https://doi.org/10.31891/2307-5732-2026-361-20
УДК 004.93

ДОРОЩУК МАКАР
Львівський національний університет імені Івана Франка
https://orcid.org/0009-0007-2877-3637
e-mail: makar.doroshchuk@lnu.edu.ua
ШЕВЧУК САВА
Львівський національний університет імені Івана Франка
https://orcid.org/0009-0006-6441-8598
e-mail: sava.shevchuk@lnu.edu.ua
ДОБУЛЯК ЛЕСЯ
Львівський національний університет імені Івана Франка
https://orcid.org/0000-0001-8665-8783
e-mail: lesia.dobuliak@lnu.edu.ua

РОЗПІЗНАВАННЯ МЕЛОДІЇ ЗА ЇЇ ФРАГМЕНТОМ ЗА ДОПОМОГОЮ МАШИННОГО НАВЧАННЯ

Розробка інтелектуальної системи для автоматичного розпізнавання музичних композицій за коротким аудіофрагментом із використанням методів глибинного навчання спрямована на розв'язання складної задачі ідентифікації мелодій у випадках, коли відсутня текстова інформація, теги або метадані. Ця проблема є особливо актуальною в сучасному цифровому середовищі, де користувачі часто стикаються з невідомою музикою на потокових платформах, у соціальних мережах або під час реальних аудіозаписів.

Запропонований підхід базується на використанні згорткових нейронних мереж (CNN) як основного механізму для вилучення та класифікації високорівневих аудіопредставлень. У межах дослідження систематично було проаналізовано різні чинники, що впливають на продуктивність і надійність системи розпізнавання. До них належали вибір формату аудіо (WAV чи MP3), оптимальна тривалість аналізованих фрагментів, набір спектральних ознак (мел-спектрограма, хрома, перетворення з постійним Q-фактором (CQT) та нормалізована хрома-енергія (CENS)), а також вплив методів аугментації даних, таких як додавання білого шуму чи зміна висоти тону.

Експериментальні результати показали, що найкращий баланс між точністю розпізнавання та обчислювальною ефективністю досягається при використанні односекундних фрагментів у форматі MP3, представлених за допомогою мел-спектрограм. Така конфігурація забезпечує високу стійкість до типових спотворень сигналу, водночас зберігаючи помірне споживання ресурсів під час навчання та розпізнавання.

Розроблена модель глибинного навчання була успішно інтегрована у Telegram-бот, який дозволяє користувачам надсилати аудіо або голосові повідомлення для ідентифікації композицій. Після отримання аудіофрагмента система здійснює його аналіз і повертає як основний результат, так і п'ять альтернативних варіантів, що забезпечує гнучкість у разі неоднозначного введення. Під час тестування особливу увагу приділяли впливу методів запису та якості передавання даних. Було встановлено, що записи, отримані за допомогою вбудованої функції голосових повідомлень Telegram, демонструють нижчу точність розпізнавання, головним чином через стиснення сигналу та появу фонових шумів.

Отримані результати підтверджують доцільність подальшого вдосконалення системи шляхом використання рекурентних або гібридних архітектур (таких як LSTM або GRU), розширення бази еталонних аудіозаписів і навчання на синтетично спотворених даних для підвищення стійкості до шуму.

Ключові слова: *розпізнавання мелодій, глибинне навчання, мел-спектрограма, аудіокласифікація, Telegram-бот.*

DOROSHCHUK MAKAR, SHEVCHUK SAVA, DOBULIAK LESIA,
Ivan Franko National University of Lviv

RECOGNIZING A MELODY BY ITS FRAGMENT USING NEURAL NETWORKS

The development of an intelligent system for automatic recognition of musical compositions from a short audio fragment using deep learning methods is aimed at addressing the complex problem of identifying melodies when textual information, tags, or metadata are unavailable. This task is particularly relevant in modern digital environments, where users frequently encounter unknown music through streaming platforms, social media, or real-life audio recordings. The proposed approach relies on convolutional neural networks (CNNs) as the core mechanism for extracting and classifying high-level audio representations.

In the course of the study, various factors influencing the performance and reliability of the recognition system were systematically examined. These included the choice of audio format (WAV versus MP3), the optimal length of analyzed fragments, the selection of spectral features (mel-spectrogram, chroma, constant-Q transform (CQT), and chroma energy normalized statistics (CENS)), as well as the effect of data augmentation techniques such as adding white noise or pitch shifting. Experimental evaluation demonstrated that the best balance between recognition accuracy and computational efficiency was achieved using one-second segments encoded in MP3 format and represented by mel-spectrograms. This configuration provided high robustness to common distortions while maintaining moderate resource consumption during training and inference.

The resulting deep learning model was successfully integrated into a Telegram bot that enables end users to send audio or voice messages for identification. Upon receiving an audio fragment, the system analyzes it and returns both the most probable match and five alternative predictions, offering flexibility in cases of ambiguous input. During testing, particular attention was paid to the influence of recording methods and data transmission quality. It was observed that recordings obtained through Telegram's built-in voice messaging feature tend to produce lower recognition accuracy, primarily due to signal compression and the introduction of background noise.

