МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
ЗАПОРІЗЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ

МАТЕМАТИЧНИЙ ФАКУЛЬТЕТ
Кафедра комп'ютерних наук

КВАЛІФІКАЦІЙНА РОБОТА МАГІСТРА

на тему: **«ДОСЛІДЖЕННЯ ТЕХНОЛОГІЙ РОЗПІЗНАВАННЯ ПСИХОЕМОЦІЙНОГО СТАНУ ЛЮДИНИ В СИСТЕМАХ МАШИННОГО СЛУХУ»**

Виконала: студентка 2 курсу, групи 8.1224
спеціальності 122 Комп'ютерні науки
(шифр і назва спеціальності)
освітньої програми Комп'ютерні науки
(назва освітньої програми)
М.Д. Нагорна
(ініціали та прізвище)
Керівник завідувачка кафедри комп'ютерних наук, професор, д.т.н. Шило Г.М.
(посада, вчене звання, науковий ступінь, прізвище та ініціали)
Рецензент професор кафедри програмної інженерії, доцент, к.ф.-м.н., Кудін О.В.
(посада, вчене звання, науковий ступінь, прізвище та ініціали)

Запоріжжя 2025

МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ

ЗАПОРІЗЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ

Факультет математичний

Кафедра комп'ютерних наук

Рівень вищої освіти магістр

Спеціальність 122 Комп'ютерні науки
(шифр і назва)

Освітня програма Комп'ютерні науки

ЗАТВЕРДЖУЮ
Завідувач кафедри комп'ютерних наук,
д.т.н., професор

_____ Шило Г.М.
(підпис)

" ____ " _____ 2025 р.

З А В Д А Н Н Я

НА КВАЛІФІКАЦІЙНУ РОБОТУ СТУДЕНТЦІ

Нагорній Марії Дмитрівні
(прізвище, ім'я та по-батькові)

1. Тема роботи Дослідження технологій розпізнавання психоемоційного стану людини в системах машинного слуху
керівник роботи Шило Галина Миколаївна, д.т.н., професор
(прізвище, ім'я та по-батькові, науковий ступінь, вчене звання)

затверджені наказом ЗНУ від « 28 » лютого 2025 року № 326-с

2. Строк подання студентом роботи 03.12.2025

3. Вихідні дані до роботи 1. Постановка задачі.
2.Перелік літератури.
.

4. Зміст розрахунково-пояснювальної записки (перелік питань, які потрібно розробити)
1. Основні теоретичні відомості та постановка задачі
2. Аналіз предметної області
3. Програмна реалізація та тестування

4. Перелік графічного матеріалу (з точним зазначенням обов'язкових креслень)
презентація

6. Консультанти розділів роботи

Розділ	Прізвище, ініціали та посада консультанта	Підпис, дата	
		завдання видав	завдання прийняв

7. Дата видачі завдання 10.03.2025

К А Л Е Н Д А Р Н И Й П Л А Н

№	Назва етапів кваліфікаційної роботи	Строк виконання етапів роботи	Примітка
1.	Розробка плану роботи.	24.03.2025	
2.	Збір вихідних даних.	21.04.2025	
3.	Обробка методичних та теоретичних	19.05.2025	
	джерел.		
4.	Розробка першого та другого розділу.	25.08.2025	
5.	Розробка третього та четвертого розділу.	24.11.2025	
6.	Оформлення та нормоконтроль	01.12.2025	
	кваліфікаційної роботи бакалавра.		
7.	Захист кваліфікаційної роботи.	17.12.2025	

Студент _____ М.Д. Нагорна
(підпис) (ініціали та прізвище)

Керівник роботи _____ Г.М. Шило
(підпис) (ініціали та прізвище)

Нормоконтроль пройдено

Нормоконтролер _____ О.Г.Спиця
(підпис) (ініціали та прізвище)

РЕФЕРАТ

Кваліфікаційна робота магістра «Дослідження технологій розпізнавання психоемоційного стану людини в системах машинного слуху»: 60 с., 21 рис., 1 табл., 24 джерела, 4 додатки.

АНАЛІЗ АУДІО, ГОЛОСОВІ ПОМІЧНИКИ, ЗГОРТКОВІ НЕЙРОННІ МЕРЕЖІ, КЛАСИФІКАЦІЯ, МАШИННЕ НАВЧАННЯ, НЕЙРОННІ МЕРЕЖІ, ОБРОБКА ГОЛОСУ, РОЗПІЗНАВАННЯ ЕМОЦІЙ, ТРАНСФЕРНЕ НАВЧАННЯ, ШТУЧНИЙ ІНТЕЛЕКТ, CNN, PYTHON, RAVDESS, WAV2VEC 2.0.

Об’єкт дослідження – процес розпізнавання психоемоційного стану людини за звуковими характеристиками її голосу.

Мета роботи – дослідження та порівняльний аналіз технологій машинного навчання для розпізнавання психоемоційного стану людини.

Методи дослідження – аналіз предметної області, визначення вимог, цифрова обробка сигналів.

У процесі дослідження систем розпізнавання емоцій було проаналізовано предметну область та визначено основні характеристики голосу, що впливають на виявлення настрою. Програмну реалізацію виконано мовою Python із використанням бібліотек Librosa для обробки звуку та TensorFlow з PyTorch для створення нейромереж. Розроблено та протестовано дві різні моделі: згорткову мережу, що працює зі спектрограмами, та трансформер. Проведено експерименти на студійних записах та на власному наборі даних українською мовою. З’ясовано, що модель Wav2Vec 2.0 значно краще справляється з шумами та розпізнаванням емоцій у новій мові, тоді як CNN підходить лише для ідеальних умов. Результати роботи можуть бути використані для створення розумних голосових помічників. Подальший розвиток передбачає покращення точності роботи з українською мовою.

