Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

G²HAN: Geometry-Guided Hierarchical AttentionNetwork for Insect Sound Classification

Suyash Patil

rishi.suyash01@gmail.com

Indian Institute of Information Technology, Bhopal

Rekha Kaushik

Indian Institute of Information Technology, Bhopal

Lalit Kumar

Manipal University Jaipur

Article

Keywords:

Posted Date: May 7th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9617709/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

G²HAN: Geometry-Guided Hierarchical Attention Network for Insect Sound Classification

Suyash P. Patil[1,*], **Rekha Kaushik**[2], **and Lalit Kumar**[3]

[1]Indian Institute of Information Technology Bhopal, Deptartment of Information Technology, Bhopal, 462003, India
[2]Indian Institute of Information Technology Bhopal, Deptartment of Information Technology, Bhopal, 462003, India
[3]Manipal University Jaipur, Department of Artificial Intelligence and Machine Learning, Jaipur, 303007, India
[*]rishi.suyash01@gmail.com

ABSTRACT

Insects produce sounds through various mechanisms and organs ranging from the vibrations of wingbeats to the sounds produced when they are feeding. These acoustic signals can be used to extract various features. This method has proven to be one of the most non invasive as well as sustainable approaches to insect classification. Traditional methods involve extracting only one kind of features and then using Machine Learning or Deep Learning methods to classify insects. While this method can be a great option for a very limited taxa of insects, it becomes inefficient while training a dataset consisting of many species and subspecies. We propose a framework that aims to solve this problem by exploring the highly nuanced relationships that these features possess by learning the feature manifold and attempting to capture relational structure among numerous features. These relationships are determined by using Graph Neural Networks(GNNs). Experiments showed that these relationships exist and they also reveal a geometrical structure. Additionally, this work also introduces a multi-modal classification framework that combines these extracted geometrical relationships in the form of embeddings as well as spectrogram features using CNN. Experimental results show that the addition of geometrical features substantially contribute to the classification performance by achieving above 97% accuracy and improving the baseline performance by over 6%.

Introduction

Insects have long been an essential part of global ecosystems, some serving as essential pollinators and nutrient cyclers, while others serve as pests that cause agricultural degradation or carriers of human diseases[1,2]. Monitoring their populations, detecting hidden insects is crucial for biodiversity conservation. Although traditional techniques such as trapping and visual inspection are laborious, invasive, and prone to sampling bias[1]. A promising alternative is to harness the unique acoustic signals[19] produced by insects, which offers a cost-effective and non-invasive way to continuously monitor insect populations across a wide range of environments[1,2]. Vocalization is a deeply rooted evolutionary trait in insects[3]. Unlike mammals or birds, insects produce sounds through highly specialized, physically distinct anatomical mechanisms, such as the scraping of stridulatory files with varying tooth densities or the scraping of abdominal tymbals[3,4]. These mechanisms produce sounds that are completely different from human speech. They are non-harmonic, highly broadband, and reach far into the ultrasonic frequencies[5]. Using these insect calls, researchers have been able to develop methods to automate the classification process and apply these methods to various use cases, such as detecting insects in stored grain, tracking mosquito presence, and assessing and monitoring biodiversity conservation and habitat restoration efforts[1,6,7].

Automated classification of insect calls has progressed beyond manual feature engineering and statistical models to highly advanced machine learning (ML) and deep learning (DL) techniques[1,2]. These techniques utilize unsupervised machine learning algorithms such as support vector machines (SVM), random forests (RF), and probabilistic neural networks (PNNs) along with mel-frequency cepstral coefficients (MFCCs) and are used in a variety of use cases, such as identifying storage pests, monitoring the health of beehives, or simply classifying singing insects. citeZhu:2011, Noda:2019, Libal:2024. Today, DL architectures, especially convolutional neural networks (CNNs) and their variations, are the gold standard in this field. This is due to their superior ability to extract hierarchical features from highly unstructured acoustic data and typically achieve classification accuracies exceeding 90% across many species.

Furthermore, the continued development of insect acoustic monitoring relies heavily on the availability of huge and diverse datasets for model pre-training. There have been some datasets like the InsectSound1000[14] and InsectSet459[16] which have contributed greatly towards this data shortage. These datasets area curation of thousands of recordings across a myriad of insect species. The sheer size of this data has been a very important resource to train deep learning models.

The purpose of these deep learning model is to identify and map a variety of features, some of the notable ones are time features, frequency features and spectral features. These features, once identified and mapped can contribute greatly towards a

model's classification performance[5,16].

This approach is biologically limiting. Insect vocal production is controlled by morphological and anatomical mechanisms, such as wing stridulation and tymbal vibration, which can produce acoustically similar signals despite being taxonomically distinct[3,4]. Therefore, acoustic similarity does not necessarily correspond to species identity. This raises a fundamental question: *Does insect acoustic data have intrinsic geometric structure beyond nominal class labels?*. In high-dimensional settings, data often lie on a structured manifold, where hidden organization reflects intrinsic generative processes. Techniques such as UMAP and t-SNE have shown that functional structure can be derived even without supervision. Despite this, the geometric approach remains poorly understood in insect acoustics, where optimizing classification methods is usually prioritized over structural understanding[1,16].