The research outcomes confirm the feasibility of further enhancement of the system through the use of recurrent or hybrid architectures such as LSTM or GRU networks, expansion of the reference audio database, and training on synthetically distorted data to improve noise tolerance

Keywords: *melody recognition, deep learning, mel-spectrogram, audio classification, Telegram bot.*

Стаття надійшла до редакції / Received 12.11.2025
Прийнята до друку / Accepted 11.01.2026
Опубліковано / Published 29.01.2026

Постановка проблеми у загальному вигляді
та її зв'язок із важливими науковими чи практичними завданнями

В дану секунду тисячі користувачів взаємодіють із величезними обсягами музичного контенту. При цьому виникає часта ситуація, коли людина чує мелодію, однак не має можливості ідентифікувати її. Стандартні методи пошуку, такі як: за текстом, тегами чи метаданими — є неефективними, коли доступний лише аудіофрагмент. Це створює потребу у побудові систем, здатних автоматично розпізнавати композицію за її коротким аудіозаписом.

Існуючі рішення, як-от Shazam, використовують спеціальні алгоритми для створення "відбитку" треку для подальшого порівняння з базою або для додавання в неї. Однак ці алгоритми мають обмеження в точності, коли у фрагментах наявні значні погіршення якості, як нестандарні умови запису чи наявні голосні шуми. Зі зростанням обчислювальних потужностей та розвитком глибинного навчання з'явилась можливість створювати більш гнучкі та стійкі моделі, які здатні адаптуватися до реальних умов і працювати з обмеженим набором даних.

Отже, актуальною є проблема розробки інтелектуальної системи, що здатна точно розпізнавати музичні композиції за їх коротким фрагментом, із використанням методів глибинного навчання та сучасних підходів до обробки аудіосигналів.

Аналіз досліджень та публікацій

Продовж останніх десятиліть досліджується проблема автоматичного розпізнавання музичних композицій. Одним із перших масштабних рішень у цій сфері став Shazam, алгоритм розпізнавання було представлено у 2003 році Ейвері Лі-Чунг Вангом. Метод Shazam ґрунтується на обчисленні спектрального відбитка аудіофрагмента за допомогою швидкого перетворення Фур'є (FFT) та побудові хеш-таблиць, що дозволяє швидко зіставляти аудіо з базою відбитків (Рис.1).[1]

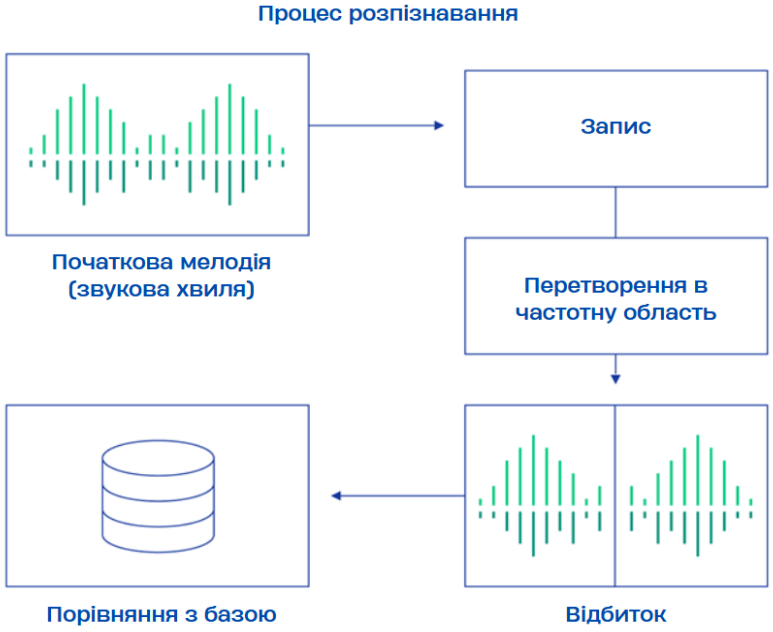

Рис. 1. Схема процес розпізнавання мелодій в Shazam

Однак підхід Shazam є чутливим до шуму та обмежений жорсткою структурою алгоритму. У зв'язку з цим, останнім часом активно досліджується використання методів машинного та глибинного навчання для задач аудіоідентифікації. Тож, згорткові нейронні мережі (CNN), які ефективно працюють із спектрограмами аудіо як зображеннями, виділяючи характерні ознаки музичних сигналів, можуть допомогти вирішити ці проблеми.

Інструменти та бібліотеки обробки аудіо, такі як LibROSA, дозволяють генерувати спектрограми, мел-спектрограми, Chroma-функції та інші ознаки, релевантні для навчання моделей. Разом із цим, бібліотека TensorFlow надає можливість будувати складні нейронні мережі з високою точністю та адаптивністю.

Незважаючи на прогрес, існує низка відкритих питань, таких як вплив того ж шуму на якість розпізнавання чи залежність результатів ідентифікації від формату та якості запису. Це підкреслює необхідність подальших прикладних досліджень у цій сфері.

