РОЗДІЛ ІІ. ІНФОРМАЦІЙНО-КОМП'ЮТЕРНІ ТЕХНОЛОГІЇ

DOI: 10.25140/2411-5363-2024-4(38)-134-149
УДК 004.2

Володимир Вікторович Казимир[1], Андрій Іванович Роговенко[2], Олексій Олександрович Карась[3]

[1]доктор технічних наук, професор, професор кафедри інформаційних та комп'ютерних систем
Національний університет «Чернігівська політехніка» (Чернігів, Україна)
E-mail: vvkazymyr@stu.cn.ua. **ORCHID:** https://orcid.org/0000-0001-8163-1119. **ResearcherID:** Q-2925-2016

[2]кандидат технічних наук, доцент кафедри інформаційних та комп'ютерних систем
Національний університет «Чернігівська політехніка» (Чернігів, Україна)
E-mail: arogovenko@gmail.com. **ORCID:** https://orcid.org/0000-0003-4594-5692. **ResearcherID:** G-3926-2014

[3]студент магістратури
Національний університет «Чернігівська політехніка» (Чернігів, Україна)
E-mail: oleksiykaras2016@gmail.com. **ORCID:** https://orcid.org/0009-0004-8862-7234. **ResearcherID:** JZT-2594-2024

ВИКОРИСТАННЯ ШТУЧНИХ НЕЙРОННИХ МЕРЕЖ ДЛЯ АУДІОКЛАСИФІКАЦІЇ НА МІНІКОМП'ЮТЕРНІЙ ПЛАТФОРМІ

У статті представлено результати аналізу досвіду практичного використання існуючих моделей штучних нейронних мереж для вирішення задач аудіокласифікації на мінікомп'ютерній платформі. Визначено основні виклики, пов'язані з обробкою шумів, недостатністю маркованих даних і адаптацією моделей до специфічних умов використання. Проаналізовано архітектурні особливості побудови типових моделей нейронних мереж для задачі аудіокласифікації. Запропоновано методику використання моделей CNN, RNN та трансформерів на мінікомп'ютері NVIDIA Jetson Nano та проведено оцінку їх продуктивності при класифікації аудіошумів БПЛА. Обґрунтовано напрямок подальших досліджень для вдосконалення методів навчання та оптимізації моделей аудіокласифікації.

Ключові слова: *штучні нейронні мережі; аудіокласифікація; мінікомп'ютер; NVIDIA Jetson Nano; БПЛА.*
Рис.: 10. Табл.: 8. Бібл.: 15.

Актуальність теми дослідження. Сучасні технології обробки аудіо сигналів значно розширюють можливості застосування штучних нейронних мереж (ШНМ) для вирішення завдань класифікації аудіо. Зростання доступності даних, обчислювальних ресурсів та вдосконалення алгоритмів глибокого навчання стимулюють дослідження та розробку нових підходів у цій сфері. Аудіо класифікація має широкий спектр застосувань, включаючи розпізнавання мови, класифікацію музичних жанрів, ідентифікацію емоцій за голосом, діагностику технічного обладнання за звуками та моніторинг навколишнього середовища, включаючи розпізнавання повітряних цілей.

Основними викликами у сфері класифікації аудіо є обробка шумів, недостатність маркованих даних та адаптація моделей до специфічних умов використання вбудованих систем. У зв'язку з цим значну увагу приділяють дослідженню продуктивності реалізації різних моделей ШНМ [1], в тому числі на графічних процесорах та FPGA. Ключовим аспектом при цьому є вибір ефективних методів навчання та інструментів для їх реалізації, що дозволяють підвищити класифікаційну здатність моделей ШНМ.

Постановка проблеми. Аудіо класифікація, як одна з ключових задач обробки аудіо сигналів, є критично важливою для таких галузей, як розпізнавання мови, музикознавство, системи безпеки, діагностика технічних систем та моніторинг середовища. Однак, не зважаючи на значний прогрес у розвитку алгоритмів глибокого навчання, існує ряд проблем, які обмежують ефективність та універсальність використання сучасних підходів до розпізнавання звуків.

Однією з основних проблем є висока чутливість моделей до шуму та варіацій у вхідних даних. Аудіо дані часто містять фонові шуми, накладення звуків або відображаються в неоднорідних умовах запису. Це ускладнює процес класифікації та вимагає застосування ефективних методів попередньої обробки сигналу.

Іншою проблемою є потреба у великих обсягах якісно маркованих даних для навчання моделей. Збір, анотація та обробка таких даних є трудомісткими та фінансово витратними, особливо для вузькоспеціалізованих завдань. Нерівномірний розподіл даних між класами також може призводити до незбалансованого навчання.

Крім того, виникає питання вибору оптимальної архітектури нейронних мереж та інструментів для конкретних умов їх використання, наприклад, як то мінікомп'ютери систем керування дронами [2]. У різних випадках можуть бути ефективними різні моделі побудови ШНМ, але недостатньо дослідженим залишається питання адаптації цих моделей до умов обмежених обчислювальних ресурсів вбудованих систем.

Таким чином, існує потреба в комплексному аналізі сучасних моделей ШНМ та методів їх навчання для вирішення задач класифікації аудіо на мінікомп'ютерних платформах, що дозволить визначити існуючі обмеження, перспективи та напрямки подальших досліджень.

Аналіз останніх досліджень і публікацій. У останні роки нейронні мережі стали основним інструментом для вирішення задач аудіо класифікації. Згорткові, або конволюційні, нейронні мережі (Convolutional Neural Networks, CNN) широко застосовуються для аналізу спектрограм аудіосигналів, демонструючи високу ефективність у класифікації звуків навколишнього середовища. Наприклад, у роботі [3] досліджено різні архітектури CNN для класифікації аудіо на великомасштабних наборах даних, а у [4] описано використання CNN для класифікації звуків навколишнього середовища, таких як гавкіт собак або шум поїздів, по спектрограмам. У [5] також обговорюється використання попередньо натренованих CNN моделей для аудіо класифікації.

Рекурентні нейронні мережі (RNN), зокрема модифікації на основі довгої короткочасної пам'яті (LSTM) та GRU, ефективно обробляють послідовні дані, що важливо для задач розпізнавання мови та музичних жанрів. У статті [6] розглядається використання RNN для класифікації звуків, зокрема пташиних та жаб'ячих записів. Попри значні досягнення, існують виклики, зокрема нестача маркованих даних для навчання моделей. У статті [7] описано використання RNN для класифікації слухових стимулів з ЕЕГ сигналів, що може бути корисним при роботі з обмеженими наборами даних.

Завдяки здатності враховувати довгострокові залежності в аудіоданих, набувають популярності трансформерні моделі, такі як Audio Spectrogram Transformer (AST), описані у [8].

Мета дослідження. Метою статті є дослідження сучасних методів навчання ШНМ та інструментів їх реалізації на мінікомп'ютерній платформі, що застосовуються для задач аудіо класифікації, з метою визначення їхніх переваг, обмежень та перспектив використання. Стаття спрямована на формування рекомендацій щодо вибору оптимальних підходів для вирішення різних завдань класифікації аудіо, враховуючи специфіку даних, ресурси та можливі виклики.

Порівняльна характеристика моделей ШНМ. Для завдань класифікації використовують багато різних моделей штучних нейронних мереж (ШНМ), але для аудіо класифікації найбільш популярні з них такі, як CNN, RNN , LSTM, GRU [9; 10].

Згорткові, або конволюційні, нейронні мережі (Convolutional Neural Networks, CNN) є одним із найбільш поширених підходів у нейронних мережах для задач класифікації, зокрема у сфері аналізу візуальних і звукових даних. Їх головна перевага полягає у здатності автоматично виділяти суттєві ознаки з вхідних даних, що робить CNN надзвичайно ефективними для задач, пов'язаних з обробкою складних структур.

