Visnyk NTUU KPI Seriia – Radiotekhnika Radioaparatobuduvannia, 2023, Iss. 93, pp. 39–51

УДК 004.93

Алгоритм оброблення аудіосигналів із використанням методу машинного навчання

Сокольський С. О., Мовчанюк А. В.

Національний технічний університет України "Київський політехнічний інститут імені Ігоря Сікорського", м. Київ, Україна

E-mail: *sokolskyi@ros.kpi.ua*, *movchanuk@rtf.kpi.ua*

Малі безпілотні літальні апарати або дрони швидко розвиваються та впроваджуються. Але це також збільшує і загрозу громадській та національній безпеці через ризики їхнього неправомірного використання. Перспективним напрямком для вирішення цієї проблеми є акустичний метод, що включає в себе аналіз звукових характеристик і доплерівського зсуву сигнатур безпілотників, використання масивів мікрофонів та методи машинного навчання. Задачею цієї статті є розроблення алгоритму для ефективного виявлення та класифікації аудіосигналів малих безпілотних літальних апаратів із використанням згорткової нейронної мережі глибокого навчання, побудова архітектури та оцінка ефективності її роботи. Перед подачею набору аудіозаписів дронів на вхід мережі, підвищили їхню якість застосувавши нормалізацію, Вінеровську фільтрацію, сегментацію — поділили аудіо на кадри тривалістю 25 мс з перекриттям 50% та віконне керування за допомогою вікна Хеммінга, оскільки у завданні оброблення аудіосигналів важливіша точність у часовій області. Отримані дані розділили у співвідношенні 60/20/20 на три набори: для навчання, валідації та тестування. Далі представили дані спрощеним набором ознак, визначивши з кожного кадру оброблених аудіосигналів мел-спектрограми, для фіксації часових та спектральних характеристик. Діапазон частот аналізу становить межі робочих частот моделі мікрофону (20 Гц — 20 кГц), частотна розділова здатність 50 Гц, а кількість робочих мел-смуг дорівнює 30. Використовуючи навчальні дані та отримані ознаки аудіосигналів, розробили архітектуру нейронної мережі для досліджень роботи алгоритму виявлення дронів. Вона складається із 10 пар шарів згортки, ReLU, пакетної нормалізації та максимального пулингу. Їхня кількість визначається розміром вікна об'єднання в часі. Наступними є шари згладжування, відсікання, повнозв'язний та Softmax. Для нормалізації вихідних даних і отримання фінальних ймовірностей застосовується шар класифікації. У якості оптимізатора для навчання моделі обрано Adam, початкова швидкість навчання дорівнює 0.001, а після проходження 75% епох поступово зменшується у 10 разів, для покращення збіжності. Точність розпізнавання даних складає 99%, оцінка F1 — 0.93, що вказує на високий рівень загальної продуктивності архітектури. Максимальна відстань ефективного виявлення дронів алгоритмом складає 200 м.

Ключові слова: дрон; малий безпілотний літальний апарат; спектр; обробка сигналу; виявлення сигналу; згорткові нейронні мережі; глибоке навчання

DOI: 10.20535/RADAP.2023.93.39-51

Вступ

Малі безпілотні літальні апарати (МЛА) або, як зараз їх часто називають, *дрони* — повітряні судна без екіпажу на борті, що керуються дистанційно та приводяться у рух за допомогою чотирьох або більше гвинтів. Швидкий розвиток та їхнє впровадження, відкрили нові можливості у різних галузях промисловості та особистому використанні. Ці компактні, маневрені та універсальні літальні апарати створені для полегшення життя людей. Вони пропонують численні переваги, наприклад, аерофотозйомку, спостереження за об'єктами та доставку вантажів у важкодоступні місця.

Однак, все частіше застосування МЛА сприяє незаконній діяльності: несанкціонованому спостереженню за важливими об'єктами інфраструктури, доставці контрабанди або бойових снарядів, що створюють величезну загрозу громадській та національній безпеці. Тому важливою ***науково-технічною проблемою*** стає ефективне виявлення МЛА, для зменшення ризиків, пов'язаних з їх неправомірним використанням.

Акустичні сенсори є перспективним напрямком для вирішення цієї проблеми, оскільки вони можуть фіксувати чіткі звукові сигнатури, які створюються пропелерами та двигунами безпілотників [1]. Аналізуючи ці сигнатури, можна відрізнити звуки дронів від фонового шуму. Такий підхід включає методи обробки сигналів: спектральний аналіз, вилучення ознак і розпізнавання образів [2]. Використовуючи масив мікрофонів, стратегічно розміщених у певній

місцевості, можна проаналізувати різницю в часі та інтенсивності звуків дронів, зафіксованих кожним мікрофоном [3]. За допомогою тріангуляції джерела звуку на основі цих відмінностей можна оцінити положення і напрямок руху МЛА. Аналізуючи доплерівський зсув в акустичному сигналі дрона, можна оцінити його швидкість і відстань. Доплерівські методи виявлення можуть бути ефективними для ідентифікації рухомих безпілотників і диференціації їх від стаціонарних або фонових звуків [4, 5].

Новим та малодослідженим підходом до виявлення МЛА є машинне навчання — нейронні мережі (НМ) глибокого навчання. Цей метод передбачає навчання алгоритмів на великому наборі аудіоданих, для вивчення і розпізнавання патернів, пов'язаних і не пов'язаних з МЛА, що дозволяє НМ класифікувати вхідні аудіосигнали в режимі реального часу, відрізняючи дрони від інших джерел шуму [6].

Застосування класичних нейронних мереж для реалізації цього завдання ускладняється великою розмірністю вхідних значень та значною кількістю нейронів у прихованих шарах, що призводить до серйозних витрат обчислювальної потужності на навчання НМ. Але згортковим нейронним мережам (ЗНМ) описані вище недоліки властиві в меншому ступені [7]. Зараз вони є потужним інструментом для вирішення різноманітних задач, включаючи класифікацію зображень [8] та аудіозаписів [9]. ЗНМ використовують алгоритми глибокого навчання для ефективного аналізу різних типів сигналу з будь-якого давача та вимагають мінімальний обсяг попередньої обробки вхідних даних через спеціальну архітектуру використання багатошарових перцептронів.

Ефективність методів може змінюватися залежно від умов навколишнього середовища, рівня фонового шуму та специфічних характеристик цілі виявлення. Відповідно ***задачею*** цієї статті є розроблення комплексного алгоритму для ефективного виявлення та класифікації аудіосигналів МЛА в режимі реального часу, використовуючи методи попереднього аналізу звукових характеристик та згорткові нейронні мережі, огляд процесу побудови архітектури, навчання і тестування моделі на спеціальних наборах даних. Оцінка ефективності і продуктивності її роботи. Також *ціллю* роботи є відповідь на питання: "Чи можлива повноцінна заміна людини-оператора, розробленою системою детекції дронів, без втрати ефективності?"

Загальна схема роботи алгоритму зображена на Рисунку 1. Проаналізуємо детально кожний етап для створення програмного забезпечення на його основі.

1 Створення набору даних

Чим більша кількість та довжина аудіофайлів, тим вища ефективність навчання ЗНМ та точність класифікації за прихованими об'єктами. У якості *цільових сигналів* використовуються аудіозаписи шумів польоту різних моделей МЛА, а також шуми навколишнього середовища: шум вітру, автомобілів, людей.

Вибір відповідного обладнання для створення бази аудіосигналів МЛА, фактори впливу зовнішніх джерел шуму навколишнього середовища, умови та план проведення експериментів детально розглянуто у [10]. У результаті аналогічного експерименту для дослідження було створено 1866 аудіозаписів зависання МЛА моделі «*DJI Mavic 3*» та 1408 фрагментів моделі «*DJI Mavic 2 Pro*», з довжиною кожного файлу рівній 1 секунді та площею перекриття 50%. Для збільшення вибірки сигнатур моделей МЛА також було згенеровано 1047 аудіофрагментів роботи моторів квадракоптера «*FPV*», 83 — «*FeiLun FX137*» і 56 — «*LH X43-W*». Але ці записи були створені у лабораторних умовах, тому їх можна використовувати лише як допоміжні сигнали для збільшення варіантів вибору при класифікації цілі.

