Journal of Korea Society of Industrial Information Systems Vol.29 No.2, Apr. 2024 :15-25
http://dx.doi.org/10.9723/jksiis.2024.29.2.015 ISSN:1229-3741

# Focal Loss와 앙상블 학습을 이용한 야생조류 소리 분류 기법

## (Wild Bird Sound Classification Scheme using Focal Loss and Ensemble Learning)

**이 재 승[1], 유 제 혁[2]***
(Jaeseung Lee and Jehyeok Rew)

**요 약** 효과적인 동물 생태계 분석을 위해서는 동물 서식 현황을 자동으로 파악할 수 있는 동물 관제 기술이 중요하다. 특히 울음소리로 종을 판별하는 동물 소리 분류 기술은 영상을 통한 판별이 어려운 환경에서 큰 주목을 받고 있다. 기존 연구들은 단일 딥러닝 모델을 사용하여 동물 소리를 분류하였으나, 야외 환경에서 수집된 동물 소리는 많은 배경 잡음을 포함하여 단일 모델의 판별력을 악화시키며, 종에 따른 데이터 불균형으로 인해 모델의 편향된 학습을 야기한다. 이에, 본 논문에서는 클래스의 데이터 수를 고려하여 페널티를 부여하는 Focal Loss를 사용한 여러 분류 모델의 예측 결과를 앙상블을 통해 결합하여 잡음이 많은 동물 소리를 효과적으로 분류할 수 있는 기법을 제안한다. 공개 데이터 셋을 사용한 실험에서, 제안된 기법은 단일 모델의 평균 성능에 비해 Recall 기준으로 최대 22.6%의 성능 개선을 달성하였다.

**핵심주제어**: 데이터 불균형, Focal Loss, 앙상블 학습, 조류 소리 분류

**Abstract** For effective analysis of animal ecosystems, technology that can automatically identify the current status of animal habitats is crucial. Specifically, animal sound classification, which identifies species based on their sounds, is gaining great attention where video-based discrimination is impractical. Traditional studies have relied on a single deep learning model to classify animal sounds. However, sounds collected in outdoor settings often include substantial background noise, complicating the task for a single model. In addition, data imbalance among species may lead to biased model training. To address these challenges, in this paper, we propose an animal sound classification scheme that combines predictions from multiple models using Focal Loss, which adjusts penalties based on class data volume. Experiments on public datasets have demonstrated that our scheme can improve recall by up to 22.6% compared to an average of single models.

**Keywords:** Data Imbalance, Focal Loss, Ensemle Learning, Bird Sound Classification

* Corresponding Author: jhrew@duksung.ac.kr
Manuscript received January 18, 2024 / revised February 26, 2024 / accepted March 06, 2024

1) 고려대학교 전기전자공학과, 제1저자
2) 덕성여자대학교 데이터사이언스학과, 교신저자

## 1. 서 론

수세기에 걸친 인간의 산업 활동은 야생동물의 서식 환경에 많은 영향을 미쳤다. 예를 들어, 척추동물의 절멸 위험종, 취약종 및 희귀종의 67%가 토지개발과 산업화로 인하여 서식지의 파괴를 겪었다 (Korea Forest Service, 2023). 이는 유전적으로 다양한 개체들 사이에 물리적 교란을 일으키고, 서식지의 파편화와 단절로 인해 동물의 고립을 초래하였다. 이러한 피해를 줄이기 위해서는, 다양한 동물들의 현황과 서식 상태를 지속적으로 관찰해야 한다. 이를 위해, 기존에는 전문가들이 직접 생태 현장을 조사했지만 효율적이지 못했기에 최근에는 음향 센서나 카메라를 통해 동물들을 식별하는 자동 야생동물 모니터링 기술이 주목받고 있다.

자동 야생동물 모니터링 기술 중 하나인 동물 소리 분류는 동물 울음소리를 이용하여 종을 판별한다. 동물 소리 분류는 작거나 야행성이거나 보호색을 띄는 등 영상 데이터만으로는 식별이 어려운 동물에게 효과적으로 적용할 수 있다. 뛰어난 판별 성능을 얻기 위해, 최근 딥러닝 기반의 동물 소리 분류 모델들이 제안되었다. 예를 들어, KW Gunawan et al.(2023)은 EfficientNet과 같은 단일 모델을 사용하여 부엉이 소리를 분류하는 방법을 제시하였다.

그러나, 동물 소리는 보통 야외 환경에서 수집되기에 많은 잡음을 포함한다. 이러한 데이터를 단일 모델로 판별하면 모델이 훈련 데이터에 포함된 노이즈까지 학습하게 되어 과적합(Overfitting) 문제가 발생하게 된다. 또한, 자주 관측되는 흔한 종의 소리들이 편향적으로 많이 수집되는 경우, 분류 모델이 그런 종에 편향된 학습을 하게 되어 과추정(Overestimation) 문제가 발생하게 된다.

이러한 문제들을 해결하기 위해, 본 논문에서는 데이터 불균형을 완화하는 Focal Loss (Lin et al., 2017)와 일반화 성능 향상을 위한 Weighted Soft Voting 기반의 앙상블 기법 (Ganaie et al., 2022)을 사용하여 잡음이 많은 야생동물 울음소리를 효과적으로 분류하는 기법을 제안한다. 먼저, 동물 소리 데이터를 2차원 데이터인 Spectrogram으로 변환하고, 클래스별 데이터 불균형을 고려하여 손실 함수로 Focal Loss를 이용한 분류 모델들을 구축한다. Focal Loss는 초모수 설정에 따라 모델의 성능 민감도가 달라지기 때문에, 훈련 셋에서의 5겹 교차검증을 통해 각 모델에 적합한 최적의 초모수를 설정한다. 최종적으로, 이전 5겹 교차검증에서 측정한 분류 모델별 정확도를 가중치로 사용한 Weighted Soft Voting을 적용한다.