In this work, we systematically analyze the geometric organization of the insect acoustic feature space, which includes over 300 features, including temporal, spectral, cepstral, filter bank, and nonlinear descriptors. Our results show that this space is not randomly distributed, but rather exhibits a stable structure at different resolutions. These results suggest that species-level separation results in a refinement of a deep hierarchical organization within the acoustic feature space, which significantly contributes to classification performance while ensuring generalization across diverse datasets. The main objective of this work is to investigate the intrinsic organization of acoustic data, going beyond the application of deep learning methods to acoustic feature representation. Although existing methods allow for fairly high-precision classification, they provide limited understanding of relationships between species or internal feature space structure[10,16]. From a biological perspective, determining these relationships is necessary because similarities about acoustic features can reflect shared behavioral or morphological traits. From a computational perspective, identifying latent structures can lead to more robust, understandable, and general representations. Furthermore, traditional learning methods treat samples individually, ignoring the relational structure in feature space. By integrating geometric relationships through similarity graphs, it becomes possible to model local dependencies and preserve variation structure during learning[15]. This justifies a unified framework that integrates feature geometry with deep spectral representations[17,18]. The main contributions of this work are:

1. **Multimodal acoustic learning framework considering geometry:** We propose a unified framework combining networks of convolutional neurons for the extraction of spectral characteristics and networks of graphical neurons for relational modeling. This allows a learning process to integrate simultaneously the local acoustic motifs and the global relationships between the samples, thus going beyond the independent classification based on the samples.

2. **Geometry-aware learning improves representation and performance:** We introduce a function of taking into account the geometry and a fusion mechanism based on the attention that guarantees structural coherence in the representation space. This improves the organization of the groups, increases the robustness and systematically improves the classification performances for all the evaluation metrics.

Collectively, these contributions establish that insect acoustic data is inherently structured and hierarchical, and that exploiting this structure enables both improved learning and deeper insight into bioacoustic phenomena.

Related Work

Both early and recent bioacoustic identification systems have demonstrated excellent classification accuracy. However, these successes are largely limited to studies focusing on a small number of species, as described in Table 1. For example, studies focusing on "distinguishing between two classes of bees: female workers and male drones" report prediction accuracies exceeding 99% using random forest (RF) and k-nearest neighbors (KNN) methods[11]. Similarly, state-of-the-art convolutional neural networks, including 1D-ResNets designed to classify disease-carrying mosquitoes and fruit flies, rely heavily on datasets containing only 3 to 10 species[20]. Even studies utilizing specialized probabilistic neural networks (PNN) or dual-tower network architectures typically limit their taxonomic scope to datasets of 32 to 50 species[8,15]. However, as the taxonomic scope expands to reflect real-world biodiversity, the performance of these predictive models rapidly declines. As the dataset size increases, the acoustic features of individual species heavily crowd the traditional feature space. For example, when attempting to classify 343 insect species, researchers found "a greater dispersion of the results due to the increase of the spectral overlap among the frequencies of the insect syllables", which means classifiers such as "SVM has difficulties to find an optimal hyperplane to properly separate the different specie sounds"[10]. This scaling bottleneck is particularly evident in large-scale comparison studies, such as the InsectSet459 dataset, which includes 459 species[16]. When existing state-of-the-art models, including EfficientNet and Transformer-based algorithms (PaSST), were applied to this large task, overall classification performance dropped to an F1 score between 56% and 58%[16]. Although these algorithms maintained high accuracy for some common species, their performance declined sharply on the "long tail" of less frequently occurring taxa. As noted in the benchmark evaluation, "the most common categories in IS459 can be classified by either algorithm with a performance above 80%, but this rapidly decreases as the less-frequent categories are included", with the F1 score for data-poor categories observed "very low indeed,

Study	Year	Dataset (Species)	Approach	Features	Performance
Marius Faiß, et al.[16]	2026	InsectSet459	CNN (EfficientNetV2-S / InsectEffNet) and Transformer (PaSST)	Mel-spectrograms (128 frequency bands)	F1-score between 56% and 58%
Béla J. Szekeres, et al.[13]	2025	Wingbeats, Fruitflies, Insects, Abuzz	CF-ResNet-9-1D (Convolutional Fourier Layer + ResNet)	Raw time-domain signals (Discrete Fourier Neural Operator)	Wingbeats: 96.01% accuracy; Fruitflies: 98.00% accuracy; Insects: 85.24% accuracy.
Urszula et al.[11]	2024	Dataset for Honey Bee Audio Detection (Zenodo)	SVM, LDA, Random Forest (RF), and k-nearest neighbors (KNN)	Mel-frequency cepstral coefficients (MFCC) (5, 10, 15, 20, and 40 coefficients)	KNN: 99.68% accuracy; RF: 99.88% accuracy; SVM: 95.32% accuracy; LDA: 95.44% accuracy.
R. Kaushik, et al.[9]	2024	USDA ARS Insect Sound Library	KNN and Principal Component Analysis(PCA)	Mel-frequency cepstral coefficient and Hilbert Huang transform	98.8% Accuracy
Hangfei He, et al.[15]	2024	InsectSet32	Dual-Tower Network (EfficientNet-b7 + Dual-frequency and Spectral Fusion Module DFSM)	Logarithmic Mel spectrograms and SpecAugment	80.26% accuracy.
Jelto Branding, et al.[12]	2023	InsectSound1000 (preliminary version)	Spectrogram-based CNN vs. Multichannel Raw Audio Model (Neural Beamforming + WaveNet)	STFT (spectrograms) vs. Raw multichannel audio	Clean data: >85% accuracy; Greenhouse noise: 57% accuracy for multichannel model (single-channel model failed).
Jaehoon Kim, et al.[7]	2021	Data from eight mosquitoes and two flies (50,000 samples)	VGG13 and Simple CNN with Activity Circadian Rhythm (Bayes' Rule)	MFCC and Activity Circadian Rhythm information	VGG13 with 1D-ConvFilter: 85.7% accuracy.
Juan J. Noda, et al.[10]	2019	SINA and Insectsingers (343 species)	Knn	Fusion of MFCC and Linear Frequency Cepstral Coefficients (LFCC)	98.07% success rate using SVM + Fusion.
Zhu, et al.[8]	2011	USDA ARS Insect Sound Library	Probabilistic Neural Network (PNN)	Mel-frequency cepstrum coefficient (MFCC)	Recognition rate above 96%.