Після запису аудіозразки необхідно структурувати, а їх колекції зберегти у базу аудіосигналів. Чітка структуризація та точне маркування даних має важливе значення для подальшого навчання та оцінки моделей машинного навчання. База аудіосигналів завжди має бути папкою даних верхнього рівня на комп'ютері та містить підпапки, кожна з яких названа на честь відповідної моделі МЛА або фонового шуму, яку вона представляє, і має складатись лише з колекції аудіофайлів, що відповідають цій назві.

Для збільшення розміру і різноманітності набору експериментальних даних, використаємо метод аугментації — застосування деформацій до колекції анотованих навчальних даних, які призводять до створення додаткових навчальних матеріалів. Ключова концепція аугментації у тому, що деформації, які застосовуються до позначених даних, наприклад, зсув висоти тону, розтягування в часі або додавання штучного шуму, не змінюють семантичного значення міток. Якщо навчати мережу додатковими спотвореними даними, то вона стане нечутливою до цих деформацій і зможе краще узагальнити дані.

Отже, деформуємо аудіофайли моделі «*DJI Mavic 3*» додавши до сигналу МЛА адитивний білий гаусівський шум із співвідношенням С/Ш від 40 дБ до 27 дБ за потужністю. У результаті кількість файлів вибірки зросла до 10000.

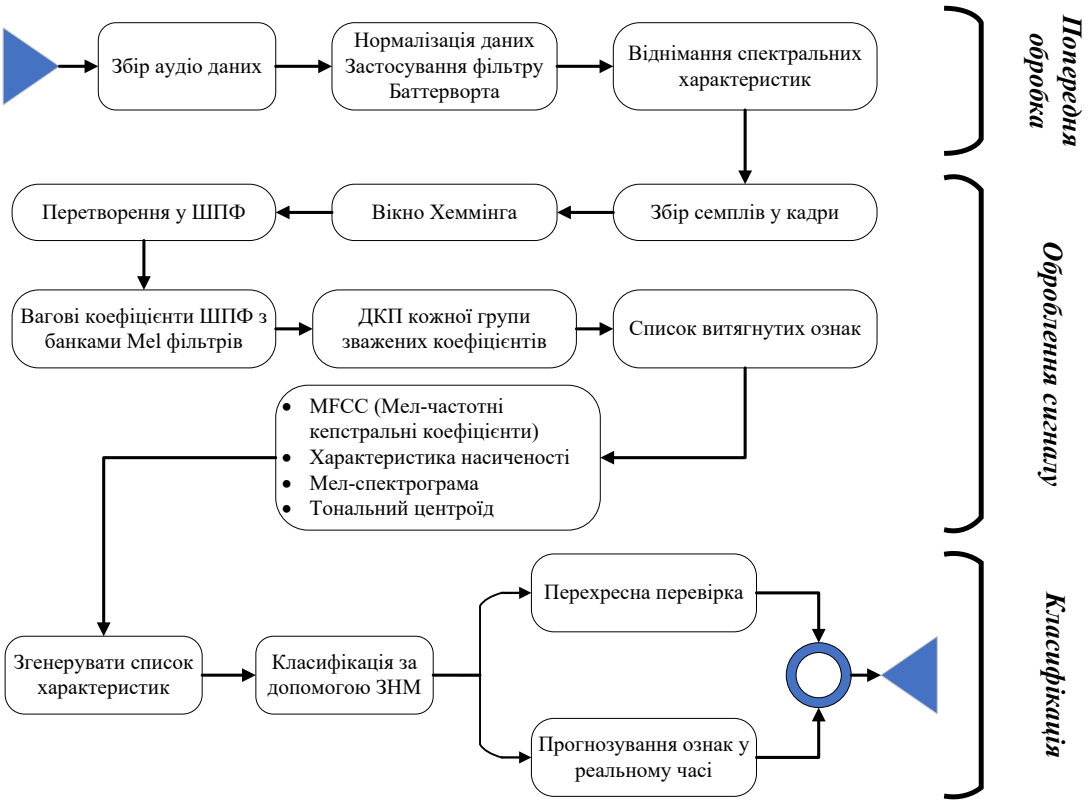

Рис. 1. Блок-схема виявлення та класифікації акустичного сигналу МЛА

2 Попередня обробка

Аудіосигнали, що надходять в режимі реального часу або завантажені з файлів, необхідно перетворити у вигляд, придатний для подачі на вхід нейронної мережі за допомогою методів попередньої обробки. Щоб уникнути зміщення, спричиненого варіаціями амплітуди вхідного сигналу, виконаємо його ***нормалізацію*** до діапазону від -1 до 1, це забезпечить постійний рівень сигналу. На короткому інтервалі часу 15-30 мс, аудіосигнали шуму МЛА є локально стаціонарними, відповідно їхній час та спектр вважаються практично однорідними на проміжку в декілька мілісекунд, тому ***зменшення рівня сторонніх шумів*** за таких умов найкраще забезпечують методи спектрального віднімання та Вінеровської фільтрації.

Метод ***спектрального віднімання*** оцінює очищений спектр аудіосигналу, шляхом віднімання спектра, що містить лише шум, із зашумленого оригінального сигналу. Перевагами цього методу є простота та обчислювальна ефективність для стаціонарного шуму або шуму з повільно змінним спектром. З мінусів виділяють можливість внесення спотворень в оброблений сигнал, а також нижчу ефективність для нестаціонарних або швидкозмінних шумів.

Вінеровська фільтрація оцінює співвідношення сигнал/шум (С/Ш) для кожної частотної складової і застосовує фільтр, який адаптується до оціненого значення С/Ш, мінімізуючи середньоквадратичну похибку між вихідним і відновленим сигналом. Він забезпечує краще зменшення рівня шуму порівняно зі спектральним відніманням, особливо для нестаціонарного шуму, а також зберігає більше характеристик вихідного сигналу. Але використання такого фільтру зазвичай вимагає знання статистики сигналу і шуму, яка не завжди може бути доступною.

Видалення періодів ***тиші*** з аудіосигналів корисне для зменшення обчислювальних ресурсів або зосередження аналізу лише на активних ділянках сигналу. Техніка ***попереднього виділення*** передбачає застосування ФВЧ, що підсилює високочастотні компоненти сигналу і зменшує енергію нижчих частот. Високочастотні компоненти, як правило, несуть більше інформації і мають більш високе співвідношення С/Ш порівняно з низькочастотними, що поліпшує їх виявлення в присутності фонового шуму. Частота зрізу фільтру має бути не менше 10 кГц. Але закладено можливість застосування більш широкого діапазону частот для виявлення МЛА меншого розміру. Надлишковість в смузі дозволить в подальшому розширити можли-

вості детекції інших типів МЛА, бо на сьогодення не відомий повний спектр апріорної інформації.

Збирання відліків (семплів) у кадри означає ***поділ*** аудіосигналу ***на менші сегменти*** або кадри фіксованої тривалості, зазвичай по 15-30 мс з деяким перекриттям (наприклад, 50%), що дозволяє якісно дослідити характеристики сигналу в часі. Аналізуючи сегменти на коротких часових інтервалах, можна зафіксувати зміни, динаміку та варіації сигналу, які можуть відбуватися в межах кожного кадру.

Під *тривалістю сегменту* розуміють довжину аудіосигналу, яка аналізується незалежно за один раз. Іншими словами, аудіосигнал розбивається на сегменти фіксованої довжини (зазвичай 1 с), і кожен сегмент аналізується окремо. Кількість відліків у кожному сегменті аудіосигналу вибирається на основі бажаної частотної роздільної здатності аналізу та довжини аудіосигналу. Вона фіксована і визначає тривалість аналізу швидкого перетворення Фур'є (ШПФ).

Тривалість кадру — довжина часового вікна, яке використовується для вилучення частотного вмісту кожного сегмента. У програмі вона дорівнює 25 мс, тож кожен сегмент тривалістю 1 с розбивається на 40 кадрів, кожен з яких триває 25 мс. Інтервал часу між моментами початку сусідніх кадрів називається *тривалістю стрибка*. Оптимальне його значення дорівнює 10 мс, відповідно, після обчислення перетворення Фур'є для кожного кадру, віконна функція зсувається вперед на 10 мс, що призводить до перекриття 15 мс між сусідніми кадрами.