제안하는 기법의 성능을 평가하기 위하여, 공공 조류 울음소리 데이터를 대상으로 다양한 실험을 수행하였다. 우선, Focal Loss의 다양한 초모수 설정에 따른 제안 기법의 분류 성능 변화를 측정하였다. 또한, 딥러닝 기반 분류 모델에서 일반적으로 활용되는 손실 함수인 Cross Entropy Loss (CE Loss)를 이용한 단일 모델과 제안 기법의 분류 성능을 비교하고, 데이터 불균형 문제를 해소하기 위해 데이터 셋의 클래스 간 균형을 맞추는 언더샘플링(Undersampling) 및 오버샘플링(Oversampling) 기법과 제안 기법의 분류 성능을 비교하였다. 마지막으로, 앙상블 기법을 구축하는 데 있어 사용하는 모델의 유형과 개수를 변경해가며 최적의 앙상블 모델 조합을 탐색하였다. Inception V3 (Szegedy et al., 2016), EfficientNet, 그리고 DenseNet (Huang et al., 2017) 등 3개의 단위 판별기 모델을 구성한 후 개수의 성능과 Voting 기반의 조합 방법을 고려하여 성능 변화를 측정하였다.

본 논문의 구성은 다음과 같다. 2장에서는 동물 소리 분류 관련 연구를 기술하고, 3장에서는 구축한 데이터 셋에 대하여 설명한다. 4장에서는 제안하는 기법에 대하여 자세히 설명하고, 5장에서는 제안하는 기법의 정량적 실험 결과를 보인다. 마지막으로 6장에서는 결론 및 향후 연구에 대해 밝히며 본 논문의 끝을 맺는다.

## 2. 관련 연구

효과적인 동물 소리 분류를 위해, 최근 딥러닝 기반의 다양한 연구들이 제안되었다(Incze et al., 2018; Koh et al., 2019; Kim et al., 2020;

Kim et al., 2020; Hidayat et al., 2021; Kahl et al., 2021; Martynov et al., 2022; Sun et al., 2022; Gunawan et al., 2023; Kim et al., 2023). 가장 일반적인 방법은 추출한 동물 소리의 음향 특징을 기반으로 소리를 분류하는 것이다(Incze et al., 2018; Kim et al., 2020; Kim et al., 2020; Hidayat et al., 2021; Martynov et al., 2022). 예를 들어, Incze et al.(2018)은 사전 학습된 모델인 MobileNet을 조류 소리로부터 추출한 음향 특징에 맞게 미세 조정하여 조류 소리를 분류하였다. Kim et al.(2020)은 음향 데이터의 시간 축에 자가주의집중을 적용하여 특징을 추출한 후 동물 소리 분류를 진행하였다. Kim et al.(2020)은 음향 데이터에서 추출한 다수의 특징을 결합하여 다양한 종의 동물 소리를 분류하는 기법을 제시하였다. Hidayat et al.(2021)은 부엉이 소리를 Mel-Frequency Cepstral Coefficient (MFCC)와 Melspetrogram로 표현한 이후, 합성곱 신경망을 이용하여 각각의 표현으로부터 추출한 특성을 결합하여 부엉이 소리를 분류하였다. Martynov et al.(2022)은 합성곱 신경망과 사전 학습된 오디오 신경망을 통해 추출한 음향 특징에 기반하여 조류 소리를 분류하는 기법을 제안하였다. 또한, 데이터 증강 기술을 활용하여 동물 소리를 분류하는 시도가 등장하였다 (Koh et al., 2019; Kahl et al., 2021; Sun et al., 2022; Kim et al., 2023). Kim et al.(2023)은 GAN (Generative Adversarial Networks) 기반의 데이터 생성 모델을 이용하여 동물 소리 데이터를 증강함으로써 데이터 불균형을 고려한 동물 소리 분류 기법을 제시하였다. Koh et al.(2019)은 조류 소리 데이터에 백색 가우시안 잡음을 추가하여 데이터를 증강한 후 조류 소리를 분류하는 기법을 제안하였다. Kahl et al.(2021)은 음향 데이터의 시간 축을 변경하여 조류 울음소리 데이터를 증강한 후 조류 소리 분류를 진행하였다. Sun et al.(2022)은 딥러닝 모델에 많은 양의 학습 데이터가 필요하다는 점에 주목하여 데이터 증강과 합성곱 신경망 기반 모델을 결합한 동물 소리 방법을 제안하였다.

하지만, 데이터 증강 방법론은 데이터의 다양성을 인위적으로 증가시킬 수 있지만, 모든 유형의 데이터 불균형 문제를 해결하는 데 충분하지는 않다. 특히, 희귀종과 같이 매우 드물게 존재하는 소수 클래스의 경우, 원본 데이터 자체의 양이 적어 충분한 양의 증강 데이터를 생성하기 어렵다. 이는 증강된 데이터가 원본 데이터의 특성을 완벽히 반영할 수 없도록 한다. 이러한 문제를 해결하기 위해서는, 학습 과정에서 불균형한 클래스에 대한 손실을 직접적으로 조정해야 한다. 즉, 잘못 분류된 샘플에 더 많은 가중치를 부여하여 모델이 소수 클래스에 속하는 샘플을 올바르게 분류하는 데 더 집중하도록 함으로써 전반적인 분류 성능의 균형을 개선해야 한다.

