Animal Species Classification from Audio Recordings: A Comparison Between Deep Variational Autoencoders and Convolutional Approaches Using MFCCs

Edwin Camilo Ramirez Sanabria
Master's in Engineering Mechatronics
University of Southern Denmark

Abstract—**This project compares two deep learning methodologies for the classification of bird species based on audio recordings. A Deep Variational Autoencoder (DVAE) with a built-in classifier is compared to a Convolutional Neural Network (CNN) that is trained directly on MFCC features. The CNN attained a test accuracy of 68.4% using 15 species from the BirdCLEF+ 2025 dataset, surpassing the DVAE's 50.51%. Although CNNs perform very well with labeled data, DVAEs retain significance due to their capacity to learn representations and employ unlabeled data, making them advantageous in monitoring environments where annotations are limited or difficult to obtain.**

I. Introduction

Monitoring biodiversity using audio recordings has become a key method in ecoacoustics, providing scalable and non-invasive instruments for tracking animal populations. A primary problem in this field is the automated classification of animal species using audio recordings. The complexity of this problem arises from the variety of vocalizations, overlapping calls, environmental noise, and the limited availability of labeled data, particularly in real-world contexts.

A key phase in developing audio categorization systems is the extraction of significant features that encapsulate the fundamental attributes of sounds. This study employs Mel-Frequency Cepstral Coefficients (MFCCs), commonly utilized in speech and bioacoustic analysis. MFCCs encapsulate audio signals in a compact format that approximates human auditory perception, making them well-suited for machine learning algorithms.

This project uses audio data from the BirdCLEF 2025 Kaggle competition, an initiative aimed at improving biodiversity research via automated species identification. The dataset was obtained from the El Silencio Natural Reserve in Colombia's Magdalena Valley, a location recognized for its natural diversity but threatened by deforestation and habitat loss [1]. The competition advocates for passive acoustic monitoring (PAM) as a scalable alternative to traditional field surveys, which are expensive and time-consuming. The comprehensive dataset comprises recordings from more than 200 avian species obtained in authentic environmental circumstances. In this study, we chose a subset of 15 species to perform a targeted assessment of deep learning models for audio classification.

The aim of this study is to evaluate two deep learning methodologies for the classification of bird species based on audio data. The initial method employs a Deep Variational Autoencoder (DVAE) with an integrated classification head. This model acquires a compressed latent representation of the MFCC input while also predicting class labels, integrating generative modeling with supervised learning. The second method employs a Convolutional Neural Network (CNN) that is trained directly on the MFCCs as two-dimensional inputs, with a purely discriminative architecture.

This comparison is motivated by the trade-offs between generative and discriminative learning methods. Convolutional Neural Networks (CNNs) exhibit great efficacy with substantial labeled datasets, whereas Deep Variational Autoencoders (DVAEs) offer more adaptable representations and leverage unlabeled data. This distinction is an essential factor in ecological monitoring, where labels frequently remain limited. Through the assessment of both methodologies on a regulated subset of avian species, I want to determine the practical advantages and constraints of each approach.

II. Methodology

A. Data Preparation

The audio recordings in the BirdCLEF 2025 dataset were originally in '.ogg' format and classified by species. All clips for this project were converted to '.wav' format to guarantee compatibility while avoiding degradation due to compression. The '.wav' format preserves signal accuracy, crucial for proper audio analysis [2]. Each file was resampled to 16,000 Hz, a common sampling rate in speech and bioacoustic applications, as it effectively balances the preservation of critical frequency information with the decrease of computational requirements [3].

Mel-Frequency Cepstral Coefficients (MFCCs) were derived from each clip to represent the spectral content in a significant way. MFCCs are extensively utilized in bioacoustic and speech-related machine learning applications because of their concise and informative representation of sound attributes.

We employed 13 coefficients and 40 mel filters per frame, a configuration frequently utilized in the literature to capture adequate spectral detail while preserving model efficiency [4]. To maintain consistent input dimensions, MFCC sequences were either padded with zeros or trimmed to a set amount of time frames. This standard procedure facilitates uniform batch processing in neural networks and prevents dimensional conflicts during training [5].

