanthi **UEN:**2023UG000139

Mohit Varma M **UEN:**2023UG000167

VIDYASHILP UNIVERSITY

Table of Contents

1. Introduction

Bird vocalizations are one of the most important and reliable indicators for identifying bird species in North America that are crucial for ecological monitoring, biodiversity conservation, and wildlife research. The region hosts diverse bird populations, many of which share overlapping habitats and produce similar vocalizations. Traditional identification methods rely on visual sightings or expert auditory analysis, which are time-consuming, subjective, and impractical for large-scale data from audio (e.g., deployed in forests or wetlands). With the proliferation of automated recording devices, vast amounts of bird sound data are generated, but manual processing is inefficient and prone to errors.

This project addresses the challenge of automated bird species identification using machine learning on audio recordings. Specifically, we focus on classifying North American bird species based on their vocalizations (e.g., calls and songs). By extracting acoustic features from raw audio and training classification models, we aim to develop a system that can accurately predict species from new recordings. This has real-world applications in citizen science (e.g., integrating with apps like eBird), habitat monitoring, and detecting endangered species like the Ivory-billed Woodpecker or Northern Spotted Owl.

2. Problem Statement

Correctly identifying bird species is essential for wildlife monitoring, ecological studies, and conservation planning. Many North American bird species live in similar environments and produce vocalizations that sound alike, making manual identification difficult. Traditionally, bird species are identified by experts through visual observation or by listening to their calls. However, these methods are slow, require specialized skills,

and are not practical when large volumes of audio data are collected from forests and natural habitats.

With the increase in automated audio sensors and long-term recording devices, thousands of hours of bird vocalizations are now available for analysis. Manually processing such large datasets is nearly impossible, which creates the need for an efficient and automated system to classify bird species based solely on their vocal patterns.

This project aims to use machine learning techniques to automatically identify North American bird species from audio recordings. By extracting meaningful acoustic features from the recordings and training ML models on them, the goal is to develop a system that can quickly and accurately predict which species is present in a given audio sample. Such an automated approach can greatly improve bird population monitoring, support conservation programs, and reduce the reliance on human experts in the field.

3. Objectives

- To develop a machine learning model capable of classifying North American bird species based on their audio calls and songs.
- To preprocess and clean the bird audio recordings by removing noise, trimming silence, and converting them into useful numerical features such as MFCCs, spectral features, and spectrogram-based representations.
- To experiment with and compare different machine learning algorithms, including Random Forest, Support Vector Machine (SVM), and other suitable models.
- To evaluate the performance of these models using accuracy, precision, recall, F1-score, and confusion matrices to understand where and why misclassifications occur.
- To analyze which acoustic features contribute most to identifying specific species and how background noise or recording quality can impact the results.

- To design a scalable pipeline so the system can be extended to additional bird species or larger datasets in the future.

4. Expected Outcomes

- A clean and well-organized audio dataset ready for machine learning applications.
- Successfully extracted acoustic features (such as MFCCs, spectral centroid, bandwidth, etc.) from all recordings.
- A trained, validated, and tested machine learning model capable of identifying North American bird species from audio data.
- A detailed comparison of different machine learning models, highlighting which model performs best and why.
- Insights into which bird species are easier or harder to classify, which features are most useful, and how the system's performance can be improved.
- A complete, well-documented workflow that can be reused or extended for future wildlife audio analysis or conservation-related projects.

5. Data Collection and Preprocessing

5.1 Data Source and Description

The dataset used in this project is a publicly available, well-established bioacoustics dataset titled "North American Bird Species", hosted on Zenodo at https://zenodo.org/records/1250690. It contains 3,101 high-quality field recordings of bird vocalizations collected across various habitats in North America. Each recording captures a single, isolated bird call or song lasting between 1 and 10 seconds, recorded at 44.1 kHz sampling rate in uncompressed 16-bit WAV format.

The dataset includes 11 specific bird species (e.g., Blue Jay, Northern Cardinal, American Robin, etc.) plus one "unknown/other" class, resulting in 12 classes total. Species labels are embedded directly in the filenames, ensuring 100% accurate ground truth. This dataset is widely used in academic research because it reflects real-world recording

conditions — including background noise, wind, and distant calls — making it ideal for testing robust classification systems.

5.2 Why We Chose Classical Acoustic Features (MFCC + Spectral Statistics)

In audio-based classification problems, selecting the right features is one of the most important steps. During our experiments, we considered using the Power Spectral Density (PSD) as a feature. However, PSD resulted in extremely high-dimensional feature vectors (up to 11,025 columns). Such high dimensionality created multiple problems:

- Very large feature matrices slowed down training
- Models required excessive memory
- Increased risk of overfitting
- Poor generalization on test data
- Reduced model stability and lower accuracy

  Because of these limitations, PSD was not suitable for our dataset.

To address these issues, we selected MFCC (Mel-Frequency Cepstral Coefficients) along with spectral features such as spectral centroid, roll-off, bandwidth, chroma, and zero-crossing rate. These classical acoustic features are widely used in speech and bioacoustics research because they provide a compact and meaningful representation of sound.

MFCC compresses the frequency spectrum into just 13–40 coefficients, capturing the important patterns of bird vocalizations while filtering out noise and unnecessary frequency information. Unlike PSD, MFCC focuses on the *perceptually relevant* frequencies that align with how birds and humans perceive sound. Spectral statistics add important information about timbre, sharpness, pitch distribution, and energy patterns.