Щоб ***отримати спектр*** сигналу, що змінюється в часі, необхідно розділити сигнал на фрейми та помножити кожен з них на віконну функцію. Вікна Ханна і Хеммінга є гарним вибором для аналізу аудіосигналів, і вибір між ними залежить від конкретних вимог програми. Якщо частотна роздільна здатність важливіша за точність у часовій області, *вікно Ханна* є кращим вибором. Але у завданні ефективного оброблення акустичних сигналів МЛА точність у часовій області важливіша, отже обираємо *вікно Хеммінга*.

Після операції згладжування вікном Хеммінга добре проявляються гармонійні складові аудіосигналу, тому за допомогою алгоритму ШПФ знайдемо амплітудний спектр та всю інформацію про фазову складову сигналу.

Оптимальну довжину ШПФ вибирають більшою або рівною довжині вікна Хеммінга, щоб запобігти спектральному витоку, коли енергія від частотної складової "витікає" в сусідні біни, а також досягти хорошої частотної роздільної здатності, зберігаючи при цьому баланс між обчислювальною ефективністю та точністю аналізу.

3 Виділення ознак

Виділення звукових ознак — це процес вилучення значущої та релевантної інформації з необроблених аудіосигналів, щоб представити їх у вигляді, який можна легко проаналізувати та обробити алгоритмами машинного навчання. Виділені ознаки використовуються як вхідні дані для моделі. Підхід глибокого навчання розглядає неструктуровані аудіопредставлення, процес виокремлення ознак відбувається автоматично. Основними представленнями в НМ є спектрограми, мел-спектрограми та мел-частотні кепстральні коефіцієнти. Розглянемо їхні переваги та недоліки.

Спектрограму одержують шляхом локального застосування ШПФ до сегментів сигналу. Вона забезпечує детальну візуалізацію частотного складу аудіосигналу в часі і може фіксувати перехідні та змінні в часі характеристики сигналу, тому їх використовують для вилучення таких характеристик, як висота, ритм і тембр музичних звуків. Вагомими недоліками є їх висока розмірність та потреба ретельного розгляду співвідношення час-частота.

Мел-спектрограма використовується як представлення ознак в аудіоаналізі, розпізнаванні мови, обробці музики та інших подібних завданнях. Кожен її стовпчик представляє величину частотного вмісту аудіосигналу в різні проміжки часу з акцентом на релевантних для сприйняття частотних мел-діапазонах. Вона зберігає більше спектральної інформації. Мел-спектрограму отримують в результаті переходу від частоти (f) до шкали мела (m) за формулою:

$$m = 2595 \cdot \lg\left(1 + \frac{f}{700}\right). \quad (1)$$

Інформацію про швидкість зміни спектральних смуг сигналу дає *кепстр* — дискретне косинусне перетворення (ДКП) логарифма спектра часового сигналу. Отриманий спектр не лежить ні в частотній, ні в часовій області [11]. *Мел-частотні кепстральні коефіцієнти* (MFCC) — це коефіцієнти, що складають мел-частотний кепстр та передають значення, які будують форманти і тембр звуку, тому вони широко використовується в задачах розпізнавання мови, ідентифікації джерел звуку, класифікації аудіо та інших завдань, які передбачають аналіз спектрального вмісту сигналу [12]. MFCC стійкі до шуму та варіацій джерел, фіксують важливу спектральну інформацію та характеристики аудіосигналу, зменшуючи при цьому розмірність простору ознак. З недоліків можна виділити втрати деяких дрібнозернистих спектральних деталей при середніх обчислювальних ресурсах, тому вони менше підходять для завдань, де важлива точна інформація про частоту. Щоб визначити MFCC, потрібно до логарифмічно масштабованих виходів банку мел-фільтрів

застосувати ДКП:

$$X(k)=\sum_{i=1}^{N-1} x(n)cos\left(\frac{\pi n}{N}\left(k-\frac{1}{2}\right)\right),\ k=0,1...N, \quad (2)$$

де $X(k)$ є вихідною послідовністю коефіцієнтів ДКП, $x(n)$ — вхідний аудіосигнал довжиною N.

Вибір кількості MFCC, що відповідають найбільш інформативним спектральним характеристикам аудіосигналу МЛА, повинен враховувати такі фактори, як характеристики сигналу, обчислювальні ресурси та компроміс між інформативністю й розмірністю. Орієнтовна загальна кількість кепстральних коефіцієнтів (n_MFCC) розраховується за наступною формулою:

$$n\_MFCC = nBands + 2 \cdot k_\delta, \quad (3)$$

де **nBands** — це кількість банків мел-фільтрів (4), а $\mathbf{k}_\delta$ — кількість додаткових дельта- або дельта-дельта-коефіцієнтів, що представляють собою відповідно першу та другу похідні від (2).

Банк мел-фільтрів — набір трикутних фільтрів, рівномірно розташованих уздовж шкали мел (Рис. 2).

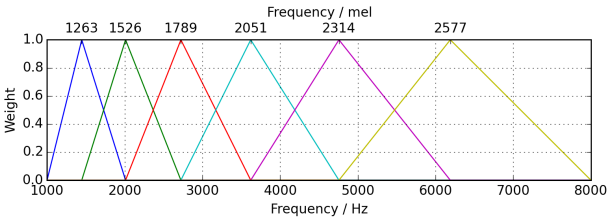

Рис. 2. Банк мел-фільтрів на шкалі мел

Кількість фільтрів мел вибирається на основі таких кроків:

1. Визначимо частотний діапазон аналізу. Він напряму залежить від характеристик аудіосигналу та дорівнює інтервалу від 0 до частоти дискретизації ($\mathbf{f_s}$) поділеній навпіл (за теоремою Найквіста).

2. Наступним кроком є визначення бажаного значення частоти роздільової здатності ($\mathbf{f_{res}}$).

3. Вираз для обчислення кількісті банків мел-фільтрів (**nBands**) випливає з нелінійного перетворення (1) і бажаної частотної роздільної здатності:

$$nBands = 2595 \cdot \frac{\lg(1+\frac{f_s}{2 \cdot 700})}{2 \cdot f_{res}} - 1. \quad (4)$$

Розрахуємо кількість фільтрів мел, що використовуються для аналізу мел-спектра аудіосигналів МЛА у розроблювальному алгоритмі. Діапазон частот аналізу знаходиться між 20 Гц та 20 кГц, що становить частотні межі роботи обраної моделі мікрофону [10]. Бажана частотна роздільова здатність 50 Гц, тому за (4) кількість мел смуг дорівнює 30.

Для зважування коефіцієнтів ШПФ кожен трикутний фільтр у банку фільтрів мел потрібно згорнути зі спектром, отриманим у результаті ШПФ. Вихідні дані банку фільтрів обчислюються як зважена сума значень коефіцієнтів ШПФ в межах частотного діапазону кожного фільтра. Щоб знайти частотний діапазон кожного мел-фільтра ($\mathbf{range_{Mel}}$), необхідно розділити шкалу мел на рівні інтервали (N):

$$range_{Mel} = \frac{m}{nBands+1} \cdot N, \quad N = 1...nBands+1. \quad (5)$$

А потім перевести мел-діапазон назад у лінійну шкалу частот, використовуючи зворотну формулу:

$$f = 700 \cdot (10^{\frac{range_{Mel}}{2595}} - 1). \quad (6)$$

Знаючи частотні межі роботи обраної моделі мікрофону, із (1) випливає, що весь мел-частний діапазон знаходиться у проміжку від *1* до *3817* Мел.