Основою роботи CNN є згорткові шари, які виконують операцію згортки над вхідними даними. Під час згортки використовуються спеціальні фільтри (ядра), що ковзають по входу, виділяючи локальні патерни, такі як текстури, межі або частотні компоненти звуку. Для введення нелінійності у модель після кожної згортки застосовується функція

активації, найпоширенішою з яких є ReLU (Rectified Linear Unit). Це дозволяє мережі моделювати складні залежності між ознаками. Щоб зменшити розмірність даних і уникнути перенавчання, у CNN використовуються шари підвибірки (Pooling Layers), найчастіше Max Pooling. Цей метод знижує обчислювальні витрати, вибираючи максимальне значення в локальних областях представлення. На завершальному етапі додаються повнозв'язні шари (Fully Connected Layers), які об'єднують виділені ознаки для виконання задачі класифікації.

У задачах аудіо класифікації CNN найчастіше використовуються для аналізу спектрограм – двовимірних представлень аудіосигналів, де кожен піксель відображає інтенсивність певної частоти в конкретний момент часу. Це дозволяє CNN ефективно виявляти ключові особливості звукових сигналів, що важливо для класифікації звуків навколишнього середовища, музичних жанрів або мовлення. Однією з переваг CNN є їхня здатність автоматично виділяти ознаки без необхідності попереднього ручного аналізу, що спрощує процес моделювання. Також ці моделі добре масштабуються на великих наборах даних і демонструють високу стійкість до шуму за умови якісного навчання. Однак CNN мають свої обмеження. Вони вимагають великої кількості маркованих даних для ефективного навчання та значних обчислювальних ресурсів, особливо через складність згорткових операцій. Крім того, через локальну природу згорткових шарів CNN можуть бути менш ефективними для задач, де важливі довгострокові залежності. Практичне використання.

Рекурентні нейронні мережі є потужним інструментом для роботи з послідовними даними, такими як аудіосигнали, текст або часові ряди. Головною особливістю RNN є здатність моделювати залежності між елементами послідовності, оскільки вони враховують попередні стани під час обробки кожного наступного елемента. Це досягається завдяки наявності рекурентних зв'язків у мережі, які дозволяють передавати інформацію між вузлами на кожному часовому кроці. Така архітектура робить RNN ефективними для задач, де контекст має велике значення, наприклад, у класифікації аудіосигналів, які змінюються в часі.

Для подолання проблеми затухання градієнтів, яка може виникати під час навчання базових RNN, були розроблені їхні вдосконалені модифікації, такі як LSTM (Long Short-Term Memory) та GRU (Gated Recurrent Unit). Ці моделі мають механізми контролю потоку інформації через спеціальні «ворота», що дозволяє їм запам'ятовувати важливу інформацію протягом тривалих інтервалів часу та ігнорувати менш значущі дані. Завдяки цій властивості LSTM та GRU часто застосовуються у задачах класифікації мовлення, розпізнавання мелодій або аналізу складних звукових сигналів. Також на особливу увагу заслуговує модифікація даного класу моделей глибокого навчання CRNN (Convolutional Recurrent Neural Network), яка поєднує в собі згортову нейронну мережу (CNN) і рекурентну нейронну мережу (RNN) для обробки послідовних даних.

У задачах аудіо класифікації RNN зазвичай працюють із сирими аудіосигналами або їхніми обробленими представленнями, такими як мел-спектрограми чи послідовності ознак. Їх застосовують, наприклад, для аналізу мови, де важливо зберігати послідовність фонем, або для розпізнавання звуків, де послідовність акустичних подій визначає належність сигналу до певного класу. Проте RNN мають певні обмеження. Вони є складними в обчисленні, особливо на довгих послідовностях, через необхідність виконувати послідовні операції на кожному часовому кроці. Крім того, ефективність RNN значною мірою залежить від налаштувань гіперпараметрів і обсягів навчальних даних.

Трансформери є сучасною архітектурою нейронних мереж, яка кардинально змінила підхід до обробки послідовних даних, зокрема тексту, аудіо та навіть зображень. В основі трансформерів лежить механізм самоуваги (self-attention), який дозволяє кожному елементу послідовності враховувати вплив усіх інших елементів незалежно від їхньої позиції. Це вирішує проблему послідовного обчислення, властиву рекурентним мережам

RNN, і дає можливість обробляти довготривалі залежності в даних значно ефективніше. Завдяки паралельній обробці даних трансформери демонструють високу швидкість роботи та масштабованість навіть на великих наборах даних.

У задачах аудіо класифікації трансформери використовуються для аналізу часових рядів або спектрограм. Наприклад, аудіосигнал розбивається на фрагменти, кожен з яких перетворюється в багатовимірне представлення через ембеддинги. Потім ці представлення проходять через кілька шарів самоуваги, що дозволяє мережі враховувати як локальні, так і глобальні залежності у звукових даних. У кінцевому результаті трансформери виділяють найважливіші ознаки сигналу, які використовуються для класифікації.

Особливістю трансформерів є використання механізму позиційного кодування (positional encoding), який додає інформацію про порядок елементів у послідовності. Це важливо для аудіо аналізу, оскільки відносний порядок частотних чи часових компонентів часто несе критично важливу інформацію. У задачах, таких як розпізнавання мовлення чи класифікація музичних жанрів, трансформери дозволяють моделювати залежності між звуковими подіями, які віддалені одна від одної в часі.

Попри свої переваги, трансформери мають певні обмеження. Вони вимагають значних обчислювальних ресурсів, особливо для роботи з довгими послідовностями, оскільки обчислювальна складність механізму уваги зростає квадратично зі збільшенням довжини вхідних даних. Проте сучасні модифікації, такі як Sparse Transformers та Linformer, частково вирішують цю проблему, зменшуючи обчислювальні витрати без суттєвих втрат точності.

Порівняльна характеристика описаних вище моделей ШНМ наведена в таблиці 1.

Таблиця 1 – Порівняльна характеристика CNN, RNN, Трансформерів

Властивість	**CNN**	**RNN**	**Трансформери**
1	2	3	4
Призначення	Найкраще підходять для аналізу двовимірних даних, таких як спектрограми.	Ефективні для роботи з послідовними даними (аудіо, текст).	Оптимальні для задач з довгостроковими послідовностям.
Обробка даних	Приймають спектрограми як вхідні дані.	Працюють напряму із аудіоданими або їхніми ознаками.	Можуть працювати як із спектрограмами, так і з аудіоданими.
Архітектура	Складається зі згорткових шарів, що автоматично виділяють просторові ознаки.	Складається з рекурентних шарів, які зберігають стан для аналізу попередніх елементів послідовності.	Використовують механізм уваги для оцінки важливості всіх елементів послідовності одночасно.
Переваги	- Висока швидкість навчання. - Менша кількість параметрів, ніж у RNN та трансформерів. - Підходять для паралельних обчислень.	- Добре враховують часову динаміку даних. - Підходять для задач, де потрібен аналіз залежностей між елементами.	- Висока ефективність у задачах із довгими послідовностями. - Паралельна обробка даних забезпечує швидке навчання.
Недоліки	- Не враховують часову залежність у сирих даних. - Залежні від попередньої обробки (створення спектрограм).	- Проблема затухаючого градієнта у базових версіях. - Високі витрати обчислювальних ресурсів.	- Висока обчислювальна складність через механізм уваги. - Потребують великих обсягів даних для навчання.
Обчислювальна складність	Низька - підходять для мобільних і вбудованих систем.	Середня - залежить від довжини послідовності.	Висока – особливо на великих наборах даних.
Придатність для шумних даних	Вимагають попередньої обробки для усунення шуму.	Можуть бути чутливими до шуму в даних, потребують очищення.	Стійкіші до шуму завдяки механізму уваги, але залежать від навчальних даних.