Table 1. Summary of Prior Work in Insect Classification using acoustic signals

below 20%"[16]. In contrast, this work shifts the focus from purely predictive modeling to structural analysis. We explicitly investigate the geometric organization of acoustic feature space and leverage this structure through a unified framework that integrates multi-scale clustering, graph-based representations, and attention-driven fusion. This approach enables the capture of both intrinsic geometry and spectral characteristics, providing a principled foundation for the methodology described in the following section.

Methodology

Overview of the Proposed Framework

This chapter expounds on the complete methodology covering both research objectives. As illustrated on Figure 1, this framework is composed of three main elements: (i) a branch of extraction of manually extracted acoustic characteristics, (ii) a branch of convolutional neural networks (CNN) based on the spectrogram and (iii) a branch of geometric modeling based

on graphs. These elements function in parallel and are interconnected by a mechanism of attention, allowing the model to simultaneously learn representations of complementary issues from multiple modalities.

Unlike conventional methods that rely only on manually extracted characteristics or on deep representations, the proposed approach combines the two by integrating the geometric relationships between the samples thanks to learning on the graphs. This results in a more expressive and contextual representation of the bioacoustic signal.

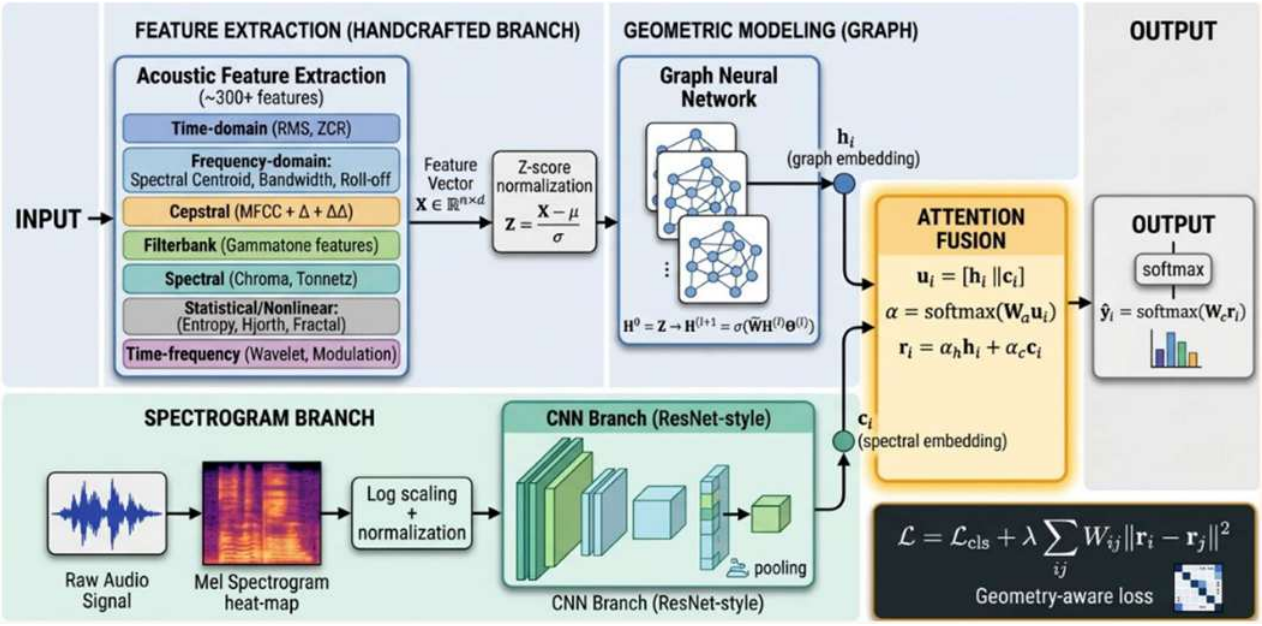


Figure 1. Overall geometry-aware multimodal acoustic learning framework.

Dataset Description

The proposed framework is evaluated on the InsectSound1000 dataset[14], which comprises over 169 000 labelled sound samples from twelve insect species. The recordings were made in an anechoic box using a four-channel low-noise measurement microphone array, with each sample stored as a four-channel wave file of 2500 ms length at a 16 kHz sampling rate and 32-bit resolution. The insect sound levels in the dataset span a wide dynamic range, from the loud buzzing of *Bombus terrestris* to sounds inaudible to human ears produced by *Aphidoletes aphidimyza*.

Acoustic Feature Extraction (Handcrafted Branch)

The handcrafted feature extraction branch, shown in Figure 2, calculates a large set of acoustic descriptors that capture multiple signal characteristics. These include time-domain features such as root mean square (RMS) energy and zero-crossing rate (ZCR), frequency-domain features such as spectral centroid, bandwidth, and roll-off, and cepstral features including MFCCs with their first- and second-order derivatives.

Furthermore, filterbank features (e.g., gammatone), spectral features (e.g., chroma and tone), and statistical or nonlinear descriptors such as entropy and Hjorth parameters are extracted. Wavelet-based features such as time-frequency representations further refine the feature space. These features are combined into a high-dimensional feature vector and subsequently normalized using z-score normalization to ensure consistent scaling across samples. This branch captures domain-specific acoustic characteristics that may not be fully represented by deep learning models alone.

Spectrogram-Based CNN Branch

The raw audio signal is converted into a Mel spectrogram that compresses the frequency axis to emulate human hearing. Logarithmic scaling and normalization are used to stabilize training and amplify small variations in spectral energy.

The spectrogram obtained is used as an input embedding to a ResNet style CNN, as illustrated on Figure 3. The remaining connections assure a stable gradient flux and allow the network to learn deeper representations without degradation. The first convolutional layers capture temporal and spectral local patterns, such as the harmonic structure and frequency transitions, while the deeper layers encode higher level abstractions based on the acoustic signature unique to each species.