Розрахуємо межі частот першого діапазону (N=*1*) у алгоритмі аналізу сигналів. Щоб визначити нижню межу, у (5) підставимо значення (1) (*1* Мел) при мінімальній робочій частоті мікрофону (*20* Гц) та потрібну кількість мел-фільтрів, обрану за (4) ($nBands$ = *30*). Отже, нижня межа дорівнює *1* Мел або *0.64* Гц. Аналогічну послідовність дій проведемо і для розрахунку верхньої межі, але вже при максимальній частоті роботи мікрофону (20 кГц). Відповідно $\mathbf{range_{Mel}}$ (N = 1) дорівнює *123* Мел, приблизно *81* Гц (6). Таким чином можна легко знайти частотний діапазон кожного мел-фільтру підставляючи у (5) бажаний номер фільтра (N).

Виходи банку фільтрів часто логарифмічно масштабуються за допомогою функції логарифмічного стиснення. Це допомагає наблизитися до нелінійного частотного сприйняття гучності і забезпечує стиснення динамічного діапазону спектра енергії.

Таким чином, найбільш ефективними методами фіксації локальних характеристик, що ляжуть у основу роботи алгоритму класифікації та розпізнавання сигналів від дронів, є *мел-спектрограми* та *MFCC*.

4 Підготовка навчальних даних

Після етапів попередньої обробки аудіосигналів, перетворення їх у формат, придатний для подачі на вхід НМ, вихідний набір даних необхідно розділити на три набори: *дані навчання, перевірки (валідації)* та *тестування*. Співвідношення розподілу залежить від розміру набору даних, тому стандартною практикою є розділення 60/20/20 або 70/15/15 [13].

Навчальний комплект застосовується для навчання НМ, яка буде класифікувати дані, за допомогою завантаження розмічених прикладів аудіошумів різних моделей МЛА. Набір валідації — для налаштування гіперпараметрів. Під час процесу навчання продуктивність моделі оцінюється на перевірочному

наборі після кожної епохи, а гіперпараметри відповідно коригуються. Дані перевірки грають вагому роль у прийнятті рішення про припинення навчання контрольованої НМ, вони допомагають збалансувати компроміс між переобладнанням і недостатнім припасуванням (нездатністю зафіксувати закономірності в даних). Дані тестування важливі для оцінки продуктивності моделі навченого класифікатора. Вони використовуються для вимірювання точності моделі та її здатності узагальнювати нові, невідомі дані.

5 Створення та навчання моделі

Для отримання архітектури ЗНМ для класифікації аудіосигналів МЛА, розташуємо послідовно три головні типи шарів: згортковий, підвибірки та повнозв'язний (Рис. 3). Проаналізуємо детально кожний шар моделі мережі.

5.1 Аналіз архітектури моделі ЗНМ

Вхідний шар (*Input Layer*) приймає вхідні дані у вигляді зображення, тому для аналізу аудіосигналів їх попередньо необхідно представити як двовимірне зображення, що і було виконано за допомогою Мел-спектрограми — часо-частотне представлення. Розмір вхідних аудіоданих тривимірний і дорівнює відношеню кількості часових інтервалів в аудіофайлі до кількості мел-фільтрів та аудіоканалів відповідно.

Обчислювальні **шари згортки** (*Convolution Layer*) особливо залежать від правильно підібраного розміру та кількості фільтрів. Розмір фільтра визначає поле шару згортки. Найпоширенішими варіантами є 3×3, 5×5 або 7×7. Менші фільтри захоплюють дрібніші деталі, тоді як більші розміри захоплюють більш глобальні особливості, тому для ефективного аналізу складових аудіосигналу краще використовувати фільтр 3×3.

Для визначення кількості фільтрів у згортковому шарі, потрібно зважати на такі фактори:

1) Кількість фільтрів відповідає кількості вихідних каналів (карт ознак), створених цим шаром. Вона залежить від розміру вхідних даних і бажаного розміру вихідних карт ознак. Кількість фільтрів у першому шарі обирають у діапазоні від 10 до 30.

2) Збільшення кількості фільтрів дозволяє моделі вловлювати дрібніші деталі та закономірності у вхідних даних. Однак більша кількість фільтрів також збільшує кількість параметрів у мережі, що призводить до довшого часу навчання та більшого ризику перенавчання, особливо якщо набір даних невеликий.

3) Кількість фільтрів збільшується в міру заглиблення в мережу — *«пірамідальна»* архітектура. Вона дозволяє мережі поступово вивчати представлення більш високого рівня, збільшуючи кількість фільтрів. Глибші шари можуть вивчати більш абстрактні ознаки, побудовані на низькорівневих ознаках, вивчених у попередніх шарах. Простою стратегією для збільшення потужності моделі є подвоєння кількості фільтрів у кожному наступному шарі згортки.

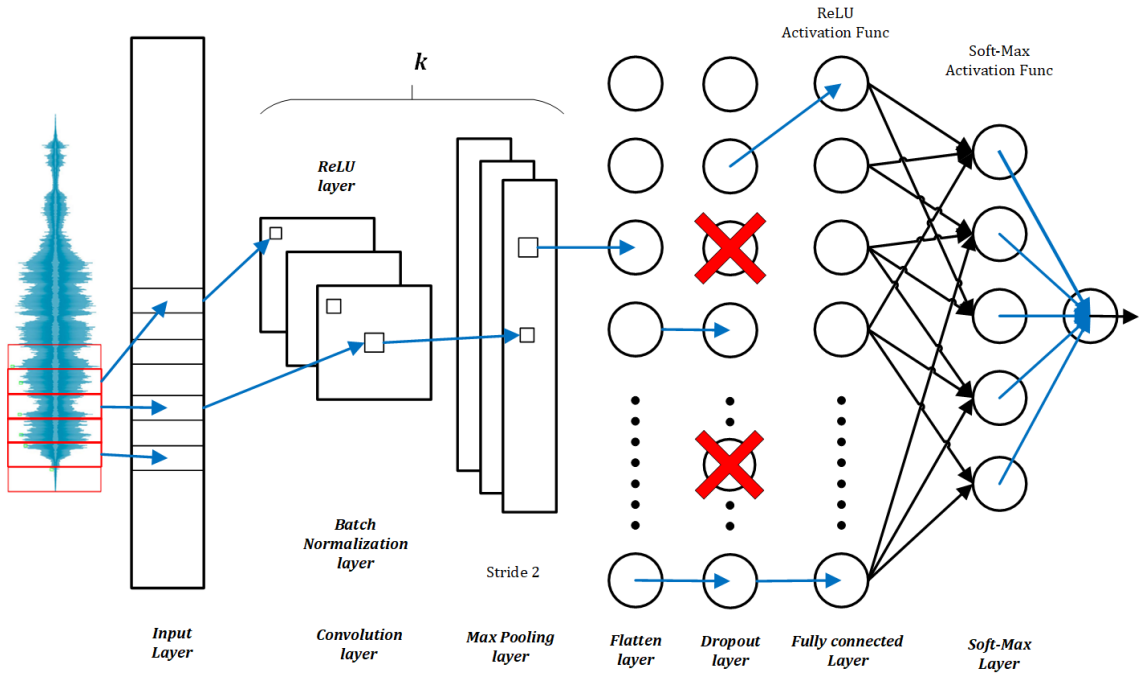

Рис. 3. Архітектура ЗНМ для класифікації аудіосигналів МЛА

Додатковим параметром при роботі з даними у шарах згортки є *падинг* — доповнення країв «фейковими» пікселями. Він може бути «однаковий» (*same*), тоді вхідні дані доповнюються нулями так, щоб вихідні дані мали ті самі просторові розміри, що й вхідні. Якщо ж він «дійсний» (*valid*), то доповнення не відбувається і вихідні карти ознак мають зменшені просторові розміри. Вибір типу заповнення залежить від бажаного розміру вихідної карти і від того, чи важливим є збереження просторової інформації.

Отже, розроблена ЗНМ буде спроектована за *«пірамідальною»* архітектурою, кожен згортковий шар буде складатись з $2n \cdot filters$ фільтрів з розміром 3×3, де n — номер згорткового шару, а $filters = 15$ — середнє значення діапазону початкової кількості каналів. Падинг встановлений на «*same*», щоб вхідні та вихідні карти об'єктів мали однакові просторові розміри.