SUMMARY

Master’s Qualifying Theses «Researching the Technologies for Recognizing Human Psycho-emotional State in Machine Listening Systems»: 60 pages, 21 figures, 1 tables, 24 references, 4 supplements

AUDIO ANALYSIS, VOICE ASSISTANTS, CONVOLUTIONAL NEURAL NETWORKS, CLASSIFICATION, MACHINE LEARNING, NEURAL NETWORKS, VOICE PROCESSING, EMOTION RECOGNITION, TRANSFER LEARNING, ARTIFICIAL INTELLIGENCE, CNN, PYTHON, RAVDESS, WAV2VEC 2.0.

Object of the study – the process of recognizing a person's psycho-emotional state based on the acoustic characteristics of their voice.

Aim of the study – research and comparative analysis of machine learning technologies for recognizing a person's psycho-emotional state.

Method of research – subject area analysis, requirements definition, digital signal processing.

In the process of researching emotion recognition systems, the subject area was analyzed and the main characteristics of the voice that influence mood detection were identified. The software implementation was performed in Python using the Librosa library for sound processing and TensorFlow with PyTorch for creating neural networks. Two different models were developed and tested: a convolutional network that works with spectrograms and a transformer. Experiments were conducted on studio recordings and on our own dataset in Ukrainian. It was found that the Wav2Vec 2.0 model copes much better with noise and emotion recognition in a new language, while CNN is only suitable for ideal conditions. The results of the work can be used to create smart voice assistants. Further development involves improving the accuracy of working with the Ukrainian language.

ЗМІСТ

Завдання на кваліфікаційну роботу……..2
Реферат 4
Summary 5
Вступ..... 7
1 Аналіз типів емоцій в аудіосистемах 9
1.1 Класифікація емоцій 9
1.2 Критерії визначення емоцій в голосі 10
1.3 Аналіз сучасних систем машинного слуху 12
1.3.1 Голосові асистенти 12
1.3.2 Спеціалізовані B2B-платформи..... 12
1.3.3 Великі мовні моделі з емоційним інтелектом..... 13
1.4 Висновки до розділу 1 14
2 Дослідження методів обробки аудіосигналів..... 16
2.1 Попередня обробка та підготовка аудіоданих 17
2.1.1 Дискретизація та квантування аудіоданих..... 17
2.1.2 Нормалізація амплітуди та зменшення шумів..... 19
2.1.3 Фреймування 20
2.1.4 Детекція мовленнєвої активності..... 20
2.2 Спектральний аналіз 21
2.2.1 Перетворення Фур’є 21
2.2.2 Мел-спектрограми 22
2.3 Кепстральний аналіз та MFCC 23
2.4 Аналіз якості голосу та просодичні характеристики 25
2.5 Висновки до розділу 2 26
3 Вибір методів машинного навчання та набору даних..... 27
3.1 Вибір набору даних..... 27
3.2 Методологія вибору архітектур машинного навчання 28

3.2.1 Класичні алгоритми машинного навчання 28
3.2.2 Згорткові нейронні мережі 30
3.2.3 Трансформери 31
3.3 Висновок до розділу 3 33
4 Програмна реалізація та експериментальне дослідження 34
4.1 Реалізація та дослідження моделі на основі згорткових нейронних мереж 34
4.1.1 Попередня обробка даних 34
4.1.2 Архітектура згорткової мережі 35
4.1.3 Навчання та аналіз результатів 36
4.2 Реалізація та дослідження моделі на основі трансферного навчання 39
4.2.1 Підготовка даних та налаштування моделі 39
4.2.2 Аналіз результатів моделювання 40
4.3 Формування контрольної вибірки україномовних даних 42
4.3.1 Методика відбору та структура даних 43
4.3.2 Попередня обробка українського контенту 44
4.4 Порівняльний аналіз ефективності моделей у крос-мовному середовищі 44
4.4.1 Програмна реалізація алгоритму тестування 45
4.4.2 Результати тестування візуальної моделі 45
4.4.3 Результати тестування трансформерної моделі 47
4.5 Висновки до розділу 4 49
Висновки 51
Перелік посилань 53

ВСТУП

Розвиток штучного інтелекту змінює взаємодію людини з комп’ютером. Якщо раніше техніка просто виконувала команди, то сьогодні від неї очікують більшого – розуміння людини. Голосові асистенти та чат-боти стають все популярнішими, але вони мають суттєвий недолік: вони добре розуміють слова, але часто ігнорують інтонацію та емоції. А для сфер на кшталт обслуговування клієнтів, онлайн-навчання чи психологічної підтримки розуміння емоційного стану користувача є не менш важливим, ніж розуміння самого тексту.

Для вирішення цієї проблеми використовують технології машинного слуху. Вони аналізують звук голосу, щоб визначити, чи людина радісна, зла або засмучена. Проте більшість існуючих рішень ідеально працюють лише на чистих студійних записах і переважно англійською мовою. Як тільки з’являється шум або змінюється мова (наприклад, на українську), такі програми починають помилятися. Саме тому актуальним є створення таких систем, які могли б працювати стабільно з різними мовами без необхідності збирати тисячі годин нових записів.

Метою роботи є покращення автоматичного розпізнавання емоцій людини шляхом дослідження та розробки програмних моделей на основі сучасних нейронних мереж і перевірки їхньої здатності працювати з українською мовою.