Також цей підхід може допомогти у розпізнаванні мелодій, які пришвидшені або, навпаки, уповільнені. Однак використання нейронних мереж має і недоліки, зокрема для реалізації такої системи щодо всіх існуючих мелодій потрібні значні ресурси як для навчання, так і для постійного донавчання з додаванням нових мелодій. Тому цей підхід поки що не застосовують повсюдно для розпізнавання мелодій, але вже використовують для не менш важливих завдань — зокрема для розпізнавання звуків [2], голосів чи навіть шумів серця [3].

Формулювання цілей статті

Метою роботи є розробка, навчання та експериментальне дослідження нейронної мережі для автоматичного розпізнавання музичних композицій за коротким аудіофрагментом, а також визначення найкращих підходів підготовки та подання аудіоданих, що забезпечують найбільшу точність та стійкість класифікації в умовах реального використання. Досягнення цієї мети передбачає визначення впливу формату аудіозапису, тривалості сегмента, вибору спектральних ознак та методів доповнення даних на результативність моделі, а також оцінку практичної ефективності системи при її інтеграції у Telgram-bot для взаємодії з користувачем.

Виклад основного матеріалу

У межах дослідження було розроблено систему для автоматичного розпізнавання музичних композицій за коротким аудіофрагментом. Рішення побудовано на основі згорткової нейронної мережі (CNN), що приймає на вхід спектрограми, отримані з аудіосигналу, та здійснює класифікацію композицій.

1. Підготовка даних

Для навчання моделі використано аудіодані, що складаються з 233 композицій одного виконавця, зібрані у форматах WAV та MP3. Мелодії було зібрано з різних джерел таких як YouTube та SoundCloud. Вибір припав на одного виконавця з декількох причин, оскільки він пише музику в одному жанрі, а його голос завжди знаходиться в одному діапазоні частот, що в свою чергу ускладнює розпізнавання мелодії для нашої моделі, і в разі успішного розпізнавання ми можемо стверджувати, що модель добре справляється із поставленою ціллю.

Для обробки аудіосигналів застосовано бібліотеку LibROSA — спеціалізовану Python-бібліотеку, розроблену для музичного та загального аудіоаналізу. Вона надає широкий набір інструментів для роботи із сигналами: завантаження аудіо, розбиття на фрейми, обчислення спектральних перетворень (STFT, мел-спектрограми, MFCC), витягнення ритмічних та тональних ознак, а також візуалізації результатів. У рамках даного дослідження за допомогою LibROSA аудіо було нарізано на фрагменти тривалістю 0.5–4 секунди, кожен з яких перетворено на спектрограми різних форматів та подано на вхід нейронної мережі.

2. Архітектура моделі

Модель реалізовано з використанням TensorFlow (Keras).

TensorFlow — відкрита бібліотека машинного навчання, розроблена Google. Він був розроблений для навчання та побудови нейронних мереж, які використовуються для обробки зображень, пошуку закономірностей і вивчення кореляцій у даних. [4]

Процес створення моделі складався з таких етапів:

Етап 1. Автоматизована підготовка та анотування аудіоданих
- зчитуються всі аудіофайли;
- перед обробкою перевіряється, чи всі файли є придатними для зчитування (уникнення битих файлів);
- назви файлів автоматично використовуються як мітки (класи), що дозволяє уникнути ручного маркування;
- для кожного нового унікального імені файлу створюється новий індекс класу.

Етап 2. Перевірка валідності аудіо
- сегментація аудіо;
- кожен аудіофайл розбивається на фрагменти фіксованої певної тривалості (уніфікація розмірів вхідних тензорів).

Етап 3. Перетворення аудіо в аудіоознаку (кожен сегмент конвертується в аудіоознаку)

Етап 4. Підготовка даних для навчання
- аудіоознаки зберігаються як багатовимірні масиви (X), а мітки класів — як цілі числа (y);
- далі використовується train_test_split (дозволяє коректно оцінювати здатність моделі до узагальнення), щоб розбити вибірку на: тренувальну (90%) та тестову (10%).

Етап 5. Вибір архітектури моделі
- застосовується базова, але не менш значуща CNN-модель, яка складається з двох згорткових шарів (Conv2D + ReLU), шару максимум-пулінгу, повнозв'язного шару (Dense + ReLU + Dropout) та вихідного шару (Dense + Softmax);
- використовується sparse_categorical_crossentropy як функція втрат (оскільки мітки представлені як цілі числа).

Етап 6. Оптимізація моделі
- для мінімізації функції втрат застосовується оптимізатор Adam, що поєднує переваги методів AdaGrad та RMSProp;
- Adam автоматично адаптує швидкість навчання для кожного параметра, що прискорює збіжність та робить навчання більш стабільним.

Етап 7. Рання зупинка (EarlyStopping). Щоб уникнути перенавчання, навчання зупиняється, якщо протягом 10 епох валідаційна помилка не покращується. Це допомагає зекономити час і зберігати найкращі ваги моделі.

3. Експерименти

Було проведено низку експериментів із метою визначити, у якому вигляді найдоцільніше подавати дані для навчання моделі. Архітектура самої моделі залишалася незмінною, змінювалися лише вхідні дані — варіювався формат аудіо, тривалість сегментів та типи обраних аудіоознак.