Закінчення табл. 1

1	2	3	4
Типові задачі	Класифікація звуків навколишнього середовища, розпізнавання музичних жанрів.	Розпізнавання мовлення, аналіз емоцій за голосом, ідентифікація звукових патернів.	Трансляція мовлення, автоматичне розшифрування звуків, класифікація складних послідовностей.
Популярні моделі	ResNet, VGGish, AudioSet, YAMNet	LSTM, GRU, RNN-based Audio Classifiers	WaveNet, Audio Spectrogram Transformer (AST), Perceiver IO
Приклади інструментів	TensorFlow, PyTorch, librosa	Keras, TensorFlow, PyTorch	HuggingFace Transformers, PyTorch, torchaudio

Методика дослідження. Для оцінки продуктивності архітектур нейронних мереж у задачах аудіо класифікації на мінікомп'ютері NVIDIA Jetson Nano, що входить до складу системи керування захисного дрона, було обрано налаштування оточення, наведено нижче. Етапи дослідження включали:

1. Збір та вибір відповідного аудіодатасету для задачі. Для дослідження було обрано датасет UrbanSound8K [11], оскільки він має записи, які розподілено по 10 класів, що добре підходить для тестування швидкодії архітектур у задачах аудіокласифікації.
2. Обробка сирих аудіофайлів для підготовки до тренування моделі. Це включає такі кроки, як нормалізація, отримання спектрограм. Для CNN та трансформерів з аудіоданих було отримано MEL-спектрограми. Для RNN було отримано MFCC.
3. Розподіл на тренувальну, валідаційну та тестову вибірки.
4. Навчання моделей на платформі Jetson Nano. Було обрано стандартні моделі ResNet-18, на основі GRU та Audio Spectrogram Transformer для CNN, RNN та трансформерів, відповідно. Модель на основі GRU має архітектуру, яка складається зі 100 нейронів в першому шарі GRU, регуляризації Dropout в 0.5 і вихідним шаром з 10 нейронами для 10 різних класів з сигмоїдною активаційною функцією. Всі моделі налаштовано на класифікацію за 10 класами.
5. Оптимізація моделей. Для оптимізації було обрано Nvidia TensorRT для покращення швидкодії. Також використовуються 128 вбудованих CUDA ядер.
6. Оцінка продуктивності моделей. Вимірювання результатів роботи моделей за такими критеріями, як точність класифікації, швидкість обробки до та після оптимізації, використання ресурсів та енергоефективність.

Визначення вхідних даних для експериментів. Для задачі аудіо класифікації звуків БПЛА вибір відповідного набору даних є критичним етапом, оскільки він визначає якість навчання моделі. Для формування тренувального, валідаційного та тестового набору даних були використані набори даних які є в публічному доступі [12; 13]. У цих наборах даних міститься три основні класи аудіосигналів: звук руху дрона Parrot Mambo, звук руху дрону Parrot Bebop та інші звуки, що не є звуками дронів.

Аудіофайли з обраних наборів даних були переконвертовані в аудіофайли з однаковими параметрами. Параметри звукових файлів що входять до тренувального, валідаційного та тестового набору даних наведені в таблиці 2.

Таблиця 2 – Параметри аудіофайлів з датасету

Частота дискретизації	Бітрейт	Кількість каналів	Формат аудіофайлу
16 КГц	16 КБ/с	моно	wav

Для дослідів було проведено розділення конвертованих аудіофайлів на кілька менших сегментів аудіофрагментів, що дозволило алгоритму глибокого навчання вивчати ознаки більш точно в порівнянні з подачею всього запису одночасно. Іншою метою сегментації була оптимізація навчання моделі для розгортання в режимі реального часу, де критичним є час, необхідний для виявлення та ідентифікації.

За основу для визначення впливу розміру аудіосегмента на загальну продуктивність було взято експериментальні дані з [12], де міститься інформація щодо різних розмірів сегментів, таких як одна секунда, дві секунди та п'ять секунд, при чому сегментація на одну секунду перевершує по ефективності інші часові інтервали.

Хоча такий підхід може призвести до втрати деяких ознак з оригінальних аудіофрагментів, на момент проведення експериментів цей метод був єдиним доступним з практично перевірених методів для глибокого навчання на аудіовході з конвертацією аудіофрагментів у спектрограми. Різноманітні ознаки потім обчислюються з отриманих спектрограм алгоритмом навчання. На рис. 1 представлено приклад спектрограми аудіофрагмента тривалістю одна секунда, що містить звук дрона та аудіофрагмент випадкового шуму, такого як звук друкування.

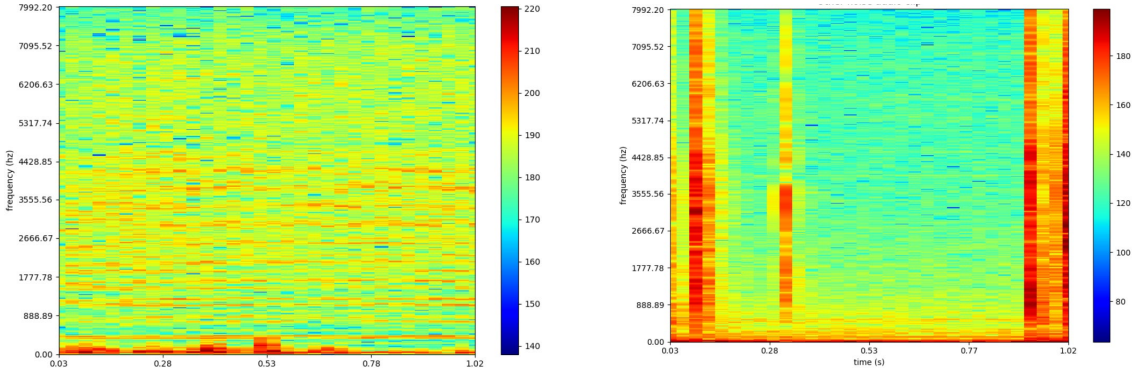

Рис. 1. Спектрограми сегмента зі звуком дрона та випадкового шуму

Для реалізації кожної з моделей CNN, RNN та CRNN було використано відкритий код, доступний у [14], який був модифікований відповідно до завдань даного дослідження. Проте треба зазначити, що код в [14] – це покращена версія прикладу розпізнавання аудіо з використанням бібліотеки TensorFlow, який наводиться в [15].

Реалізація моделі на основі CNN архітектури. Модель CNN дозволяє ефективно аналізувати аудіодані у вигляді спектрограм. Завдання моделі полягало у класифікації звуку по трьох категоріях: дрон Mambo, дрон Bebop та інші звуки.

У процесі дослідження використовувалась конволюційна модель ШНМ аудіорозпізнавання з CNN архітектурою, представленою на рис. 2.

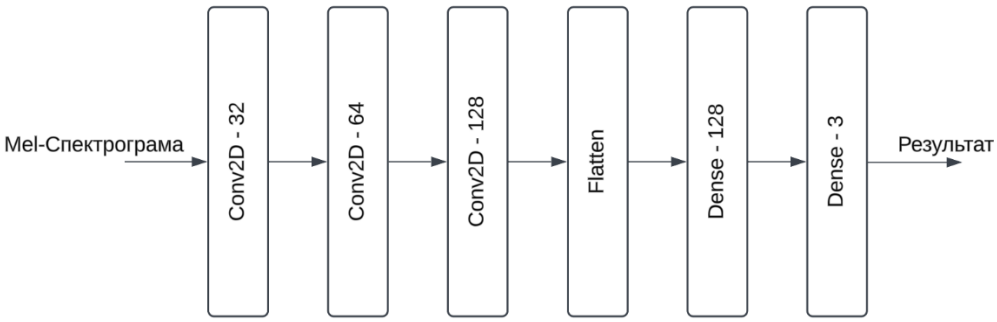

Рис. 2. Архітектура моделі на основі CNN

Перший конволюційний шар в 32 нейрони, який отримує дані у вигляді MEL-спектрограм, виявляє локальні патерни у спектрограмі, такі як особливі частоти чи періодичності. ReLU допомагає моделі вчитися нелінійним залежностям, усуваючи негативні значення. У цьому шарі створюється 32 вихідні карти ознак, кожна з яких відповідає певному фільтру. Також використовується шар пулінгу MaxPooling, який:

- зменшує розмірність вихідних даних, скорочуючи їх у 2 рази (з 128×128 до 64×64);

- видаляє шум, залишаючи найважливіші ознаки, що були виявлені конволюційними фільтрами;
- зменшує обчислювальні витрати, зберігаючи головну інформацію.