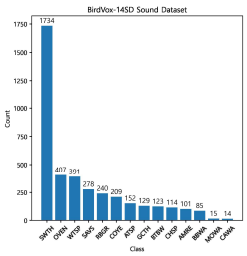
이와 함께, 아직까지 동물 소리 수집 시 흔히 발생하는 데이터 불균형 문제와 잡음의 영향을 함께 고려한 연구는 수행되지 않았다. 이에, 본 논문에서는 Focal Loss를 통해 클래스 가중치를 설정함으로써 데이터 불균형 상황에서의 학습 과정을 보다 직접적으로 최적화하고, 앙상블 학습을 통해 잡음이 많은 동물 소리를 효과적으로 분류할 수 있는 방법을 제안한다.

## 3. 데이터 셋 구축

### 3.1 BirdVox-14SD 데이터 셋

BirdVox-14SD 데이터 셋 (Cramer et al., 2020)은 바람 소리, 물결 소리, 차량 소음 등 다

(1) Class Distribution




(2) Class Name

American Tree Sparrows (ATSP)
Chipping Sparrows (CHSP)
Savannah Sparrow (SAVS)
White-Throated Sparrow (WTSP)
Rose-Breasted Grosbeak (RBGR)
Gray-Cheeked Thrush (GCTH)
Swainson's Thrush (SWTH)
American Redstart (AMRE)
Bay-Breasted Warbler (BBWA)
Black-Throated Blue Warbler (BTBW)
Canada Warbler (CAWA)
Common Yellowthroat (COYE)
Mourning Warbler (MOWA)
Ovenbird (OVEN)

Fig. 1 Summary of BirdVox-14SD Dataset

양한 잡음이 존재하는 야외 환경에서 수집한 조류 울음소리 데이터 셋으로, 총 14종의 조류 클래스의 음향 클립을 제공한다. Fig. 1은 BirdVox-14SD 데이터 셋의 요약을 나타낸다. 본 논문에서는 음향 길이 0.5초로 구성된 총 3,992개의 조류 울음소리 데이터를 이용하였다. 데이터 셋 전체의 통일성을 위하여, 모델을 구축하는 데 있어 음향 길이 0.5초 전체의 샘플을 이용하였다. 또한, 클래스별 샘플 개수가 최소 14개에서 최대 1,734개로 데이터 불균형이 존재한다고 볼 수 있다.

### 3.2 데이터 전처리

동물 소리는 주파수, 지속시간, 속도와 같은 고유한 음향 패턴을 가지고 있다 (Kim et al., 2023). 이러한 특성을 딥러닝 모델이 효과적으로 학습하기 위해서는, 진폭만 고려하는 1차원 데이터인 Waveform을 진폭과 주파수를 모두 고려하는 2차원 데이터인 Spectrogram으로 변환해야 한다 (Kim et al., 2022). Fig. 2는 BirdVox-14SD 데이터 셋에서 동물 소리 데이터의 Waveform과 Spectrogram 예시를 보여준다. 동물 소리는 특정 주파수 범위에서의 강한 에너지 변화와 지속 시간으로 특징지어질 수 있다. 이에, 동물 소리 데이터를 Spectrogram으로 변환한다면 동물 소리의 특징적인 주파수 대역과 강도 변화를 시간 축 상에서 명확히 관찰할 수 있게 된다. 이러한 패턴은 Spectrogram 상에서 시각적으로 구분될 수 있기에 동물 소리가 나타나는 구간과 그렇지 않은 구간을 효과적으로 판별할 수 있고, 이를 기반으로 동물 소리 분류 모델을 구축할 수 있게 된다.

본 논문에서는 각 동물 소리 파형에 대해 단기간 푸리에 변환 연산을 적용하여 다양한 주파수를 가지는 주기 함수들로 분해하며, 이를 Spectrogram으로 나타낸다. 연산 파라미터는 NFFT, Hop Length, Sample Rate 등을 사용하였다. 변환된 모든 Spectrogram은 고정된 크기를 갖도록 제로 패딩(Zero-Padding) 되었다. 결과적으로, 모든 동물 소리 데이터는 128×128 크기의 1채널 Spectrogram으로 표현되었다.

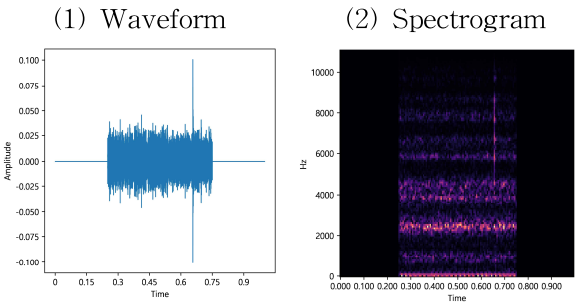



Fig. 2 Representation of Animal Sound Data

## 4. 제안하는 기법

### 4.1 분류 모델 선정

본 논문에서는 동물 소리 분류를 위해 입력으로 2차원 이미지 데이터와 유사한 형태를 가지는 Spectrogram을 이용하기 때문에, 최근 이미지 분류 분야에서 좋은 성능을 보이고 있는 합성곱 신경망 기반의 모델들을 이용할 수 있다. 이 모델들은 다수의 컨볼루션 레이어를 통해 이미지의 특징을 효과적으로 추출할 수 있는 특성이 있다 (Jeong et al., 2021). 본 논문에서는 대표적인 이미지 분류 모델인 Inception V3, EfficientNet, DenseNet을 이용하였다. Inception V3는 여러 크기의 컨볼루션 필터를 병렬적으로 사용하여 다양한 스케일의 특징을 추출하는 모델이며, EfficentNet은 복합 계수를 사용하여 네트워크의 깊이, 너비, 해상도를 균형 있게 확장하는 모델이다. DenseNet은 각 레이어가 이전 모든 레이어로부터 추가적인 정보를 받음으로써 특성 정보를 보다 효율적으로 사용하는 모델이다. 세 모델 모두 깊은 네트워크 구조를 이용하여 복잡한 이미지 특성을 추출하고 학습할 수 있으며, 네트워크의 효율성과 성능을 최적화하기 위하여 여러 기법을 도입하였다.