Для досягнення мети необхідно вирішити такі завдання:

- проаналізувати, як емоції проявляються в голосі та як комп’ютер може обробляти звукові сигнали;
- дослідити сучасні нейронні мережі (згорткові мережі та трансформери) і те, як їх використовують для аналізу звуку;
- розробити власні програмні моделі для розпізнавання емоцій та навчити їх на стандартних наборах даних;
- створити власний тестовий набір записів українською мовою, щоб перевірити роботу моделей у реальних умовах;
- порівняти ефективність різних моделей і з’ясувати, яка з них краще справляється зі зміною мови та наявністю побутових шумів.

Об’єктом дослідження є процес розпізнавання психоемоційного стану людини за звуковими характеристиками її голосу.

Предметом дослідження є методи та алгоритми машинного навчання, які дозволяють аналізувати аудіосигнали та класифікувати емоції.

Наукова новизна отриманих результатів полягає в тому, що було вдосконалено підхід до порівняння систем машинного слуху. Зокрема, додано етап тестування на реальних записах українською мовою, на якій моделі не навчалися. Це дозволило на практиці довести, що новітні моделі на базі трансформерів значно краще підходять для створення універсальних систем, ніж популярні візуальні методи, оскільки вони здатні розпізнавати емоції незалежно від мови.

Структурно робота складається зі вступу, чотирьох розділів, висновків та списку використаних джерел.

1 АНАЛІЗ ТИПІВ ЕМОЦІЙ В АУДІОСИСТЕМАХ

1.1 Класифікація емоцій

Завдяки передовим технологіям, зокрема штучному інтелекту та машинному навчанню, ми спостерігаємо значний прорив у розумінні та аналізі людської поведінки. Одним з найскладніших, але й найцікавіших напрямків досліджень є розпізнавання емоцій.

Створення систем, здатних розпізнавати емоції, відкриває широкі перспективи, особливо в сфері цифрової взаємодії. Ці системи можуть не лише виконувати команди користувачів, але й адаптуватися до їх емоційного стану. Це робить взаємодію більш природною та індивідуальною, забезпечуючи унікальний досвід для користувачів.

Емоції відіграють ключову роль у різних сферах життя. Наприклад, у бізнесі аналіз емоцій клієнта дозволяє створювати більш персоніфіковані набори товарів чи послуг, ґрунтуючись на його реальному настрої в момент спілкування. Такий підхід допомагає компаніям краще розуміти потреби людини і робити сервіс не просто автоматизованим, а більш чуйним та доречним.

В умовах інформаційного перенасичення, коли обробка тисяч годин аудіоданих (дзвінків, інтерв’ю, подкастів) вручну є неможливою, автоматизовані системи стають безальтернативним рішенням для аналітики.

Однак перед розробкою таких систем виникає питання: як перетворити суб’єктивні переживання людини на зрозумілі для комп’ютера дані?

Історично моделювання емоцій базується на двох підходах: категоріальному та вимірному.

Категоріальний підхід ґрунтується на припущенні, що існує обмежений набір базових, універсальних для всіх культур емоцій. Фундаментальною тут вважають роботу Пола Екмана, написану у 1992 р., який виділив шість основних станів: гнів, огида, страх, щастя, сум та здивування. У сучасних дослідженнях

машинного навчання ця модель залишається домінуючою завдяки своїй чіткій структурі, що дозволяє звести задачу розпізнавання до класичної мультикласової класифікації [1]. Більшість популярних датасетів (наприклад, RAVDESS, TESS, EMODB) використовують саме категоріальну розмітку, що дозволяє звести задачу машинного навчання до класичної багатокласової класифікації.

Водночас модель Екмана часто критикують за її обмеженість: вона погано описує складні, змішані (наприклад, ностальгія, розчарування) або ледь виражені емоційні стани.

Вимірна модель розглядає емоції не як окремі категорії, а як точки в багатовимірному просторі. Найвідоміший приклад – циркумплексна модель Джеймса Рассела, запропонована у 1980 році. Хоча сучасні огляди підтверджують, що цей підхід точніше описує складні стани [2], його головним недоліком залишається складність розмітки даних: асесори часто дають суперечливі оцінки за шкалами, що ускладнює навчання нейромереж.

У цій роботі перевага надається категоріальній моделі, оскільки вона найкраще відповідає структурі більшості еталонних датасетів.

1.2 Критерії визначення емоцій в голосі

Виявлення емоцій значною мірою спирається на голос, адже він надає комплексну інформацію: від змісту сказаного до таких нюансів, як інтонація, темп мови та паузи.

Те, як ми себе почуваємо психологічно, відбивається у нашому голосі завдяки вокальній просодії. Це стосується не стільки слів, скільки того, як ми їх говоримо: з якою інтонацією, в якому ритмі, з якою швидкістю та гучністю. Ці голосові характеристики змінюються залежно від нашого внутрішнього стану і є головними підказками про наші емоції. Тому, коли ми аналізуємо мову, щоб зрозуміти емоції, ми вимірюємо певні акустичні показники (кореляти), які пов'язані з цими елементами просодії.

Ключові звукові ознаки, що вказують на емоції:

а) частота основного тону (F0):
   – емоції високого збудження (гнів, щастя, страх) вирізняються високим середнім значенням F0 та широким динамічним діапазоном;
   – спокійні стани (сум, нудьга) характеризуються нижчим голосом і меншою зміною його висоти, тобто мова стає більш монотонною (мала варіативність F0) (див. рис. 1.1) [3];

б) гучність:
   – емоції, що вимагають активної реакції, як правило, виражаються голосніше, з вищою середньою гучністю та помітними сплесками гучності;
   – стани, що вимагають меншої активності (сум, нудьга), навпаки, звучать тихіше;

в) швидкість та темп мовлення:
   – емоції, пов’язані з високим рівнем збудження, часто проявляються у швидшому темпі мовлення на коротших паузах між словами
   – сум і нудьга, навпаки, зазвичай супроводжуються повільнішим темпом мовлення та довшими паузами.