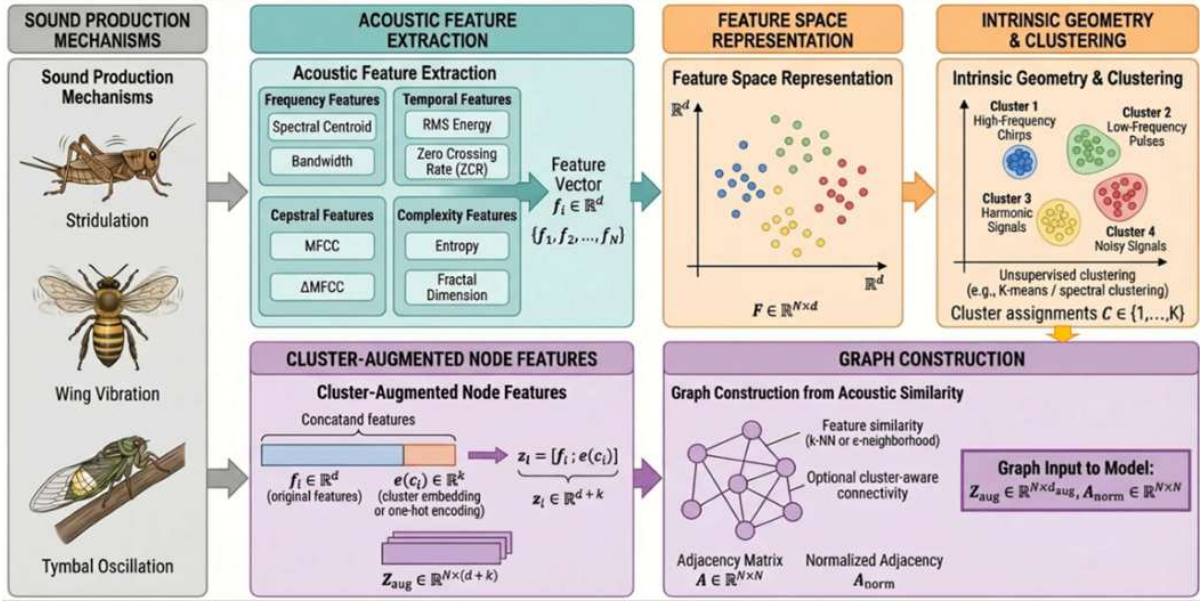


Figure 2. Handcrafted acoustic feature extraction pipeline

A global pooling layer creates a compact embedding:

$$\mathbf{c}_i \in \mathbb{R}^{d_c} \tag{1}$$

which represents the spectral features of the input signal.

Geometric Modeling via Graph Neural Network

To capture relationships between samples, handcrafted feature vectors are used to construct a graph representation. Each sample corresponds to a node, and edges represent similarity between samples in the feature space.

Edge weights are computed using a similarity function:

$$w_{ij} = \exp\left(-\frac{\|\mathbf{f}_i - \mathbf{f}_j\|^2}{2\sigma^2}\right) \tag{2}$$

where $\mathbf{f}_i, \mathbf{f}_j \in \mathbb{R}^{d_f}$ represent the handcrafted feature vectors of samples i and j, respectively, and σ is a scaling parameter that controls the sensitivity of the similarity function. The graph is processed using an attention-based Graph Neural Network (Fig. 2). Unlike standard aggregation methods, attention allows the model to assign different importance to neighboring nodes.

Each node updates its representation by aggregating information from its neighbors:

$$\mathbf{h}_i = \sum_{j \in \mathcal{N}_i} \alpha_{ij} \mathbf{W} \mathbf{h}_j \tag{3}$$

where $\mathbf{h}_i$ denotes the updated embedding of node i, $\mathcal{N}_i$ represents its set of neighboring nodes, α_{ij} are learned attention coefficients that quantify the relative importance of neighbor j in the aggregation process, and $\mathbf{W}$ is a trainable linear transformation matrix.

The resulting embedding:

$$\mathbf{h}_i \in \mathbb{R}^{d_g} \tag{4}$$

captures global structural relationships between samples.

Attention-Based Multimodal Fusion

The CNN and GNN branches create complementary representations: $\mathbf{c}_i$ captures local spectral patterns, while $\mathbf{h}_i$ encodes relational context. To optimally combine them, an attention-based fusion mechanism is used. Instead of assigning equal importance to them, the model learns adaptive weights:

$$\mathbf{z}_i = \alpha_c \mathbf{c}_i + \alpha_g \mathbf{h}_i \tag{5}$$

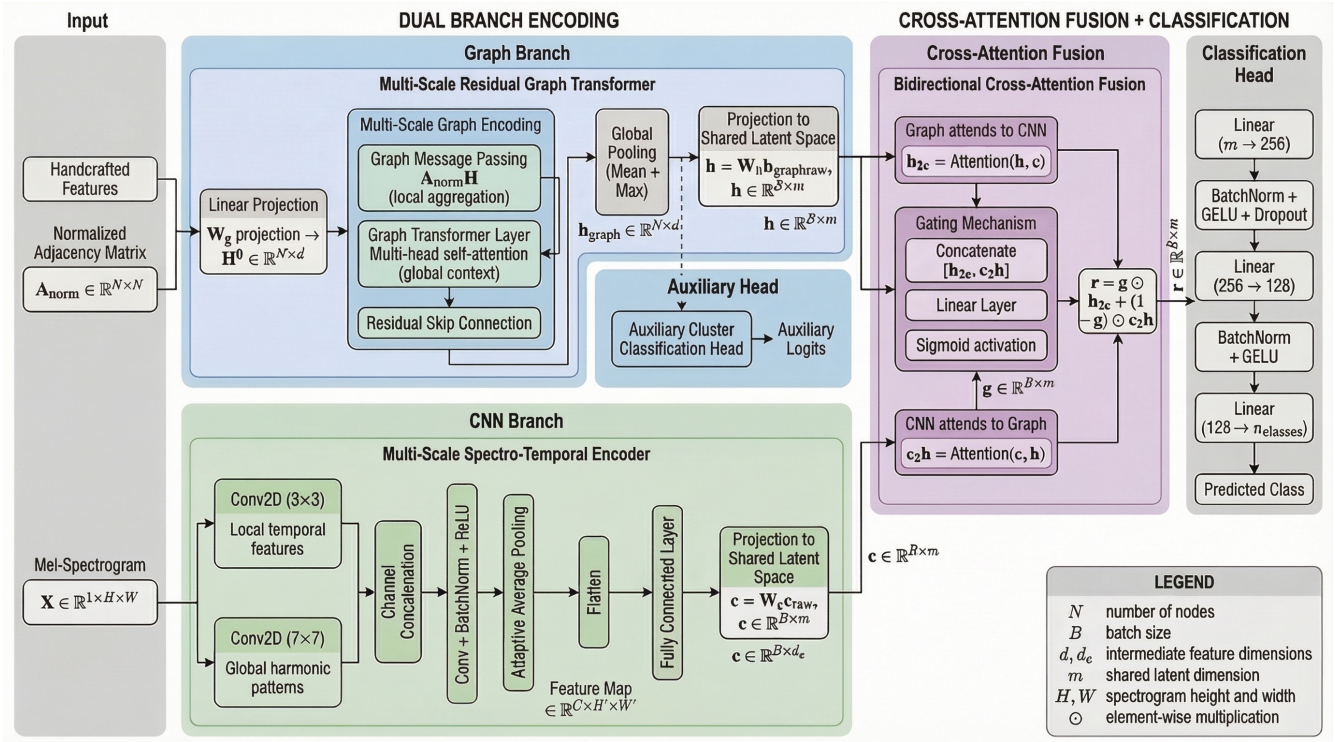


Figure 3. Spectrogram-based ResNet branch for deep feature extraction from matching spectrograms.

where α_c and α_g are learned coefficients. This allows the model to dynamically prioritize spectral or relational information based on the characteristics of each sample.

Geometric Loss Function

In addition to the standard classification loss, a geometric regularization term is also introduced to maintain the relationship between samples. The overall loss is defined as:

$$\mathscr{L} = \mathscr{L}_{CE} + \lambda \mathscr{L}_{geom} \tag{6}$$

Where $\mathscr{L}_{CE}$ represents the cross-entropy loss and $\mathscr{L}_{geom}$ penalizes the significant difference between vector representations of similar samples. This encourages the model to maintain consistency in the learned feature space, thus ensuring that acoustically similar samples maintain a close representation.

Training Strategy

The model is trained end-to-end, optimizing the CNN, GNN, and fusion components simultaneously. In each training iteration, spectrogram features and manually extracted features are processed in parallel by the CNN and GNN branches. The resulting lexical embeddings are merged and passed to a classifier, whereupon the total loss is calculated. Optimization is performed using the Adam optimizer, whose training hyperparameters are summarized in Table 2. The graph structure remains fixed during training, ensuring stable relational learning.

Parameter	Value
Optimizer	Adam
Learning rate	10^{-3}
Batch size	32
Dropout	0.5
Early stopping	Yes

Table 2. Training configuration

Evaluation Protocol

The model is tested and evaluated using stratified 5-fold cross-validation to mitigate the effects of an unbalanced data by making sure the classes are homogeneously distributed for each subset. to visualize the results, a confusion matrix analysis is provided to give a robust estimate of generalization performance. Metrics such as accuracy, precision, recall, and F1 score are used to evaluate the model's classification performance. To assess statistical significance, paired t-tests are performed between folds when comparing different model configurations.

Graph and Clustering Evaluation:

The quality of the graph structure is evaluated using:

- Adjusted Rand Index (ARI), which measures how aligned the clustering assignments and ground truth labels are.
- Silhouette Score, which measures cluster compactness and separation

A high ARI and silhouette score implies that the groupings are well defined which by extension implies that a geometric structure exists. These metrics ensure that the constructed graph reflects meaningful relationships in the data.

Ablation Experiments

To evaluate the contribution of each component, four configurations are considered:

1. CNN-only: Spectrogram-based model without graph information
2. GNN-only: Graph-based model using handcrafted features
3. CNN+GNN: 1 and 2 without attention mechanism
4. CNN+GNN+Attention: 1 and 2 with attention mechanism

Each configuration is evaluated using the same cross-validation setup to ensure fair comparison.

Results and Analysis

Evidence of Geometric Structure in Feature Space

Clustering Method	Best K	ARI	Silhouette Score
Hierarchical Clustering (Ward)	6	0.81	0.62
Spectral Clustering	4	0.68	0.54
Gaussian Mixture Model	3	0.61	0.49
DBSCAN	2	0.52	0.43