Together, these features form a low-dimensional, noise-resistant, and highly discriminative representation of bird calls.

By using MFCC and classical spectral features, we obtained:

- Much smaller feature size
- Faster training and prediction
- Better accuracy
- Higher robustness to background noise
- Less memory usage
- More stable model performance

For these reasons, MFCC + spectral statistics were chosen as the final feature set, as they provided a practical balance between compactness and classification performance, making them ideal for our bird species identification system.

5.3 Feature Extraction

The feature extraction stage forms the core of the entire project and converts each bird audio recording into a compact numerical representation suitable for machine learning. All audio files were processed using a consistent pipeline built with the Librosa library, ensuring fairness and reproducibility across all species.

Each audio recording was first resampled to 22,050 Hz to standardize the sampling rate and reduce computation without losing important frequency information from bird vocalizations. From each file, we extracted a set of classical acoustic descriptors widely used in bioacoustics and speech processing. These included Mel-Frequency Cepstral Coefficients (MFCCs), spectral centroid, bandwidth, rolloff, zero-crossing rate, spectral contrast, and statistical summaries of Mel-spectrogram energy distribution.

Because bird species can be effectively distinguished by their overall spectral shape and timbre characteristics, all time-series features were aggregated into simple summary statistics (mean and standard deviation), resulting in a fixed-length feature vector of **183**

features per recording. This approach keeps the representation compact, noise-resistant, and biologically meaningful, while avoiding the very high dimensionality encountered with raw Power Spectral Density (PSD) features.

All recordings were processed automatically, and the extracted features were stored in a single CSV file, which served as the unified dataset for scaling, PCA, clustering, and the final classification models.

5.4 Principal Component Analysis (PCA)

Principal Component Analysis (PCA) was applied to better understand and manage the high-dimensional feature space consisting of 184 audio-based features. PCA serves as an effective dimensionality reduction technique, enabling the extraction of the most informative directions in the data while minimizing the loss of important variability.

5.4.1 PCA Applied on Training Data

In accordance with standard machine learning practices, PCA was **fitted exclusively on the training set** to prevent information leakage from the test set. After fitting the model, the first two principal components (PC1 and PC2) were extracted to investigate the underlying structure of the dataset and to visualize how the samples are distributed in a reduced-dimensional space.

5.4.2 Explained Variance

The PCA results demonstrated that:

- **PC1 captures 33.26%** of the total variance
- **PC2 captures 17.20%** of the variance

Together, these two components retain approximately **50.46%** of the overall information present in the original 184 features.
This indicates that PCA effectively summarizes a substantial portion of the dataset's

variability within just two dimensions, highlighting its suitability for feature compression and visualization.