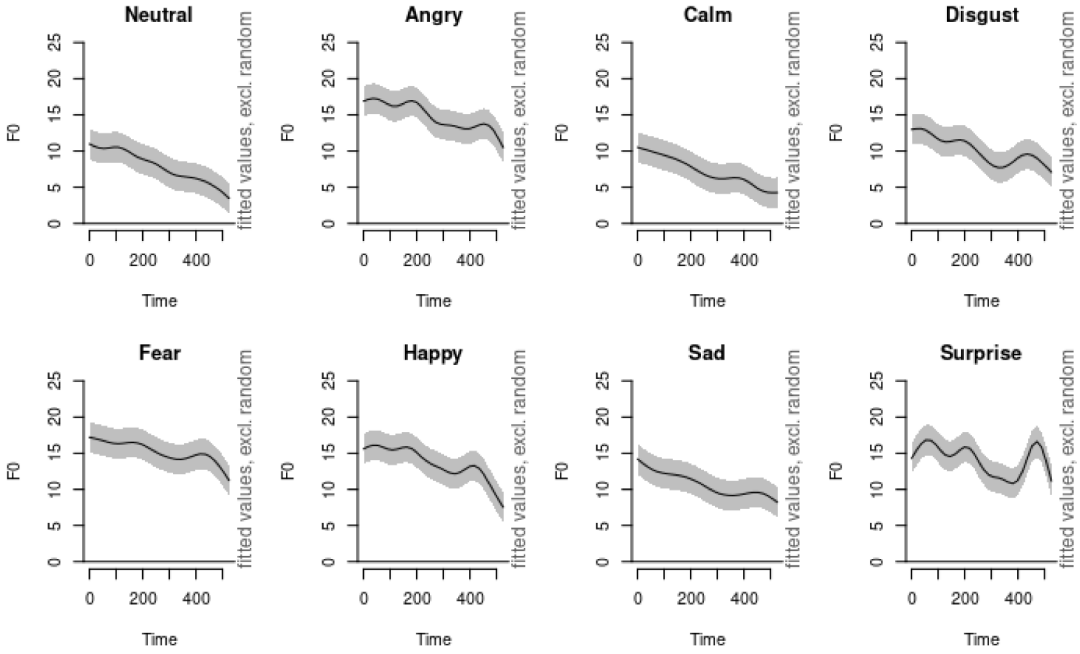


Рисунок 1.1 – F0-контури емоційних типів

1.3 Аналіз сучасних систем машинного слуху

Сучасний ринок технологій машинного слуху розділений на два сегменти: масові споживчі продукти, де аналіз емоцій виконує допоміжну роль, та спеціалізовані B2B-платформи, орієнтовані на глибокий емпатичний аналіз.

1.3.1 Голосові асистенти

Найбільш поширеними системами, які щоденно обробляють мільйони годин аудіоданих, є голосові асистенти, такі як Amazon Alexa та Google Assistant. Їхня пріоритетна мета – високоточне автоматичне розпізнавання мовлення та розуміння природної мови.

У цих екосистемах розпізнавання емоцій не є явною функцією для користувача, проте слугує інструментом покращення UX.

Система може виявити роздратування (наприклад, різке підвищення гучності та F0), коли користувач стикається з повторними помилками розпізнавання. Це дозволяє асистенту змінити свій сценарій відповіді (наприклад, перепросити або запропонувати альтернативу) [4].

Також аналіз тону допомагає системі відрізнити пряму команду, адресовану пристрою, від фонової розмови між людьми.

1.3.2 Спеціалізовані B2B-платформи

На відміну від голосових асистентів, такі компанії, як Affectiva (зараз частина Smart Eye), розробили технологію Emotion AI як прямий комерційний продукт для бізнес-клієнтів (B2B) [5].

На відміну від масових рішень, B2B-платформи фокусуються на кількісній оцінці емоцій. Їхні рішення використовуються в аналізі медіаконтенту (реакція аудиторії на рекламу), моніторингу стану водія та аналітиці кол-центрів.

Affectiva Voice AI реалізована через API, що повертає набір метрик, отриманих з аудіопотоку. Загалом, принцип роботи ґрунтується на гібридній моделі, що поєднує:

– категоріальний підхід, спрямований на розпізнавання дискретних емоцій (наприклад, гнів, радість, сміх тощо);
– вимірний підхід, який передбачає безперервне оцінювання емоційних характеристик за шкалами збудження та валентності.

Такий підхід дозволяє клієнтам отримувати деталізовані дані про емоційну динаміку, що є критично важливим для їхніх бізнес-задач.

1.3.3 Великі мовні моделі з емоційним інтелектом

Hume AI представляє новітнє покоління систем, які намагаються подолати обмеження класифіка ції на 6-7 базових емоцій. Їхній підхід базується на критиці обмеженості дискретних моделей та робить акцент на гранулярному моделюванні вокальних експресій.

Моделі Hume AI навчені розпізнавати не категорії, а інтенсивність десятків емоційних станів (наприклад, захоплення, розгубленість, презирство, задоволення).

Ключова інновація Hume AI полягає в поєднанні аналізу та генерації. Їхній продукт Empathic Voice Interface є діалоговою системою, що аналізує тон голосу користувача і генерує відповідь з відповідною емпатичною інтонацією. Таким чином, система намагається створити повний цикл сприйняття-реакція на просодичному рівні [6].

Для систематизації відмінностей розглянутих рішень складено порівняльну таблицю (таблиця 1.1).