Вплив аудіоформату (WAV проти MP3).

Аудіоформатом називається метод, що застосовується програмним забезпеченням для збереження аудіоданих, імпортованих з Інтернету або компакт-диска, у вигляді аудіофайлів.[5]

Wav скорочено від waveform аудіоформат без втрат. Іншими словами, він не застосовує жодного стиснення, навіть неруйнівного кодування. Файл WAV є найближчою цифровою версією аналогового сигналу. [6]

MP3 - це формат файлів і розширення для аудіофайлів, стиснутих за допомогою алгоритму MP3. Стиснення MP3 знижує точність звукових сигналів та їхніх компонентів до рівня, що не перевищує можливості слуху більшості людей. [6]

Навчальна точність – це точність з якою правильно визначаються мітки на навчальній вибірці.

Валідаційна точність – це точність з якою правильно визначаються мітки на навчальній вибірці.

Таблиця 1

Точність моделі залежно від формату аудіо

Формат сегменту	Навчальна точність (%)	Валідаційна точність (%)
WAV	90	63
MP3	88	60

Експерименти показали (Табл. 1.), що точність розпізнавання відрізняється. Зокрема, система працювала краще, коли працювала з файлами у форматі .WAV: навчальна вибірка має 90% точності класифікації, а точність тестової вибірки становила 63%. Порівняно з цим, результати були дещо нижчими при використанні того самого датасету, але у форматі .MP3: точність на навчальній вибірці була 88% та точність на тестовій вибірці — 60%.

Такі результати виникли через те, що .WAV не має стискання сигналу — він зберігає повну інформацію про аудіосигнал без втрат, що надає більше інформації для навчання. Натомість .MP3 — це формат з втратами, який під час стиснення відкидає частину спектру звуку, менш помітну для людського вуха, але потенційно важливу для нейронної мережі.

Звернувши увагу на розмір даних, можна помітити, що датасет у форматі .MP3 має розмір, у 8 разів менший, ніж той самий набір у форматі .WAV. Швидкість обробки, швидкість завантаження та економія дискового простору — є значною перевагою. Коли точність не є надзвичайно важливою та похибкою в декілька відсотків можна знехтувати або наявні ресурси є обмежені, MP3 може бути кращим варіантом.

Вплив довжини сегмента (від 0.1 до 4 секунд) [2].

У багатьох задачах машинного навчання, де опрацьовується аудіо (наприклад, класифікація звуків, розпізнавання мовлення, виявлення емоцій), важливо, щоб вхідні дані були однакової форми — тобто мали однакову кількість характеристик і однакову "довжину" у часовій площині. Але аудіофайли-композиції бувають різної тривалості: деякі — короткі, інші — значно довші.

Щоб зробити дані придатними до аналізу нейронною мережею, застосовується сегментування аудіо. Це означає, що кожен аудіофайл ділиться на частини фіксованої тривалості.

Кожен сегмент — це окремий приклад навчальної вибірки. Якщо аудіо має певну мітку, тобто назву композиції, то кожен сегмент успадковує цю мітку.

Проте варто зазначити, що в розбитті є й недоліки, зокрема розділення довших подій: ми неначе розриваємо речення, через що може втрачатися контекст і зміст.

Таблиця 2

Точність моделі залежно від довжини фрагменту аудіо

Тривалість сегменту	Навчальна точність (%)	Валідаційна точність (%)
5 с	95	53
3 с	92	59
2 с	91	64
1 с	93	71
0.5 с	87	73
0.1 с	65	73

Можна помітити (Табл. 2), що чим коротший сегмент, тим нижча точність навчання, але вища точність на валідаційній вибірці. При 0.1 с — навчальна точність падає до 65%, але валідаційна все ще висока (73%), що свідчить про сильний регуляризаційний ефект.

Попри високі показники валідації для коротких сегментів, ці результати можуть бути оманливими, а ось основні причини цього:

– Data leakage: короткі фрагменти з одного й того ж треку (наприклад, відрізані з 0.1 с чи 0.5 с інтервалом) є надто схожими. Під час поділу на train/test однакові або майже ідентичні сегменти можуть опинитися в обох множинах. У результаті модель не узагальнює, а просто впізнає знайомі фрагменти, що дає штучно завищену валідаційну точність;

- менш інформативні, але прості сегменти: короткі уривки (0.1 с чи 0.5 с) часто містять лише один звук чи шум (наприклад, удар барабана або частину голосу). Це полегшує класифікацію, але не гарантує розуміння мелодії;
- збільшення кількості прикладів: зі зменшенням довжини сегмента зростає кількість навчальних прикладів, що може допомагати уникати перенавчання. Але ця перевага супроводжується зростанням надмірності (повторення схожих даних) і не завжди призводить до кращого узагальнення.

Отож, незважаючи на високу валідаційну точність для 0.1-секундних та 0.5-секундних сегментів, ці результати не можна вважати цілком достовірними через ризик data leakage і знижену якість узагальнення. З урахуванням балансу між точністю, інформативністю сегментів і ризиком переоцінки, оптимальним варіантом можна вважати сегмент довжиною 1 секунду. Він забезпечує високу валідаційну точність (71%) та має достатню довжину, щоб зберігати мелодійну структуру та контекст, що дає можливість тренувати модель без значного ризику перенавчання.

