Issue 14, 2023

https://doi.org/10.23939/sisn2023.14.297

УДК 004.9

ІДЕНТИФІКАЦІЯ ГОЛОСІВ ПТАХІВ ЗА ДОПОМОГОЮ ЗГОРТКОВИХ НЕЙРОННИХ МЕРЕЖ З ВИКОРИСТАННЯМ STFT ТА MEL СПЕКТРОГРАМ

Оксана Гонсьор[1], **Юрій Гонсьор**[2]

[1] Національний університет "Львівська політехніка", кафедра спеціалізованих комп'ютерних систем, вул. С. Бандери, 12, Львів Україна

[2] Національний університет "Львівська політехніка", кафедра систем штучного інтелекту, вул. С. Бандери, 12, Львів, Україна

[1] E-mail: oksana.y.honsor@lpnu.ua, ORCID: 0000-0003-0895-5859

[2] E-mail: yurii.honsor.knm.2020@lpnu.ua ORCID: 0009-0001-7486-6689

Загрози для клімату та глобальні зміни в екологічних процесах залишаються актуальною проблемою у всьому світі. Тому важливий постійний моніторинг цих змін, зокрема із використанням нестандартних підходів. Це завдання можна виконати на основі дослідження інформації про міграцію птахів. Одним із ефективних методів дослідження міграції птахів є слуховий метод, який потребує вдосконалення. Ось чому побудова моделі на основі методів машинного навчання, яка допоможе точно ідентифікувати наявність голосів птахів у аудіофайлі з метою дослідження міграцій птахів з певної території, є актуальною проблемою. У цій роботі розглянуто способи побудови моделі машинного навчання на основі аналізу спектрограм, яка допоможе точно ідентифікувати наявність голосів птахів в аудіофайлі з метою дослідження міграції птахів по визначеній території. Дослідження передбачає збирання та аналіз аудіофайлів, які можна використати для виявлення характеристик, відповідно до яких звук файлів буде ідентифікуватись як голоси птахів або відсутність звуку у файлі. Продемонстровано використання моделі CNN для класифікації наявності голосів птахів у аудіофайлі. Аналіз ефективності та точності моделі CNN в класифікації звуків у аудіофайлах показав, що краще використовувати Mel-спектрограми, ніж STFT-спектрограми, для дослідження та класифікації наявності звуків птахів у середовищі. Точність класифікації моделі, тренованої на основі Mel-спектрограм, становила 72 %, що на 8 % вище, ніж точність моделі, натренованої на STFT-спектрограмах.

Ключові слова: машинне навчання; ідентифікація звуку; спектрограма; згорткова нейронна мережа.

Вступ

Дослідження міграції птахів – важлива складова для аналізу глобальних екологічних процесів. За допомогою міграції птахів можна проаналізувати різноманітні екологічні чинники, що охоплюють велику ділянку землі, такі як: температура, забруднення повітря, забруднення в річках та озерах. Одним із ефективних методів дослідження міграції птахів є слуховий метод, що використовує параболічний відбивач із прикладеним мікрофоном задля підсилення голосових сигналів птахів та їх запису. Саме тому актуальна побудова моделі на основі методів машинного навчання, яка до-

поможе точно ідентифікувати наявність голосів птахів у аудіофайлах з метою дослідження міграцій птахів з певної території. Важливий технологічний аспект – вивчення методів аналізу голосів птахів у середовищі за допомогою моделі CNN та встановлення його переваг та недоліків.

Постановка проблеми

Використання слухового методу ідентифікації голосів птахів із обширної території передбачає генерування великої кількості аудіофайлів, що породжує проблему аналізу наявності звуків птахів у цих аудіофайлах, який виконують люди. Попри те, що у людини слух доволі розвинений для класифікації звуків, висока ймовірність неточності в ідентифікації голосу птаха серед інших шумів, а також людський фактор. Тому ефективним вирішенням цієї проблеми є створення моделі машинного навчання для аналізу аудіофайлів, що дасть можливість швидше та точніше встановлювати наявність чи відсутність голосів певних птахів у певному файлі.

Об’єктом дослідження є модель машинного навчання для точної та швидкої класифікації аудіофайлу та наявність чи відсутність звуків птахів у цьому файлі з методою вивчення міграцій птахів.

Предметом дослідження є ефективність та точність моделі CNN для класифікації гоосів птахів у аудіофайлах та вибір найкращого класифікатора для цього типу файлів та моделі.

Аналіз останніх досліджень та публікацій

Фундаментальні поняття щодо основних принципів організації та роботи CNN висвітлено в працях [1–7], в яких ґрунтовно описано використання такого типу нейронної мережі для класифікації. Хоча праці стосуються здебільшого класифікації об’єктів на основі зображень, в них досконало описано математичний апарат згорткових нейронних мереж, принцип побудови необхідної моделі та детальний огляд усіх її елементів, а саме: попереднє опрацювання даних, необхідні нейронні шари для побудови, функції активації нейронів, використання додаткових шарів для регуляризації CNN та вибір оптимізатора для мережі.

Попереднє опрацювання аудіофайлів для класифікації за допомогою CNN ґрунтовно описано в [8–11]. Робота дає чітке уявлення про особливості використання методів перетворення звукових файлів на спектрограму. Тут чітко описаний математичний апарат методів цього перетворення для подальшої екстракції ознак, а також порівняння цих методів. Також автор порівнює використання спектрограм із відтінками сірого (greyscale) та кольором та зазначає, що використання кольорових спектрограм для класифікації краще у середньому на 21 %.

В [12] розглянуто використання конкретно CNN (DCNN) для аналізу аудіофайлів, а саме звуків різних машин на будівельному майданчику. Описано архітектуру нейронної мережі, що буде враховано для побудови архітектури NN для цього дослідження. Також у цій статті викладено порівняльний аналіз класифікації аудіофайлів різними моделями ШІ: Random Forest, MLP, k-NN, SVM та DCNN, яка забезпечила точність близько 97 %.

Праці [13–17] висвітлюють особливості використання нейронних мереж глибинного навчання для розпізнавання різних звуків у середовищі (звуки натовпу, дорожнього руху, аплодисментів та музики). У цій статті наведено порівняння глибинних нейронних мереж на основі двох гіперпараметрів: кількості нейронів у прихованому шарі та вибору класифікатора. Проаналізувавши наведені дані, отримані внаслідок тренування цих моделей, ми зробили висновок, що використання DNN дає менш точну класифікацію об’єктів на основі аудіофайлів, адже середня похибка натренованої нейронної мережі коливається в межах 14,76–17,07 %, а це гірше, ніж точність, описана в праці [12]. Особливості використання MEL-спектрограм детально розглянуто в [18]. Також важлива базова інформація щодо глибокого навчання нейромереж для класифікації аудіо за допомогою Tensorflow, висвітлена в [19, 20].

Формулювання цілі статті

Мета цієї роботи – побудова моделі, яка допоможе точно ідентифікувати наявність голосів птахів у аудіофайлах, щоб дослідити міграції птахів із цієї території. Для досягнення цієї мети потрібно виконати такі завдання:

· проаналізувати методи класифікації аудіофайлів на підставі їх спектрограм;

· виявити основні переваги та недоліки використання CNN для аналізу спектрограм аудіофайлів;

· опрацювати набір аудіофайлів для отримання спектрограм з метою подальшого тренування моделі машинного навчання;

· побудувати модель машинного навчання для аналізу спектрограм;

· порівняти використання різних оптимізаторів для моделі;

· виконати тестування моделі та оптимізувати її до бажаної точності;

· на основі аналізу ефективності отриманої моделі зробити висновки та сформувати рекомендації щодо напрямів подальших досліджень.