### 4.2 Focal Loss

정확한 동물 소리 분류 모델을 구축하기 위해서는 모델이 학습하는 데이터의 분포를 균일하

게 만들어주어야 한다. 그러나 울음소리는 보통 정해진 장소에 설치된 음향 센서를 통해 수집되기 때문에, 희귀한 종들에 비해 흔히 나타나는 종들의 소리가 편향적으로 수집된다. 종마다 샘플 개수가 다른 불균형 데이터 셋을 이용하면 이를 학습한 분류 모델이 데이터 수가 많은 종의 울음소리에 편향되기에 모델의 성능이 하락하게 된다 (Fernández et al., 2018). 분류 모델의 성능을 향상시키기 위해서는 데이터 불균형 문제에 적합한 손실 함수를 이용하여, 샘플 개수가 적은 클래스에 대해서도 모델이 효율적으로 학습할 수 있도록 해야 한다.

본 논문에서는 동물 소리 데이터 셋에서의 데이터 불균형 문제를 해결하기 위해 CE Loss의 변형된 손실 함수인 Focal Loss를 이용하였다. CE Loss는 잘못 분류한 클래스에 대하여 페널티를 부여하는 손실 함수다. 반면, Focal Loss는 클래스들의 샘플 개수에 따라 분류가 쉬운 클래스와 어려운 클래스를 고려하여 페널티를 부여하기 때문에, 본 논문의 데이터 불균형 문제에 적합한 손실 함수다.

식 (1)과 (2)는 각각 CE Loss와 Focal Loss를 나타낸다. Focal Loss는 CE Loss에 $(1-p_i)^{\gamma}$를 곱함으로써 클래스 샘플 개수가 적어 분류가 어려운 경우에 사용되는 Loss의 가중치를 높인다. 이때, $\gamma \geq 0$의 값을 잘 조절해야 좋은 성능을 얻을 수 있다. 본 논문에서는 Focal Loss를 통해 최적의 성능을 얻기 위해 훈련 셋에서의 5겹 교차검증을 통해 초모수인 $\gamma$ 값을 선택하였다.

$$CE(P,y) = -\sum_{i=1}^{n} y_i \log(P_i) \tag{1}$$

$$FL(P,y) = -\sum_{i=1}^{n} y_i (1-p_i)^{\gamma} \log(P_i) \tag{2}$$

### 4.3 Voting 기반의 앙상블 기법

Fig. 3은 본 논문에서 제안하는 기법의 전체적인 구성을 나타낸다. 본 논문에서는 동물 소리의 다양한 특성을 효과적으로 파악하여 뛰어

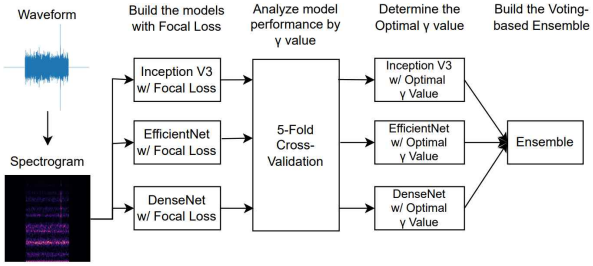



Fig. 3 Overview of the proposed scheme

난 판별력을 가질 수 있는 앙상블 기법을 이용한다. 이를 위해, 다양한 $\gamma$값으로 설정된 Focal Loss를 기반으로 4.1에서 선택한 분류 모델들을 구축한다. 이후, 훈련 셋에서의 5겹 교차검증을 통해 $\gamma$값에 따른 분류 모델의 성능을 측정한다. 이 과정을 통해 각 모델에서 최적의 성능을 도출할 수 있는 $\gamma$값을 결정한다. 최종적으로, 학습된 모델들의 예측 결과를 결합하는 앙상블 기법을 구축한다 (Lee et al., 2020).

Voting 기반의 앙상블 기법은 서로 다른 알고리즘을 가진 분류 모델들의 예측 결과 중 투표를 통해 최종 예측 결과를 도출한다. 구체적인 기법으로는 Hard Voting, Soft Voting, Weighted Soft Voting 등이 있다 (Mohammed et al., 2023).

Hard Voting은 여러 모델이 예측한 결과 중 다수의 분류 모델이 결정한 클래스 레이블을 최종 결과로 결정한다. Hard Voting은 식 (3)과 같이 정의되며, 간단하고 구현이 쉽다는 장점이 있지만, 모든 분류 모델이 동등한 중요도를 가진다고 가정하기에 모든 분류 모델이 비슷한 성능을 보여야 최적의 예측 결과를 얻을 수 있다. 각 분류 모델의 성능이 다를 경우, 성능이 좋지 않은 분류 모델이 앙상블의 전체적인 성능에 오히려 부정적인 영향을 끼칠 수 있다는 단점이 있다 (Mohammed et al., 2023).

$$\hat{y}_{Ensemble} = \underset{i}{argmax}\left(\sum_{j=1}^{n} I(\hat{y_j} = i)\right), \quad i \in \{1, \text{Number of Classes}\} \tag{3}$$

Soft Voting은 클래스별로 모델들이 예측한 확률값들을 평균화하고, 이중 가장 높은 확률값

을 가지는 클래스를 최종 결과로 결정한다. Soft Voting은 식 (4)와 같이 정의되며, 각 분류 모델의 예측 결과를 단순히 다수결로 선택하는 Hard Voting과 달리, 확률값을 이용하여 더욱 정교한 예측을 할 수 있어 일반적으로 Hard Voting보다 더 좋은 성능을 보이는 것으로 알려져 있다 (Mohammed et al., 2023).