Другий конволюційний шар в 64 нейрони знаходить складніші ознаки на основі вхідних даних із першого шару пулінгу. На даному шарі збільшується кількість фільтрів, оскільки модель починає навчатися більш детальним характеристикам, таким як поєднання частот чи перехідні процеси в спектрограмі. Ще один шар пулінгу зменшує розмір вихідних даних із 64×64 до 32×32 та знижує ризик перенавчання, зменшуючи розмірність, але залишаючи найважливіші ознаки.

Третій конволюційний шар в 128 нейронів знаходить високорівневі ознаки, такі як складні патерни в спектрограмі, які можуть бути характерними для звуків дронів, оскільки на цьому етапі модель потребує більше потужності для аналізу комплексних характеристик, збільшується кількість фільтрів. Останній шар пулінгу зменшує розмір вихідних даних із 32×32 до 16×16.

Шар згортання перетворює двовимірні карти ознак в одновимірний вектор, що підготовляє дані для передачі в повнозв'язні шари. Повнозв'язні шари складаються з першого шару, який виявляє залежності між високорівневими ознаками та забезпечує моделі можливість виконувати більш складні класифікаційні завдання; та другого шару, який перетворює вихідні значення у ймовірності для кожного класу.

Результати навчання даної моделі наведено на рисунку 3.

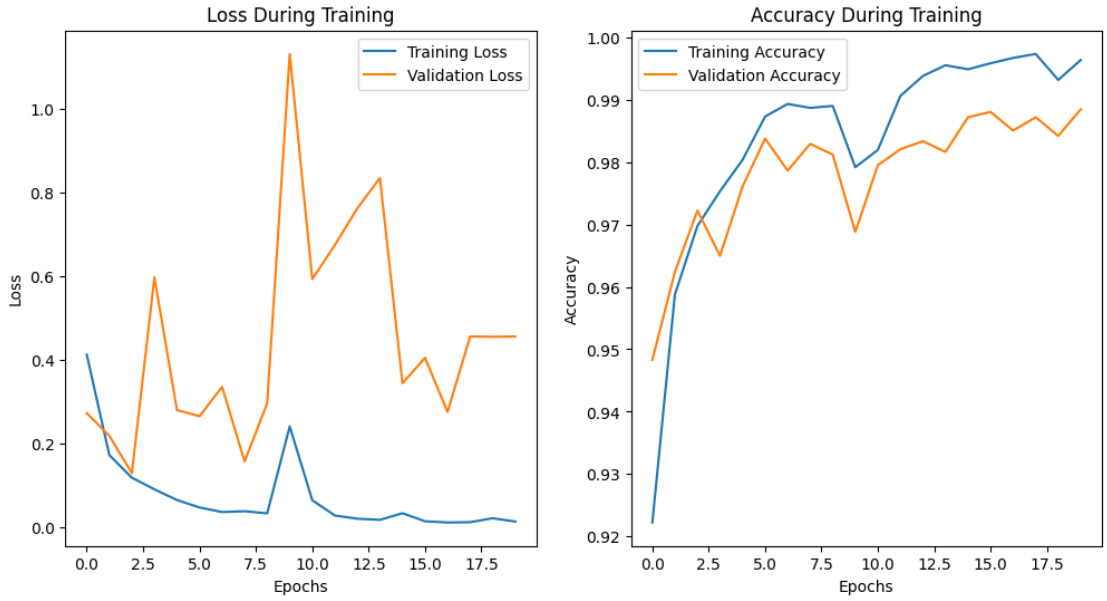

Рис. 3. Результат навчання моделі CNN

Під час навчання моделі тренувальна втрата поступово знижується, що вказує на ефективне навчання моделі. Валідаційна втрата має значні коливання, особливо на початку навчання, але згодом стабілізується. Це може свідчити про можливе перенавчання, яке вдалося уникнути до кінця епох. Тренувальна точність поступово зростає і досягає високих значень, що вказує на те, що модель добре "запам'ятала" тренувальні дані. Валідаційна точність також демонструє хороші результати, досягаючи 98-99%. Це свідчить про те, що модель здатна узагальнювати знання на нових даних, хоча деякі коливання можуть вказувати на невеликі проблеми з узагальненням.

Результати тестування моделі наведено в табл. 3.

Таблиця 3 - Результати тестування моделі CNN

Клас	Точність, %	Відклик	F-міра
Mambo	93	0,90	0,91
Bebop	96	0,95	0,96
Інші звуки	99	1	1

Під час тестування моделі метрики F-міри для всіх класів перевищують 0.90, що підтверджує гарний баланс між точністю та відкликом.

Реалізація моделі на основі GRU архітектури. Рекурентна нейронна мережа (GRU) була побудована для класифікації аудіозаписів у три класи: дрон Mambo, дрон Bebop та інші звуки, що не є дронами. Модель використовує часові ряди характеристик звуку, таких як мел-кепстральні коефіцієнти (MFCC), для виявлення звукових патернів, що характерні для кожного класу.

Було створено модель RNN з архітектурою, представленою на рис. 4.

Перший GRU шар у 128 нейронів зберігає залежності між часовими кроками в даних і повертає повну послідовність, щоб наступний шар GRU міг обробляти весь часовий ряд. На цьому кроці використовується стандартна активація для GRU та регуляризація, яка випадковим чином відключає 30 % нейронів у шарі GRU під час навчання для запобігання перенавчанню.

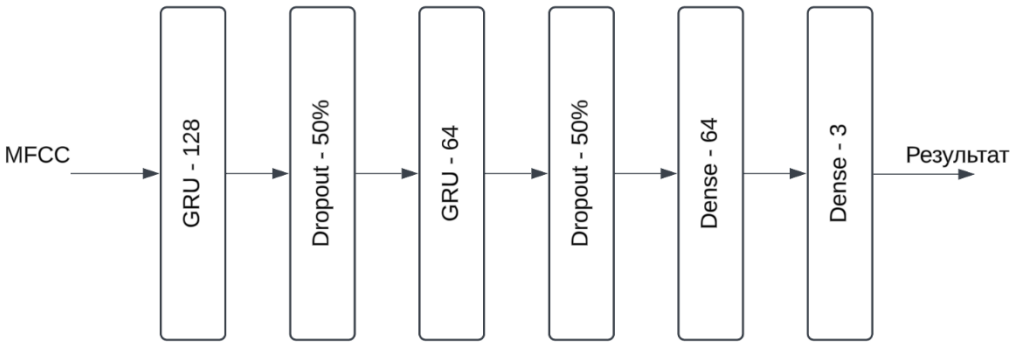

Рис. 4. Архітектура моделі на основі GRU

Другий GRU шар в 64 нейрони зменшує розмірність послідовності та надає більш узагальнені патерни. Він повертає лише останній стан GRU, представляючи всю послідовність у вигляді одного вектора. Також використовується регуляризація для другого GRU шару, яка запобігає перенавчанню.

Повнозв'язний шар з 64 шарами та вихідний повнозв'язний шар з 3 шарами, що є вихідним шаром для класифікації, повертає ймовірності належності до кожного з трьох класів. Використовується функція активації Softmax для інтерпретації виходів як ймовірностей.