Вплив спектральних ознак (mel-spectrogram, chroma, CQT, CENS).

Спектрограма Мела схожа на звичайну спектрограму, але частотна шкала перетворюється на шкалу Мела, яка краще відображає слухове сприйняття людини.

Різниця між спектрограмою та Мел-спектрограмою полягає в тому, що Мел-спектрограма перетворює частоти в мел-шкалу. Згідно з Каліфорнійським університетом, мел-шкала – це «перцептивна шкала висот звуків, які слухачі оцінюють як однакові за відстанню один від одного» [7].

Хроматичні функції, також відомі як профілі класів висоти звуку, є потужними інструментами для аналізу гармонійного та тонального складу музики. Вони особливо корисні для завдань, пов'язаних з аналізом висоти звуку та гармонії, таких як розпізнавання акордів, виявлення тональностей та оцінка музичної схожості. [7]. Хроматичні характеристики відображають інтенсивність кожного з 12 класів висоти звуку (до, до-дієти, ре, ре-дієти, мі, фа, фа-дієти, соль, соль-дієти, ля, ля-дієти, сі) в аудіосигналі, незалежно від октави. Це робить їх особливо придатними для аналізу гармонічних та мелодійних характеристик музики.

Ключові поняття тут є клас висоти тону – усі висоти тону, що мають однакову назву ноти (наприклад, усі ноти «до» в різних октавах) та октавна інваріантність – функції Chroma підсумовують енергію по всіх октавах для кожного класу висоти звуку.

У математиці та обробці сигналів перетворення з постійною Q-фактором та перетворення зі змінною Q-фактором, відомі як CQT та VQT, перетворюють ряд даних у частотну область. Вони пов'язані з перетворенням Фур'є та дуже тісно пов'язані зі складним вейвлет-перетворенням Морле. Його конструкція підходить для музичного представлення [8].

Ознаки CENS – це варіант ознак кольоровості, які є більш стійкими до змін динаміки та тембру. Вони включають додаткові етапи обробки, зокрема локальну нормалізацію енергії та квантування [6].

Функції CENS часто забезпечують плавніше відображення гармонійного вмісту, що може бути корисним для таких завдань, як зіставлення аудіо та ідентифікація кавер-версій пісень.

Таблиця 3

Точність моделі залежно від спектральних ознак

Ознака	Точність Train (%)	Точність Validation (%)
Mel-спектрограма	90	63
Chroma	64	33
CQT (Constant-Q)	93	53
CENS	59	45

За результатами експериментів (Табл. 3) можна помітити, що:

- Mel-спектрограма відображає енергетичний спектр сигналу у логарифмічній шкалі частот (наближеній до людського слуху), що дає повну картину спектрально-часових змін. Вона найкраще зберігає інформацію, критичну для класифікації як музичних, так і немузичних звуків. Навчання на таких ознаках дало високу точність на тренуванні (90%) і порівняно високу на валідації (63%), що свідчить про хорошу узагальнюваність.
- Chroma відображає лише тональну (гармонічну) структуру сигналу — тобто, які висоти присутні, без інформації про енергетику чи спектральну текстуру. Є ефективними в задачах музичної гармонічної класифікації, але слабкими для загальних звуків або шумів. Як результат цього видно низьку точність (64% та 33%) та велике падіння при переході від тренування до валідації.
- CQT (Constant-Q Transform) має нерівномірну шкалу частот, що імітує музичну логарифмічну шкалу (як клавіші на фортепіано). Дуже висока точність на train (93%) свідчить про високу модельну здатність, але низька валідація (53%). Ймовірно, велика роздільність у нижніх частотах може спричиняти надмірне запам’ятовування особливостей, незначущих у загальному випадку.

– CENS є статистичним агрегатом Chroma-фіч у часовому вікні. Має найнижчу точність на train і на val (59% / 45%), що вказує на слабку придатність для задачі класифікації звуків у загальному випадку.

Таким чином, Mel-спектрограма є найбільш придатною ознакою для загальної аудіокласифікації, особливо у випадках, коли джерела сигналів не обмежуються лише музикою (наприклад, мова, побутові шуми, природні звуки тощо).

Chroma та CENS підходять для специфічних музичних задач (наприклад, виявлення акордів або тональностей), але не забезпечують достатньої глибини для задач загальної класифікації.

CQT має потенціал, але потребує регуляризації та кращих стратегій боротьби з перенавчанням.

4. Доповнення даними

Доповнення даних – метод, який використовують в машинному навчанні для штучного збільшення розміру та/або різноманітності навчального набору даних. Його суть у застосуванні певних перетворень або модифікацій наявних даних, генеруючи нові зразки, які зберігають ту саму мітку або клас, що й вхідні дані. Доповнені дані покращують продуктивність моделі, зменшуючи перенавчання, покращуючи узагальнення та підвищуючи стійкість моделі [9].