$$\hat{y}_{Ensemble} = \underset{i}{argmax}\left(\frac{1}{n}\sum_{j=1}^{n} P(y=i)\right), \quad i \in \{1, \text{Number of Classes}\} \tag{4}$$

Weighted Soft Voting은 각각의 모델이 예측한 확률값에 대하여 가중치를 부여한 다음, 이를 평균화하였을 때 가장 높은 확률값을 가지는 클래스를 최종 결과로 결정한다. 이때 각각의 모델에 대해 얼마만큼의 가중치를 부여할 것인지를 결정하는 것이 중요한데, 일반적으로 훈련 셋에서 각 모델이 도출한 분류 성능에 비례하여 가중치를 부여하게 된다. 식 (5)와 같이 표현되는 Weighted Soft Voting은 가중치를 조정하여 각 분류 모델의 기여도를 조절할 수 있기 때문에, 앙상블 기법의 성능을 최적화하는 데 유용하게 이용될 수 있다 (Mohammed et al., 2023).

$$\hat{y}_{Ensemble} = \underset{i}{argmax}\left(\frac{\sum_{j=1}^{n}(TrainAcc_j)\, I(\hat{y_j}=i)}{\sum_{j=1}^{n}(TrainAcc_j)}\right), \quad i \in \{1, \text{Number of Classes}\} \tag{5}$$

본 논문에서는 Focal Loss에서의 $\gamma$값을 선택하기 위해 진행한 5겹 교차검증에서 측정한 분류 모델별 정확도를 가중치로 설정한 Weighted Soft Voting 기반의 앙상블 기법을 사용하였다. 본 기법이 여러 Voting 기반 앙상블 기법 중 가장 우수한 분류 성능을 보였다.

## 5. 실험 및 결과

### 5.1 실험 환경

제안하는 기법의 효과성을 검증하는 실험을 위해, 동물 소리 데이터 셋을 8:2의 비율로 훈련 셋과 평가 셋으로 분할하였다. 그리고 평가 지표로는 대표적인 분류 성능 평가 지표인 Precision, Recall, F1 Score, Accuracy를 선택하였다 (Aggarwal, 2014). 실험에 사용한 모든 소리 분류 모델들은 훈련 셋에서 최대 50 Epoch 훈련하였다. 본 실험은 Intel i7-9700 CPU와 NVIDIA GeForce RTX 2080 Ti GPU 환경에서 수행하였다. 개발 언어는 Python 3.7.13과 PyTorch 1.13.1 (Paszke et al., 2019) 프레임워크를 사용하였다.

### 5.2 Focal Loss의 $\gamma$값에 따른 성능 민감도 분석

Focal Loss의 $\gamma$값은 [0.5, 5.0] 범위 내에서 선택하여야 좋은 성능을 보인다고 알려져 있다 (Lin et al., 2017). 그러므로, 본 실험에서는 Focal Loss의 $\gamma$값을 다양하게 설정하여 모델의 분류 성능 변화를 측정하였다. Table 1은 5겹 교차검증을 수행하여 $\gamma$값에 따라 달라지는 모델의 Accuracy를 측정한 결과를 나타낸다. 표에서 볼 수 있듯이, 모든 모델은 $\gamma$값이 2.0일 때 가장 좋은 분류 성능을 보였으므로, 해당 값으로 설정된 Focal Loss를 사용하여 최종적인 앙상블을 위한 기반 모델들을 구축하였다.

Table 1 Comparison of 5-Fold Cross-Validation about Gamma Value of Focal Loss

| Model | ɣ=0.5 | ɣ=1.0 | ɣ=2.0 | ɣ=5.0 |
|---|---|---|---|---|
| Inception V3 | 0.3820 | 0.8001 | **0.8017** | 0.7735 |
| EfficientNet | 0.3820 | 0.7156 | **0.7225** | 0.7059 |
| DenseNet | 0.2024 | 0.8487 | **0.8490** | 0.8477 |

### 5.3 동물 소리 분류 모델 성능 평가

동물 소리 분류 모델을 구축하는 데 있어서의 모델의 유형과 손실 함수를 변경해가며 분류 모델의 성능을 평가하였다. Table 2는 사용된 분류 모델들의 Precision, Recall, F1 Score 및 Accuracy를 나타낸다. Precision과 Recall 측면에서 CE Loss 기반의 모델들에 비해 Focal Loss 기반 모델들이 더 좋은 성능을 달성하였다.

Table 2 Comparison of the performance of classification model

| Model | Precision | Recall | F1 Score | Accuracy |
|---|---|---|---|---|
| Inception V3 w/ CE Loss | 0.7287 | 0.6788 | 0.6946 | 0.8711 |
| EfficientNet w/ CE Loss | 0.6093 | 0.5989 | 0.5945 | 0.8223 |
| DenseNet w/ CE Loss | 0.7285 | 0.7968 | 0.7467 | 0.8736 |
| Inception V3 w/ Focal Loss | 0.6706 | 0.7499 | 0.6904 | 0.8548 |
| EfficientNet w/ Focal Loss | 0.6542 | 0.7002 | 0.6571 | 0.8323 |
| DenseNet w/ Focal Loss | 0.7311 | 0.7393 | 0.7330 | 0.8974 |
| Hard Voting-based Ensemble | 0.7371 | 0.7978 | 0.7569 | 0.8999 |
| Soft Voting-based Ensemble | 0.7314 | 0.8475 | 0.7646 | 0.8986 |
| Proposed Scheme | **0.7411** | **0.8478** | **0.7724** | **0.9024** |

한편, 제안 기법은 다른 모델들과 비교해 모든 평가지표면에서 가장 우수한 수치를 달성하였다.