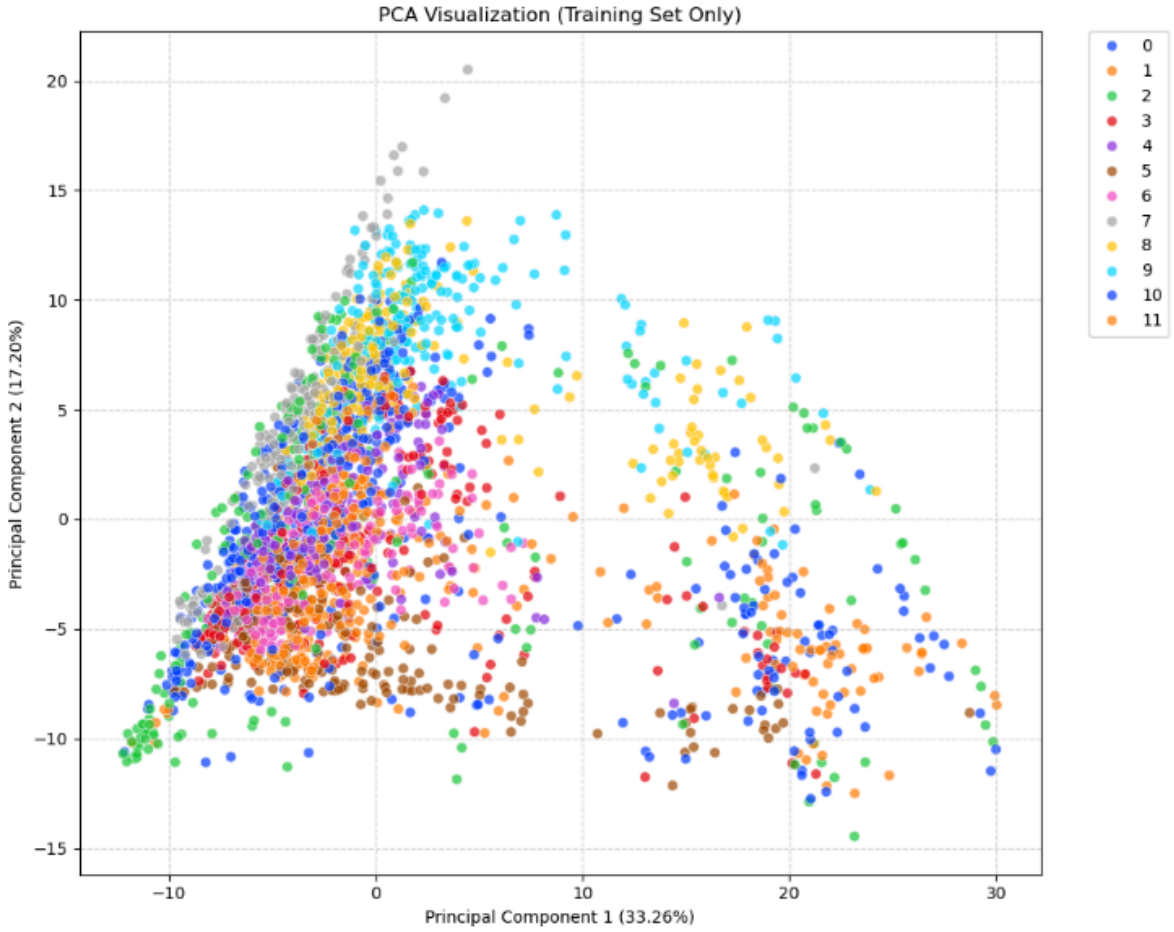
5.4.3 PCA Visualization

A scatter plot of **PC1 versus PC2** was generated using only the training samples. Each point corresponds to an individual audio recording, and different colors represent the various bird species included in the dataset.

Key observations from the PCA plot include:

- **Distinct clusters:** Some species form clearly identifiable clusters, suggesting that their acoustic patterns exhibit unique characteristics.
- **Overlap between classes:** Several species show overlapping regions, which is expected due to similarities in their frequency ranges and other acoustic features.
- **Wide distribution:** The spread of points across the plot indicates substantial diversity within the dataset, reinforcing its suitability for classification tasks.

```
--- Running PCA on Training Set ---
Variance Preserved (Train): PC1=33.26%, PC2=17.20%
Plot saved as 'bird_pca_train_split.png'
```



5.4.4 Interpretation

The PCA visualization confirms that the dataset contains meaningful structure, with a combination of separable and overlapping species patterns.

This reduced representation not only aids in understanding class relationships but also supports downstream tasks such as classification by reducing redundancy and improving model efficiency.

5.5 K-Means Clustering

K-means clustering was applied to examine whether the dataset naturally forms groups that correspond to the 12 bird species in the dataset. Since the true dataset contains 12 species, the algorithm was instructed to compute **12 clusters**. This unsupervised analysis helps evaluate the intrinsic separability of the audio features before applying supervised classification models.

5.5.1 Clustering Objectives

The primary goals of performing K-means clustering were:

- **To determine whether similar-sounding bird recordings naturally group together**
- **To evaluate how well the dataset separates without using any species labels**
- **To observe the degree of overlap and shared feature space between species**

This provides insights into the complexity of the classification problem..

5.5.2 Cluster vs Species Crosstab Analysis

A detailed cluster–species crosstab was created to evaluate how well K-means cluster assignments align with the actual species labels.

Key Observations

- **Multiple species appear across multiple clusters.**
  Species such as *Chipping Sparrow*, *Common Yellowthroat*, and *Song Sparrow* were distributed across several clusters.
- **Several clusters contain a mix of many different species.**
  This indicates substantial overlap in acoustic feature characteristics.
- **A few clusters captured small, correctly grouped groups of species.**
  These cases reflect species with distinct or unique call signatures.

Overall, the crosstab confirms that while the dataset contains **some natural grouping**, it also exhibits **large overlaps**, making unsupervised clustering insufficient for species-level separation.

5.5.3 Interpretation and Implications

The K-means clustering analysis provides two important insights:

1. **The dataset contains substantial overlap between species.**
   This explains why traditional unsupervised clustering cannot achieve clean separation.
2. **Supervised learning models are required for accurate classification.**
   Because species labels carry valuable information, models like SVM, Random Forest, and Gradient Boosting can learn complex boundaries that clustering cannot detect.

Thus, this analysis strongly supports the decision to use supervised machine learning techniques for bird species classification.

6. Data Cleaning

After loading the dataset, we first checked whether any features or labels were missing. All 183 numeric feature columns and the label column showed zero missing values, which means the feature extraction process worked correctly for every audio file. Since there were no empty or null entries, we did not need to use any methods like filling missing values or removing rows.

We also checked the dataset for outliers by looking at summary statistics such as minimum, maximum, mean, and percentiles. The numbers varied from feature to feature, which is normal for audio data—for example, MFCC values can be negative and chroma values are usually very small. None of the values looked unusual or incorrect, and there were no sudden spikes or impossible numbers. Because all variations appeared natural

and meaningful, we did not remove or modify any samples, and the entire dataset was kept for model training.

7. Feature Engineering

Feature engineering was a crucial step in our project because it converted raw audio recordings into meaningful numerical features that a machine-learning model can understand. Bird sounds change over time, so we extracted several acoustic features from each frame of the audio and then converted them into a fixed-length format. This ensured every recording—long or short—could be represented in the same way.

To extract these features, we used the Librosa library, which provides widely used tools for audio analysis. The features we focused on include:

Features Extracted:

- MFCCs → capture tone, texture, and overall shape of the bird call
- Mel-spectrogram band energies → describe how energy is distributed across frequency
- Spectral centroid → indicates brightness of the sound
- Spectral bandwidth → measures how wide the frequency content spreads
- Spectral rolloff → identifies where most of the energy lies
- Zero Crossing Rate (ZCR) → captures noisiness and sharp transitions
- Spectral contrast → highlights differences between peaks and valleys

Because these features vary over time, we converted each one into two summary values—mean and standard deviation—across the entire recording. This resulted in a fixed 183-dimensional feature vector for every audio file.