Table 3. Clustering evaluation for graph construction

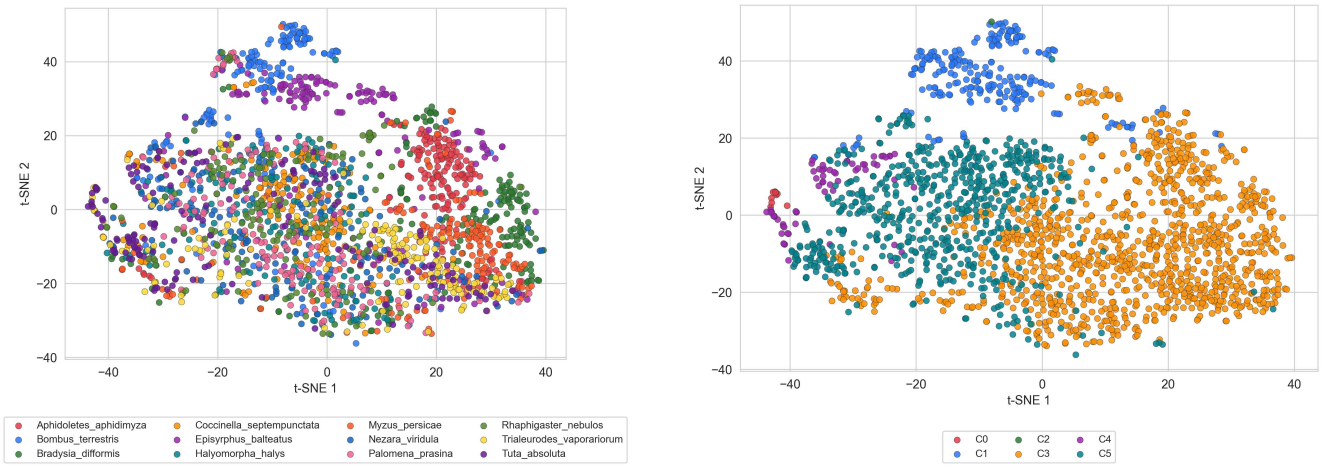
The results presented in the table 3 show that the space of learned characteristics presents a geometrically coherent and well-structured organization. The hierarchical clustering on Ward's method gave the best results, with an ARI of 0.91 and a silhouette score of 0.62, indicating a strong correlation between the characteristics and the body of the representation.

It is worth noting that the relatively low number of clusters (K = 6) means that the space of characteristics is primarily structured around high-level acoustic similarities, rather than strictly divided into classes with clearly defined boundaries. This observation suggests that many species possess similar acoustic characteristics and naturally cluster in the same region of the integration space. This behavior is confirmed by the visualization t-SNE presented in the figure 4. Although the space of global characteristics presents significant differences between the classes, the geometric representation shows a clustering more structured and compact. More precisely, the assignments of the clusters correspond to the regions observed in the integration space, indicating that the hierarchical clustering is determined by the intrinsic geometry of the data.

Although a certain improvement is observed with the lower dimension projections. A high ARI score indicates that the grouping of samples is consistent within the space of characteristics of origin, confirming that the learned representations capture a significant relational structure. Globally, these results support the hypothesis that the acoustic data of insects lie on a structured variety, where the geometric relationships between the samples provide valuable information beyond the independent representations of the characteristics.

Embedding Visualization and Effect of Geometry-Aware Learning

To further analyze the learned representations, t-SNE is applied to a stratified subset of the dataset for visualization. Figure 4 compares embeddings learned without and with geometry-aware loss.



(a) Raw feature space colored by ground truth class labels

(b) Geometry-aware clustering (k = 6) (ARI = 0.81)

Figure 4. t-SNE visualization using a stratified subset (166 samples per class, 12 species) for balanced representation and clarity. A subset is used due to the large dataset size (166,000 samples), which would otherwise cause overcrowding.

Without geometric regularization, the representations exhibit significant overlap between classes, indicating poor separability. Meanwhile, the proposed framework produces relatively compact and better-separated groups, suggesting that geometric learning contributes towards the induced structure of the feature space. This separation reduces overlap between similar acoustic species and facilitates more efficient classification.

Effect of Geometry-Aware Loss on Classification Performance

To quantify the impact of geometry-aware learning, we compare models trained with and without the geometry-aware loss term.

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
Without Geometry Loss	94.1 ± 1.4	94.5 ± 1.3	94.1 ± 1.4	94.0 ± 1.3
With Geometry Loss	96.8 ± 1.0	97.0 ± 1.1	96.2 ± 1.4	96.66 ± 1.2

Table 4. Effect of geometry-aware loss on classification performance

The integration of a weighting function into the geometry systematically improves all the evaluation metrics, increasing the precision and the F1 score by around 3%. More precisely, the performances pass from 94.1% to 96.8%, demonstrating that the application of a geometric coherence in the integration space simultaneously strengthens the ability of discrimination and the generalization.

Overall Classification Performance

Table 5 summarizes the overall classification performance of the proposed model.

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
Proposed Model	97.0 ± 1.0	97.3 ± 0.9	97.0 ± 1.0	96.9 ± 0.9

Table 5. Overall classification performance (5-fold cross-validation)

The proposed model achieves strong and consistent performance across all evaluation metrics, with an overall accuracy of 97.0%. The low standard deviation across folds indicates stable generalization performance. These results highlight the effectiveness of integrating spectral representations with graph-based relational modeling and attention-based fusion.

Ablation Study

In order to evaluate the contribution of each architectural component, we have conducted an ablation study comparing four configurations: (i) CNN alone, (ii) GNN alone, (iii) CNN + GNN and (iv) CNN + GNN with an attention mechanism. The results are summarized in the Table 6 and visualized in the Figure 5. The CNN effectively captures the patterns within the individual spectra, while the GNN reinforces the global structure by modeling the relationships between the samples or at the level of characteristics. Using attention mechanism further improves the performances, reaching a precision of 97.8% and an F1 score of 97.86%. This improvement highlights the importance of the adaptive weighting of characteristics, which allows the model to focus on the most informative components of the fused representation. The attention module improves the space of combined characteristics by suppressing unnecessary or redundant motifs, which leads to a more robust classification.

Globally, the results confirm that each component contributes significantly to the final performances. More precisely, the transition from a module unique model to a framework hybrid CNN+GNN gives the best results, while the addition of an attention mechanism continues to improve the performances, leading to the best results globally.

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
CNN-only	59.59	59.60	59.52	59.56
GNN-only	52.36	52.41	52.45	52.44
CNN+GNN	90.10	90.68	91.2	90.94
CNN+GNN+Attention	97.8	98.1	97.6	97.86

Table 6. Ablation study comparing different model configurations

The CNN fulfills practically the same function, but lacks contextual reasoning. The model based only on GNN exploits the relational structure, but is limited by the expressivity of the characteristics extracted manually. The combination of the two approaches improves performances, demonstrating their complementarity. The complete model produces the best results, proving that fusion based on attention balances spectral and spatial differences. This is supported by the The normalized confusion matrix shown in Fig. 6 provides a detailed view of class-wise performance. The model achieves high classification accuracy across most species, with strong diagonal dominance indicating correct predictions.