B. Model Architectures

Two neural network architectures were implemented using PyTorch.

1) Deep Variational Autoencoder (DVAE): The initial model deployed was a Deep Variational Autoencoder (DVAE), including components of unsupervised and supervised learning. The architecture contains three key elements: an encoder, a latent representation layer, and a decoder, as well as an auxiliary classification head.

The encoder is a unidirectional Gated Recurrent Unit (GRU) that sequentially processes input sequences of MFCCs. GRUs are a type of recurrent neural network (RNN) designed to remember important information over time in sequences, like Long Short-Term Memory (LSTM) units, but they have a simpler way of managing this. Compared to regular RNNs, GRUs help prevent the issue of vanishing gradients and are more efficient than LSTMs because they have fewer parameters. GRUs are an appropriate selection for extracting temporal data from audio, maintaining a lightweight model that is easy to train.

The encoder generates a fixed-length hidden state that contains the temporal dynamics of the input sequence. The hidden state is then sent through two separate fully connected layers to find the values for a Gaussian distribution in the latent space: a mean vector μ and a log-variance vector $\log \sigma^2$. The latent space dimensionality was established at 64, achieving a compromise between expressive capacity and regularization. This size is often used in similar sequence modeling tasks that try to reduce structured features like MFCCs while keeping important information [6].

To add randomness while keeping backpropagation intact, the reparameterization trick was used: a latent vector z was created by using the formula $z = \mu + \sigma \cdot \epsilon$, where ϵ comes from a standard normal distribution. This method facilitates the network's acquisition of a smooth and continuous latent space, hence allowing gradient descent optimization [6].

In addition to the generative path, a classification head was attached directly to the sampled latent vector. This head consisted of two fully connected layers with ReLU activation and a softmax output layer, producing probability scores for each of the 15 bird species. This setup allowed the encoder to learn latent features that are both generative (supporting reconstruction) and discriminative (supporting classification).

The decoder was developed utilizing a GRU network. The sampled latent vector z is received and initially projected into a hidden state. A zero-initialized sequence is subsequently input into the GRU decoder, which reconstructs the MFCC sequence gradually, frame by frame. A linear layer at each time step's output translates the decoder output into the MFCC space.

By working together on reconstruction loss, KL divergence, and classification loss, the DVAE learns important hidden features that help in generating signals and predicting labels.

2) Convolutional Neural Network (CNN): The second model utilized was a Convolutional Neural Network (CNN) designed for supervised classification based on MFCC features. This approach analyzes MFCC tensors as 2D pictures, with the vertical axis representing frequency bands (13 MFCCs) and the horizontal axis indicating time frames.

The model architecture includes three convolutional blocks. Each block includes a two-dimensional convolutional layer, succeeded by batch normalization and a ReLU activation algorithm. These layers help identify specific features in the MFCC input, such as changes in frequency and timing, which usually represent different bird sounds. After each convolutional block, a 2×2 max pooling layer was used to reduce the detail and processing power needed while keeping the most important features.

After the convolutional layers, I transformed the output into a one-dimensional vector and transmitted it through two fully linked layers. The first dense layer reduces the size of the data and adds complexity using a ReLU activation, while the next layer produces probabilities for the 15 bird species using a SoftMax function.

This design was chosen because the CNN can recognize both quick and slow frequency patterns using spatial filters, which has proven effective in classifying audio [7]. The use of convolutional layers allows the model to adapt to small changes in frequency and time, which is important because natural bird calls can vary a lot.

In contrast to the DVAE, the CNN model is entirely discriminative and does not acquire a generative latent space. Its effectiveness comes from directly improving classification performance using labeled data, without needing to reconstruct the input features.

C. Training Procedure

Both models were executed with PyTorch and trained independently on the identical dataset split. The training set contains 80% of the available labeled samples from 15 bird species, and the remaining 20% was used for testing purposes.

CNN Training: The CNN model was trained for 100 epochs using the Adam optimizer with a learning rate of 10^{-4}. We applied a step-based learning rate scheduler, which reduces the learning rate by 0.5 every 30 epochs to enhance convergence. The employed loss function was categorical cross-entropy, which is appropriate for multi-class classification problems. Cross-entropy loss measures how different the predicted probabilities are from the actual outcomes, giving bigger penalties to the model when it gives low probabilities to the correct class, which encourages accurate and confident predictions [7].