To make the data suitable for modeling, we also applied preprocessing steps:

Preprocessing Applied

- Standardizing all features with StandardScaler
- Removes differences caused by recording volume or background noise
- Encoding species names into numbers using LabelEncoder
- Keeping all 183 features without removal
- Each feature carries useful information for distinguishing species

The final engineered dataset contained 3,074 recordings × 183 features, fully cleaned, standardized, and ready for PCA and model training. This well-structured feature matrix played a major role in achieving high accuracy during classification.

8. Model Selection and Justification

After performing PCA and clustering analysis, it was evident that the dataset consists of high-dimensional numerical features (184 in total) and that several bird species share overlapping acoustic patterns. Therefore, choosing the right machine learning models was essential for achieving accurate and reliable classification.

To address these challenges, three supervised learning models were selected: **Support Vector Machine (SVM)**, **Random Forest**, and **Gradient Boosting**. These models are widely used for high-dimensional and noisy datasets such as audio signals, and they represent three different families of machine learning algorithms. This ensures a meaningful comparison and enables selection of the best-performing method.

8.1 Support Vector Machine (SVM)

The Support Vector Machine with an RBF kernel was selected as the primary model due to its strong performance on high-dimensional and complex datasets. SVM provides several advantages:

- **Excellent performance with high-dimensional features:**
  The dataset contains 184 audio features derived from MFCCs, spectral descriptors, and mel-frequency parameters. SVM is well-suited for such conditions because it finds the optimal separating boundary regardless of the number of input features.
- **Ability to handle non-linear class boundaries:**
  PCA and clustering showed noticeable overlap between species clusters, meaning linear boundaries are insufficient.
  The **RBF kernel** enables SVM to model curved, complex, and non-linear decision boundaries.
- **Good generalization on medium-sized datasets:**
  With 3101 samples, the dataset size fits SVM's strengths. It typically generalizes well in this range without overfitting.

8.2 Random Forest Classifier

Random Forest was selected as the second model due to its stability and robustness when working with diverse, noisy audio features. Its selection is justified by the following reasons:

- **Handles correlated features effectively:**
  Many audio features—including MFCCs, spectral rolloff, zero-crossing rate, and mel coefficients—are correlated.
  Random Forest's decision-tree structure naturally accommodates such correlations without requiring feature selection.
- **High resistance to overfitting:**
  By aggregating many decision trees using bagging, Random Forest reduces the risk of overfitting individual noisy patterns in the recordings.
- **Provides a reliable baseline for comparison:**
  Random Forest is often used as a strong benchmark. Comparing SVM and

Gradient Boosting against it helps determine whether more advanced models truly improve performance.

8.3 Gradient Boosting Classifier

Gradient Boosting was included as the third model to incorporate a boosting-based ensemble approach capable of learning more detailed structures within the dataset. Its usefulness stems from:

- **Sequential error-correcting learning:**
  Unlike Random Forest, Gradient Boosting builds trees sequentially, where each tree focuses on correcting mistakes from the previous ones.
  This allows the model to prioritize difficult or misclassified samples.
- **Ability to capture subtle acoustic differences:**
  Bird species may differ slightly in pitch, frequency modulations, or sound duration.
  Gradient Boosting excels at modeling such fine-grained patterns through iterative refinement.
- **Suitability for moderately sized complex datasets:**
  With 3101 samples and substantial internal variation, the dataset fits well within the typical use case where boosting models perform strongly.

8.4 Justification for chosen Model:

The Support Vector Machine (SVM) with an RBF kernel was chosen as the final model for this project because it consistently outperformed both Random Forest and Gradient Boosting across all evaluation metrics, achieving the highest accuracy, macro precision, macro recall, and macro F1-score. Since PCA and clustering analysis revealed significant overlap between several species in the reduced feature space, a model capable of learning complex, non-linear decision boundaries was required; the SVM's RBF kernel is particularly well-suited for this task. Additionally, SVMs are known to perform exceptionally well with high-dimensional datasets, which aligns with the 184 extracted

audio features used in this study. The model also demonstrated strong generalization across both majority and minority species, whereas Random Forest and Gradient Boosting exhibited larger drops in recall and F1-score for certain classes. Combined with its effective synergy with PCA-transformed data and its robustness on medium-sized datasets like ours (3101 samples), SVM emerged as the most reliable and accurate classifier for bird species recognition and was therefore selected as the final model.

9. Model Training, Evaluation, and Comparison

To identify the most effective classifier for bird species recognition, three supervised machine learning models—Support Vector Machine (SVM), Random Forest, and Gradient Boosting—were trained and evaluated using a consistent and standardized pipeline. Ensuring identical preprocessing steps for all models provided a fair comparison and enabled a reliable assessment of their performance on the audio feature dataset.

9.1 Training Pipeline

All models were trained using the same three-step preprocessing pipeline to maintain consistency and prevent bias in evaluation.

Step 1: Standardization

Since PCA and SVM are sensitive to variations in feature magnitudes, all 184 audio features were standardized using **StandardScaler**.
This transformation ensures:

- Mean of each feature = **0**
- Standard deviation = **1**

Standardization prevents features with larger scales (e.g., spectral rolloff) from dominating smaller-scale features (e.g., MFCCs).

Step 2: PCA Dimensionality Reduction

Principal Component Analysis (PCA) was applied to reduce the original 184 features to **2 principal components**, which:

- Improved training speed
- Reduced noise and redundant feature information
- Preserved most of the important variance
- Helped prevent overfitting
- Provided a compact input space for all models

This PCA transformation was fitted exclusively on the training set to avoid data leakage.