Таблиця 1 – Порівняльна характеристика сучасних систем машинного слуху

Система	Amazon Alexa	Affectiva	Hume AI
Тип системи	голосовий асистент	B2B платформа	велика мовна модель з емоційним інтелектом
Цільова аудиторія	звичайні люди	компанії	звичайні люди
Основна мета	зрозуміти слова і виконати команду	виміряти реакцію людини	відчути настрій та підтримати розмову
Роль розпізнавання емоцій	неявна функція для покращення відповіді	комерційний продукт для виявлення емоцій	живий діалог з людиною
Принцип розпізнавання емоція	чують, якщо людина почала кричати або говорити різко	аналізують міміку, голос	виявляють найтонші нотки: зітхання, розгубленість, смішок
Результат	асистент змінює відповідь	компанія розуміє, чи працює її реклама або чи засинає водій	ШІ відповідає з відповідною до вашого настрою емоцією

1.4 Висновки до розділу 1

У даному розділі проведено аналіз предметної області та теоретичних основ розпізнавання емоційного стану людини. Встановлено, що емоції відіграють критичну роль у сучасних інтерфейсах, перетворюючи їх з інструментів виконання команд на адаптивних асистентів.

Аналіз існуючих рішень показав, що ринок рухається від простих класифікаторів до складних генеративних систем (як Hume AI). Однак більшість доступних рішень або є закритими комерційними продуктами, або не адаптовані

для роботи з різними мовними групами. Теоретичний огляд моделей емоцій дозволив обрати категоріальний підхід (модель Екмана) як базовий для даної роботи, оскільки він забезпечує найкращу сумісність з існуючими розміченими датасетами. Визначені акустичні кореляти (F0, енергія, темп) стануть основою для побудови системи ознак у наступних розділах.

2 ДОСЛІДЖЕННЯ МЕТОДІВ ОБРОБКИ АУДІОСИГНАЛІВ

Первинний сигнал являє собою послідовність амплітудних значень, що містить значну кількість шуму та надлишкової інформації. Для того, щоб модель машинного навчання могла здійснити класифікацію емоцій, первинний аудіосигнал потребує багатоступеневої трансформації. Цей процес перетворення сирих даних на інформативний вектор ознак включає ряд послідовних етапів: дискретизацію та квантування для оцифрування звуку, попередню обробку (нормалізація амплітуди, шумозаглушення, фреймування, детекція мовленнєвої активності), а також глибинний аналіз частотних характеристик методами спектрального (STFT, Мел-спектрограми), кепстрального (MFCC) та просодичного аналізу (рис. 2.1).

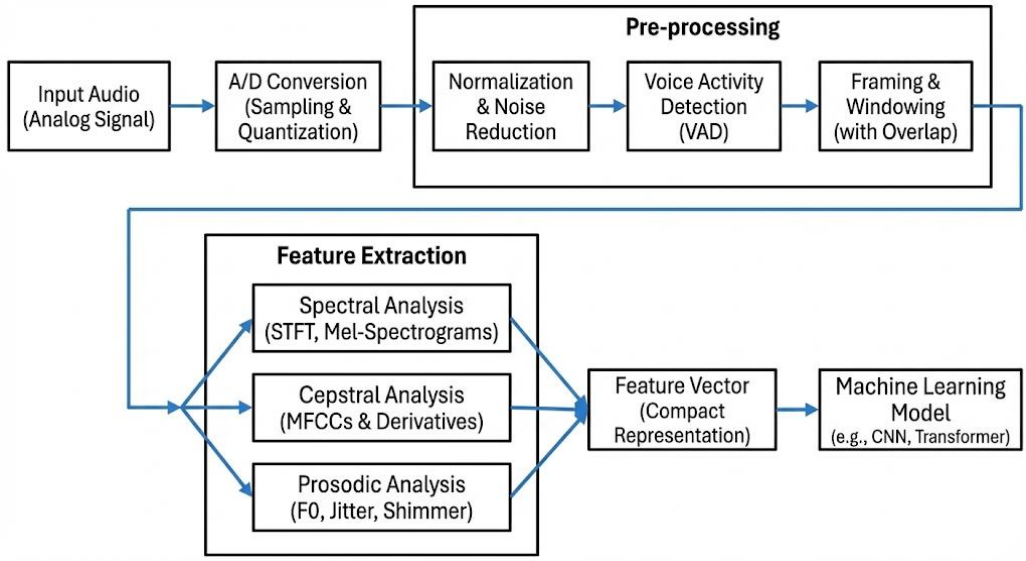


Рисунок 2.1 – Структурна схема етапів обробки аудіосигналу

Тому ключовим етапом обробки аудіосигналів є виділення ознак. Цей процес передбачає перетворення вхідного сигналу у компактне інформаційно насичене представлення, яке фокусується на акустичних корелятах емоцій.

2.1 Попередня обробка та підготовка аудіоданих

Перед тим як застосовувати алгоритми глибокого навчання, «сирі» акустичні дані повинні пройти етап уніфікації та очищення. Це зумовлено тим, що якість вхідних даних є вирішальним фактором для ефективності моделі: наявність шумів, різниця у форматах запису або технічна тиша можуть призвести до формування хибних патернів під час навчання. Процес попередньої підготовки формує фундамент для подальшого виділення ознак [7].

2.1.1 Дискретизація та квантування аудіоданих

В основі будь-якої системи машинного слуху лежить перехід від фізичного сприйняття звуку до його цифрової обробки. Звукова хвиля, що потрапляє на мікрофон, є аналоговим сигналом, тобто безперервною зміною звукового тиску. Щоб комп'ютерні алгоритми, особливо нейронні мережі, могли обробляти цей сигнал, його необхідно перетворити на дискретний набір числових значень. Цей процес, який називається оцифруванням або аналого-цифровим перетворенням, включає два ключові етапи: дискретизацію в часі та квантування за амплітудою.