Data augmentation was implemented during training to enhance robustness and mitigate overfitting. The process involved incorporating Gaussian noise, time masking, and frequency masking directly on the MFCC tensors. Gaussian noise involves the incorporation of stochastic noise characterized by a normal distribution into the input features to replicate noise and augment the model's capacity to work in environments with noise. Time masking randomly selects a temporal interval inside the MFCC sequence and obscures it, simulating conditions where specific segments of the audio are absent or obstructed. Frequency masking involves obscuring a range of frequency channels to simulate the loss of specific frequency data caused by environmental influences or recording constraints. These augmentation strategies enhance the model's resilience to variations typically encountered in real-world recordings [8].

DVAE Training: The DVAE model was trained for 1000 epochs using the Adam optimizer with a learning rate of 10^{-3}. The overall loss function was made up of three parts: (1) mean squared error (MSE) to ensure the MFCC input is accurately reconstructed, (2) Kullback-Leibler (KL) divergence to keep the latent space similar to a standard Gaussian distribution, and (3) cross-entropy loss to make sure the latent representation helps in classifying species. The KL divergence was gradually increased from 0 to 1 over the first 100 epochs to avoid too much regularization early on and to help the model learn the generative structure steadily. A higher fixed weight was given to the classification loss to highlight discriminative performance. This combination allows the model to learn latent variables that are both generative and beneficial for supervised tasks. No early stopping or validation checks were used during training because a set number of epochs helped keep the latent space organized without relying on possibly unreliable validation feedback.

D. Evaluation Metrics

The model performance was evaluated largely through classification accuracy, which is defined as the ratio of correct predictions to the total number of test samples. This statistic offers a broad assessment of each model's capacity to differentiate between bird species using their MFCC representations. To complement this metric, I created confusion matrices, allowing the assessment of the classes to discover distinct patterns of misclassification. This approach is especially pertinent in bioacoustic activities, where specific species may provide acoustically similar vocalizations, resulting in overlapping frequency characteristics.

Additionally, we monitored the classification loss across the epochs throughout the training of both models. This exercise allowed me to observe convergence and training stability. For the DVAE, I individually documented the reconstruction loss, KL divergence, and classification loss to examine the interaction between generative and discriminative elements.

Finally, we used t-SNE to map the DVAE's hidden vectors into two dimensions, making it easier to visually check how the learned representations are grouped by species. This approach facilitated the assessment of the architecture and distinctiveness of the latent space acquired by the encoder.

III. Results

A. How well did the models work?

The CNN classifier obtained a test accuracy of 68.77%, while the DVAE+MLP pipeline achieved 50.21%. The anticipated performance disparity arises from the fact that the CNN is specifically designed for supervised classification, whereas the DVAE must simultaneously balance generative reconstruction and regularization objectives.

The training loss curves further support this behavioral disparity. The CNN's classification loss (Figure 1) diminished steadily, achieving convergence within the initial 60 epochs.

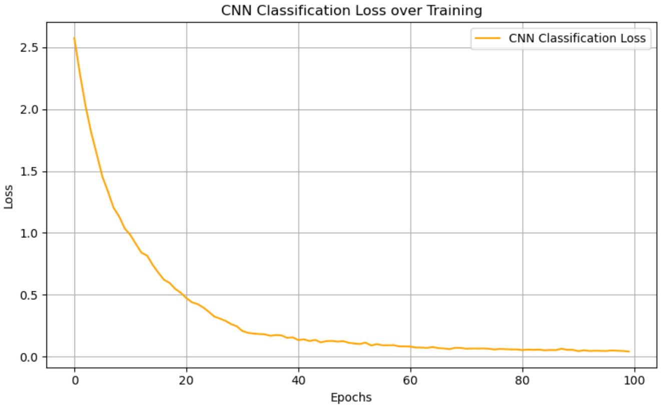


Fig. 1. CNN training loss over epochs.