Виклад основного матеріалу

Згорткові нейронні мережі (*Convolutional neural networks, CNN*) – це клас глибинних штучних нейронних мереж прямого поширення, який успішно застосовували для аналізу візуальних зображень. Основними шарами в CNN є згорткові, або конволюційні шари. Параметри цих шарів складаються з набору фільтрів для навчання (або ядер), які мають невеличке рецептивне поле, але простягаються на всю глибину вхідної ємності. Під час прямого проходу кожен фільтр здійснює згортку за шириною та висотою вхідної ємності, обчислюючи скалярний добуток даних фільтра та входу і формуючи двовимірну карту збудження цього фільтра. В результаті мережа навчається, які фільтри активуються, коли вона виявляє певний конкретний тип ознаки у певному просторовому положенні на вході.

Спектрограма сигналу – це візуальне зображення спектра частот сигналу в часі. Найпоширенішим відображенням спектрограми є двовимірна діаграма: на горизонтальній осі відкладено час, вертикальної осі – частота; третій вимір із зазначенням амплітуди на певній частоті в конкретний момент часу представлено інтенсивністю або кольором кожної точки зображення.

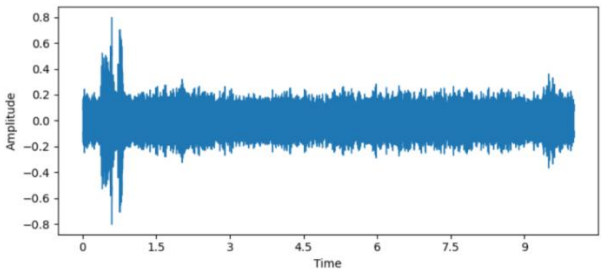


Рис. 1. Приклад звукової хвилі для аудіофайлу зі звуком птахів

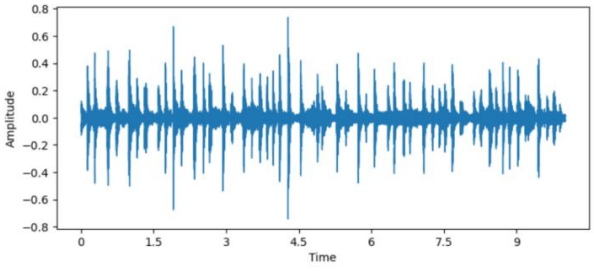


Рис. 2. Приклад звукової хвилі для аудіофайлу без звуку птахів

Використання CNN для аналізу аудіофайлів, а саме їх спектрограм, є оптимальним вибором між точністю та швидкістю аналізу аудіофайлу.

Для побудови нейронної мережі використано датасет із платформи Kaggle [21], який містить 7690 аудіофайлів у форматі .wav. Інформація про наявність чи відсутність звуку птахів у аудіофайлах була збережена в окремому файлі metadata.csv. Для побудови моделі дані із датасету були використані для тренування та оцінювання ефективності нейронної мережі.

Опис методів дослідження

Основною операцією CNN є конволюція.

Конволюція – математична операція двох функцій $f(t)$ та $g(t)$, що дає змогу отримати третю функцію.

Для неперервного випадку формула конволюції виглядає так:

виконується аналіз бази відомих фактів (retrieve).

$$(f \times g)(t) = \int_{-\infty}^{+\infty} f(t-\tau)g(\tau)\,d\tau, \quad (1)$$

де $(f \times g)(t)$ – результат конволюції у момент часу t; $f(t-\tau)$ – значення функції f у момент часу $t-\tau$; $g(\tau)$ – значення функції g у момент часу τ.

Інтегрування здійснюється за всіма можливими значеннями τ.

Для дискретного випадку, що застосовується для опрацювання зображень, формула конволюції зображення, що подається двовимірною матрицею розміром M×N, та ядра конволюції, що представляється двовимірною матрицею M×N, матиме такий вигляд:

$$(f * g)[i,j] = \sum_{m=-\infty}^{\infty} \sum_{n=-\infty}^{\infty} f[m,n] \times g[i-m, j-n], \quad (2)$$

де $(f \times g)[i,j]$ – значення пікселя результуючого зображення після конволюцій матриць f та g у точці (i, j); *f[m,n]* – значення пікселя у матриці f в точці (m, n); *g[i–m, j–n]* – значення пікселя у матриці g у точці $(i–m, j–n)$.

Підсумування здійснюється за усіма можливими значеннями m та n.

Під час аналізу зображень конволюція використовується для накладання певної матриці конволюції (ядра) на початкове зображення із метою виділення певних ознак на зображенні (ліній, границь, кутів), розмиття чи збільшення чіткості зображення тощо.

У разі використання двовимірної матриці конволюції розмірністю $M_x \times M_y$ на зображення, подане у вигляді двовимірної матриці розміром $N_x \times N_y$, застосовується формула конволюції для дискретного випадку, що має вигляд

$$C[i,j] = \sum_{u=0}^{m_x-1} \sum_{v=0}^{m_y-1} A[i+u, j+v] \times B[u,v], \quad (3)$$

де *C[i, j]* – значення пікселя результуючого зображення після конволюції початкового зображення A та матриці конволюції B у точці *(i,j)*; A$[i+u, j+v]$ – значення пікселя у матриці зображення A у точці $(i+u, j+v)$; $B[u, v]$ – значення матриці конволюції у точці (u, v). Підсумування здійснюється для кожного пікселя (i,j) початкового зображення за усіма значеннями *(u,v)* матриці конволюції.

Завантаження та обробка аудіофайлу

Аудіофайл являє собою запис звукової хвилі. Оскільки це хвиля, її амплітуда змінна у кожний проміжок часу, тому під час завантаження аудіофайлу в програму ця звукова хвиля записується значенням амплітуди – сили звуку, в кожний проміжок часу у вигляді масиву. Кількість значень у масиві варіюється залежно від розміру аудіофайлу та частоти дискретизації сигналу.

Частота дискретизації в цьому випадку є важливим параметром, адже саме від неї залежить кількість значень, якими буде задаватись кожна секунда аудіофайлу.

Для прикладу, якщо тривалість аудіофайлу становитиме 100 секунд, а частота дискретизації – 44100, то загалом дискретна множина міститиме 441000 значень амплітуди сигналу.

Ще однією важливою характеристикою звукового сигналу є частота звуку у певний проміжок часу. Частота звуку – це кількість коливань певної точки середовища, в якому поширюються звукові хвилі, за секунду. Саме за допомогою цієї характеристики людина відрізняє усі звуки. Чим вища частота звуку – тим вищий тон або висота звуку.

У цій роботі використано обрізання частот, нижчих від 1 кГц, адже звук більшості птахів варіюється від 1 до 10–15 кГц. Частоти, нижчі за 1 кГц, можна прийняти за шум, що лише спотворюватиме подальшу класифікацію.