Щоб дозволити мережі вивчати більш складні представлення, введемо нелінійність за допомогою шару з функцією активації (передавальною функцією). Вона визначає функціональні можливості НМ: які нейрони будуть активовані, яка інформація буде передаватися наступним шарам та метод навчання мережі [14]. В алгоритмі глибокого навчання краще використовувати функцію випрямленої лінійної одиниці **ReLU**, яка є обчислювально ефективною, оскільки просто встановлює в нуль від'ємні значення виходу попереднього шару, а додатні залишає незмінними. Це перетворення не змінює розмірність даних та дозволяє створювати більш потужні та виразні моделі. Активація ReLU може сприяти розрідженим представленням, що може допомогти зробити мережу більш стійкою до шуму і зменшити перенавчання, зосередившись на найбільш релевантних ознаках.

Для прискорення навчання, збіжності та зменшення чутливості до мережевої ініціалізації, між згортковими шарами та передавальною функцією, застосуємо **шар пакетної нормалізації** (*Batch Normalization Layer*). Це покращує стійкість моделі до ініціалізації, градієнтний потік та нівелює проблему внутрішнього коваріаційного зсуву — зміні розподілу входів шару, яка відбувається під час навчання, коли оновлюються параметри попередніх шарів, що робить процес навчання більш повільним та складним.

Шар максимального об'єднання (*Max Pooling Layer*) застосовується для поступового зменшення просторового розміру вхідних даних, зменшуючи кількість параметрів і обчислювальну складність, зберігаючи при цьому найважливіші ознаки. Він працює з кожною картою ознак незалежно, витягує максимальне значення в межах певного вікна або розміру пулу, допомагає виділити найбільш домінуючі ознаки, роблячи мережу більш стійкою до невеликих варіацій вхідних даних. Операція максимального об'єднання застосовується з фіксованим кроком, що визначає величину зміщення вікна об'єднання після кожної операції. Найпоширенішим підходом є використання розміру об'єднання 2×2 або 3×3 і кроку *страйдингу* 2 — пропуск кожного другого елементу, що призводить до зменшення просторових розмірів карт удвічі. Перевагами використання цього шару є:

1. Зменшення розмірності, що допомагає контролювати надмірне припасування і зменшує обчислювальну складність наступних шарів.

2. Інваріантність трансляції, якщо об'єкт або ознака трохи зміщується в межах вікна об'єднання, максимальне значення все одно фіксується, зберігаючи важливу інформацію.

3. Стійкість до невеликих просторових змін.

Однак максимальне об'єднання може призвести до втрати просторової інформації, оскільки зберігається лише максимальне значення в межах кожного вікна об'єднання. У деяких випадках ця втрата інформації може бути компромісом між зменшенням обчислювальної складності та збереженням дрібних деталей. Для усунення цього недоліку можна використовувати середнє або адаптивне об'єднання.

Згорткові та шари максимального об'єднання чергують між собою для ефективного вивчення ознак вищої складності та абстракції, а також зменшення просторових розмірів. Визначення кількості таких пар (k) в архітектурі ЗНМ (Рис. 3) залежить від різних факторів. Якщо задача вимагає врахування складних і дрібнозернистих особливостей, то мережа з більшою кількістю згорткових шарів може бути корисною, тому завдання виявлення об'єктів або сегментація, виграють від глибини архітектури. Більший розмір і різноманітність наборів даних дозволяють створювати глибші архітектури, оскільки вони надають більше прикладів, на яких мережа може вчитися. Однак, маючи обмежені дані, важливо уникати надмірного пристосування надто складних моделей. Також глибокі архітектури вимагають і більше пам'яті та обчислювальної потужності апаратного забезпечення для навчання. Важливо досягти балансу між складністю моделі та доступними ресурсами. Оптимальним є поступове зменшення просторових розмірів шляхом застосування максимального об'єднання після кількох шарів згортки. Це допомагає охопити та узагальнити локальні особливості, зберігаючи при цьому керований розмір карт об'єктів.

Ліміт кількості пар шарів задає параметр *timePoolSize* у функції створення останнього шара максимального пулинга і дорівнює кількості часових інтервалів в аудіофайлі поділеній на 10. Вона необхідна для підтвердження того, що зона об'єднання часових інтервалів менша за вхідну спектрограму, щоб уникнути потенційної невідповідності розмірів під час навчання. Велике значення змінної призведе до агресивного даунсемплінгу, зменшу-

ючи часову роздільну здатність вхідних даних. Це може бути корисно, коли часова динаміка даних не є критично важливою або при роботі з довгими послідовностями для зменшення обчислювальних вимог. З іншого боку, мале значення зберігає більше часової інформації і корисне, коли для задачі важливі дрібнозернисті часові патерни.

Розраховане значення кількості пар шарів згортки та максимального об'єднання розроблювальної моделі архітектури НМ дорівнює 10, при цьому розмір вихідних даних згорточного шару > *timePoolSize*.

Шар згладжування (*Flatten Layer*) бере вихідні дані останнього шару максимального об'єднання і перетворює частотно-часову репрезентацію аудіосигналів у одновимірний (згладжений) вектор, ефективно «вирівнюючи» просторову структуру та зменшуючи кількість параметрів, що дозволяє приєднати його до наступних шарів, які потребують одновимірного входу. При цьому він зберігає інформацію про вихідні просторові виміри. Це допомагає зменшити надмірну підгонку і підвищити обчислювальну ефективність.

Шар відсікання (*Dropout Layer*) випадковим чином встановлює частину вхідних одиниць на нуль при кожному оновленні, фактично «відсікаючи» ці одиниці. Це запобігає груповій адаптації нейронів і змушує мережу покладатися на підмножину одиниць для прогнозування, таким чином запобігаючи надмірній залежності від специфічних особливостей та запам'ятовуванню навчальних даних.

Наступним кроком є створення **повнозв'язного шару** (*Fully Connected layer*), де кожен нейрон з'єднаний з кожним нейроном попереднього шару. Він помножує вхідні елементи на вагову матрицю, а потім додає вектор зміщення. Шар визначає кількість вихідних класів або категорій у задачі класифікації і обчислюється шляхом підрахунку числа унікальних категорій в масиві даних навчальної вибірки.

Повністю зв'язаний шар виконує лінійне перетворення вхідних даних, після чого застосовується активаційна функція для введення нелінійності. Це дозволяє мережі вивчати складні взаємозв'язки між вхідними характеристиками та цільовими класами.

Використовуючи **шар Softmax** нормалізуємо вихідні значення повнозв'язного шару для отримання прогнозованих ймовірностей, що вказують на ймовірність належності заданих вхідних даних до кожного класу. Функція *Softmax* підносить вхідні значення до ступеня експоненти і нормалізує їх за сумою всіх експоненційованих значень. Це гарантує, що отримані значення додатні і в сумі дорівнюють 1, представляючи ймовірності. Клас з найвищою ймовірністю обирається як передбачуваний для вхідних даних.

Отримане значення максимальної ймовірності пропускаємо крізь **шар класифікації** (*Classification layer*), який приписує максимальне значення ймовірності відповідній мітці класу для заданих вхідних даних на основі активації попередніх шарів. Таким чином на виході ми отримуємо мітку класу, а не ймовірність.

Classification Layer використовується разом з функцією активації *softmax* і функцією *перехресних ентропійних втрат*. Перша гарантує, що вихідні значення представляють ймовірності класів, а друга вимірює розбіжність між прогнозованими ймовірностями класів і справжніми мітками. Під час навчання мережа намагається мінімізувати втрати, налаштовуючи свої параметри для підвищення точності прогнозів.

Повна архітектура спроектованої згорткової нейронної мережі для проведення експериментальних досліджень роботи алгоритму виявлення та класифікації аудіосигналів МЛА, зображена на Рис. 4.

5.2 Навчання нейронної мережі

Процес навчання ЗНМ включає в себе подачу виділених ознак аудіозразків навчальних даних у модель, застосування оптимізаторів для оновлення коефіцієнтів ваги шляхом зворотного розповсюдження та обчислення зміщення для мінімізації функції втрат. Виберемо оптимізатор для алгоритму навчання НМ.