The DVAE model (Figure 2) showed large changes in its reconstruction loss, particularly in the first training phase, which suggests that its data generation process was unstable. Despite the gradual reduction in reconstruction errors, they persisted in exhibiting noise throughout the epochs. The classification loss remained continuously low but plateaued early, indicating that the encoder did not succeed in learning. These suggest that the latent space has limited ability to express information, likely because of the needs for accurate reconstruction, KL regularization, and distinguishing between classes.

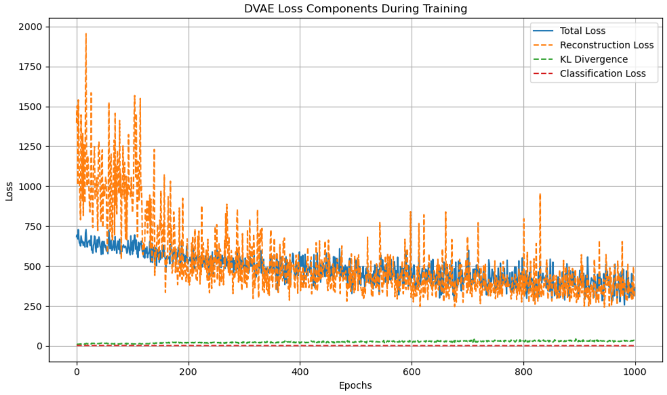


Fig. 2. DVAE loss components: total, reconstruction, KL divergence, and classification loss.

B. Examples where the models worked well

The CNN exhibited consistently robust performance throughout the majority of classes. This is evident in its confusion matrix (Figure 3), which exhibits significant diagonal dominance. Species such as *whtdov* and *yeofly1* were accurately categorized in most instances, indicating that their MFCC patterns were distinctive and effectively represented by the convolutional filters.

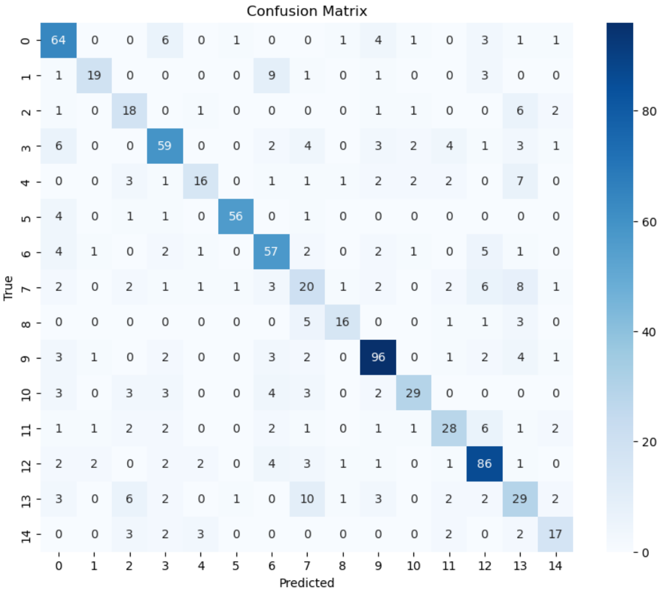


Fig. 3. Confusion matrix for CNN-based classification on the test set.

Likewise, the DVAE successfully classified specific species, especially those exhibiting more pronounced spectral signatures. The t-SNE projection of its latent space (Figure 4) demonstrates partial clustering for certain classes, suggesting that the encoder has discerned the underlying structure in the input, despite the classifier head's inability to completely utilize it.