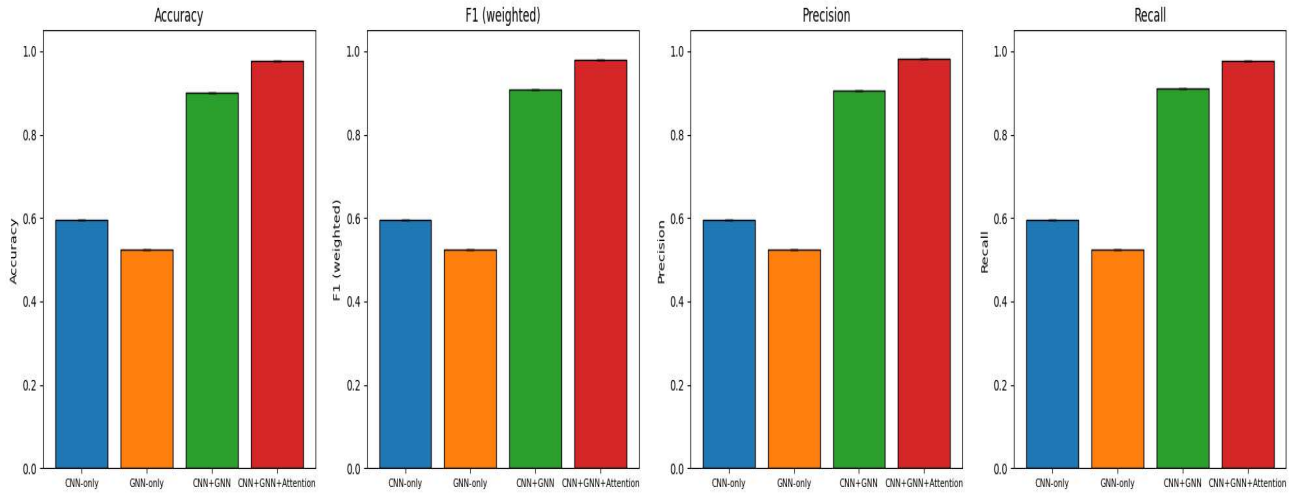


Figure 5. Performance comparison of different model configurations, highlighting the contribution of each component.

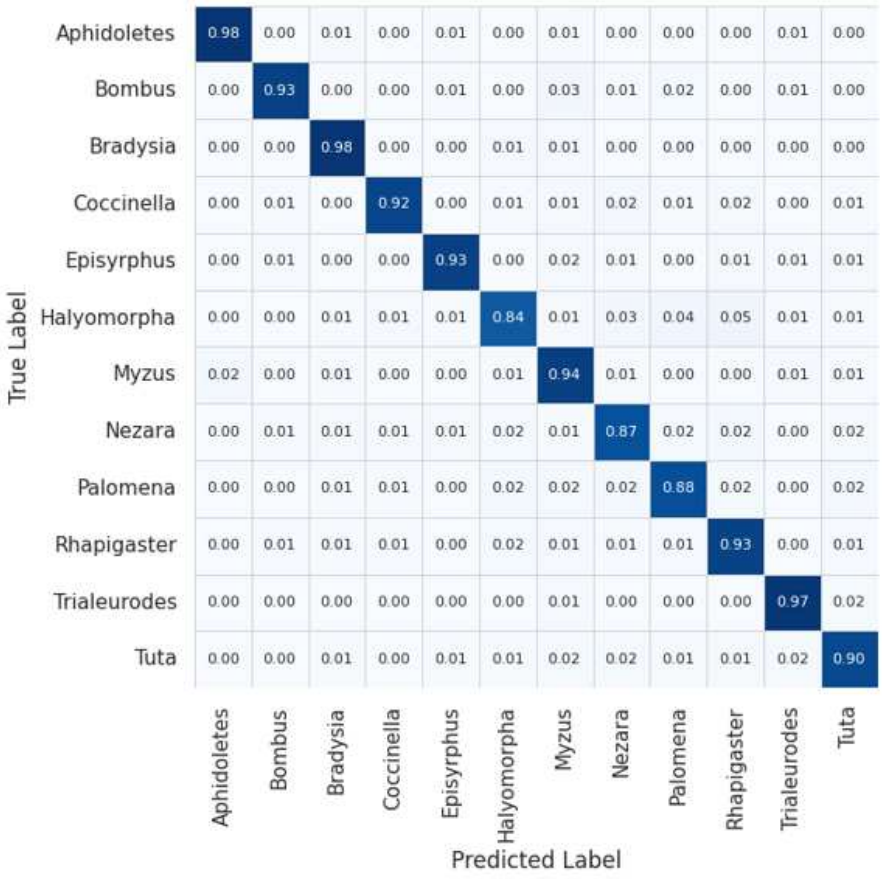


Figure 6. Normalized confusion matrix aggregated across 5-fold cross-validation.

Discussion

The focus of this study was to show that the extremely high dimensional feature space of insect acoustic signals possesses an intrinsic geometric structure that significantly contributes towards a model's classification performance. While the general approach is to treat each audio sample individually, the proposed work explores and reveals the highly nuanced relationships among these acoustic signals produced by a myriad of insect species which makes the acoustic signals more structured and contextually relevant. Upon analysing and visualizing these relationships, the high dimensional feature space was represented as groupings in the form of clusters. However, it is important to note that the number of groupings/clusterings is not equal to the actual number of species listed in the data. This is because the groupings are purely based on the relationships between the learned features.