Оскільки на вхід подається великий набір аудіосигналів МЛА, рекомендується використовувати оптимізатори *Adam* або *RMSProp*. Вони адаптують швидкість навчання для кожного параметра на основі величини останніх градієнтів, що допомагає швидше збігатися і переміщатися через плоскі ділянки ландшафту втрат. Але Adam має перевагу, оскільки використовує оцінки не тільки першого (середнє значення), а і другого (дисперсія) моменту. Також він добре працює з наборами даних, які мають рідкісні градієнти, що часто трапляється в аудіоданих. Adam ефективно адаптує швидкість навчання до різної величини градієнта, дозволяючи моделі прогресувати навіть при рідкісних його оновленнях [15].

Зрештою, для поставленої задачі найкращим оптимізатором є **Adam**, оскільки він видає хороші результати вже за мінімального підгону параметрів. Але продуктивність залежить від правильно підібраних гіперпараметрів: розміру партії, швидкості навчання, кількості ітерацій та епох. Розрахуємо їхні значення для отримання максимальної продуктивності моделі.

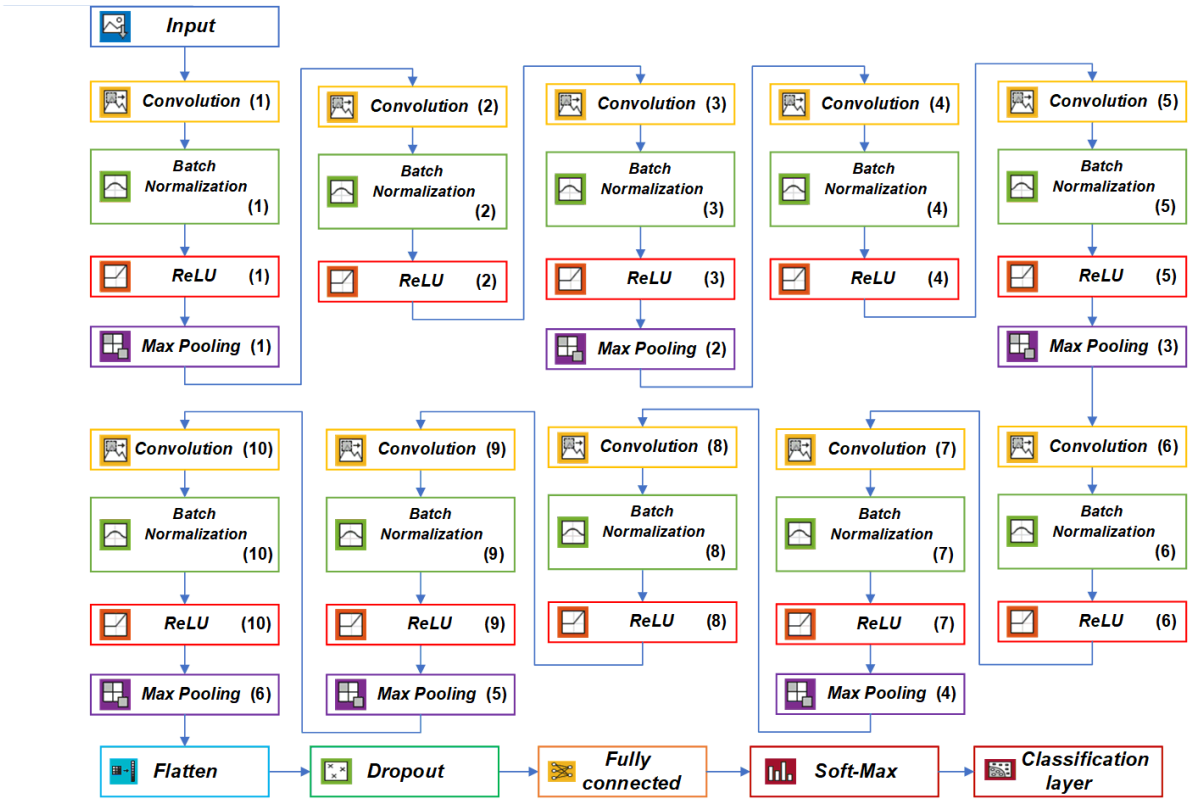

Рис. 4. Повна архітектура розробленої ЗНМ для класифікації аудіосигналів у Matlab

Міні-партія — це підмножина навчальної вибірки, яка обробляються при кожній ітерації для оцінки градієнта функції втрат та оновлення ваг моделі. Більший її розмір може призвести до швидшого навчання, оскільки паралельно обробляється більше прикладів. Однак це не завжди призводить до кращих результатів і може сповільнити навчання через обчислювальні обмеження. Важливо знайти баланс між швидкістю навчання та продуктивністю моделі. При навчанні на менших наборах даних або при роботі з моделями, які схильні до перенавчання, корисно використовувати менші розміри міні-партій. Вони можуть створювати більше шуму в оцінках градієнта, що матиме регуляризуючий ефект і покращує узагальнення. Однак, дуже малі розміри можуть призвести до нестабільності та повільнішої збіжності. На практиці починають з розмірів у діапазоні від 32 до 256.

Кількість *ітерацій* дорівнює частоті валідації даних НМ і визначається як загальна кількість навчальних даних поділена на розмір міні-партії. Навчання з великою кількістю ітерацій займає більше часу і вимагає значних обчислювальних витрат, у той час як навчання з малою кількістю ітерацій може призвести до того, що модель не буде пристосована. Тому важливо збалансувати доступні обчислювальні ресурси та бажаний час навчання. Якщо модель під час навчання не збігається, потрібно збільшувати кількість *епох*. Однак, якщо модель вже збіглася, навчання на великій кількості епох вже не дасть значних покращень. Для уникнення цього ефекту, впровадимо ранню зупинку. Якщо продуктивність мережі не покращується або починає погіршуватися після певної кількості епох, навчання зупиняється достроково.

Також при надмірному пристосуванні моделі можна застосувати метод перехресної валідації — розбиття набору даних на декілька підмножин для навчання моделі на різних згинах. Це дає можливість спостерігати продуктивність на різних епохах та визначати їхню кількість для вищого рівня узагальнення.

Відповідно до характеристик набору даних та архітектури мережі, налаштуємо *швидкість навчання* так, щоб після проходження 75% епох вона зменшилась у 10 разів. Це дозволить моделі робити менші оновлення в міру наближення до оптимального рішення, покращуючи збіжність. Почнемо із загальноприйнятого значення швидкості навчання 0.001.

6 Оцінка роботи та оптимізація моделі

Оцінимо продуктивність роботи моделі НМ на навчальному та валідаційному наборах. Мережа достатньо точна на них, однак навчальні, валідаційні та тестові дані мають схожий розподіл, який не завжди відображає реальні умови.

Повний час навчання системи склав 57 хв при швидкості 42 ітерацій на епоху. Загальна кількість ітерацій дорівнює 2300, а епох — 25. На основі порівняння передбачених і фактичних міток, відсоток помилок валідації дорівнює 0.3 % , а навчання — 0.297%. За отриманими даними, побудуємо матрицю помилок, яка показує кількість правильних і неправильних прогнозів, зроблених моделлю, порівняно з фактичними результатами (Рис. 5).

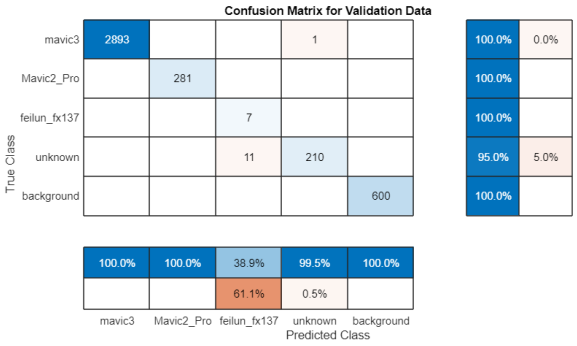

Рис. 5. Матриця помилок для набору даних валідації

Отримана матриця має розмір 5 на 5 елементів, основна діагональ якої відповідає правильно визначеним класам з навчальної вибірки, інші ж значення — усім комбінаціям помилкових рішень. Видно, що за отриманої кількості даних, помилка виникає при розпізнаванні моделі «*Feilun Fx137*».