단일 모델은 조류 울음소리에서의 많은 배경 잡음으로 인해 판별력이 떨어진 모습을 보였다. 반면, 제안 기법은 Weighted Soft Voting을 통해 각 분류 모델의 성능을 고려한 가중치에 기반하여 유연하게 여러 분류 모델의 예측 결과들을 결합함으로써 개별 모델이 조류 울음소리에서 잡음의 영향을 받는 것을 최소화하였다. 이를 통해, 제안 기법은 데이터 노이즈에 대한 강인성과 일반화 성능을 향상시킬 수 있었다. 그 결과, 단일 모델의 평균 성능보다 제안된 앙상블 기법이 Recall 기준으로 최대 22.6% 향상된 분류 성능을 보였고, 앙상블 기법 중에서는 제안된 기법이 타 Voting 기반 앙상블 기법보다 Recall 기준으로 최대 6.2% 향상된 분류 성능을 달성할 수 있었다.

### 5.4 데이터 샘플링 기반 동물 소리 분류 기법 성능 평가

학습에 참여하는 데이터의 샘플링 빈도를 다르게 주는 기법인 언더샘플링과 오버샘플링을 이용하여 동물 소리 분류 기법의 성능을 평가하였다. 언더샘플링이란 다수 클래스의 샘플 수를 줄여 전체 데이터 셋의 클래스 간 균형을 맞추는 방법이며, 오버샘플링이란 소수 클래스의 샘플 수를 증가시켜 데이터 셋의 클래스 간 균형을 맞추는 방법이다. 본 실험에서는 오디오 신호의 시간 축을 일정한 비율로 변형하는 기법인 시간 워핑(Time Warping)을 이용한 오버샘플링을 진행하였다 (Kahl et al., 2021). 구체적으로, 시간 워핑 구현 방법론 중 하나인 위상 보코더(Phase Vocoder)를 이용하였다 (Prusa et al., 2022). 위상 보코더 기법은 오디오 신호의 피치를 보존하면서 시간 축을 조정할 수 있는 알고리즘이다. 이 방법은 고속 푸리에 변환을 통해 오디오를 주파수 영역으로 변환하고, 주파수 성분의 위상을 조정하여 신호의 속도를 변화시킨다. 이후 역 고속 푸리에 변환을 통해 변형된 주파수 영역의 신호를 다시 시간 영역으로 변환하여 최종적으로 조정된 오디오 신호를 얻는다. 본 실험에서는 오디오 샘플에서 임의로 지정된 주파수 영역에 대하여 0.5배에서 2.0배 사이의 속도 변환을 적용하여 오버샘플링을 진행하였다. Table 3, 4는 각각 언더샘플링과 오버샘플링을 이용한 분류 기법의 Precision, Recall, F1 Score 및 Accuracy를 나타낸다. 언더샘플링과 오버샘플링 모두 제안 기법과 비교해 낮은 분류 성능을 보였다.

언더샘플링은 다수 클래스와 소수 클래스 사이의 많은 샘플 수 차이로 인해 판별력이 크게 떨어진 모습을 보였다. 가장 적은 샘플 수가 있는 소수 클래스는 Canada Warbler (CAWA)로, 14개의 샘플 수가 존재한다. 다른 모든 클래스의 샘플 수를 가장 적은 샘플 수가 있는 소수 클래스에 맞춰 줄였기에 학습 데이터의 양이 크게 줄어들었다. 많은 양의 학습 데이터가 필요

Table 3 Comparison of the performance of classification scheme - Undersampling

| Model | Precision | Recall | F1 Score | Accuracy |
|---|---|---|---|---|
| Inception V3 | 0.2528 | 0.2418 | 0.1439 | 0.1139 |
| EfficientNet | 0.1970 | 0.2582 | 0.1298 | 0.1039 |
| DenseNet | 0.2506 | 0.2743 | 0.0946 | 0.1001 |

Table 4 Comparison of the performance of classification scheme - Oversampling

| Model | Precision | Recall | F1 Score | Accuracy |
|---|---|---|---|---|
| Inception V3 | 0.7131 | 0.8024 | 0.7388 | 0.8974 |
| EfficientNet | 0.6956 | 0.7489 | 0.7084 | 0.8861 |
| DenseNet | 0.7243 | 0.7546 | 0.7356 | 0.8986 |

Table 5 Comparison of the performance of classification scheme - Hard Voting-based Ensemble

| Model | Precision | Recall | F1 Score | Accuracy |
|---|---|---|---|---|
| Inception V3 + DenseNet | 0.7317 | 0.7375 | 0.7305 | 0.8849 |
| Inception V3 + EfficientNet | 0.6897 | 0.6996 | 0.6824 | 0.8473 |
| EfficentNet + DenseNet | 0.7145 | 0.7102 | 0.7050 | 0.8748 |
| All Models | **0.7371** | **0.7978** | **0.7569** | **0.8999** |

Table 6 Comparison of the performance of classification scheme - Soft Voting-based Ensemble

| Model | Precision | Recall | F1 Score | Accuracy |
|---|---|---|---|---|
| Inception V3 + DenseNet | 0.7312 | 0.7571 | 0.7402 | **0.9011** |
| Inception V3 + EfficientNet | 0.6904 | 0.7955 | 0.7216 | 0.8773 |
| EfficentNet + DenseNet | 0.6938 | 0.7795 | 0.7195 | 0.8849 |
| All Models | **0.7314** | **0.8475** | **0.7646** | 0.8986 |

Table 7 Comparison of the performance of classification scheme - Weighted Soft Voting-based Ensemble

| Model | Precision | Recall | F1 Score | Accuracy |
|---|---|---|---|---|
| Inception V3 + DenseNet | 0.7318 | 0.7545 | 0.7397 | 0.9011 |
| Inception V3 + EfficientNet | 0.7053 | 0.8250 | 0.7398 | 0.8849 |
| EfficentNet + DenseNet | 0.7075 | 0.7455 | 0.7215 | 0.8911 |
| All Models | **0.7411** | **0.8478** | **0.7724** | **0.9024** |

한 딥러닝 모델의 특성상, 데이터 불균형 문제를 해결하기 위해 언더샘플링을 한 것이 오히려 분류 성능의 하락을 유발하였다.