Додавання шуму, навіть якщо він є штучним, дозволяє імітувати поширені фонові звуки, що виникають у реальних життєвих ситуаціях. Такий підхід моделює умови, у яких мелодія може програватися з іншого пристрою в шумному середовищі, а також допомагає відтворити перешкоди, що не є частиною самої мелодії, але присутні під час її прослуховування в побуті.

Навчання моделі в умовах, максимально наближених до реальних, підвищує її здатність правильно сприймати та обробляти аудіосигнали, що безпосередньо впливає на точність та ефективність розпізнавання у практичному використанні.

Враховуючи великий обсяг аудіоданих і обмеження обчислювальних ресурсів, було обрано саме цю техніку. Вона є оптимальним компромісом між ресурсозатратністю, ефективністю та реалістичністю, забезпечуючи кращу адаптацію моделі до умов реального застосування.

Шум – це категорія звуків, що характеризується руйнівними або небажаними звуками, які заважають нашій здатності ефективно чути або спілкуватися. Зазвичай це характеризується нерегулярними або хаотичними коливаннями в повітрі чи іншому середовищі [10].

Шум може мати різні форми, включаючи, але не обмежуючись гучними звуками, випадковими частотами, фоновим гуркотом, механічним гулом або будь-якими іншими чутними перешкодами, які перешкоджають сприйняттю звуків чи сигналів.

5. Застосування

Узагальнюючи результати всіх проведених експериментів і застосованих методів оптимізації, була побудована модель, що базується на аналізі фрагментів аудіофайлів формату .MP3, оскільки різниця в точності не значна, а от різниця в розмірі може бути критична під час навчання моделі із обмеженими ресурсами.

Кожен аудіозапис попередньо розбивався на сегменти тривалістю по 1 секунді, що дозволяло моделі краще адаптуватися до короткочасних змін у звуковому сигналі. Як основну аудіоознаку було використано мел-спектрограму, яка є інформативним представленням частотного спектра з урахуванням особливостей людського слуху. З метою підвищення стійкості моделі до впливу шумових факторів у реальних умовах, до навчального набору було також включено модифіковані версії аудіо, викривлені шумом. Це дозволило покращити загальну узагальнювальну здатність системи.

Аналіз експериментів показав, що найкращі результати валідації спостерігалися на середніх епохах, після чого модель демонструвала ознаки перенавчання. Тому доцільним є коригування стратегії зупинки навчання — зменшення параметра early_stopping.patience з 10 до 5, що дозволить ефективніше зберігати оптимальні ваги моделі.

Кінцева версія моделі, розробленої у межах дослідження, продемонструвала високі результати точності, що свідчить про її ефективність у задачі розпізнавання мелодій.

Для спрощення взаємодії з навченим класифікатором аудіо було реалізовано Telegram-бота(який було реалізовано за допомогою бібліотеки Aiogram), який дозволяє користувачам швидко й інтуїтивно отримувати результат класифікації аудіозаписів. Бот підтримує надсилання голосових повідомлень або аудіофайлів у будь-якому популярному форматі (MP3, WAV, M4A, OGG, AAC тощо). Після отримання повідомлення, бот автоматично виконує попередню обробку: зчитує аудіо, конвертує його у формат MP3 з частотою дискретизації 44.1 кГц, після чого передає файл до попередньо натренованої моделі для аналізу.

Модель здійснює розпізнавання мелодії та повертає користувачеві:

– основну назву мелодії, яку вона ідентифікувала як найбільш ймовірну;
– п’ять альтернативних варіантів, які мають найвищу схожість за результатами класифікації (тобто ті, що йдуть наступними за впевненістю моделі). Це реалізовано з метою підвищення надійності: навіть у випадку помилки класифікації користувач отримує можливість самостійно вибрати схожий варіант.
– надсилає всю композицію у форматі MP3, що дає змогу користувачеві одразу прослухати та зіставити результат без необхідності залишати середовище чату.

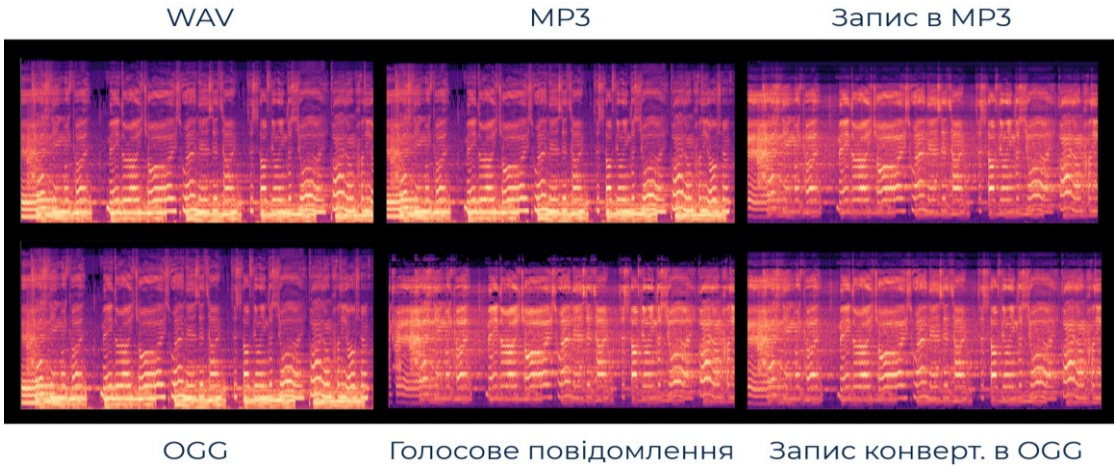

Рис. 2. Мел-спектрограми однієї мелодії, але в різних форматах

Спостерігається, що:

- фрагменти, вирізані з оригінальної доріжки у різних форматах (WAV, MP3, OGG на Рис. 2.) мають хоч і помітні, але незначні відмінності та загалом добре розпізнаються моделлю;
- фрагмент, записаний на диктофон телефону та збережений у форматі MP3 (Запис в MP3 на Рис. 2.), продемонстрував ще більші викривлення, однак мелодія все ще успішно розпізнається;
- фрагмент, отриманий через голосові повідомлення в Telegram (Голосове повідомлення на Рис. 2.), не був розпізнаний правильно — він навіть не потрапив до п'ятірки найподібніших. З мел-спектрограми видно, що частоти вище 12 кГц майже повністю відсутні, а сама спектрограма виглядає змазаною;
- фрагмент, записаний на диктофон телефону у форматі MP3, а потім переконвертований у OGG (Запис конверт. в OGG на Рис. 2.), вже викликав труднощі з розпізнаванням. Проте як другий варіант система все ж запропонувала правильну композицію.

Результати перевірок показали, що використання Telegram для запису музичних фрагментів через голосові повідомлення було помилкою. Формат .OGG спотворює якість і додає шуми, а низький бітрейт та вбудовані фільтри (еквалайзер із зрізом частот, компресор, плавне нарощування гучності) ще більше викривлюють сигнал. Це ускладнює розпізнавання мелодій, змушуючи або ускладнювати модель для роботи з сильно спотвореними даними, або переносити її на іншу платформу, наприклад, власний сайт.

Висновки з даного дослідження і перспективи подальших розвідок у даному напрямі

У межах дослідження було реалізовано систему автоматичного розпізнавання музичних фрагментів із використанням методів глибинного навчання.

Система ефективно працює з короткими фрагментами тривалістю, демонструючи стійкість до змін формату аудіо та умов запису. Інтеграція моделі в Telegram-бот підтвердила її прикладну цінність і готовність до використання в реальних умовах.

Основні переваги розробленого підходу:

- використання мел-спектрограм для ефективної подачі звукових даних;
- гнучкість моделі при роботі з різними форматами аудіо;
- простота розгортання у вигляді інтерактивного сервісу.

Подальші напрями дослідження включають:

Хоча згорткова нейронна мережа (CNN), застосована у дослідженні, продемонструвала певну ефективність у задачі розпізнавання мелодій, її архітектура має обмеження, пов'язані з обробкою послідовностей. Зокрема, CNN не враховує порядок фрагментів у часовому контексті, що може бути критично важливим у випадку розпізнавання мелодій.

Тому одним із потенційних напрямів удосконалення є перехід до рекурентних моделей (Recurrent Neural Network), таких як LSTM (Long Short-Term Memory) або GRU (Gated Recurrent Unit), які краще адаптовані до задач, де порядок і динаміка у часі відіграють ключову роль. Альтернативно, можна модифікувати існуючу CNN, інтегрувавши в неї компоненти, чутливі до порядку.

Окремим технічним викликом стали голосові повідомлення в Telegram, які значно спотворюють якість аудіосигналу. Це негативно впливає на точність розпізнавання. Вирішенням цієї проблеми може стати перехід на іншу платформу для збору аудіо, наприклад, власний вебсайт, який дозволить більш контрольовано записувати та передавати звук до моделі. Інший підхід — навчання моделі на штучно викривлених аудіофайлах, які імітують спотворення, що виникають при передачі через Telegram.

Крім того, використання стисненого формату MP3, попри втрати якості, дає змогу значно зменшити обсяг даних, що відкриває можливість масштабного розширення аудіобази композицій. Розширення датасету — ще один важливий напрям для підвищення узагальнюваності моделі, що, ймовірно, покращить результати розпізнавання в умовах реального застосування.

Література

1. Jovan Jovanovic How does Shazam work? Music Recognition Algorithms, Fingerprinting, and Processing - [Електронний ресурс] : [веб-сайт]. – Режим доступу: https://www.toptal.com/algorithms/shazam-it-music-processing-fingerprinting-and-recognition - (дата звернення: 21.09.2025).

2. Nicholas Renotte. Build a Deep Audio Classifier with Python and Tensorflow - [Електронний ресурс] : [*YouTube*.]. – Режим доступу: https://www.youtube.com/watch?v=ZLIPkmmDJAc - (дата звернення: 21.09.2025).