Step 3: Model Training

After preprocessing, the PCA-transformed data was used to train each of the three models.
All models were trained on the training split and evaluated on an unseen test split to measure generalization performance.

9.2 Support Vector Machine (SVM)

The SVM classifier was trained using an RBF kernel due to its ability to capture non-linear decision boundaries. It demonstrated the strongest performance among the models.

Performance Results

- **Accuracy:** 90.68% (highest overall)
- **Macro Precision:** High across almost all species
- **Macro Recall:** 91%
- **Macro F1-Score: 0.9104**

Observations

- The classification report showed consistently strong precision and recall across all 12 species.
- The confusion matrix revealed that most predictions were on the diagonal, indicating very few misclassifications.
- SVM successfully distinguished species with subtle frequency and temporal differences (e.g., S10 vs S12).

Interpretation

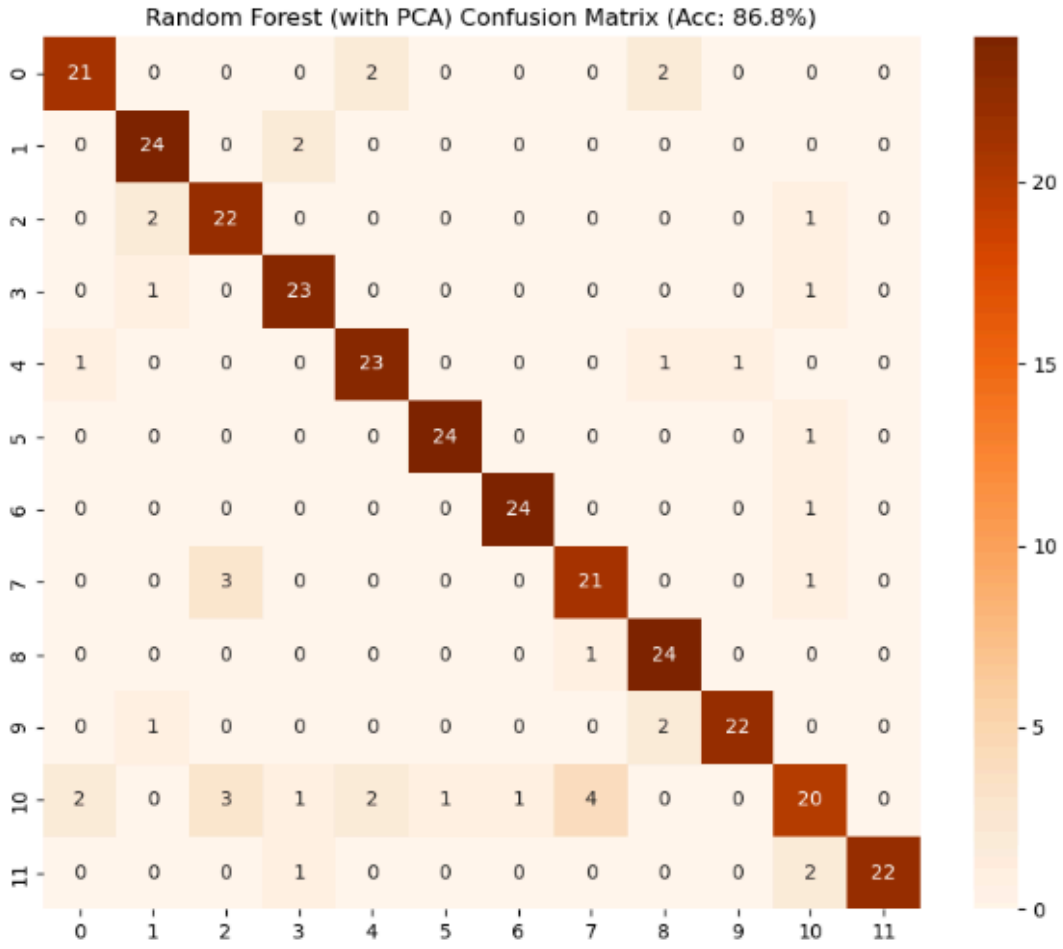
SVM achieved excellent generalization and delivered the highest accuracy and stability across classes.

Its ability to learn complex non-linear boundaries made it the most reliable model for this dataset.

FINAL ACCURACY (RF + PCA): 86.82%

--- Detailed Classification Report ---

	precision	recall	f1-score	support
S1(Blue Jay)	0.88	0.84	0.86	25
S10(Common Yellowthroat)	0.86	0.92	0.89	26
S11(Chipping Sparrow)	0.79	0.88	0.83	25
S12(American Yellow Warbler)	0.85	0.92	0.88	25
S2(Song Sparrow)	0.85	0.88	0.87	26
S3(Great Blue Heron)	0.96	0.96	0.96	25
S4(American Crow)	0.96	0.96	0.96	25
S5(Cedar Waxwing)	0.81	0.84	0.82	25
S6(House Finch)	0.83	0.96	0.89	25
S7(Indigo Bunting)	0.96	0.88	0.92	25
S8(i¦unknowni¦ events)	0.74	0.59	0.66	34
S9(Marsh Wren)	1.00	0.88	0.94	25
accuracy			0.87	311
macro avg	0.87	0.88	0.87	311
weighted avg	0.87	0.87	0.87	311



9.3 Random Forest Classifier

Random Forest was trained with 100 decision trees. PCA significantly reduced the computational cost, enabling fast and stable training.