Для обрізання необхідних частот зазвичай використовують такі фільтри: фільтр Бесселя. фільтр Чебишова, фільтр Кайзера та фільтр Баттерворта.

Основне призначення цих фільтрів у обробленні звуку – виділення із вхідних частот бажаних та послаблення (заглушення) інших частот.

Фільтр Баттерворта – тип фільтра з максимально плоскою амплітудно-частотною характеристикою в області пропускання.

Фільтр Баттерворта має певні переваги над іншими фільтрами, а саме:

1. Рівномірна амплітудно-частотна характеристика: фільтр Баттерворта забезпечує рівномірне загасання амплітуди в усьому діапазоні частот, тобто він не спотворює амплітуди сигналу в заданому діапазоні частот.
2. Широкий діапазон настроювання: фільтр Баттерворта можна налаштувати на різні параметри загасання і частоти зрізу, що дає змогу гнучко контролювати його характеристики залежно від необхідних вимог.
3. Лінійна фазова характеристика: фільтр Баттерворта має лінійну фазову характеристику. Це означає, що фазова затримка сигналу залежить від частоти лінійним способом. Це дає змогу зберігати відносини фаз між різними компонентами сигналу.

Амплітудно-частотна характеристика фільтра Баттерворта виражається такою формулою:

$$G^2(\omega) = \frac{G_0^2}{1 + \left(\frac{\omega}{\omega_c}\right)^{2n}}, \tag{4}$$

де n – порядок фільтра Баттерворта; ω_c – частота зрізу (границя пропускання сигналу певної частоти); G_0 – коефіцієнт посилення постійної складової (визначає посилення чи послаблення сигналу постійної складової (нульової частоти)).

Наступним кроком є накладання цього фільтра на оригінальний сигнал для заглушення неінформативних частот.

Побудова спектрограми

Загалом є низка методів для класифікації звуків методами машинного навчання, а саме: метод опорних векторів, наївний баєсівський класифікатор, дерево рішень, глибинні нейронні мережі, гаусові суміші моделей, згорткові нейронні мережі.

Згороткові нейронні мережі (CNN) широко використовують для розпізнавання звуків. Саме цим методом у роботі виконано класифікацію наявності звуків птахів у аудіофайлі. Ідея методу ґрунтується на аналізі звуків у форматі зображень їх спектрограм. CNN використовують згорткові шари для виявлення просторових шаблонів у спектрограмах.

Спектрограма сигналу – це візуальне зображення спектра частот сигналу в часі. Спектрограма являє собою графік амплітуди звуку певної частоти у дискретний проміжок часу. На рис. 3 наведено приклад перетворення аудіофайлу на спектрограму.

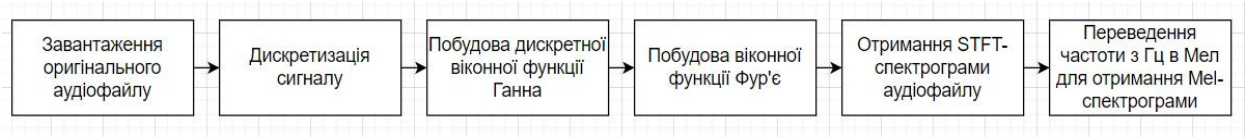


Рис. 3. Етапи перетворення аудіофайлу на спектрограму

Як видно з рис. 3, після завантаження оригінального аудіофайлу та його дискретизації необхідно застосувати віконну функцію Ганна для перетворення сигналу на спектрограму. Переваги

використання цієї функції – згладженість, контроль витоку спектра та широкий головний пелюсток. Це забезпечує хорошу просторову роздільну здатність, що особливо корисно для виявлення вузькосмугових компонентів.

Віконна функція Ганна має вигляд

$$w(n) = 0{,}5\left(1 - \cos\left(\frac{2\pi n}{N-1}\right)\right) = \sin^2\left(\frac{\pi n}{N}\right), \text{де } 0 \le n \le N, \quad (5)$$

де N – ширина вікна.

Наступний крок – віконне перетворення Фур'є (трансформація Фур'є), що застосовується для визначення синусоїдної частоти та вмісту фази локальної секції сигналу, що має властивість змінюватись у часі.

Віконне перетворення Фур'є можна обчислити за формулою:

$$\boldsymbol{F[m,\omega]} = \sum_{\boldsymbol{n=-\infty}}^{\boldsymbol{+\infty}} \boldsymbol{f[n] \times w[n-m] e^{-i\omega n}}, \quad (6)$$

де $F[m,\omega]$ – значення спектра для частоти w та точки m; $f[n]$ – значення дискретного сигналу в точці n; $w[n-m]$ – значення віконної функції у точці $n-m$; ω – частота сигналу.

Для отримання STFT-спектрограми аудіофайлу необхідно виконати таку операцію:

$$spectrogram\{x(t)\}[m,\omega] = |F[m,\omega]|^2, \quad (7)$$

де $F[m, \omega]$ – віконне перетворення Фур'є.

Для кращого аналізу спектрограм використовують Mel-спектрограму. По суті, це та сама STFT-спектрограма, але виражена в Mel-одиницях – психофізичній одиниці висоти звуку, яку застосовують у музичній акустиці. Формула для переходу від частоти сигналу до мел така

$$m \approx 1127 \times \ln\left(1 + \frac{f}{700}\right). \quad (8)$$

Експериментальні дослідження

Для виконання цієї роботи використано 1935 зразків датасету із звуком птахів та 2435 семплів без звуку птахів. Довжина кожного аудіофайлу – 10 секунд.

1. Попереднє оброблення аудіофайлу.

Оскільки зазвичай звуку птахів притаманні високі частоти, з аудіофайлу можна виокремити лише їх для подальшої класифікації. Це здійснюється за допомогою створення фільтра Баттерворта для високих частот із бібліотеки ***scipy***, після чого відбувається накладання цього фільтра на звуковий сигнал за допомогою методу ***lfilter*** цієї ж бібліотеки, тому сигнал зберігає лише високі частоти для подальшої роботи з аудіфайлом.

2. Конвертація аудіофайлів у спектрограму.

Під час цього кроку кожен аудіофайл конвертується у спектрограму. Перетворення відбувається методом STFT, після чого накладається Mel-шкала. Ці перетворення здійснюються за допомогою методу melspectrogram бібліотеки librosa. Після цього кожне зображення зберігається для подальшого навчання CNN на них.

3. Попереднє оброблення зображень.

Усі зображення конвертуються до розміру 256 на 256 пікелів та використовують три канали кольору (RGB).

Зображення довільно поділяють на тренувальну та тестувальну вибірки у співвідношенні 7:3 для подальшого навчання нейронної мережі.

4. Побудова моделі CNN

За допомогою бібліотеки tensorflow.keras будують згорткову нейронну мережу, яка загалом містить в собі такі типи шарів:

· Conv2D – шар, що створює згорткове ядро для конволюції вхідних даних задля утворення вихідного тензора.