Якість оцифрування на етапі збору даних визначає максимальну точність, яку може досягти майбутня модель машинного слуху. Якщо під час оцифрування втрачаються важливі частоти або динамічні характеристики звуку, їх неможливо буде відновити, навіть використовуючи найсучасніші алгоритми.

Дискретизація – фіксація миттєвого значення амплітуди сигналу через рівні проміжки часу (рис. 2.2). Головним параметром є частота дискретизації (sample rate), тобто кількість таких вимірів за одну секунду.

Згідно з теоремою фундаментальною теоремою відліків, яка є базовим принципом цифрової обробки сигналів у сучасних системах штучного інтелекту, для коректного відновлення аналогового сигналу частота дискретизації має бути як мінімум удвічі вищою за найвищу частоту спектру цього сигналу [8]. Оскільки мовленнєвий діапазон людини, що несе основне інформаційне навантаження,

лежить у межах від 300 Гц до 3400 Гц (стандарт телефонії), історично для передачі голосу використовували частоту 8 кГц.

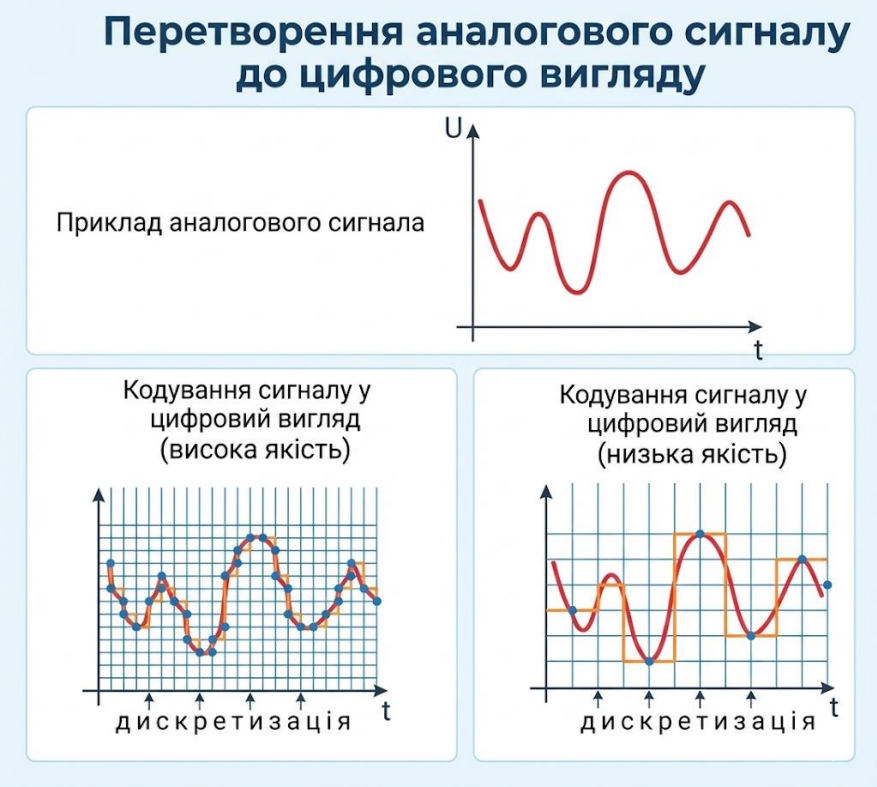


Рисунок. 2.2 – Принцип дискретизації аналогового сигналу

Проте для задачі розпізнавання емоцій стандарту у 8 кГц недостатньо. Наприклад, різкість у голосі при гніві або шепіт при страху формуються саме у верхньому спектральному регістрі. Обрізання цих частот робить голос «пласким» і емоційно невиразним, що суттєво ускладнює роботу класифікатора [9].

На практиці для сучасних систем афективних обчислень стандартом стала частота 16 кГц або 22.05 кГц. У такому діапазоні, з одного боку, ми захоплюємо частотний діапазон до 8–11 кГц, що покриває 99% акустичних ознак мовлення, а з іншого – не перевантажуємо систему зайвими даними, як це було б при використанні музичного стандарту 44.1 кГц.

Якщо дискретизація «розбиває» сигнал по горизонталі (у часі), то квантування відповідає за точність вимірювання по вертикалі (амплітуда).

Кожному виміряному значенню присвоюється найближче ціле число з фіксованого набору. Кількість таких рівнів визначається розрядністю.

Найбільш розповсюдженим є 16-бітне квантування, що дає 65 356 рівнів гучності та динамічний діапазон близько 96 дБ.

Використання меншої розрядності (наприклад, 8 біт) призводить до появи «шуму квантування». Для нейромережі такий шум може стати такою ознакою, яку вона помилково охарактеризує як хрипкість голосу або дихання, що призведе до хибних результатів.

2.1.2 Нормалізація амплітуди та зменшення шумів

Аудіозаписи в датасетах часто зроблені за різних умов: на різній відстані від мікрофону, з різним рівнем підсилення та гучністю голосу. Без попередньої обробки класифікатор може помилково інтерпретувати технічно тихий запис як прояв суму або невпевненості, а гучний – як гнів.

Для нівелювання цього фактору застосовують нормалізацію. Найчастіше використовують нормалізацію за піковим рівнем (peak normalization) або за середньоквадратичним значенням (RMS). Це дозволяє привести всі файли до єдиного динамічного діапазону, зберігаючи при цьому гучність всередині фраз, що є важливим для емоційної виразності.

Одночасно з цим відбувається процес зменшення шуму. На практиці аудіозаписи часто містять небажані звуки, такі як гудіння техніки, вуличний шум або відлуння приміщення. Ці перешкоди спотворюють аудіосигнал і можуть ускладнити роботу алгоритмів. Для усунення цієї проблеми використовуються методи фільтрації, які дозволяють відокремити постійний фоновий шум від мови [10]. Таке очищення сигналу зберігає лише важливу інформацію, що суттєво покращує точність подальшого розпізнавання емоцій.