Використовуючи матрицю помилок, визначимо повну оцінку якості роботи НМ, розрахувавши метрики оцінювання: precision, recall (Табл. 1) і оцінка F1, окремо для кожного класу та загальне значення.

Табл. 1 Значення метрик оцінювання ефективності

Класи	Точність	Recall
Mavic 3	0.9997	1.0000
Mavic 2 Pro	1.0000	1.0000
Feilun Fx137	1.0000	0.3889
unknown	0.9502	0.9953
background	1.0000	1.0000
Загалом	0.99	0.877

Точність («*precision*») — здатність моделі правильно відрізняти відповідний клас від інших класів. Вона розраховується як відношення істинно позитивних зразків — значення основної діагоналі матриці помилок, до суми істинно позитивних і хибнопозитивних зразків — сума значень відповідного рядка матриці.

Відновлюваність («*recall*»), також відома як чутливість або частота істинних спрацьовувань, обчислюється як відношення кількості істинно позитивних зразків до суми істинно позитивних і хибнонегативних зразків — сума значень відповідного стовпця матриці.

Аналізуючи результати Таблиці 1 видно, що загальна точність розпізнавання НМ вхідних аудіоданих складає 99.7 %, що є відмінним результатом для відносно невеликої кількості даних.

Оцінка *F1* — забезпечує збалансовану метрику оцінки — враховує помилки типу I (хибнопозитивні) та типу II (хибнонегативні):

$$F1 = \frac{2 \cdot precision \cdot recall}{precision + recall}. \quad (7)$$

За виразом (7) значення показника F1 дорівнює 0.93, що вказує на дуже високий рівень загальної продуктивності архітектури мережі з точки зору правильної ідентифікації позитивних зразків.

7 Тестування моделі

Протестуємо роботу моделі на наборі даних, які мережа не бачила в процесі навчання. Нажаль, через військовий стан виникає проблема наявності у вільному доступі та використанні різноманітних МЛА, тому у якості тестових даних будуть використовуватись аудіозаписи зависання лише моделі «*Mavic 3*», що зроблені на висоті 5 м та відстані до мікрофону від 10 до 300 м, аналогічно до плану та умов експерименту, проведеному у [10]. Подамо 20 фрагментів відповідних аудіосигналів МЛА на вхід навченої моделі ЗНМ і визначимо значення ймовірностей правильного спрацювання алгоритму (Рис. 6).

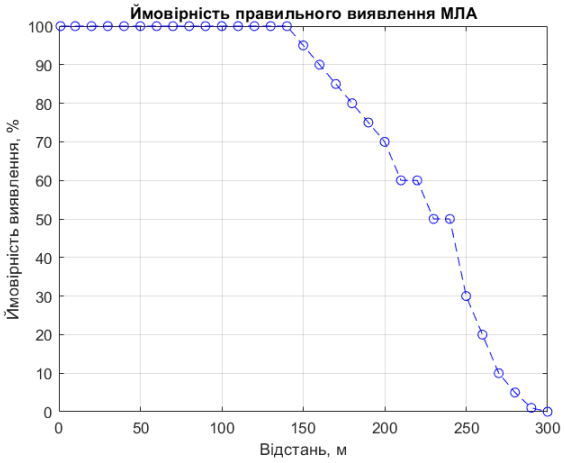

Рис. 6. Графік ймовірності правильного виявлення МЛА алгоритмом

На графіку видно, що 100% ймовірність виявлення безпілотників відбувається на дистанції до 140 м. Далі вона починає зменшуватись в залежності від збільшення дистанції до об'єкту.

Створимо критерій для поняття «*ефективне детектування*» — це відстань, на якій безпомилкове виявлення дронів відбувається з ймовірністю не менше 70%. Відповідно ефективна дистанція роботи алгоритму складає 200 м.

Для підтвердження якості алгоритму, використаємо метод експертних оцінок. Оскільки під час

роботи над пристроєм у його структурі не використовувались складні акустичні системи, то у ролі «експертів» будуть виступити 5 людей, які будуть визначати наявність дрону «Mavic 3» за таких самих умов, що й НМ. Отримані дані експерименту занесемо в Таблицю 2. Їхній аналіз дасть відповідь на питання: «чи зможе розроблена НМ замінити оператора у завданні ефективного виявлення МЛА?»

Табл. 2 Результати виявлення МЛА

Відст., м	**№1**	**№2**	**№3**	**№4**	**№5**	**НМ**
70	+	+	+	+	+	+
80	+	+	+	+	+	+
90	+	+	+	+	+	+
100	+	+/−	+	+/−	+	+
110	+/−	+/−	+	+/−	+	+
120	+/−	−	+	−	+	+
130	−	−	+/−	−	−	+
140	−	−	−	−	−	+
150	−	−	−	−	−	+
160	−	−	−	−	−	+

Аналізуючи отримані дані, можна зробити висновок, що розроблений алгоритм виявлення малих безпілотних літальних апаратів, завдяки вражаючому радіусу ефективного виявлення безпілотників в 200 метрів, перевершує діапазон людського слуху, який обмежений максимальною відстанню в 120 метрів.

Також, на відміну від людини, розроблений алгоритм працює не втомлюючись, забезпечуючи постійну пильність та своєчасне виявлення дронів у різних умовах середовища. Він може ефективно обробляти величезні обсяги аудіоданих, швидко розрізняючи звичайні навколишні звуки та виразні акустичні сигнатури МЛА. Крім того, алгоритм має значну перевагу над людиною з точки зору надійності та послідовності. Він усуває потенційну можливість людської помилки, суб'єктивності та обмежень, пов'язаних з людським сприйняттям. Автоматизуючи процес виявлення, він зменшує ймовірність помилкових спрацьовувань або пропущених виявлень, гарантуючи вищий рівень точності та загальної продуктивності. Як наслідок, він може замінити людський персонал в операціях спостереження за МЛА, пропонуючи підвищену ефективність, надійність і безпеку для громадян, об'єктів критичної інфраструктури та національної оборони.

Також на практиці, ефективна дистанція виявлення МЛА «*Mavic 3*» прототипом пристрою із програмним алгоритмом, вийшла майже у 3 рази більшою, ніж розраховане у [1] теоретичне значення для схожої за характеристиками моделі квадрокоптера «*DJI Mini 2 Fly More Combo*» іншим видом акустичного детектора (75 м).

В подальшому планується модернізувати пристрій, шляхом використання системи направлених мікрофонів. Це дозволить детекторам не тільки виявляти МЛА акустичним методом, а і визначати напрям підльоту об'єктів до цілі.

Висновки

Було створено алгоритм для виявлення та класифікації аудіосигналів МЛА із використанням методу машинного навчання.

1. Для підвищення якості набору аудіозаписів дронів, застосували методи попередньої обробки: нормалізацію, зменшення рівня шуму за допомогою Вінеровської фільтрації, сегментацію (поділили аудіосигнал на кадри тривалості 25 мс з перекриттям 50%) та керування вікном Хеммінга, оскільки у завданні оброблення акустичних сигналів точність у часовій області важливіша.

2. Для представлення даних спрощеним набором акустичних ознак, з кожного кадру оброблених аудіосигналів, визначили MFCC та мел-спектрограми для фіксації часових та спектральних характеристик. Діапазон частот аналізу знаходиться в межах від 20 Гц до 20 кГц, що становить межі робочих частот обраної моделі мікрофону, а бажана частотна розділова здатність 50 Гц, тому кількість робочих смуг мел дорівнює 30.

3. Для подачі на вхід НМ, дані розділили на три окремі набори: для навчання, перевірки (валідації) та тестування у співвідношенні 60/20/20.

4. Використовуючи навчальні дані та виділені ознаки аудіосигналів, розробили архітектуру моделі глибокого навчання на базі ЗНМ для проведення експериментальних досліджень роботи алгоритму виявлення та класифікації аудіосигналів МЛА. Вона складається із вхідного шару, 10 пар шарів згортки, ReLU, пакетної нормалізації та шарів максимального об'єднання. Їхня кількість визначається розміром вікна об'єднання вздовж часового виміру. Далі ідуть шари згладжування, відсікання, повнозв'язний та Softmax. Для нормалізації вихідних даних і отримання прогнозованих фінальних ймовірностей застосовується шар класифікації.