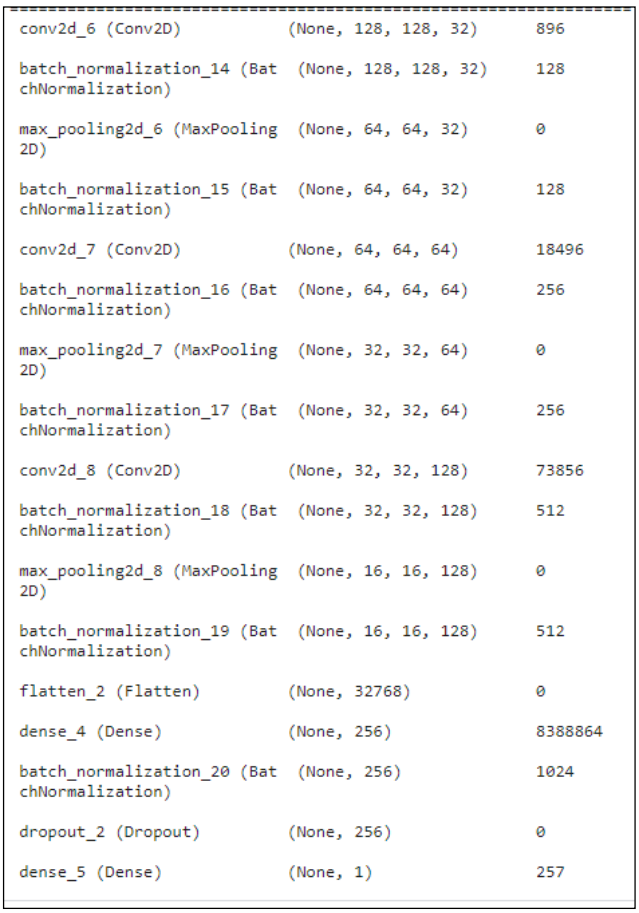
· BatchNormalization – шар, що нормалізує дані на вході.

· MaxPooling2D – шар дискретизації на основі вибірки, мета якого – зменшити вибірку вхідного зображення (чи вихідних даних попереднього шару конволюції).

· Flatten – шар для перетворення багатовимірної структури даних на одновимірну.

· Dense – повнозв'язний нейронний шар.

· Dropout – шар регуляризації, задля уникнення перенавчання із вимиканням частини нейронів (рис. 4).



conv2d_6 (Conv2D)	(None, 128, 128, 32)	896
batch_normalization_14 (BatchNormalization)	(None, 128, 128, 32)	128
max_pooling2d_6 (MaxPooling2D)	(None, 64, 64, 32)	0
batch_normalization_15 (BatchNormalization)	(None, 64, 64, 32)	128
conv2d_7 (Conv2D)	(None, 64, 64, 64)	18496
batch_normalization_16 (BatchNormalization)	(None, 64, 64, 64)	256
max_pooling2d_7 (MaxPooling2D)	(None, 32, 32, 64)	0
batch_normalization_17 (BatchNormalization)	(None, 32, 32, 64)	256
conv2d_8 (Conv2D)	(None, 32, 32, 128)	73856
batch_normalization_18 (BatchNormalization)	(None, 32, 32, 128)	512
max_pooling2d_8 (MaxPooling2D)	(None, 16, 16, 128)	0
batch_normalization_19 (BatchNormalization)	(None, 16, 16, 128)	512
flatten_2 (Flatten)	(None, 32768)	0
dense_4 (Dense)	(None, 256)	8388864
batch_normalization_20 (BatchNormalization)	(None, 256)	1024
dropout_2 (Dropout)	(None, 256)	0
dense_5 (Dense)	(None, 1)	257

Рис. 4. Архітектура CNN

Ця мережа містить вхідний шар, що приймає дані, розміром 256×256×3, три конволюційні шари нейронів, з функцією активації ReLU, три шари об'єднання розміром 2×2, сім шарів нормалізації батчів, один повнозв'язний нейронний шар розміром 256 нейронів із функцією активації ReLU та вихідний шар розміром 1 нейрон із функцією активації сигмоїда. Хоч по суті класів два (наявність та відсутність звуку), достатньо використовувати один вихідний шар із сигмоїдою для визначення ймовірності наявності звуку птахів на спектрограмі.

Для тренування використано loss-функція binary_crossentropy, оптимізаторами є Adam та RMSprop (для порівняння), основною метрикою – точність (accuracy). CNN тренується 30 епох, чого достатньо для забезпечення необхідної точності.

Метод для класифікації аудіо за допомогою натренованої CNN передбачає кілька етапів (рис. 5):

1. Завантаження та попередня обробка аудіофайлу.
2. Перетворення зображення на Mel-спектрограму.
3. Попереднє опрацювання зображень.
4. Передбачення натренованою CNN спектрограми.

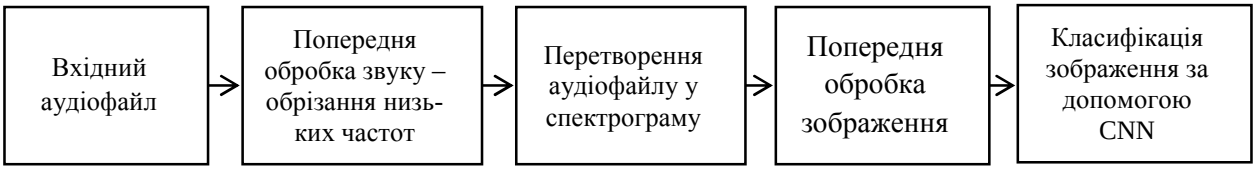


Рис. 5. Алгоритм класифікації нового аудіофайлу на наявність звуку птахів

Програмну реалізацію здійснено за допомогою середовища програмування Kaggle Notebook, що містить в собі Jupiter Notebook. Це дало змогу ефективно поєднувати код, що виконується, та паралельно зберігати виведення інших блоків коду.

Завантаження аудіофайлу відбувається за допомогою методу load бібліотеки librosa. Цей метод приймає шлях до необхідного файлу та частоту дискретизації звукового сигналу.

Для дослідження вибрано файл з id 19037. У цьому файлі є звуки птахів.

amplitude_time_series, sample_rate = librosa.load(id(19037), sample_rate = 22050)

На рис. 6 подано графік часового ряду амплітуди аудіофайлу, який демонструє, коли звук співу птахів є голоснішим, а коли тихішим.

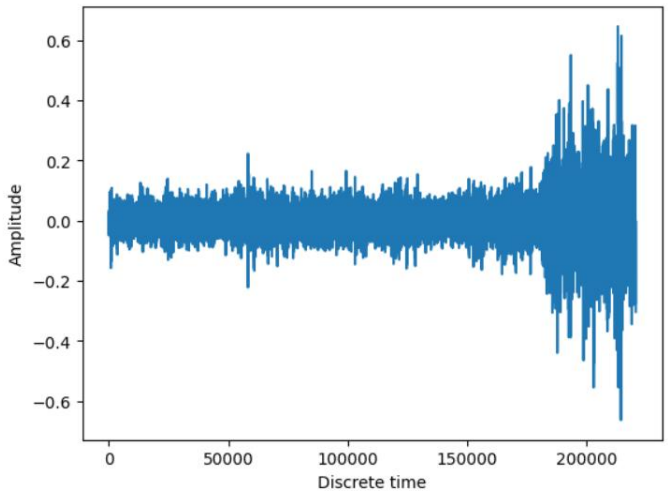


Рис. 6. Графік часового ряду амплітуди аудіофайлу

Відомо, що кількість значень масиву амплітудно-частотного ряду є добутком тривалості аудіофайлу на встановлену частоту дискретизації.