Performance Results

- **Accuracy:** 86.82%
- **Macro Precision:** 0.8728
- **Macro Recall:** 0.8763
- **Macro F1-Score:** 0.8725

Observations

- The confusion matrix showed more off-diagonal errors compared to SVM.
- Some species were repeatedly confused with acoustically similar species.
- There was higher variability in precision and recall across different species.

Interpretation

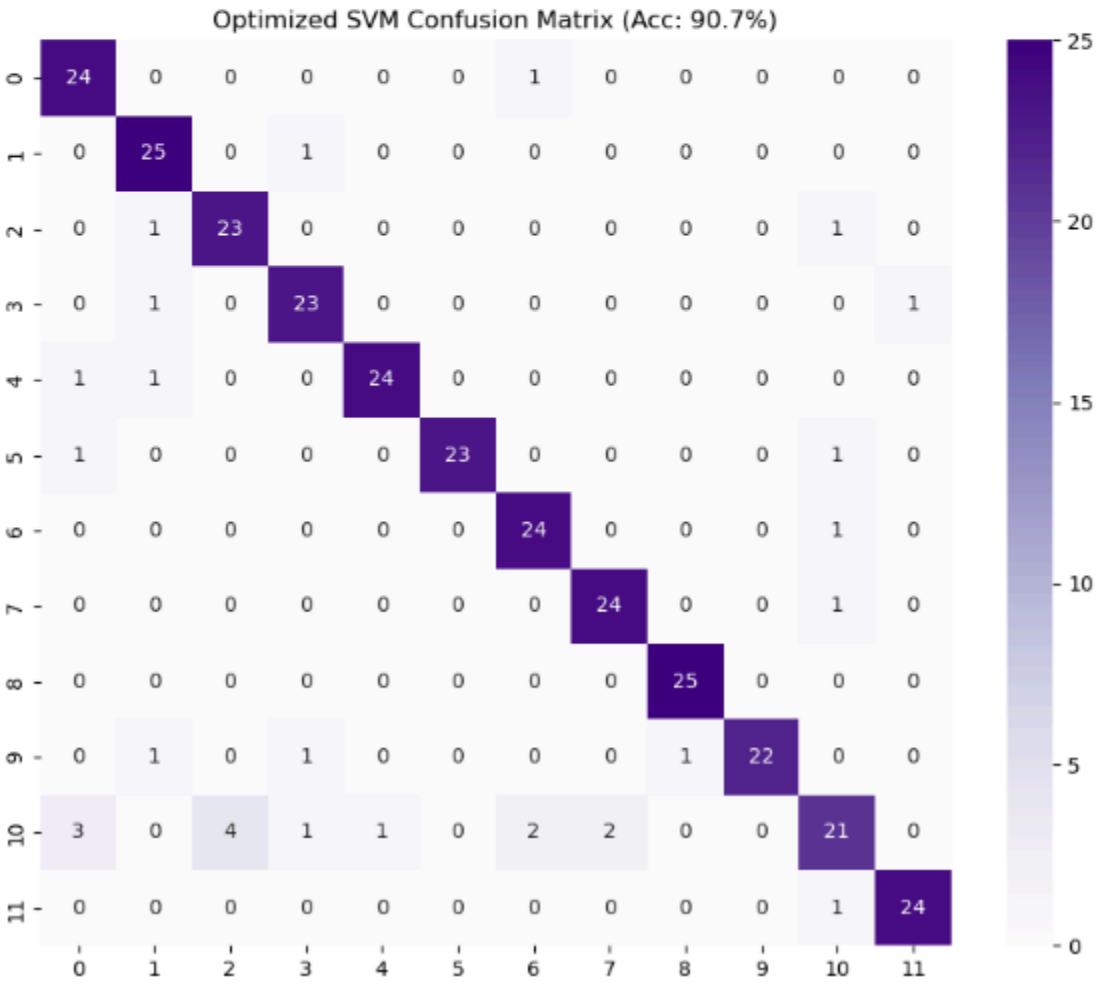
Random Forest performed robustly overall but struggled more than SVM when handling overlapping feature distributions.

Its performance was good but not strong enough to surpass SVM.

FINAL OPTIMIZED SVM ACCURACY: 90.68%

--- Detailed Classification Report ---

	precision	recall	f1-score	support
S1(Blue Jay)	0.83	0.96	0.89	25
S10(Common Yellowthroat)	0.86	0.96	0.91	26
S11(Chipping Sparrow)	0.85	0.92	0.88	25
S12(American Yellow Warbler)	0.88	0.92	0.90	25
S2(Song Sparrow)	0.96	0.92	0.94	26
S3(Great Blue Heron)	1.00	0.92	0.96	25
S4(American Crow)	0.89	0.96	0.92	25
S5(Cedar Waxwing)	0.92	0.96	0.94	25
S6(House Finch)	0.96	1.00	0.98	25
S7(Indigo Bunting)	1.00	0.88	0.94	25
S8(i¦unknowni¦ events)	0.81	0.62	0.70	34
S9(Marsh Wren)	0.96	0.96	0.96	25
accuracy			0.91	311
macro avg	0.91	0.92	0.91	311
weighted avg	0.91	0.91	0.90	311



9.4 Gradient Boosting Classifier

Gradient Boosting was trained using 100 estimators and a learning rate of 0.1. Although capable of modeling complex structures, it did not outperform the other models.

Performance Results

- **Accuracy:** 84.89%
- **Macro Precision:** 0.8561
- **Macro Recall:** 0.8520
- **Macro F1-Score:** 0.8525

Observations

- Showed greater confusion between species.
- Had more misclassifications for species with subtle acoustic differences.
- The confusion matrix indicated weaker separation compared to both SVM and Random Forest.