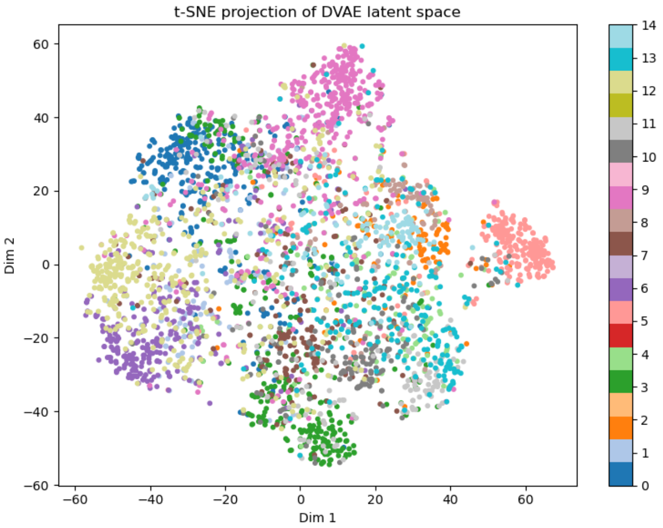
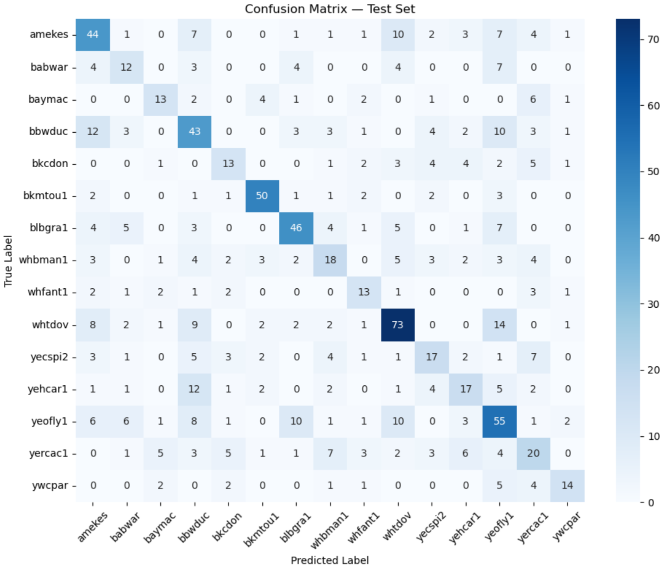


Fig. 4. t-SNE projection of DVAE latent space colored by class label.

C. Model limitations and failure cases

Despite its better accuracy, the CNN still has its limitations. It couldn't tell apart some species that sound similar, as shown by the confusion matrix, especially *blbgra1* vs. *yehcar1* and *yeraca1* vs. *whbman1*. Those confusions indicate that local time-frequency characteristics alone may be inadequate for complete discrimination.

The DVAE's main limitation lies in its inability to form well-separated latent clusters. Its confusion matrix (Figure 5) reveals widespread misclassification, and the t-SNE visualization confirms overlap between classes. This is likely due to the trade-off between learning a reconstructive latent space and achieving class separability. Additionally, the high variance in the reconstruction loss suggests that the decoder struggles with consistency, which may hinder the encoder's capacity to focus on class-relevant features.



	amekes	babwar	baymac	bbwduc	bkcdon	bkmtou1	blbgra1	whbman1	whfant1	whtdov	yecspi2	yehcar1	yeofly1	yercac1	ywcpar
amekes	44	1	0	7	0	0	1	1	1	10	2	3	7	4	1
babwar	4	12	0	3	0	0	4	0	0	4	0	0	7	0	0
baymac	0	0	13	2	0	4	1	0	2	0	1	0	0	6	1
bbwduc	12	3	0	43	0	0	3	3	1	0	4	2	10	3	1
bkcdon	0	0	1	0	13	0	0	1	2	3	4	4	2	5	1
bkmtou1	2	0	0	1	1	50	1	1	2	0	2	0	3	0	0
blbgra1	4	5	0	3	0	0	46	4	1	5	0	1	7	0	0
whbman1	3	0	1	4	2	3	2	18	0	5	3	2	3	4	0
whfant1	2	1	2	1	2	0	0	0	13	1	0	0	0	3	1
whtdov	8	2	1	9	0	2	2	2	1	73	0	0	14	0	1
yecspi2	3	1	0	5	3	2	0	4	1	1	17	2	1	7	0
yehcar1	1	1	0	12	1	2	0	2	0	1	4	17	5	2	0
yeofly1	6	6	1	8	1	0	10	1	1	10	0	3	55	1	2
yercac1	0	1	5	3	5	1	1	7	3	2	3	6	4	20	0
ywcpar	0	0	2	0	2	0	0	1	1	0	0	0	5	4	14

Fig. 5. Confusion matrix for DVAE+MLP classification on the test set.