Оскільки частота дискретизації – 22050, а тривалість кожного аудіофайлу – 10 секунд, отримано масив дискретного сигналу завдовжки 220500.

На рис. 7 подано приклади значень цього масиву, його розмір, а також обчислено мінімальне (за допомогою функції np.min) та максимальне (за допомогою функції np.max) значення амплітуди цього масиву.

```
Amlpitude Time Series: [ 0.01047631  0.01933496  0.02669679 ... -0.25903526 -0.26455885
 -0.30332947]
Size of series: 220500
Max amplitude: 0.6452135443687439, Min amplitude: -0.6631531715393066
```

Рис. 7. Відображення значень ряду амплітуди та розміру цього ряду

Оскільки птахам здебільшого притаманний звук частотою від 1 до 10–15 кГц, доцільно обрізати інші частоти, накладаючи на аудіофайл фільтр Баттерворта, та не брати до уваги нижчі частоти,

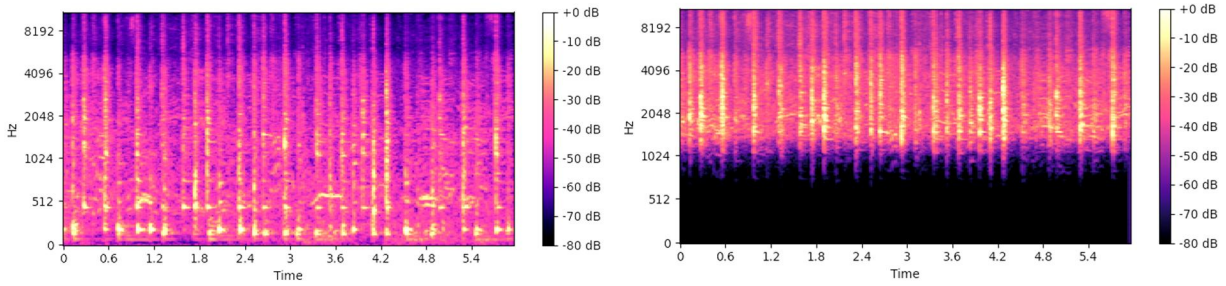
які, ймовірно, не відповідають співу птахів, завдяки чому шум значно менше впливатиме на точність класифікації файлу.

Для обрізання частот скористаємось двома методами бібліотеки scipy:

– butter – для створення фільтра Баттерворта;

– lfilter – для накладання цього фільтра на вхідний аудіофайл.

Після застосування цього фільтра можна переконатись у заглушенні необхідних частот, побудувавши спектрограму, на якій чорним кольором відображено заглушені частоти (рис. 8).



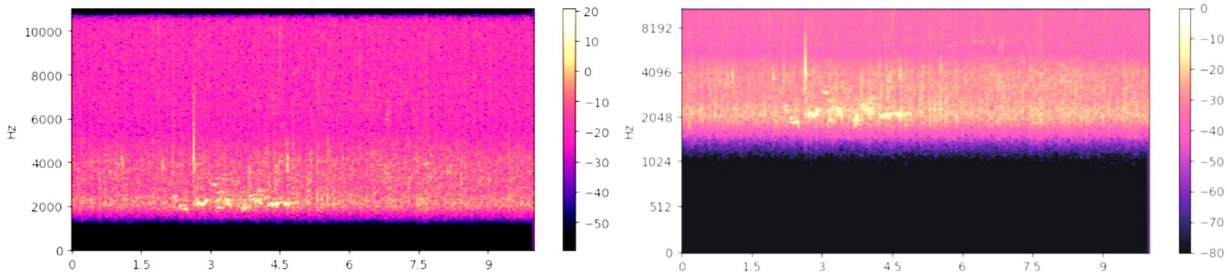
а *б*

Рис. 8. Спектрограма, утворена із сигналу з повною частотою (а) та з обрізаними низькими частотами (б)

Оскільки CNN – нейронна мережа для класифікації зображень, необхідно перетворити звуковий сигнал на спектрограму – візуальне зображення амплітуди звуку певного спектра частот сигналу на дискретний проміжок часу.

Спектрограму побудуємо на основі Mel-спектрограми. Це перетворення сигналу за допомогою STFT із подальшою конвертацією одержаних значень за допомогою Mel-шкали.

Для отримання STFT-спектрограми із вхідного часового ряду амплітуд застосовують функцію stft бібліотеки librosa. Щоб одержати Mel-спектрограму із вхідного часового ряду амплітуд, застосовують функцію melspectrogram бібліотеки librosa.features. На рис. 9 наведено STFT (рис. 9, *а*) та Mel (рис. 9, *б*) спектрограми відповідно.



а *б*

Рис. 9. STFT-спектрограма (а) та Mel-спектрограма (б)

Loss-функцією вибрано Binary_crossentropy, яка підходить для класифікації об’єкта серед двох класів.

Порівняно два оптимізатори для CNN, а саме: адаптивну оцінку моменту (Adam) та середньоквадратичне поширення (RMSProp). Оскільки точність майже не змінювалась у разі зміни оптимізатора (в межах 2 %), можна зробити висновок, що для цих задачі та датасету вибір оптимізатора не впливає на результат, тому надалі буде використовувався лише Adam.

Усі зображення датасету поділено на батчі по 32 зображення в кожному.

Розподіл датасету на батчі дає змогу обробляти декілька зображень водночас, використовуючи векторизацію та оптимізовані обчислення на GPU. Його виконують, розділяючи загальний набір даних на рівні групи.

Оскільки тренувальний датасет містить 3059 зображень (70 % від всього датасету), після поділу його на батчі по 32 зображення утвориться 96 груп зображень, з яких усі будуть однаковими, окрім останнього, який міститиме залишкову кількість зображень від цілочислового поділу.

Аналогічно валідаційний датасет ділять на батчі по 32 зображення в кожному. Валідаційний датасет складається із 1311 зображень (30 % датасету), відповідно унаслідок поділу утвориться 41 група, в якій аналогічно остання матиме неповну кількість зображень, як залишок від цілочислового поділу.

Далі необхідно виокремити за допомогою CNN ті ділянки спектрограми, які відповідають звукам птахів у аудіофайлі. Основними ознаками є частота та амплітуда сигналу. Для пошуку цих ознак використовують конволюційні матриці. Ці ядра конволюції для пошуку ознак впродовж тренування CNN формуються так:

1. Ініціалізація – початкові ядра конволюції ініціалізуються випадковими значеннями або ж попередньо навченими, якщо модель попередньо навчена. У цьому випадку вони ініціалізуються випадковими значеннями, адже модель CNN не є попередньо навченою.

2. Прямий прохід (*forward pass*): вхідне зображення проходить через згортковий шар нейронної мережі. Кожне ядро конволюції застосовують до вхідних даних за допомогою конволюції з відповідною частиною вхідних даних. Це створює карту ознак або активаційну карту для кожного ядра.

3. Зворотне проходження (*backward pass*): оскільки навчання CNN полягає у виявленні оптимальних матриць конволюції, під час зворотного проходження обчислюють втрати за допомогою loss-функції (різницю між прогнозованими і правильними значеннями), після чого оновлюють параметри нейронної мережі, зокрема ядра конволюції. Для оновлення значень ядер використовують метод градієнтного спуску.