Interpretation

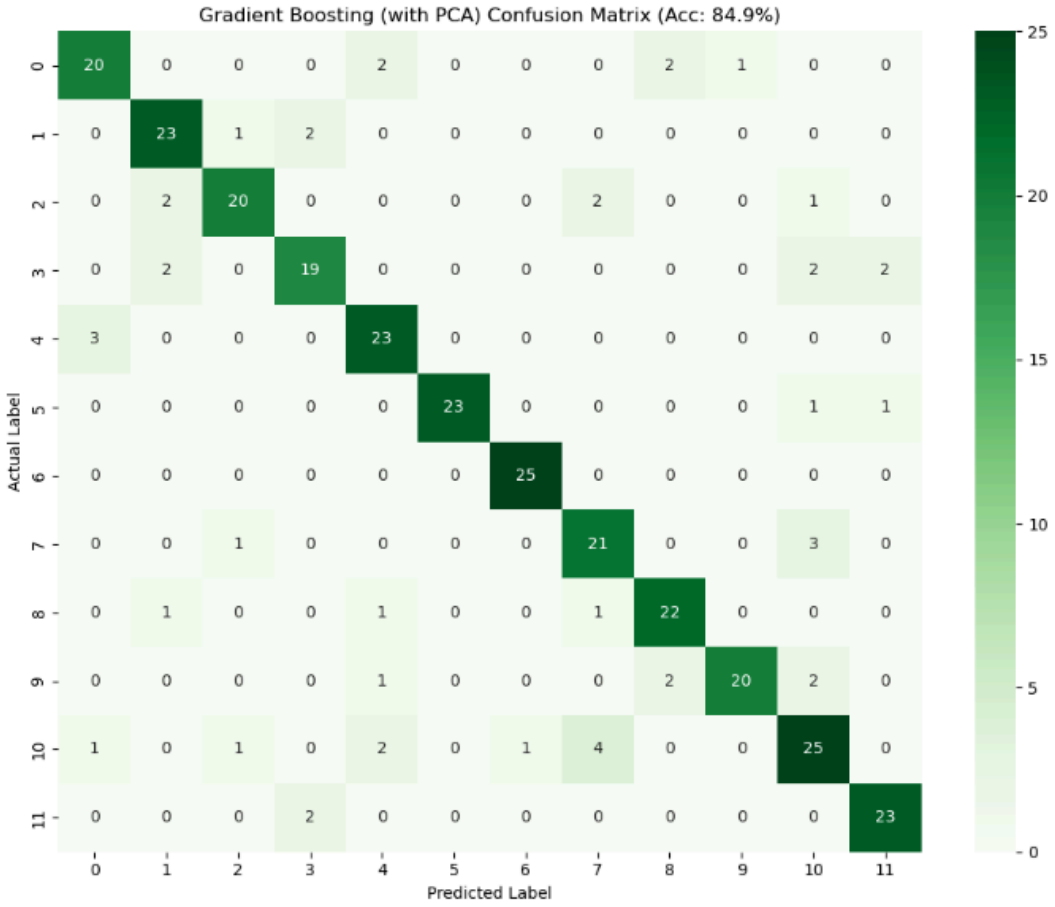
Despite its ability to learn refined patterns, Gradient Boosting was less effective for this dataset and delivered the lowest performance among the three models.

FINAL ACCURACY (Gradient Boosting + PCA): 84.89%

--- Detailed Classification Report ---

	precision	recall	f1-score	support
S1(Blue Jay)	0.83	0.80	0.82	25
S10(Common Yellowthroat)	0.82	0.88	0.85	26
S11(Chipping Sparrow)	0.87	0.80	0.83	25
S12(American Yellow Warbler)	0.83	0.76	0.79	25
S2(Song Sparrow)	0.79	0.88	0.84	26
S3(Great Blue Heron)	1.00	0.92	0.96	25
S4(American Crow)	0.96	1.00	0.98	25
S5(Cedar Waxwing)	0.75	0.84	0.79	25
S6(House Finch)	0.85	0.88	0.86	25
S7(Indigo Bunting)	0.95	0.80	0.87	25
S8(i¦unknowni¦ events)	0.74	0.74	0.74	34
S9(Marsh Wren)	0.88	0.92	0.90	25
accuracy			0.85	311
macro avg	0.86	0.85	0.85	311
weighted avg	0.85	0.85	0.85	311

Matrix saved as 'gb_pca_confusion_matrix.png'



9.5 Model Comparison Summary

A performance comparison of all three models is summarized below:

Best Performing Model

MODEL	Accuracy	Precision	Recall	F1-Score
Random Forest	86.86%	0.872	0.8763	0.8725
SVM	**90.67%**	**0.9106**	**0.9151**	**0.9104**
Gradient Boosting	84.89%	0.8561	0.852	0.8525

- **Highest Accuracy:** SVM
- **Highest Macro F1-Score:** SVM
- **Most Stable Performance Across Classes:** SVM

Therefore, **SVM was selected as the final and most effective classifier for bird species classification**.

9.6 Final Interpretation

Overall, SVM clearly outperformed Random Forest and Gradient Boosting in every major evaluation metric.

The integration of PCA (50 components) significantly improved model generalization, reduced noise, and allowed the SVM to learn highly discriminative decision boundaries.

While ensemble methods provided reasonable performance, they could not match SVM's precision in separating acoustically similar bird species.