This strongly supports relational modeling. For example, graphical neural networks are designed to handle precisely this kind of inter-sample dependency, as evidenced by t-SNE graphs. With geometric constraints, clusters are quite tight; without these constraints, classes merge. Combining spectral and relational features generally improved classification while maintaining a largely consistent structure. The CNN branch detects temporal and spectral patterns. The GNN branch handles context through sample similarity. An attention-based merging layer determines the most relevant stream, sample by sample. All metrics improved, showing that these two branches are not redundant: each brings something to the table that the other does not. Once the Geometric consistency was enforced with the neighboring samples in feature space, the model went beyond simply learning a discriminatory cue. It learned the regularity of the structure itself. This is especially important when classes overlap acoustically. Cross-validation results were stable, eliminating the possibility of any noise effects.

Future work will aim to address these shortcomings. Temporal modeling or the integration of multi-modal data on habitat and morphological features can help distinguish species with highly similar songs. Other variation learning methods can also uncover unsuspected geometric properties. The main idea is simple: bioacoustic data have a structure that transcends the information provided by individual samples. Viewing them as a variety, where the relationships between samples provide the real information, leads to better representations and more efficient classifiers. Ignoring this structure means missing out on optimal performance.

References

1. Kohlberg, A. B., Myers, C. R. & Figueroa, L. L. From buzzes to bytes: A systematic review of automated bioacoustics models used to detect, classify and monitor insects. *Journal of Applied Ecology* **61**, 1199–1211 (2024).
2. Kyalo, H. *et al.* Automatic synthesis of insects bioacoustics using machine learning: a systematic review. *International Journal of Tropical Insect Science* **45**, 101–120 (2025).
3. Mason, A. Insect acoustics: Listening in on the deep past. *Current Biology* **34**, R26–R28 (2023).
4. Baker, E. & Chesmore, D. Standardisation of bioacoustic terminology for insects. *Biodiversity Data Journal* **8**, e54222 (2020).
5. Faiß, M. & Stowell, D. Adaptive representations of sound for automatic insect recognition. *PLoS Computational Biology* **19**, e1011541 (2023).
6. Rathore, D. S., Ram, B., Pal, B. L. & Malviya, S. Analysis of classification algorithms for insect detection using matlab. *2nd International Conference on Advanced Computing and Software Engineering (ICACSE)* **1**, 300–304 (2019).
7. Kim, J., Oh, J. & Heo, T.-Y. Acoustic classification of mosquitoes using convolutional neural networks combined with activity circadian rhythm information. *International Journal of Interactive Multimedia and Artificial Intelligence* **7**, 59–66 (2021).
8. Zhu, L.-Q. Insect sound recognition based on mfcc and pnn. *2011 International Conference on Multimedia and Signal Processing* 42–46 (2011).
9. Rekha Kaushik, J. S. Classification of insect's acoustic signals using a hybrid approach: Mel-frequency hilbert-huang transformation. *IJAACS* **17**, 489–501 (2024).
10. Noda, J. J., Travieso-González, C. M., Sánchez-Rodríguez, D. & Alonso-Hernández, J. B. Acoustic classification of singing insects based on mfcc/lfcc fusion. *Applied Sciences* **9**, 4097 (2019).
11. Libal, U. & Biernacki, P. Mfcc-based sound classification of honey bees. *International Journal of Electronics and Telecommunications* **70**, 849–853 (2024).
12. Branding, J., von Hörsten, D., Wegener, J. K. & Hartung, E. Towards noise robust acoustic insect detection: from the lab to the greenhouse. *KI - Künstliche Intelligenz* **37**, 157–173 (2023).
13. Béla J. Szekeres, J. B., Máté Natabara Gyöngyössy. Applying fourier neural operator to insect wingbeat sound classification: Introducing cf-resnet-1d. *Ecological Informatics* **86**, 103055 (2025).
14. Branding, J., von Hörsten, D., Böckmann, E., Wegener, J. K. & Hartung, E. Insectsound1000 an insect sound dataset for deep learning based acoustic insect recognition. *Scientific Data* **11**, 475 (2024).
15. He, H. *et al.* Enhancing insect sound classification using dual-tower network: A fusion of temporal and spectral feature perception. *Applied Sciences* **14**, 3116 (2024).
16. Faiß, M., Ghani, B. & Stowell, D. A dataset of insect sounds from 459 species for bioacoustic machine learning. *Scientific Data* **13**, 499 (2026).
17. Castro-Ospina, A. E., Solarte-Sanchez, M. A., Vega-Escobar, L. S., Isaza, C. & Martínez-Vargas, J. D. Graph-based audio classification using pre-trained models and graph neural networks. *Sensors* **24**, 2106 (2024).
18. Singh, S., Steinmetz, C. J., Benetos, E., Phan, H. & Stowell, D. ATGNN: Audio tagging graph neural network. *IEEE Signal Processing Letters* **31**, 825–829 (2023).
19. Cominelli, S., Bellin, N., Brown, C. D., Rossi, V. & Lawson, J. Acoustic features as a tool to visualize and explore marine soundscapes: Applications illustrated using marine mammal passive acoustic monitoring datasets. *Ecology and Evolution* **14**, e10951 (2024).
20. Szekeres, B. J., Gyöngyössy, N. M. & Botzheim, J. A resnet-9 model for insect wingbeat sound classification. *2023 IEEE Symposium Series on Computational Intelligence (SSCI)* **23**, 587–592 (2023).

Author contributions statement

All authors contributed to the conceptualization of the study. S.P. implemented the model, conducted the evaluation, performed statistical analysis, and wrote the initial draft of the manuscript. R.K. and L.K. contributed to statistical analysis and editing of the manuscript. All authors reviewed and approved the final version of the manuscript.

Competing Interests

The authors declare no conflicts of interest.

Code Availability

The code used in this study is available from the corresponding author upon reasonable request.

Data Availability

The datasets used in this study are publicly available from their respective sources. In particular, the InsectSound1000[14] dataset is available at: https://www.openagrar.de/receive/openagrar_mods_00091171.