오버샘플링은 증강된 데이터가 원본 데이터의 특성을 명확히 반영하지 못해 판별력이 크게 오르지 못한 모습을 보였다. 데이터 셋에서 드물게 존재하는 소수 클래스 자체의 원본 샘플 수가 적기에, 소수 클래스의 데이터를 증강하는 속도 변환 과정에서 원본 데이터의 중요한 특성이 왜곡되어 판별에 핵심적인 증강 데이터를 다수 확보할 수 없었으므로 분류 성능 개선에 있어 더 나은 영향을 미치지 못하였다. 결과적으로, 본 연구에서는 언더샘플링이나 오버샘플링과 같은 변중성을 제한하는 방법보다 Focal Loss를 통해 모델을 학습하고 분류하는 방법이 Table 2와 같이 Precision, Recall, F1 Score, Accuracy 평가에서 더 나은 성능을 보였다.

### 5.5 앙상블 학습 기반 동물 소리 분류 기법 성능 평가

Voting 기반의 앙상블 기법을 구축하는 데 있어 사용한 모델의 유형과 개수를 변경해가며 분류 기법의 성능을 평가하였다. Table 5, 6, 7은 각각 Hard Voting, Soft Voting, 그리고 Weighted Soft Voting을 이용한 앙상블 기반 분류 기법들의 Precision, Recall, F1 Score 및 Accuracy를 나타낸다. Hard Voting과 Weighted Soft Voting 기반 앙상블 기법은 주어진 모델을 모두 이용하였을 때 타 모델 조합과 비교해 모든 평가지표 면에서 가장 우수한 수치를 달성하였다. Soft Voting은 주어진 모델을 모두 이용하였을 때 Precision, Recall과 F1 Score 측면에서 타 모델 조합과 비교해 더 좋은 성능을 달성하였다.

모든 모델을 이용하여 조합한 경우, 조류 울음소리에서의 잡음의 영향을 최소화하였기에 가장 우수한 판별력을 달성한 모습을 보였다. 두 모델을 이용하여 조합한 경우 역시 잡음의 영향을 줄일 수 있었으나, 그 정도가 모든 모델을 이용하여 조합한 경우에 비해 적은 모습을 보였다. 이를 통해, 제안 기법은 각 모델의 예측을 종합함으로써 조류 울음소리에서의 잡음에 의한 오류를 감소시키고, 전반적인 분류 정확도를 향상시키는 데 기여할 수 있었다. 그 결과, 모든 모델을 이용하여 조합하였을 때, 타 모델 조합보다 Recall 기준으로 최대 13.7% 향상된 분류 성능을 달성할 수 있었다.

## 6. 결 론

본 논문에서는 Focal Loss와 앙상블 기반의 동물 소리 분류 기법을 제안하였다. 제안하는 기법의 효용성 검증을 위하여, 공개 조류 울음소리 데이터 셋을 기반으로 Focal Loss에 대한 제안 기법의 성능 민감도를 분석하였고, 다양한 평가지표를 기준으로 다른 동물 소리 분류 모델들과 성능을 비교하였다. 또한, 데이터 샘플링을 통해 데이터 불균형을 해소하는 기법과 분류 성능을 비교함과 동시에, 동물 소리에서의 잡음의 영향력을 최소화하기 위한 최적의 앙상블 모델 조합을 탐색하였다. 실험 결과, 제안 기법은 최적의 Focal Loss를 설정하여 데이터 불균형 문제를 해소할 수 있었으며, Weighted Soft Voting 기반의 앙상블 기법을 통해 많은 잡음을 포함하고 있는 동물 소리를 다른 비교 모델들보다 더욱 효과적으로 분류할 수 있었다.

Focal Loss는 데이터 불균형 문제를 해소하는 데 효과적이지만, 균형 데이터에서도 Focal Loss에 기반하여 일부 클래스에 대한 가중치를 변경한다면 모델이 특정 클래스에 대한 성능을 더욱 개선할 수 있다 (Lin et al., 2017). 따라서 데이터의 상황에 적합한 손실 함수를 선택하고 이를 조정할 필요가 있다. 향후 연구에서는 자동화된 앙상블 기법을 개발하여 사용자가 별도의 초모수와 가중치를 조정하는 과정이 없이도 최적의 앙상블 기법을 구축할 수 있도록 할 계획이다.

## References

Aggarwal, C. C. (2014). Data Classification:

Algorithms and Applications, *CRC Press*.

Cramer, A., Lostanlen, V., Farnsworth, A., Salamon, J. and Bello, J. P. (2020). Chirping up the Right Tree: Incorporating Biological Taxonomies into Deep Bioacoustic Classifiers. *IEEE International Conference on Acoustics, Speech, and Signal Processing*. May. 04-08, Barcelona, Spain, pp. 901-905.

Fernández, A., García, S., Galar, M., Prati, R. C., Krawczyk, B. and Herrera, F. (2018). Learning from Imbalanced Data Sets, *Springer*.