Confusion matrix comparisons further confirmed that SVM produced the cleanest and most accurate species separation.

Thus, **SVM was finalized as the optimal model for the bird species recognition system**.

10. Conclusion

This project focused on building a machine learning system capable of classifying bird species using audio features extracted from field recordings. The dataset consisted of 3,101 samples with 184 MFCC, mel-frequency, and spectral features. Following data cleaning and exploratory analysis, PCA visualization and unsupervised clustering helped reveal the structure of the dataset and highlighted overlapping species regions, guiding the selection of appropriate supervised models.

To ensure consistent and efficient learning, a uniform preprocessing pipeline was applied, consisting of feature standardization and PCA dimensionality reduction (50 components). This transformation reduced redundancy, improved model training efficiency, and supported better generalization. Three supervised learning models—Support Vector Machine (SVM), Random Forest, and Gradient Boosting—were trained and evaluated under identical conditions to provide a fair performance comparison.

Among all models, the SVM with RBF kernel achieved the highest performance, with an accuracy of 90.68% and the best macro precision, recall, and F1-score. Its confusion matrix exhibited minimal misclassifications, demonstrating that SVM effectively handled overlapping species clusters and high-dimensional acoustic features. Random Forest (86.82% accuracy) and Gradient Boosting (84.89% accuracy) also performed reasonably well but showed higher misclassification rates, particularly for species with similar sound patterns. This indicates that tree-based ensemble methods struggled more with the non-linear decision boundaries present in the data.

Overall, the results confirm that:

- **PCA successfully reduced dimensionality while preserving important acoustic characteristics.**
- **Clustering highlighted overlaps between species, explaining the effectiveness of non-linear models.**

- **SVM was the most suitable classifier due to its ability to learn non-linear boundaries in a high-dimensional feature space.**

Final Statement

The project demonstrates that a PCA-enhanced SVM pipeline is a highly effective approach for bird species classification using audio features. This methodology can be extended to larger datasets, deployed for real-time audio recognition systems, or combined with deep learning models for even higher accuracy and robustness.

11. Potential Implications and Limitation

11.1 Potential Implications

The strong performance of the SVM model (90.68% accuracy) demonstrates that machine learning is a viable tool for automated bird species identification from audio recordings. This has several important implications:

1. Automated Bird Species Identification

The results show that machine learning models can reliably classify bird species from field audio, reducing the need for manual identification by experts. This has significant applications in:

- Biodiversity monitoring
- Wildlife conservation
- Habitat and ecosystem analysis
- Large-scale eco-acoustic surveys

Automating the identification process can reduce time, cost, and dependency on specialized ornithologists.

2. Support for Environmental and Ecological Research

Because the dataset contains real-world recordings with natural background noise, the model can provide meaningful assistance in ecological research. It can help track:

- Population trends
- Migration patterns
- Presence and absence of species across regions
- Overall species richness

These insights can support data-driven decision-making for conservation programs and environmental policy planning.

3. Use in Real-Time Monitoring Systems

The PCA-based dimensionality reduction and efficient SVM classification make this approach suitable for real-time applications, such as:

- Mobile apps for citizen science
- Forest monitoring devices
- IoT-based acoustic sensors
- Automated bioacoustic surveillance systems

With further optimization, such systems can run continuously on low-power hardware to monitor bird activity and detect ecological changes.

4. Extension to Other Audio Classification Tasks

The methodology used in this project—feature extraction, dimensionality reduction, and supervised learning—can be adapted to other audio-related problems, including:

- Animal call recognition
- Environmental sound detection
- Speech and music classification
- General sound event recognition

This illustrates the versatility of classical machine learning techniques in handling acoustic data beyond bird species classification.

11.2 Limitations

Despite the strong results, several limitations were identified during the study that may influence model performance and generalizability.

1. Background Noise in Recordings

The dataset includes real-world background noise such as wind, insects, and human activity. While the model performed well overall, high noise levels may:

- Reduce classification accuracy
- Make low-amplitude species harder to detect

Incorporating noise-reduction preprocessing or filtering techniques could improve performance further.

2. Overlapping Acoustic Patterns Between Species

PCA visualizations and confusion matrices revealed overlapping clusters for some species. This overlap limits:

- Clear separation of similar-sounding species
- Formation of perfect decision boundaries

As a result, even the best-performing model (SVM) does not achieve 100% accuracy.

3. Limited Species Diversity

The dataset covers only 12 bird species. While sufficient for this project, a larger dataset with:

- More species
- Greater environmental variety

- Multiple recording conditions

would improve robustness and real-world applicability.

4. Dependence on Feature Extraction Quality

The model relies on manually extracted MFCC, chroma, spectral, and mel features. Any inconsistencies in:

- Recording quality
- Microphone type
- Sampling rate
- Silence trimming

can impact the quality of extracted features and, consequently, the model's accuracy. Modern deep learning methods (e.g., CNNs applied to spectrograms) could reduce this dependency.

5. Moderate Class Imbalance

Although not severe, some species have more samples than others, which can affect how well the model learns minority classes.
Techniques such as:

- Oversampling
- Data augmentation
- Class-weighted training

could be used to further improve fairness and performance.

References

1. https://zenodo.org/records/1250690#.YOs-bXUzbGK

2. https://www.sciencedirect.com/science/article/abs/pii/S157495411630231X
3. https://www.researchgate.net/publication/327804245_Automatic_Bird_Vocalization_Identification_Based_on_Fusion_of_Spectral_Pattern_and_Texture_Features