2.1.3 Фреймування

Голос людини – дуже динамічний потік: інтонація, гучність змінюються миттєво. Якщо спробувати проаналізувати довгий шматок запису як єдине ціле, результат буде неточний, адже за цей час емоція могла змінитися кілька разів. Тому стандартним підходом є розділення аудіопотоку на дуже короткі відрізки – фрейми.

Оптимальною тривалістю фрейму є 20-30 мілісекунд. Саме за такий проміжок часу голосовий апарат людини фізично не встигне кардинально перебудуватися.

Щоб не втратити важливу інформацію на стиках між цими шматочками, їх нарізають не впритул один до одного, а з перекриттям (overlap), зазвичай на 50% (рис. 2.3) [11]. Це забезпечує плавність аналізу і гарантує, що жоден нюанс інтонації не загубиться при переході від одного фрагмента до іншого.

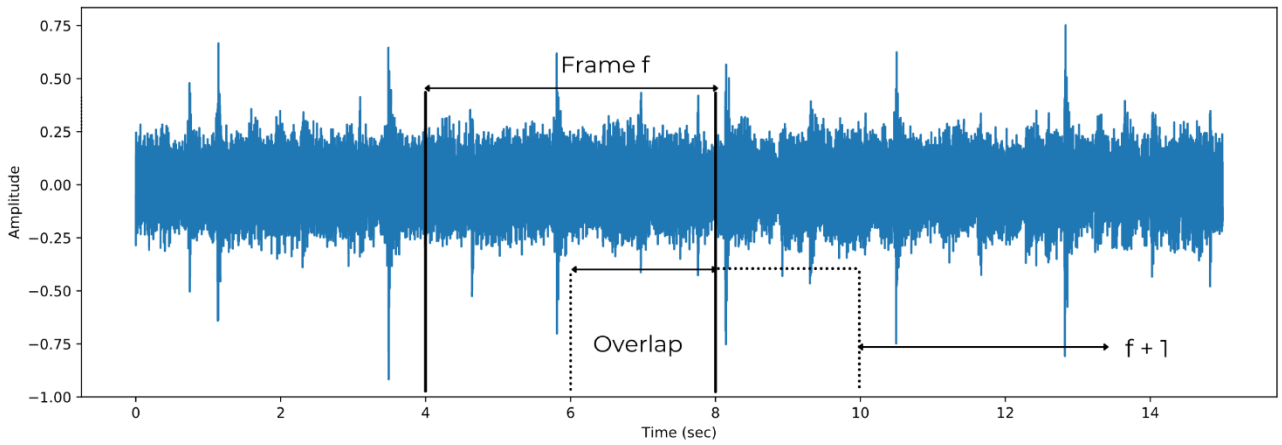


Рисунок 2.3 – Процес фреймування сигналу з перекриттям

2.1.4 Детекція мовленнєвої активності

Значну частину будь-якого аудіозапису складають моменти тиші або вдихів між словами. Ці ділянки не несуть інформації про емоційний стан, але створюють зайве навантаження на обчислювальну систему і можуть погіршувати статистику при навчанні моделі.

Алгоритми детекції мовленнєвої активності (VAD) автоматично розмічають сигнал, відокремлюючи мовлення від тиші. Найпростіші реалізації VAD базуються на енергетичному критерії: фрейми з енергією нижче заданого рівня відкидаються. Більш складні методи аналізують частотний спектр та періодичність сигналу, щоб відрізнити голос від гучного фонового шуму [12]. Видалення пауз дозволяє сфокусувати увагу нейромережі виключно на емоційно забарвлених сегментах аудіопотоку.

2.2 Спектральний аналіз

Аналіз аудіосигналу виключно в часовій області надає лише поверхневе уявлення про його структуру. Але емоційне забарвлення голосу приховане у частотах. Саме там проявляється тембр, інтонаційні акценти та інші емоційні забарвленності. Наприклад, щоб відрізнити смуток від спокою необхідно бачити, як енергія голосу розподіляється між низькими та високими частотами.

2.2.1 Перетворення Фур’є

Перетворення Фур’є є своєрідним математичним фундаментом переходу від часового представлення звуку до частотного. Його принцип полягає у декомпозиції складного звукового сигналу на сукупність простих синусоїд, що відрізняються за частотою та амплітудою. Оскільки мовлення, як ми вже з'ясували, є нестаціонарним явищем, ми вдаємося до застосування короткочасного перетворення Фур'є (STFT).

Кожен підготовлений кадр обробляється алгоритмом для отримання його частотного спектра. Якщо розташувати отримані спектри послідовно один за одним у хронологічному порядку, ми отримаємо спектрограму (рис. 2.4).

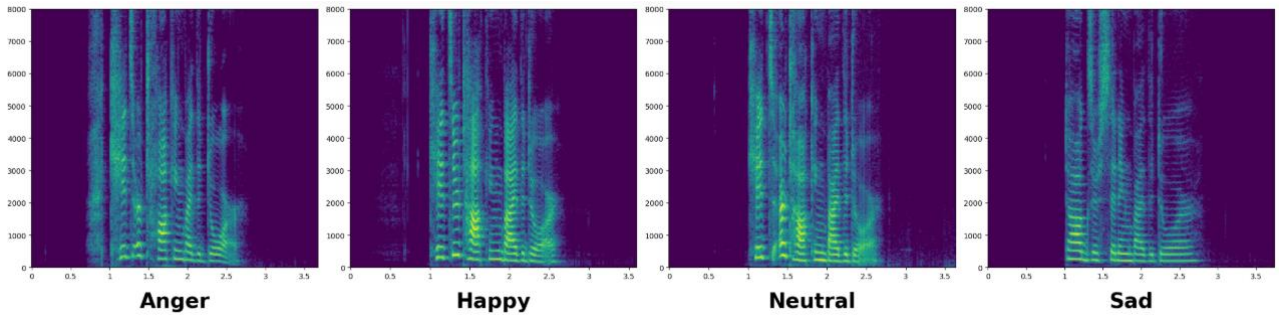


Рисунок 2.4 – Приклади спектрограм емоцій

2.2.2 Мел-спектрограми

Проте звичайна лінійна спектрограма погано відповідає людському сприйняттю, оскільки наше вухо розрізняє низькі частоти набагато краще, ніж високі.