4. Повторення проходу: процес прямого та зворотного проходу повторюється протягом певної кількості епох навчання (CNN у цій роботі тренується протягом 30 епох) для поліпшення ядер конволюції, і відповідно здатності мережі виконувати коректну класифікацію.

На рис. 10 наведено приклад матриці конволюції після тренування.

```
[-0.03755007, -0.1368099 ,  0.03560786, -0.06365225,
 -0.12015828, -0.05399318, -0.03378767, -0.06931988,
 -0.03071228, -0.13304812,  0.04215126, -0.09484857,
 -0.13503404,  0.08916029,  0.14188696,  0.11267205,
 -0.02037111,  0.01001463,  0.00123424, -0.0447947 ,
  0.01973012, -0.01274301,  0.03657272, -0.05352908,
  0.00310243,  0.01999148, -0.01047826,  0.07803536,
 -0.04200397, -0.15908046,  0.04779833, -0.0151548 ],
[ 0.00912207,  0.04088343, -0.12898225,  0.06137509,
 -0.1925271 ,  0.08243312, -0.04367905,  0.08229449,
 -0.03308559, -0.03219797,  0.10791407,  0.01673812,
 -0.14803284,  0.1679846 , -0.13851133,  0.13150571,
 -0.10954688, -0.12912565, -0.07910062, -0.06995527,
 -0.06783991,  0.09085149,  0.03590787, -0.00769596,
  0.01466693, -0.01600416, -0.11011907, -0.05252029,
 -0.00747553, -0.1376736 , -0.01561829, -0.0721685 ],
[-0.04660718, -0.08319906, -0.09755186, -0.07049508,
  0.08904252,  0.06073883, -0.03166704, -0.06159271,
 -0.0772192 , -0.02078658,  0.03024908,  0.08552631,
  0.02959098,  0.0766892 , -0.02942621,  0.05348112,
 -0.09633254,  0.06215728, -0.1119996 ,  0.10013463,
 -0.03700884,  0.04635266, -0.03154969, -0.14336938,
  0.03010832, -0.04268707,  0.11363497, -0.12258115,
  0.02468136, -0.13216355,  0.07192035,  0.13748866]],
```

Рис. 10. Матриця конволюції першого шару тренованої CNN

Під час тренування CNN випробовували обидва типи спектрограм: отриману за допомогою методу STFT, та спектрограму, переведену опісля у шкалу Mel (див. рис. 9), і досліджували точність моделі.

Навчання на основі STFT-спектрограм

Під час навчання нейронної мережі обчислюють точність моделі кожної епохи як відношення правильно класифікованих семплів до усіх семплів. На рис. 11 наведено історію точності моделі упродовж проходження кожної епохи.

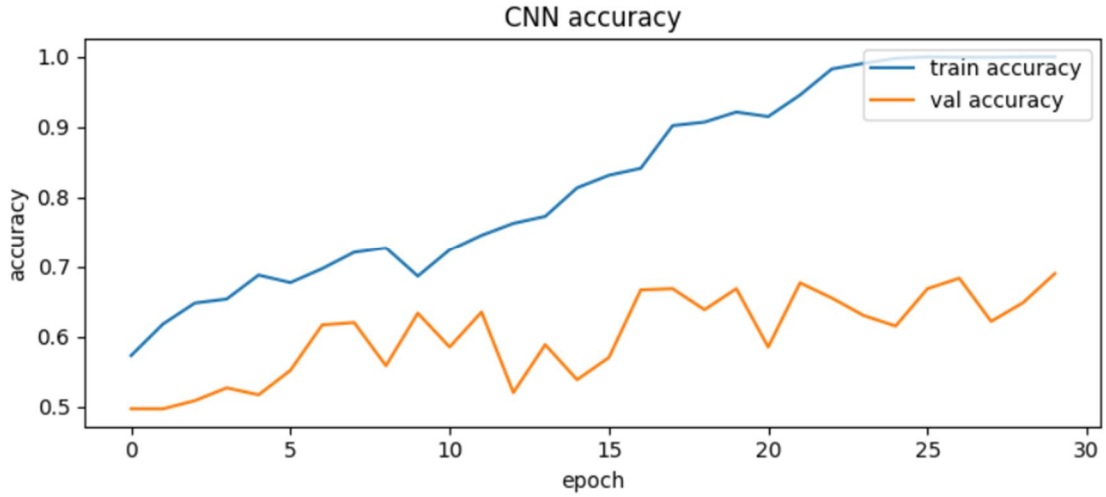


Рис. 11. Історія точності під час тренування CNN на STFT-спектрограмах

Під час навчання нейронної мережі як критерій коригування ваг моделі використовують функція втрат, яка відображає значення помилки між класифікованими значеннями та справжніми. На рис. 12 відображено історію значень функції втрат для кожної епохи. Як бачимо, втрати постійно зменшуються, а отже, модель навчається.

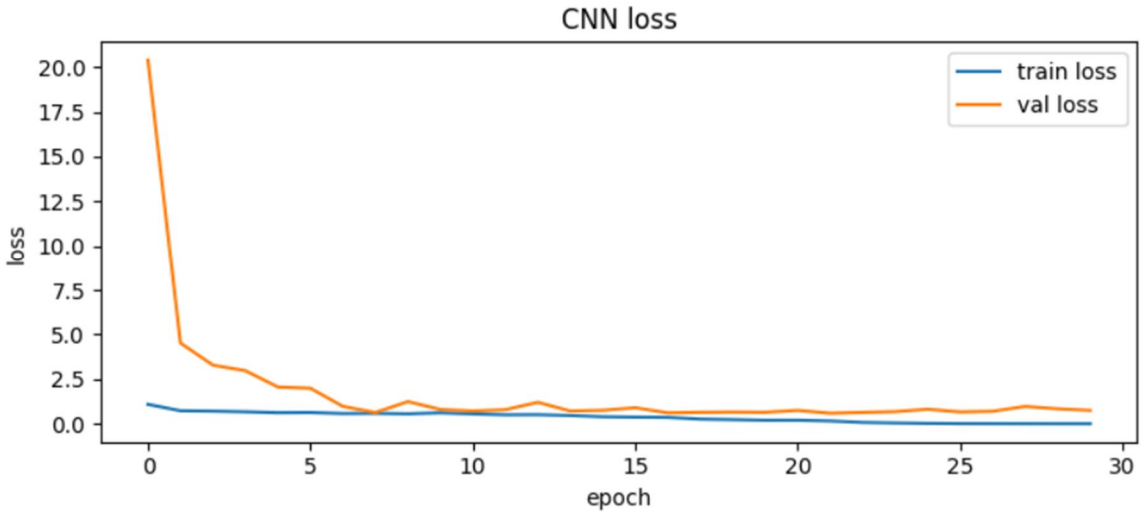


Рис. 12. Історія втрат CNN під час тренування на STFT-спектрограмах

Остаточну оцінку точності та втрат обчислюють методом evaluate бібліотеки keras, як відношення правильно класифікованих семплів до усіх семплів. Також цим методом відображають значення функції втрат (*loss function*) – оцінки помилки між прогнозованими значеннями та фактичними.