Під час навчання моделі втрати на тренувальних даних поступово знижуються, як це представлено на рисунку 4, що свідчить про успішне навчання моделі. Валідаційні втрати коливаються після 5-ї епохи, що може свідчити про невелике перенавчання моделі на тренувальних даних. Тренувальна точність поступово зростає, досягаючи значення близько 99%. Валідаційна точність також зростає, але дещо повільніше і стабілізується на рівні близько 96–97%. Це свідчить про хорошу узагальнюючу здатність моделі.

Рис. 4. Результати навчання моделі GRU

Під час тестування моделі точність класифікації звуків дронів Mambo і Bebop дещо нижча відносно інших моделей, що може бути спричинено недостатньою кількістю тренувальних даних для кожного з цих класів. Спостерігається певне перенавчання моделі, що виражається в коливанні валідаційних втрат. Результати тестування моделі GRU представлені в таблиці 4.

Таблиця 4 – Результати тестування моделі на основі GRU

Клас	Точність, %	Відклик	F-міра
Mambo	83	0,83	0,83
Bebop	83	0,94	0,88
Інші звуки	99	0,98	0,99

Результати показують, що модель на основі GRU здатна класифікувати звуки дронів і розпізнавати інші звуки з достатньою точністю. Однак якість класифікації звуків класів Mambo і Bebop потребує покращення. Це може бути досягнуто за рахунок збільшення розміру датасету, оптимізації моделі та застосування додаткових технік обробки даних.

Реалізація моделі LSTM архітектури. Дана модель аудіокласифікації спеціально адаптована для роботи з часовими рядами, такими як аудіодані. Вона використовує мел-кепстральні коефіцієнти (MFCC) як вхідні ознаки, що дозволяє ефективно виділяти важливу інформацію з аудіозаписів. Архітектуру створеної моделі LSTM представлено на рис. 5.

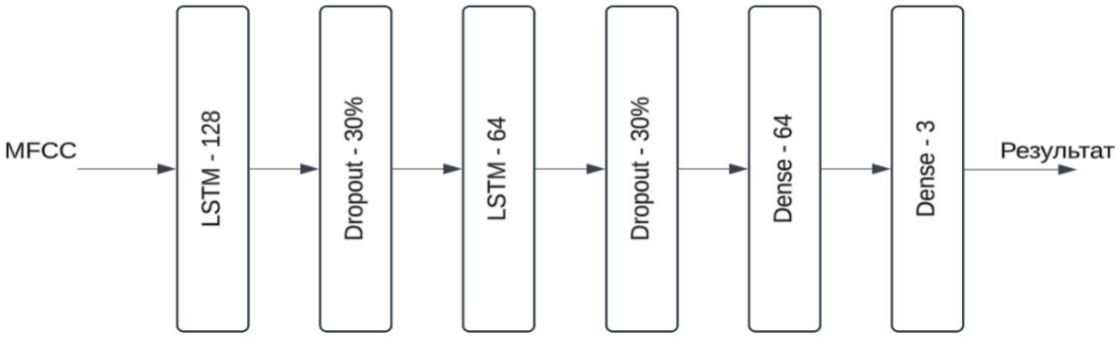

Рис. 5. Архітектура моделі LSTM

Перший шар із 128 нейронів аналізує кожен часовий крок і виділяє довготривалі залежності в аудіосигналі. Застосовується Dropout (0.3), щоб уникнути перенавчання. Він

передає повну послідовність виходів до наступного шару. Шар із 64-ма LSTM нейронів передає лише остаточний вихід до наступного шару. Цей шар конденсує інформацію, отриману від попереднього шару, у компактне представлення. Застосовується Dropout (0.3). Останніми шарами є два шари, перший з яких повнозв'язний шар з 64 нейронами та функцією активації relu. Цей шар обробляє вектор ознак, виділений LSTM-шарами. Другий вихідний шар із трьома нейронами та функцією активації softmax формує ймовірності для кожного з трьох класів.

Під час навчання моделі і тренувальна, і валідаційна втрати поступово зменшуються до стабільних значень, як це показано на рис. 6.

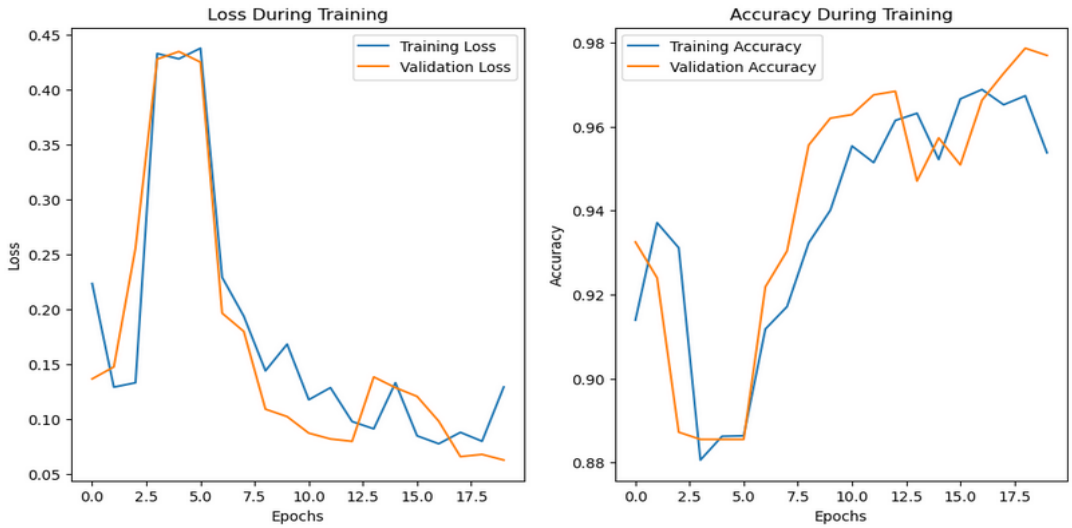

Рис. 6. Результати навчання моделі LSTM

Значення втрат для тренувального і валідаційного набору є близькими, що свідчить про відсутність перенавчання. Для перших кількох епох спостерігається нестабільність втрат, що є типовим для рекурентних моделей при роботі з часовими рядами. Точність для тренувальних і валідаційних даних зростає до 98 %. Тренувальна і валідаційна точності мають схожу динаміку, що свідчить про ефективне навчання без значного перенавчання.

Під час тестування, результати якого представлені у таблиці 5, для класів Mambo і Bebop результати були високі, хоча відклик для класу Mambo (0.71) свідчить про можливість пропуску частини цього класу. Загалом модель добре узагальнює дані і здатна класифікувати звуки трьох класів з високою ефективністю.

Таблиця 5 – Результати тестування моделі LSTM мережі

Клас	Точність, %	Відклик	F-міра
Mambo	92	0,71	0,80
Bebop	92	0,94	0,93
Інші звуки	98	1,00	0,99

Реалізація моделі на основі CRNN. Гібридна модель CRNN поєднує потужність конволюційних мереж CNN для виділення просторових ознак і рекурентних нейронних мереж, а саме LSTM, які аналізують часову структуру звукових даних. Ця архітектура дозволяє ефективно працювати з мел-спектрограмами – часово-частотними представленнями звуку. Модель було створено за архітектурою, представленою на рис. 7.