Ganaie, M. A., Hu, M., Malik, A. K., Tanveer, M. and Suganthan, P. N. (2022). Ensemble Deep Learning: A Review. *Engineering Applications of Artificial Intelligence*, 115. https://doi.org/10.1016/j.engappai.2022.105151

Gunawan, K. W., Hidayat, A. A., Cenggoro, T. W. and Pardamean, B. (2023). Repurposing Transfer Learning Strategy of Computer Vision for Owl Sound Classification. *Procedia Computer Science*, 216, 424-430. https://doi.org/10.1016/j.procs.2022.12.154

Hidayat, A. A., Cenggoro, T. W. and Pardamean, B. (2021). Convolutional Neural Networks for Scops Owl Sound Classification. *Procedia Computer Science*, 179. https://doi.org/10.1016/j.procs.2020.12.010

Huang, G., Liu, Z., Maaten, L. and Weinberger, K. Q. (2017). Densely Connected Convolutional Networks. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Jul. 21-26, Honolulu, HI, USA, pp. 4700-4708.

Incze, A., Jancso, H., Szilagyi, Z., Farkas, A. and Sulyok, C. (2018). Bird Sound Recognition Using a Convolutional Neural Network. *IEEE 16th International Symposium on Intelligent Systems and Informatics*, Sep. 13-15, Subotica, Serbia, pp. 295-300.

Jeong H., Go, J. and Shin, C. (2021). Abnormal Detection with Microscope through Deep Learning. *Journal of Korea Society of Industrial Information Systems*, 26(2), https://doi.org/10.9723/jksiis.2021.26.2.001

Kahl, S., Wood, C., Eibl, M. and Klinck, H. (2021). BirdNET: A Deep Learning Solution for Avian Diversity Monitoring. *Ecological Informatics*, 61, https://doi.org/10.1016/j.ecoinf.2021.101236

Kim, C., Cho, Y., Jung, S., Rew, J. and Hwang, E. (2020). Animal Sounds Classification Scheme based on Multi-Feature Network with Mixed Datasets. *KSII Transactions of Internet and Information Systems*, 14(8), 3384-3398, https://doi.org/10.3837/tiis.2020.08.013

Kim, E., Moon, J., Shim, J. and Hwang, E. (2023). DualDiscWaveGAN-Based Data Augmentation Scheme for Animal Sound Classification. *Sensors*, 23(4), https://doi.org/10.3390/s23042024

Kim, J., Seok, C., Kim, M. and Kim, S. (2022). A System for Recommending Audio Devices based on Frequency Band Analysis of Vocal Component in Sound Source. *Journal of Korea Society of Industrial Information Systems*, 27(6), 1-12, https://doi.org/10.9723/jksiis.2022.27.6.001

Kim, J., Lee, Y., Kim, D. and Ko, H. (2020). Temporal Attention based Animal Sound Classification. *The Journal of the Acoustical Society of Korea*, 39(5), 406-413, https://doi.org/10.7776/ASK.2020.39.5.406

Koh, C., Chang, J., Tai, C., Huang, D., Hsieh, H. and Liu, Y. (2019). Bird Sound Classification using Convolutional Neural Networks. *Conference and Labs of the Evaluation Forum*, Sep. 9-12, Lugano, Switzerland, 2380

Korea Forest Service (2023). Changes in Forests due to Climate Change,

https://www. forest.go.kr/ (Accessed on Jan. 03rd, 2024)

Lee, W., Kim, Y., Kim, J. and Lee, C. (2020). Forecasting of Iron Ore Prices using Machine Learning. *Journal of Korea Society of Industrial Information Systems*, 25(2), 57-72, https://doi.org/10.9723/jksiis.2020.25.2.057

Lin, T., Goyal, P., Girshick, R., He, K. and Dollár, P. (2017). Focal Loss for Dense Object Detection. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, Oct. 22-29, Venice, Italy, pp. 2980-2988.

Martynov, E. and Uematsu, Y. (2022). Dealing with Class Imbalance in Bird Sound Classification. *Conference and Labs of the Evaluation Forum*, Sep. 5-8, Bologna, Italy, pp. 2151-2158.

Mohammed, A. and Kora, R. (2023). A Comprehensive Review on Ensemble Deep Learning: Opportunities and Challenges. *Journal of King Saud University - Computer and Information Science*, 35(2), 757-774, https://doi.org/10.1016/j.jksuci.2023.01.014

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L, Bai, J., and Chintala, S. (2019). PyTorch: An Imperative Style, High-Performance Deep Learning Library. *Proceedings of the Advances in Neural Information Processing Systems*, Dec. 08-14, Vancouver, BC, Canada, pp. 8026-8037.

Prusa, Z. and Holighaus, N. (2022). Phase Vocoder Done Right. *arXiv*, arXiv:2202.07382, https://doi.org/10.48550/arXiv.2202.07382

Sun, Y., Maeda, T. M., Solís-Lemus, C., Pimentel-Alarcón, D. and Buřivalová, Z. (2022). Classification of Animal Sounds in a Hyperdiverse Rainforest using Convolutional Neural Networks with Data Augmentation. *Ecological Indicators*, 145. https://doi.org/10.1016/j.ecolind.2022.109621

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J. and Wojna, Z. (2016). Rethinking the Inception Architecture for Computer Vision. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Jun. 26-Jul 01, Las Vegas, NV, USA, pp. 2818-2826.

**이 재 승 (Jaeseung Lee)**

- 고려대학교 통계학과 학사
- (현재) 고려대학교 전기전자공학과 석사과정
- 관심분야: 통계적 머신러닝, 인공지능, 그래프 신경망

**유 제 혁 (Jehyeok Rew)**

- 정회원
- 경희대학교 전자전파공학과 학사
- 고려대학교 전기전자컴퓨터공학과 박사
- 덕성여자대학교 데이터사이언스학과 조교수
- 관심분야: GIS, 정보검색, 빅데이터, 인공지능