Нижче наведено дані, які демонструють, що точність нейронної мережі, тренованої на STFT-спектрограмах, становить 63 %, а функція втрат – 0,88.

```
{'loss': 0.8863168954849243, 'accuracy': 0.6324999928474426}
```

Аналогічно відобразимо історію точності кожної епохи тренування нейронної мережі на Mel-спектрограмах (рис. 13):

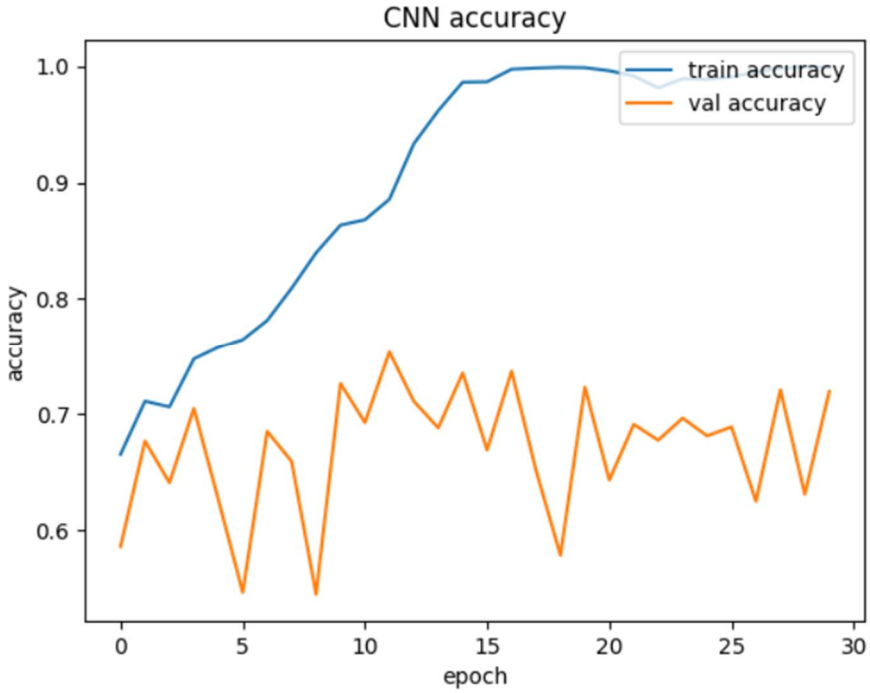


Рис. 13. Історія тренування CNN із застосуванням Mel-спектрограм

Історію втрат кожної епохи тренування нейронної мережі на Mel-спектрограмах наведено на рис. 14.

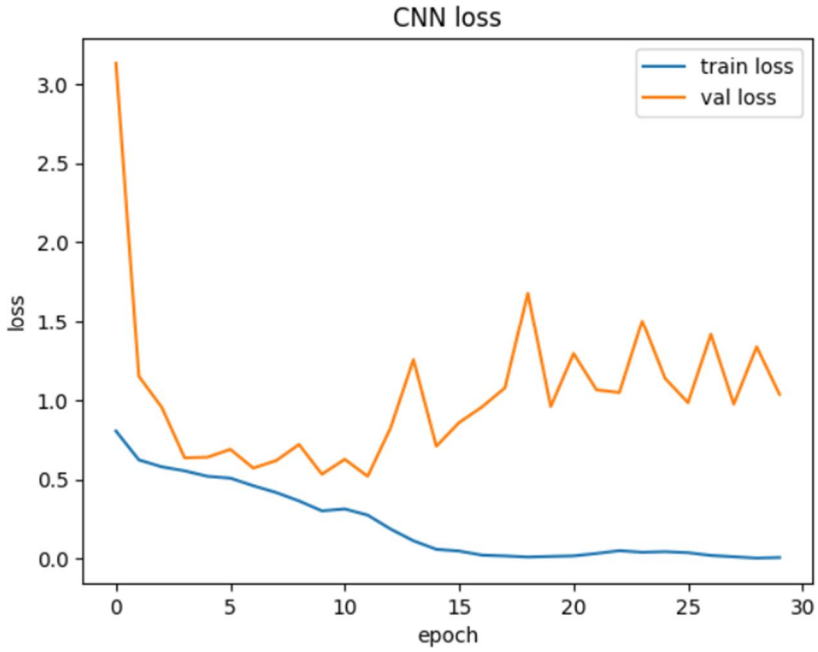


Рис. 14. Історія втрат CNN під час тренування на Mel-спектрограмах

Наведемо дані, які демонструють, що точність нейронної мережі, тренованої на Mel-спектрограмах, становить 72 %, а функція втрати – 1,03

```
{'loss': 1.0359022617340088, 'accuracy': 0.719298243522644}
```

Після цього виконано класифікацію аудіофайлів на основі CNN, натренованої на датасеті. Оскільки передбаченням нейронної мережі є ймовірність того, що спектрограма, а відповідно, і аудіофайл, містить звук птахів, то кінцевим значенням класифікації файлу є округлення цієї ймовірності до цілого числа методом np.round().

Якщо для конкретного зразка CNN прогнозує наявність звуку птахів з ймовірністю понад 0,5 (50 %), то цей зразок класифікують як такий, що містить звук птахів. Якщо ймовірність є меншою за 0,5, зразок класифікується як такий, у якому немає звуків птахів (0 або 1 для відсутності або наявності звуку птахів у файлі, відповідно).

Під час класифікації десяти випадково відібраних аудіофайлів із датасету правильно ідентифіковано усі десять, серед яких три зі звуком птахів та сім без звуку. Середній час класифікації одного зразка становив 0,428 секунди.

Висновки

Проаналізувавши отримані теоретичні та практичні результати, можна зробити такі висновки:

1. Побудовано та випробувано модель класифікації аудіофайлів на основі зображень їх спектрограм за допомогою згорткової нейронної мережі. Точність класифікації цього методу становить 72 %, що менше, ніж очікували під час побудови класифікаційної моделі. Проте швидкість для передбачення аудіофайлу цією моделлю становить менше ніж 0,5 секунди, що є доволі непоганим результатом.
2. Використання Mel-спектрограм дає кращі результати, ніж STFT-спектрограм, для дослідження класифікації наявності звуків птахів у середовищі. Точність класифікації моделі, тренованої на основі Mel-спектрограм, становила 72 %, що на 8 % краще, ніж точність моделі, натренованої на STFT-спектрограмах.
3. Щоб підвищити точність цього методу, необхідно отримувати зразки звуків ділянки, на якій цей метод буде застосовуватись, або ж створити генералізований датасет з набагато якіснішими аудіофайлами.
4. Створення моделі класифікації аудіофайлів за допомогою конволюційних нейронних мереж, що потребують вхідних дані у вигляді зображень, є не найточнішим методом для аналізу аудіофайлів, оскільки виникають похибки під час перетворення часових рядів амплітуди на спектрограму сигналу. Проте це один із хороших методів, для того, щоб почати аналіз датасету аудіофайлів, адже датасет відрізняється від використаного у цьому дослідженні, який матиме кращі зразки звукових даних та кращі показники точності моделей.

Отримані результати можуть стати корисними для розроблення систем контролю міграцій птахів на великих ділянках території. Окрім того, методику цього дослідження можна вдосконалити, використовуючи декілька аудіофайлів для класифікації наявності птахів у середовищі. Отже, ймовірно, що точність методу і моделі зросте, проте може дещо знизитися швидкість у разі комплексного оброблення багатьох аудіофайлів водночас. Це дослідження буде корисним із поєднанням методів відстеження птахів у середовищі (наприклад, ансамбль прямого спостереження та слухового методу).