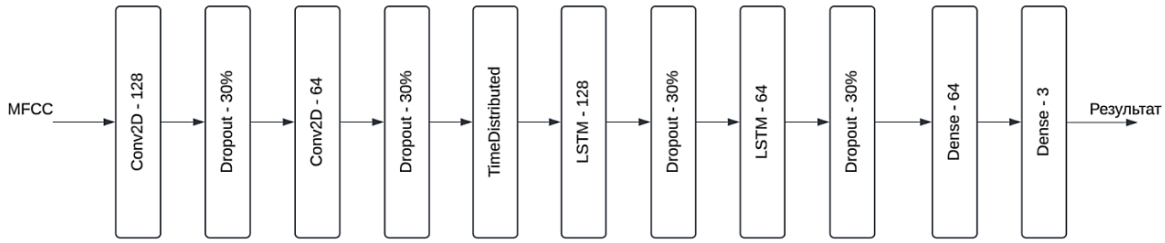

Рис. 7. Схема архітектури гібридної моделі на основі CNN та LSTM

У CNN частині використовуються фільтри розміром 3×3 для виявлення локальних патернів у спектрограмах (наприклад, гармоніки або частотні ряди). Вихідні дані кожного фільтру нормалізуються, зменшуючи ризик перенавчання і прискорюючи навчання. Просторовий розмір спектрограми скорочується для зменшення обчислювальної складності, залишаючи найважливіші ознаки. Після обробки в CNN результати перетворюються для передачі до рекурентної частини.

RNN частина, що складається з двох шарів LSTM (128 та 64 вузли). Перша LSTM (128) аналізує часові залежності у вхідних ознаках, при цьому забезпечується передача повної часової послідовності до наступного шару. Використовується Dropout (0.3) регуляризація для уникнення перенавчання. Друга LSTM (64) завершує аналіз часових залежностей, повертаючи лише остаточний.

Вихідна частина також складається з двох шарів: Dense (64 вузли, ReLU) забезпечує інтеграцію ознак із попередніх шарів для підготовки до класифікації, а Dense (3 вузли, Softmax) – це фінальний класифікатор на три класи.

Результати навчання моделі наведено на рис. 8, а результати тестування представлені в табл. 6.

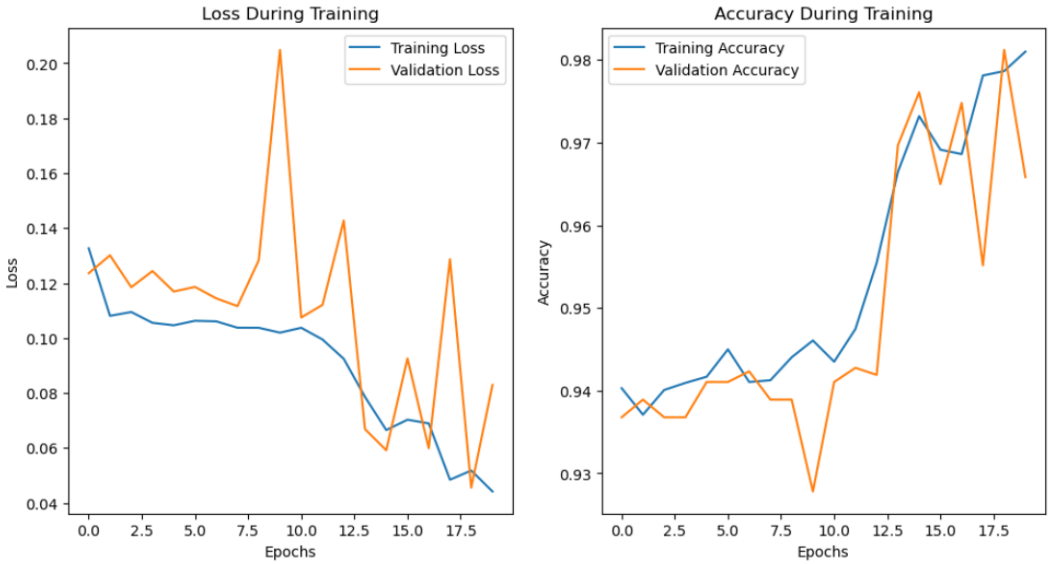

Рис. 8. Результати навчання гібридної моделі на основі CNN та LSTM

Як видно з рис. 10, спостерігається плавне зниження тренувальної втрати з кожною епохою, що свідчить про поступове навчання моделі. На етапах від 6 до 10 епох валідаційна втрата показує значні коливання, що свідчить про нестабільність моделі в узагальненні даних. Після 15 епох валідаційна втрата стабілізується та стає співмірною з тренувальною втратою. Незначні коливання валідаційних втрат можуть свідчити про те, що модель могла трохи перенавчитися на певних етапах, але після стабілізації вона демонструє адекватне узагальнення.

Таблиця 6 – Результати тестування гібридної моделі на основі CNN та LSTM

Клас	Точність, %	Відклик	F-міра
Mambo	83	0,76	0,78
Bebop	85	0,73	0,82
Інші звуки	92	0,93	0,90

Під час тестування (табл. 6) модель показала результати, які підтверджують її здатність ефективно класифікувати звуки. Однак результати для класів дронів Mambo і Bebop потребують покращення.

Реалізація моделі на основі архітектури трансформера. Модель для аудіокласифікації на основі трансформера ефективно працює з послідовними даними. Ця модель використовує механізм самоуваги (self-attention) для врахування довготривалих залежностей у вхідних даних. Архітектура моделі трансформера представлена на рис. 9.

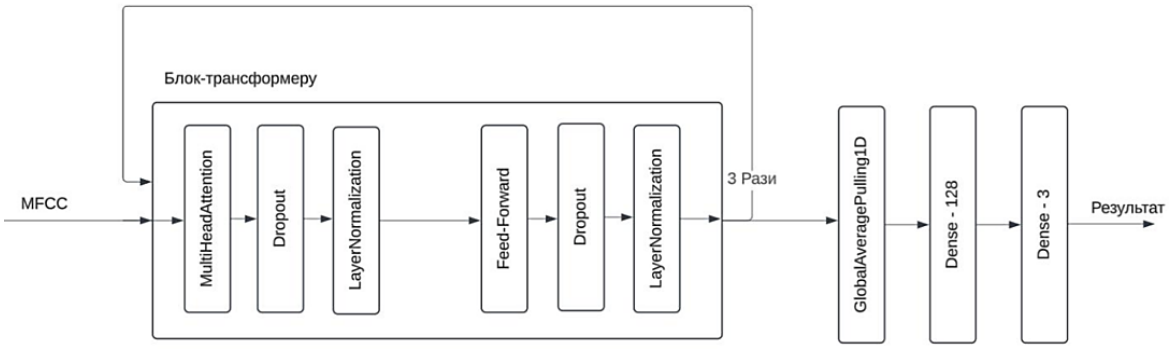

Рис. 9. Схема архітектури моделі трансформера

У цій моделі вхідний шар очікує на послідовність ознак, які отримані після попередньої обробки аудіо у вигляді MFCC. Механізм MultiHeadAttention - це ключовий компонент трансформера, який дозволяє моделі зважувати вплив різних частин послідовності одна на одну. Він використовує декілька "голів" уваги (Multi-head), щоб паралельно обчислювати різні аспекти залежностей між елементами послідовності. Шари нормалізації (LayerNormalization), застосовуються після кожного кроку обчислення самоуваги та повнозв'язного шару. Вони допомагають стабілізувати навчання та прискорити його. Повнозв`язні Feed-Forward шари які забезпечують нелінійність, обчислюючи проміжні ознаки, складається з двох послідовних шарів: один для збільшення розмірності, а інший — для її зменшення до початкової. Глобальний шар агрегації (GlobalAveragePooling1D) узагальнює інформацію з усього тимчасового контексту, зводячи послідовність до одного вектора. Вихідний шар складається з повнозв'язного шару з функцією активації softmax. Цей шар забезпечує прогноз ймовірностей належності до одного з трьох класів (дрон типу Mambo, дрон типу Bebop, інші звуки).