The DVAE has advantages like interpretability and generative modeling capabilities. However, this is accompanied by a reduction in discriminative precision for this job.

IV. Discussion and Conclusions

A. Key Findings

The findings indicate that the CNN outperformed the DVAE+MLP pipeline by almost 20 percent (68.4% vs. 50.21% accuracy). This disparity is anticipated due to the CNN's exclusively discriminative aim and comprehensive training approach. On the other hand, the DVAE was required to simultaneously achieve three learning objectives: recreating MFCC sequences, regularizing the latent space, and predicting species labels. The confusion matrices corroborate this difference: the CNN consistently identified the proper class for the majority of species, whereas the DVAE exhibited significant confusion, particularly among acoustically similar avian species. Furthermore, the training loss curves indicated that the CNN achieved rapid and stable convergence, while the DVAE's reconstruction loss exhibited persistent volatility throughout training. Despite a relatively modest classification loss, it plateaued early, suggesting inadequate class separation in the encoder's latent space.

The t-SNE visualization supported this finding: the DVAE latent space had some clusters that overlapped, showing that its need to recreate inputs limited its ability to organize the space in a way that clearly separates different classes. Nonetheless, the DVAE maintained its use as a representation learner, and its organized latent space may be advantageous in more extensive downstream applications.

B. Limitations

Despite its enhanced performance, the CNN lacks generative functionality. Its dependence on labeled data constrains its applicability in low-resource ecological contexts. In contrast, although the DVAE exhibited inadequate results in this task, its architecture is intrinsically better for semi-supervised learning.

Furthermore, the DVAE's MLP classifier was trained on the encoder's latent representations, potentially constraining its integration with the overall architecture. Joint training or alternative loss configurations may be investigated to improve the alignment between encoder objectives and classifier effectiveness.

C. Future Work

Several directions are open for extending this study. For the DVAE, incorporating attention mechanisms in the encoder or decoder, experimenting with β-VAE variants, or conditioning the decoder on labels could improve class separability. Incorporating unlabeled data would enable the evaluation of the model's semi-supervised capabilities.

The CNN could gain advantages from architectural improvements such as temporal convolutional networks, residual blocks, or self-attention mechanisms. Additional enhancement techniques such as pitch shifting or blending with natural ambient noises may enhance robustness.

References

[1] Kaggle, "Birdclef 2025 - bird sound classification," https://www.kaggle.com/competitions/birdclef-2025/data, 2025, accessed: 2025-05-29.

[2] N. Khairunnisa *et al.*, "Classification of bird sound using high-and low-complexity convolutional neural networks," *International Journal of Applied Engineering Research*, vol. 17, no. 1, pp. 1–8, 2022. [Online]. Available: https://www.iieta.org/download/file/fid/70786

[3] I. Aljubayri, "Comparative analysis of different sampling rates on environmental sound classification using the urbansound8k dataset," *Journal of Computer and Communications*, vol. 11, no. 6, pp. 19–27, 2023. [Online]. Available: https://www.scirp.org/journal/paperinformation?paperid=125679

[4] S. Ali, S. Tanweer, S. S. Khalid, and N. Rao, "Mel frequency cepstral coefficient: A review," in *Proceedings of the International Conference on Intelligent Computing and Control Systems (ICICCS)*. EAI, 2020, pp. 1234–1240. [Online]. Available: https://eudl.eu/pdf/10.4108/eai.27-2-2020.2303173

[5] Y. Li, Z. Wang, and L. Zhao, "A comparative study of padding strategies in audio classification using convolutional neural networks," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2021, no. 1, pp. 1–14, 2021. [Online]. Available: https://asmp-eurasipjournals.springeropen.com/articles/10.1186/s13636-021-00200-1

[6] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[7] S. Hershey, S. Chaudhuri, D. P. Ellis, J. F. Gemmeke, A. Jansen, R. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold *et al.*, "Cnn architectures for large-scale audio classification," in *ICASSP*. IEEE, 2017, pp. 131–135.

[8] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "Specaugment: A simple data augmentation method for automatic speech recognition," in *Proc. Interspeech*, 2019, pp. 2613–2617.