Список літератури

1. Ghosh A., Sufian A., Sultana F., Chakrabarti A. & Debashis De. (2020). Fundamental Concepts of Convolutional Neural Network. *Recent Trends and Advances in Artificial Intelligence and Internet of Things,* 519–567. DOI:10.1007/978-3-030-32644-9_36.

2. Krizhevsky A., Sutskever I., & Hinton G. E. (2012). Imagenet classification with deep convolutional neural networks. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, 1097–1105.

3. Sultana F., Sufian A., & Dutta P. (2019). A review of object detection models based on convolutional neural network. *CoRR,* abs/1905.01614. DOI:10.1007/978-981-15-4288-6_1.

4. Sultana F., Sufian A., & Dutta P. (2018). Advancements in image classification using convolutional neural network. In *2018 Fourth International Conference on Research in Computational Intelligence and Communication Networks (ICRCICN),* 122–129.

5. Everingham M., Van Gool L., Williams C. K. I., Winn J. & Zisserman A. (2010). The pascal visual object classes (VOC) challenge. *International Journal of Computer Vision, 88*(2), 303–338. DOI:10.1007/s11263-009-0275-4.

6. Szegedy C., Liu W., Jia Y., Sermanet P., Reed S., Anguelov D., Erhan D., Vanhoucke V., & Rabinovich A. (2015). Going deeper with convolutions. *In The IEEE Conference on ComputerVision and Pattern Recognition (CVPR).* DOI: 10.48550/arXiv.1409.4842.

7. Shelhamer E., Long J., & Darrell T. (2015). Fully convolutional networks for semantic segmentation. *IEEE Trans. Pattern Anal. Mach. Intell*., 39(4), 640–651. DOI: 10.1109/CVPR.2015.7298965.

8. Dennis J. W. (2014). Sound event recognition in unstructured environments using spectrogram image processing. *Doctoral thesis, Nanyang Technological University, Singapore*. DOI: 10.32657/10356/59272

9. Mesaros A., Heittola T., Eronen A., & Virtanen T. (2010). Acoustic event detection in real life recordings. *Proceedings of the European Signal Processing Conference (EUSIPCO),* 1267–1271.

10. Tsau E., Chachada S., & Kuo C.-C. J. (2012). Content/Context-Adaptive Feature Selection for Environmental Sound Recognition. *Proceedings of the Asia Pacific Signal & Information Processing Association (APSIPA).*

11. Zhang Z. and Schuller B. Semi-supervised learning helps in sound event classification. *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), IEEE*, 333–336. March, 2012.

12. Maccagno A., Mastropietro A., Mazziotta U., Scarpiniti M., Lee Y.-Ch. & Uncini A. (2021). A CNN Approach for Audio Classification in Construction Sites. *Progresses in Artificial Intelligence and Neural Systems*, 371–381. DOI: 10.1007/978-981-15-5093-5_33.

13. Ekpezu A., Wiafe I., Katsriku F. & Yaokumah W. (2021). Using deep learning for acoustic event classification: The case of natural disasters. *The Journal of the Acoustical Society of America, 149(4): 292.* DOI: 10.1121/10.0004771.

14. Khamparia A., Gupta D., Nguyen N. G., Khanna A., Pandey B., & Tiwari P. (2019). Sound classification using convolutional neural network and tensor deep stacking network, *IEEE Access, 7(1),* 7717–7727. DOI: 10.1109/ACCESS.2018.2888882.

15. Zhang, T., Lee, Y.-C., Scarpiniti, M., Uncini, A. (2018). A supervised machine learning-based sound identification for construction activity monitoring and performance evaluation. *Proceedings of 2018 Construction Research Congress (CRC 2018), New Orleans, Louisiana, USA*, 358–366.

16. Kons Z., Toledo-Ronen O. (2013). Audio Event Classification Using Deep Neural Networks. *Proc. Interspeech 2013*, 1482–1486. DOI: 10.21437/Interspeech.2013-384.

17. Lee H., Grosse R., Ranganath R., & Ng A.Y. (2011). Unsupervised Learning of Hierarchical Representations with Convolutional Deep Belief Networks. *Communications of the ACM, Vol. 54, No. 10,* 95–103. DOI: 10.1145/2001269.2001295.

18. Gartzman D. Getting to Know the Mel Spectrogram. *Towards Data Science*. August, 2019. Retrieved from: https://towardsdatascience.com/getting-to-know-the-mel-spectrogram-31bca3e2d9d0 (date of access: 20.09.2023)

19. Papia Nandi – CNNs for audio classification. A primer in deep learning for audio classification using TensorFlow. *Towards Data Science.* Murch, 2021. Retrieved from: https://towardsdatascience.com/cnns-for- audio-classification-6244954665ab (date of access: 16.09.2023)

20. Chollet, F. Deep Learning with Python (2018), v. 361, New York: Manning.

21. SHANTAMVIJAYPUTRA - Bird Voice Detection Dataset. Retrieved from: https://www.kaggle.com/datasets/vshantam/bird-voice-detection (date of access: 15.05.2023)

IDENTIFICATION OF BIRDS' VOICES USING CONVOLUTIONAL NEURAL NETWORKS BASED ON STFT AND MEL SPECTROGRAM

Oksana Honsor[1], **Yuriy Gonsor**[2]

[1] Lviv Polytechnic National University, Specialized Computer Systems Department,
12, S. Bandery str., Lviv, Ukraine
[2] Lviv Polytechnic National University, Artificial Intelligence Systems Department,
12, S. Bandery str., Lviv, Ukraine
E-mail: oksana.y.honsor@lpnu.ua, ORCID: 0000-0003-0895-5859
E-mail: yurii.honsor.knm.2020@lpnu.ua ORCID: 0009-0001-7486-6689

Threats to the climate and global changes in ecological processes remain an urgent problem throughout the world. Therefore, it is important to constantly monitor these changes, in particular, using non-standard approaches. This task can be implemented on the basis of research on bird migration information. One of the effective methods of studying bird migration is the auditory method, which needs improvement. That is why building a model based on machine learning methods that will help to accurately identify the presence of bird voices in an audio file for the purpose of studying bird migrations from a given area is an urgent problem. This paper examines ways of building a machine learning model based on the analysis of spectrograms, which will help to accurately identify the presence of bird voices in an audio file for the purpose of studying the migration of birds in a certain area. The research involves the collection and analysis of audio files that can be used to identify characteristics that will identify the sound of the files as birdsong or the absence of sound in the file. The use of the CNN model for the classification of the presence of bird voices in an audio file is demonstrated. Special attention is paid to the effectiveness and accuracy of the CNN model in the classification of sounds in audio files, which allows you to compare and choose the best classifier for a given type of file and model. Analysis of the effectiveness and accuracy of the CNN model in the classification of sounds in audio files showed that the use of Mel-spectrograms is better than the use of STFT-spectrograms for studying the classification of the presence of bird sounds in the environment. The classification accuracy of the model trained on the basis of Mel spectrograms was 72 %, which is 8 % better than the accuracy of the model trained on STFT spectrograms.

Key words: machine learning; sound identification; spectrogram; convolutional neural network.