Наведена модель трансформера навчалася протягом 20 епох, що дозволило досягти високих результатів точності та низьких втрат, як це видно з графіків на рис. 10. Під час навчання втрати зменшувалися стабільно, починаючи з початкового значення близько 0,12, досягнувши мінімального значення менше 0.02. Валідаційна втрата демонструє певні коливання, але в цілому підтримує низькі значення, що свідчить про ефективну генералізацію моделі. Навчальна точність зросла до 99%, що свідчить про те, що модель добре адаптується до навчальних даних. Валідаційна точність також знаходиться на рівні 99%, що підтверджує високу якість моделі на валідаційних даних.

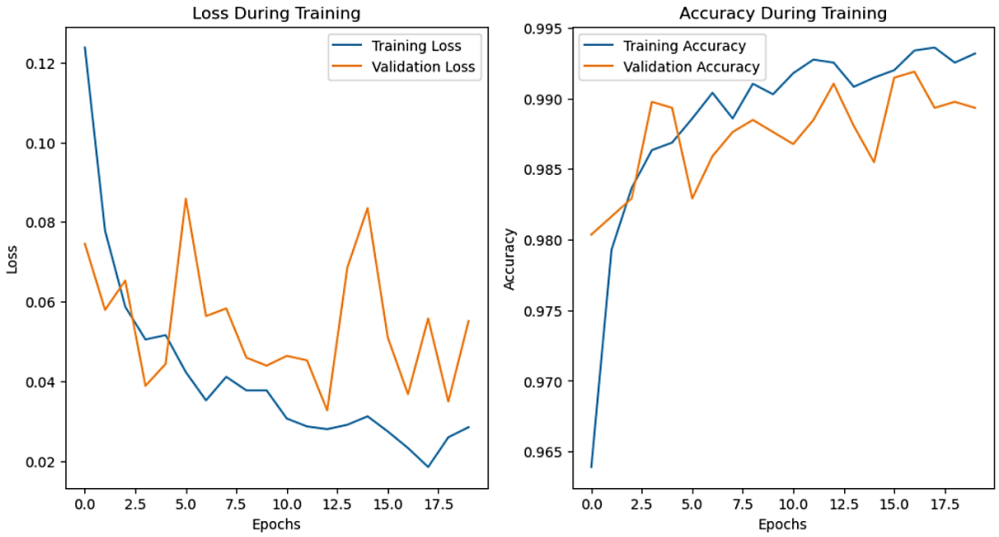

Рис. 10. Результати навчання моделі на основі трансформера

Тестування моделі проводилося на вибірці даних, яка не використовувалася в навчанні або валідації. Модель демонструє показники для трьох класів, які наведено в табл. 7.

Таблиця 7 – Результати тестування моделі на основі трансформера

Клас	Точність, %	Відклик	F-міра
Mambo	97	0,93	0,95
Bebop	95	0,95	0,95
Інші звуки	99	1,00	0,99

Модель на основі трансформера показує стабільну продуктивність з точністю класифікації понад 95 % для всіх класів. Високі значення F-міри свідчать про ефективність моделі у врахуванні як точності, так і відклику. Це робить модель придатною для завдань аудіокласифікації в умовах реального часу, таких як виявлення дронів у шумовому середовищі.

Аналіз результатів дослідження.

Результати порівняльного аналізу досліджених ШНМ при вирішенні задачі аудіо класифікації наведено в таблиці 8.

Таблиця 8 – Результати аналізу архітектур ШНМ при аудіо класифікації

Модель	Час обробки до оптимізації, мс	Час обробки після оптимізації, мс	Точність, %	Використання пам'яті GPU, МБ	Енергоспоживання, Вт
CNN	42	17	89.3	256	6
GRU	81	30	85	512	7.7
AST	125	55	92	1024	9

CNN демонструють високу ефективність у виділенні локальних частотних ознак з аудіосигналів, таких як спектральні патерни або гармоніки. Вони є особливо корисними для задач, де звуки мають виразні частотні характеристики, наприклад, шум моторів дронів.

RNN, включаючи їх модифікації (LSTM, GRU), спеціалізуються на обробці послідовних даних, таких як аудіосигнали. Вони здатні враховувати часові залежності, що є критичним для звуків дронів, де важливий контекст змін частот.

Поєднання CNN і RNN в моделі трансформера дозволяє використовувати сильні сторони обох архітектур: CNN виявляє локальні ознаки з аудіоданих, а RNN аналізує їх часовий контекст. Це забезпечує високу точність і універсальність трансформерів для задач аудіокласифікації. До того ж використовуючи механізм уваги, вони дозволяють ефективно моделювати як локальні, так і глобальні залежності в сигналі. Вони мають високу масштабованість і здатність обробляти великі набори даних.

Висновки. На основі проведеного аналізу ШНМ для задач аудіо класифікації встановлено, що використання різних архітектур, таких як CNN, RNN та трансформери, дозволяє досягти високої ефективності для специфічних задач класифікації аудіосигналів. Для оптимального вирішення завдань, пов'язаних з обробкою аудіо на обмежених обчислювальних ресурсах, рекомендовано застосовувати комбіновані підходи, що враховують переваги кожної архітектури.

Використання платформи NVIDIA Jetson Nano забезпечує ефективну реалізацію обчислювальних задач для нейронних мереж, завдяки підтримці оптимізованих бібліотек та можливості роботи з паралельними обчисленнями.

Результати тестування нейронних мереж на Jetson Nano демонструють високу ефективність, зокрема моделі CNN, яка забезпечує швидкість обробки та точність понад 90 % навіть для обмежених обчислювальних ресурсів. Завдяки оптимізації за допомогою Nvidia TensorRT досягнуто зменшення часу обробки даних на 50 %.

Подальші дослідження передбачають адаптацію запропонованих моделей для роботи в умовах реального часу, інтеграцію додаткових джерел даних, таких як візуальні сигнали, та вдосконалення алгоритмів обробки аудіо для підвищення стійкості до шумів та інших викликів, характерних для реальних середовищ.

Список використаних джерел

1. Хома, Ю. В. Порівняльний аналіз програмно-апаратного забезпечення алгоритмів глибокого навчання / Ю. В. Хома, А. Я. Бенч // Комп'ютерні системи і мережі. – 2019. – Т.1. – № 1. – С. 97-102.

2. Казимир, В. В. Проєктування системи керування дрона у складі захисної мультиагентної системи / В. В. Казимир, А. І. Роговенко, О. О. Карась // Технічні науки т атехнології. – 2024. – № 2(36). – С. 102-115.

3. Hershey, S. CNN architectures for large-scale audio classification / S. Hershey [et al.] // 2017 IEEE international conference on acoustics, speech and signal processing (icassp). – IEEE, 2017. – С. 131-135.

4. Palanisamy, K. Rethinking CNN models for audio classification [Electronic resource] / K. Palanisamy, D. Singhania, A. Yao. – Accessed mode: https://arxiv.org/pdf/2007.11154.pdf.

5. Gong Y. AST: Audio Spectrogram Transformer [Electronic resource] / Yuan Gong, Yu-An Chung, James R. Glass // arXiv.org. – 2021. – Accessed mode: https://arxiv.org/abs/2104.01778.

6. Nandi P. Recurrent Neural Nets for Audio Classification [Electronic resource] / Papia Nandi // Medium. – 2024. – Accessed mode: https://towardsdatascience.com/recurrent-neural-nets-for-audio-classification-81cb62327990.

7. Moinnereau, M. A. Classification of auditory stimuli from EEG signals with a regulated recurrent neural network reservoir / Marc-Antoine Moinnereau, Thomas Brienne, Simon Brodeur, Jean Rouat, Kevin Whittingsta, Eric Plourde // arXiv.org. – 2018. – Accessed mode: https://arxiv.org/pdf/1804.10322.

8. Yuan Gong. AST: Audio Spectrogram Transformer / Yuan Gong, Yu-An Chung, James R. Glass // INTERSPEECH 2021 30 August – 3 September, 2021, Brno, Czechia. – Pp. 571- 575.