5. Найкращим оптимізатором для ефективного навчання моделі є Adam, оскільки він видає хороші результати вже за мінімального підгону параметрів.

6. Для підвищення точності і здатності до узагальнення моделі, застосували методи оптимізації: налаштування гіперпараметрів, швидкості навчання та методів регуляризації. Розмір міні-партії дорівнює 128, що забезпечує найкращий компроміс між використанням пам'яті ПК, швидкістю навчання та продуктивністю моделі. Швидкість навчання дорівнює 0.001, а після проходження 75% епох вона зменшується у 10 разів. Це дозволить моделі робити менші оновлення в міру наближення до оптимального рішення, покращуючи збіжність. Повний час

навчання системи складає 57 хв при швидкості 92 ітерацій на епоху. Загальна кількість ітерацій дорівнює 2300, а епох — 25.

7. Загальна точність розпізнавання НМ вхідних аудіоданих складає 99%. Оцінка F1 навченої моделі дорівнює 0.93, що демонструє високий рівень загальної продуктивності архітектури мережі з точки зору правильної ідентифікації позитивних зразків.

8. Радіус ефективного виявлення розробленого алгоритму складає 200 м, що перевершує діапазон людського слуху (120 м). Також, на відміну від людини, створена система забезпечує постійну пильність, може ефективно обробляти величезні обсяги аудіоданих, швидко розрізняючи звичайні навколишні звуки та виразні акустичні сигнатури МЛА, а також усуває потенційну можливість людської помилки, суб'єктивності та обмежень, пов'язаних з людським сприйняттям. Як наслідок, розроблена система може замінити людський персонал в операціях спостереження за МЛА, пропонуючи підвищену ефективність, надійність і безпеку для громадян, об'єктів критичної інфраструктури та національної оборони.

9. Надалі планується модернізувати пристрій, шляхом використання системи направлених мікрофонів. Це дозволить детекторам не тільки виявляти МЛА акустичним методом, а і визначати напрям підльоту об'єктів до цілі.

References

[1] Park S., Kim H. T., Lee S., Joo H. and Kim H. (2021). Survey on Anti-Drone Systems: Components, Designs, and Challenges. *IEEE Access*, Vol. 9, pp. 42635-42659. DOI: 10.1109/ACCESS.2021.3065926.

[2] Kozeruk S. O., Korzhyk O. V. (2022). Detection, Localization and Identification of Small Aircraft by Acoustic Radiation. *Visnyk NTUU KPI Seriia - Radiotekhnika Radioaparatobuduvannia*, Iss. 89, pp. 29-38. DOI:10.20535/RADAP.2019.76.15-20.

[3] Junfeng Guo, Ishtiaq Ahmad and KyungHi Chang (2020). Classification, positioning, and tracking of drones by HMM using acoustic circular microphone array beamforming. *EURASIP Journal on Wireless Communications and Networking*, Iss. 9, pp. 29-38. DOI:10.1186/s13638-019-1632-9.

[4] Yousaf, J., Zia, H., Alhalabi, M. et al. (2020). Drone and Controller Detection and Localization: Trends and Challenges. *EAppl. Sci.*, Vol. 12(24), pp. 1-22. DOI:10.3390/app122412612.

[5] Mazumder J., Raj A. B. (2020). Detection and Classification of UAV Using Propeller Doppler Profiles for Counter UAV Systems. *2020 5th International Conference on Communication and Electronics Systems (ICCES)*, pp. 221–227. DOI:10.1109/ICCES48766.2020.91380.

[6] Al-Emadi S., Al-Ali A. and Al-Ali A. (2021). Audio-Based Drone Detection and Identification Using Deep Learning Techniques with Dataset Enhancement through Generative Adversarial Networks. *Sensors*, Vol. 21(15), pp. 1-26. DOI:10.3390/s21154953.

[7] Subbotin S. O. (2020). *Neironni merezhi: teoriia ta praktyka: navch. posibnyk* [Neural Networks: Theory and Practice]. Zhytomyr: O. O. Yevenok, 184 p.

[8] Mahdavi F., Rajabi R. (2020). Drone Detection Using Convolutional Neural Networks. *2020 6th Iranian Conference on Signal Processing and Intelligent Systems (ICSPIS)*, pp. 1-5. DOI:10.1109/ICSPIS51611.2020.9349620.

[9] Zeghidour N., Xu Q., Liptchinsky V., et al. (2019). *Fully Convolutional Speech Recognition*, Vol. 2. pp. 1-5. DOI: 10.48550/arXiv.1812.06864.

[10] Sokolskyi S. O., Movchaniuk A. V. (2023). Electro-Acoustic Path of the Detector for Detection of Small Unmanned Aerial Vehicles. *Visnyk VPI*, Iss. 2, pp. 135-144. DOI:10.31649/1997-9266-2023-167-2-135-144.

[11] Singh J. (2019). An introduction to audio processing and machine learning using Python. *Opensource.com*, accessed on: Aug 19, 2023.

[12] Pratheeksha N. (2018). The dummy's guide to MFCC. *Medium.com*, accessed on: Sep 3, 2023.

[13] Ignatenko G. S., Lamchanovskii A. G. (2019). Classification of audio signals using neural networks. *International Scientific Journal «Young Scientist»*, Iss. 48(286), pp. 23-25.

[14] Nair Vinod; Hinton Geoffrey E. (2010). Rectified Linear Units Improve Restricted Boltzmann Machines. *27th International Conference on International Conference on Machine Learning*, pp. 807–814. DOI:10.5555/3104322.3104425.

[15] Diederik P. Kingma, Jimmy Lei Ba (2015). Adam: A method for stochastic optimization. *3rd International Conference for Learning Representations*, pp. 1-15. DOI:10.48550/arXiv.1412.6980.

Algorithm for Processing Audio Signals Using Machine Learning

Sokolskyi S. O., Movchanyuk A. V.

Small unmanned aerial vehicles (UAVs) rapidly develop and are implemented in various industries to make people's lives easier. However, there are potential risks in their use, such as unauthorized surveillance of critical infrastructure objects and the delivery of explosive devices, which poses a significant threat to public and national security. The acoustic method promises direction for solving this issue by analyzing the sound characteristics and Doppler shift signatures of UAVs, using microphone arrays and machine learning techniques. The aim of this article is to develop an algorithm for effective detection and classification of drone audio signals using a deep learning convolutional neural network (CNN), constructing its architecture, and evaluating its performance. Before submitting the drone audio dataset into the neural network, the quality of the audio recordings is improved through normalization, Wiener filtering, and segmentation. The audio is segmented into frames with a duration of 25 ms and a 50% overlap, applying Hamming windowing for better accuracy in the time domain, as temporal precision is crucial in audio signal processing. The obtained data is divided into three sets in a 60/20/20 ratio: for training, validation, and testing purposes. Next, the data is represented by a simplified set of features, extracting mel-spectrograms from each frame of the processed audio signals to capture their

temporal and spectral characteristics. The frequency range of analysis corresponds to the working frequency limits of the microphone model (20 Hz - 20 kHz), with a frequency resolution of 50 Hz and 30 working mel frequency bands. Using the training data and the extracted audio features, a neural network architecture is developed to investigate the performance of the drone detection and classification algorithm. It consists of 10 pairs of convolutional layers, ReLU activation, batch normalization, and max-pooling layers. The number of these layers is determined by the size of the pooling window along the time dimension. This follows by flattening, dropout, fully connected, and Softmax layers. A classification layer is applied to normalize the output data and obtain final probabilities. The Adam optimizer is chosen for model training. Based on the dataset set, the initial learning rate is set to 0.001, gradually decreasing by a factor of 10 after 75% of the epochs to enhance convergence. The accuracy of the input data recognition reaches 99%, and the F1 score of the trained model is 0.93, indicating a high level of overall architecture performance. The maximum distance of effective detection of drones by the algorithm is 200 m.

Keywords: drone; small unmanned aerial vehicle; spectrum; signal processing; signal detection; convolutional neural networks; deep learning