3. Jonathan Rubin, Rui Abreu, Anurag Ganguli, Saigopal Nelaturi, Ion Matei, Kumar Sricharan Recognizing Abnormal Heart Sounds Using Deep Learning - [Електронний ресурс] : [веб-сайт]. – Режим доступу: http://eztuir.ztu.edu.ua/handle/123456789/5319 - (дата звернення: 21.09.2025).

4. Учасники проектів Вікімедіа. TensorFlow - [Електронний ресурс] : [веб-сайт]. – Режим доступу: https://uk.wikipedia.org/wiki/TensorFlow - (дата звернення: 21.09.2025).

5. Що таке аудіоформат? [Електронний ресурс] : [веб-сайт] – Режим доступу: https://helpguide.sony.net/gbmig/44142764/v1/ua/contents/09/02/01/01.html - (дата звернення: 21.09.2025).

6. WAV vs MP3: Detailed Comparison For Users. *Online Audio Converter - Convert audio files to MP3, WAV, MP4, M4A, OGG or iPhone Ringtones* - [Електронний ресурс] : [веб-сайт]. - Режим доступу: https://online-audio-converter.com/blog/wav-vs-mp3 - (дата звернення: 21.09.2025).

7. Audio Features. - [Електронний ресурс] : [веб-сайт]. - Режим доступу: https://ravinkumar.com/GenAiGuidebook/audio/audio_feature_extraction.html - (дата звернення: 21.09.2025).

8. Contributors to Wikimedia projects. Constant-Q transform - Wikipedia. *Wikipedia, the free encyclopedia*. - [Електронний ресурс] : [веб-сайт]. - Режим доступу: https://en.wikipedia.org/wiki/Constant-Q_transform - (дата звернення: 21.09.2025).

9. José miguel Herrera. Audio Data Augmentation: Techniques and Methods. - [Електронний ресурс] : [веб-сайт]. - Режим доступу: https://blog.pangeanic.com/audio-data-augmentation-techniques-and-methods - (дата звернення: 21.09.2025).

10. What are the Different Types of Noise?. *Acoustiblok UK*. - [Електронний ресурс] : [веб-сайт]. - Режим доступу: https://www.acoustiblok.co.uk/what-are-the-different-types-of-noise/ - (дата звернення: 21.09.2025).

References

1. Jovan Jovanovic How does Shazam work? Music Recognition Algorithms, Fingerprinting, and Processing - [Elektronnyi resurs] : [veb-sait]. – Rezhim dostupu : https://www.toptal.com/algorithms/shazam-it-music-processing-fingerprinting-and-recognition - (Data zvernennia 21.09.2025).

2. Nicholas Renotte. Build a Deep Audio Classifier with Python and Tensorflow- [Elektronnyi resurs] : [veb-sait]. – Rezhim dostupu: https://www.youtube.com/watch?v=ZLIPkmmDJAc - (Data zvernennia 21.09.2025).

3. Jonathan Rubin, Rui Abreu, Anurag Ganguli, Saigopal Nelaturi, Ion Matei, Kumar Sricharan Recognizing Abnormal Heart Sounds Using Deep Learning - [Elektronnyi resurs] : [veb-sait]. – Rezhim dostupu: http://eztuir.ztu.edu.ua/handle/123456789/5319 - (Data zvernennia 21.09.2025).

4. Wikimedia project contributors. TensorFlow - [Elektronnyi resurs] : [veb-sait]. – Rezhim dostupu: https://uk.wikipedia.org/wiki/TensorFlow - (Data zvernennia 21.09.2025).

5. What is an audio format?- [Elektronnyi resurs] : [veb-sait]. – Rezhim dostupu: https://helpguide.sony.net/gbmig/44142764/v1/ua/contents/09/02/01/01.html - (Data zvernennia 21.09.2025).

1. WAV vs MP3: Detailed Comparison For Users. *Online Audio Converter - Convert audio files to MP3, WAV, MP4, M4A, OGG or iPhone Ringtones* - [Elektronnyi resurs] : [veb-sait]. – Rezhim dostupu: https://online-audio-converter.com/blog/wav-vs-mp3 - (Data zvernennia 21.09.2025).

2. Audio Features. - [Elektronnyi resurs] : [veb-sait]. – Rezhim dostupu: https://ravinkumar.com/GenAiGuidebook/audio/audio_feature_extraction.html - (Data zvernennia 21.09.2025).

3. Contributors to Wikimedia projects. Constant-Q transform - Wikipedia. *Wikipedia, the free encyclopedia*. - [Elektronnyi resurs] : [veb-sait]. – Rezhim dostupu: https://en.wikipedia.org/wiki/Constant-Q_transform - (Data zvernennia 21.09.2025).

4. José miguel Herrera. Audio Data Augmentation: Techniques and Methods. - [Elektronnyi resurs] : [veb-sait]. – Rezhim dostupu: https://blog.pangeanic.com/audio-data-augmentation-techniques-and-methods - (Data zvernennia 21.09.2025).

5. What are the Different Types of Noise?. *Acoustiblok UK*. - [Elektronnyi resurs] : [veb-sait]. – Rezhim dostupu: https://www.acoustiblok.co.uk/what-are-the-different-types-of-noise/ - (Data zvernennia 21.09.2025).