9. Ramzan, F. A Deep Learning Approach for Automated Diagnosis and Multi-Class Classification of Alzheimer's Disease Stages Using Resting-State fMRI and Residual Neural Networks / F. Ramzan, M. U. Khan, A. Rehmat, S. Iqbal, T. Saba, A. Rehman, Z. Mehmood // Journal of Medical Systems. – 2019. – Vol. 44(2). DOI:10.1007/s10916-019-1475-2.

10.Liu, Y. A Convolutional Recurrent Neural-Network-Based Machine Learning for Scene Text Recognition Application / Yiyi Liu, Yuxin Wang, Hongjian Shi // Symmetry. – 2023. – Vol. 15, № 4. – P. 849. DOI: https://doi.org/10.3390/sym15040849.

11.Papers with Code - UrbanSound8K Dataset [Electronic resource] // The latest in Machine Learning : Papers With Code. – Accessed mode: https://paperswithcode.com/dataset/urbansound8k-1.

12.Piczak, K. J. ESC: Dataset for Environmental SoundClassification / K. J. Piczak // Proceedings of the 23rd Annual ACMConference on Multimedia, Brisbane, Australia: ACMPress, Oct. 13, 2015. – Pp. 1015–1018. – Accessed mode: https://www.karolpiczak.com/papers/Piczak2015-ESC-Dataset.pdf.

13.Al-Emadi, S. Saraalemadi/DroneAudioDataset [Electronic resource] // GitHub. – 2018). – Accessed mode: https://github.com/saraalemadi/DroneAudioDataset.

14.Zhang, Y. Hello Edge: Keyword Spotting on Microcontrollers [Electronic resource] / Y. Zhang, N. Suda, L. Lai, V. Chandra // arXiv.org. – Accessed mode: https://arxiv.org/abs/1711.07128.

15. 2018. [Online]. Available: https://www.tensorflow.org/tutorials/sequences/audiorecognition.

References

1. Khoma, Yu.V., Bench, A.Ya. (2019). Porivnialnyi analiz prohramno-aparatnoho zabezpechennia alhorytmiv hlybokoho navchannia [Comparative analysis of software and hardware of deep learning algorithms]. *Kompiuterni systemy i merezhi – Computer systems and networks, 1(1),* 97-102.

2. Kazymyr, V.V., Rohovenko, A.I., Karas, O.O. (2024). Proiektuvannia systemy keruvannia drona u skladi zakhysnoi multyahentnoi systemy [Designing a drone control system as part of a protective multiagent system]. *Tekhnichni nauky ta tekhnolohii – Technical sciences and technologies, 2(36),* 102-115.

3. Hershey, S. et al. (2017). CNN architectures for large-scale audio classification. *2017 IEEE international conference on acoustics, speech and signal processing (icassp)* (pp. 131-135).

4. Palanisamy, K., Singhania, D., Yao, A. (2020). Rethinking CNN models for audio classification. *arxiv.org.* https://arxiv.org/abs/2007.11154.

5. Gong, Y., Chung, Y., & Glass, J.R. (2021). AST: Audio Spectrogram Transformer. *www.isca-archive.org.* https://www.isca-archive.org/interspeech_2021/gong21b_interspeech.pdf.

6. Nandi, P. (2024). Recurrent Neural Nets for Audio Classification. *towardsdatascience.com.* https://towardsdatascience.com/recurrent-neural-nets-for-audio-classification-81cb62327990.

7. Moinnereau, M. A., et al. (2018). Classification of auditory stimuli from EEG signals with a regulated recurrent neural network reservoir. *arxiv.org.* https://arxiv.org/abs/1804.10322.

8. Gong, Y., Chung, Y., & Glass, J.R. (2021). AST: Audio Spectrogram Transformer. INTERSPEECH 2021. 30 August – 3 September, 2021. Brno, Czechia (pp. 571-575).

9. Ramzan, F., Khan, M. U., Rehmat, A., Iqbal, S., Saba, T., Rehman, A., Mehmood, Z. (2019). A Deep Learning Approach for Automated Diagnosis and Multi-Class Classification of Alzheimer's Disease Stages Using Resting-State fMRI and Residual Neural Networks. Journal of Medical Systems, 44. doi:10.1007/s10916-019-1475-2.

10. Liu, Y., Wang, Y., Shi, H. A (2023). Convolutional Recurrent Neural-Network-Based Machine Learning for Scene Text Recognition Application. *Symmetry, 15(4),* 849.

11. Salamon, J., et al. (2024). Urbansound8k. *paperswithcode.com.* https://paperswithcode.com/dataset/urbansound8k-1.

12. Piczak, K. J. (2015). ESC: Dataset for Environmental SoundClassification. Proceedings of the 23rd Annual ACMConference on Multimedia, Brisbane, Australia: ACMPress, Oct. 13. (pp. 1015–1018). https://www.karolpiczak.com/papers/Piczak2015-ESC-Dataset.pdf.

13. Al-Emadi, S. (2018). Saraalemadi/DroneAudioDataset. https://github.com/saraalemadi/DroneAudioDataset.

14. Zhang, Y., Suda, N., Lai, L., and Chandra, V. (2017). Hello edge: Keyword spotting on microcontrollers. *arxiv.org.* http://arxiv.org/abs/1711.07128.

15. 2018. [Online]. Available: https://www.tensorflow.org/tutorials/sequences/audio recognition.

Отримано 21.12.24

UDC 004.2

Volodymyr Kazymyr [1], Andrii Rohovenko[2], Oleksii Karas[3]

[1] Doctor of Sciences, Professor, Professor of the Department of Information and Computer Systems
Chernihiv Polytechnic National University (Chernihiv, Ukraine)
E-mail: vvkazymyr@gmail.com. **ORCHID:** https://orcid.org/0000-0001-8163-1119. **ResearcherID:** Q-2925-2016

[2] PhD in Technical Sciences, associate professor of the Department of Information and Computer Systems
Chernihiv Polytechnic National University (Chernihiv, Ukraine)
E-mail: arogovenko@gmail.com. **ORCID:** https://orcid.org/0000-0003-4594-5692. **ResearcherID:** G-3926-2014

[3] student of the Department of Information and Computer Systems
Chernihiv Polytechnic National University (Chernihiv, Ukraine)
E-mail: oleksiykaras2016@gmail.com. **ORCID:** https://orcid.org/0009-0004-8862-7234. **ResearcherID:** JZT-2594-2024

APPLICATION OF ARTIFICIAL NEURAL NETWORKS FOR AUDIO CLASSIFICATION ON AN EMBEDDED PLATFORM

The paper investigates the application of state-of-the-art artificial neural networks (ANNs) for audio classification tasks on embedded systems, specifically the NVIDIA Jetson Nano. The authors focus on evaluating the performance of Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and Transformers in classifying unmanned aerial vehicle (UAV) noise.

Key challenges in this domain, such as high levels of noise and variability in input data, limited labeled audio datasets, and the need to adapt ANNs to resource-constrained embedded systems, are discussed.

The study delves into the architectural characteristics of these ANN models, highlighting the strengths of CNNs for spectrograms, RNNs with LSTM and GRU for sequential data, and Transformers for their ability to handle long-range dependencies.

A methodology for implementing these models on the Jetson Nano platform is proposed, considering its hardware constraints.

Experimental results demonstrate the effectiveness of different architectures for UAV noise classification, with CNNs excelling in spectrogram analysis, while RNNs and Transformers proving more suitable for raw audio or sequential feature processing.

The authors outline directions for future research, including the development of optimised training methods for small datasets and the adaptation of advanced noise classification approaches to resource-constrained platforms.

Recommendations are provided for selecting the appropriate architecture based on task-specific requirements and constraints.

Keywords: *artificial neural networks, audio classification, embedded systems, NVIDIA Jetson Nano, UAV.*

Fig.: 10. Table: 8. References: 15.