Щоб наблизити комп'ютерний аналіз до людського сприйняття, використовують Mel-шкалу (рис. 2.5).

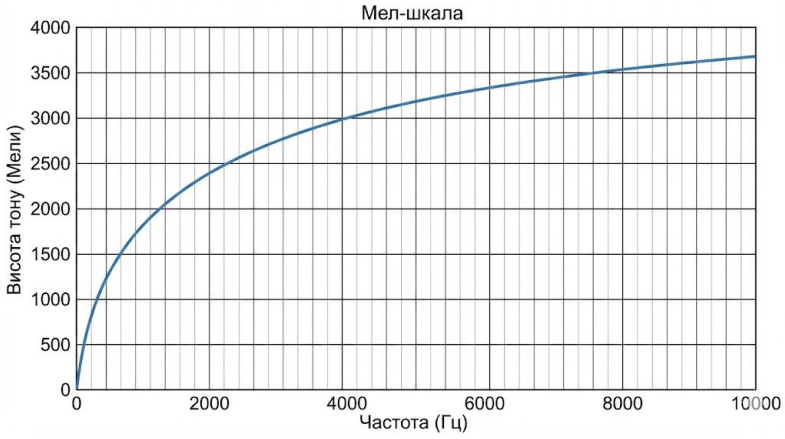


Рисунок 2.5 – Вигляд мел-шкали

В Mel-спектрограмі (рис. 2.6) вісь Y відображає частоти в Mel-шкалі, а колір – їхню інтенсивність. Сучасні нейромережі, зокрема CNN, часто використовують Mel-спектрограми як вхідні дані, обробляючи їх як зображення для пошуку патернів емоцій [13].

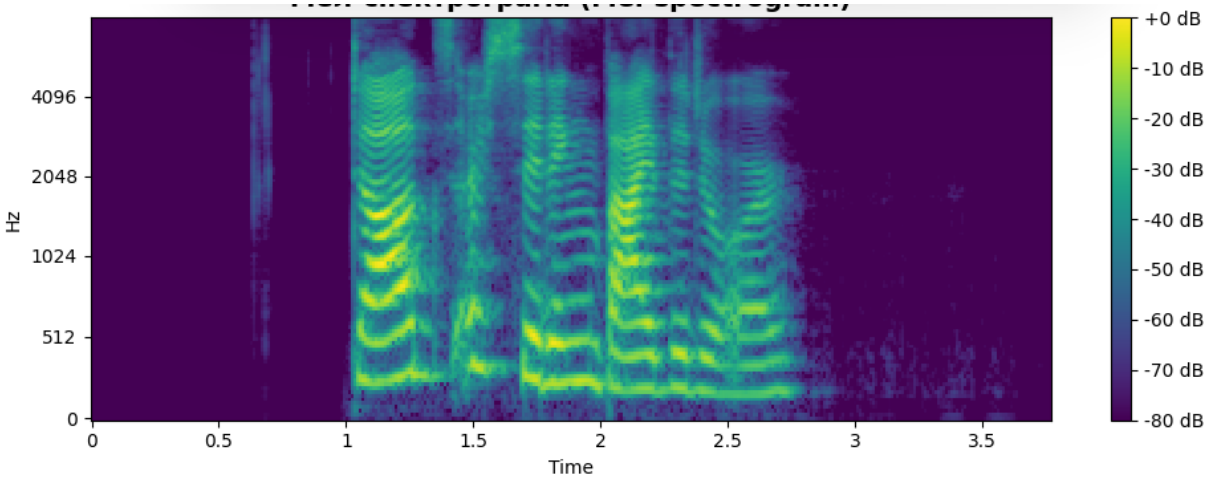


Рисунок 2.6 – Приклад Mel-спектрограми

2.3 Кепстральний аналіз та MFCC

Якщо спектральний аналіз дає загальну картину розподілу енергії частот, то кепстральний аналіз дозволяє зазирнути у фізику формування звуку (рис. 2.7). Саме на даному етапі у задачах розпізнавання емоцій ми намагаємось відокремити індивідуальні особливості голосу від того, як людина говорить у конкретний момент.



Рисунок 2.7 – Алгоритм отримання кепстральних коефіцієнтів

Не зважаючи на появу нових архітектур, у задачах обробки мовлення ключовими залишаються мел-кепстральні коефіцієнти. Їхня ефективність зумовлена тим, що вони найкраще моделюють специфіку людського слуху.

В основі методу лежить фізична модель Source-filter theory (рис. 2.8). Згідно з нею, мовленнєвий сигнал формується двома незалежними процесами: голосові зв’язки генерують вихідний звук, що визначає висоту тону, а фільтр (глотка, ротова та носова порожнини) модулює цей звук, формуючи унікальний тембр.

При розпізнаванні емоцій саме фільтр несе основне смислове навантаження. MFCC дозволяють алгоритмічно відокремити цю тембральну складову від висоти тону.

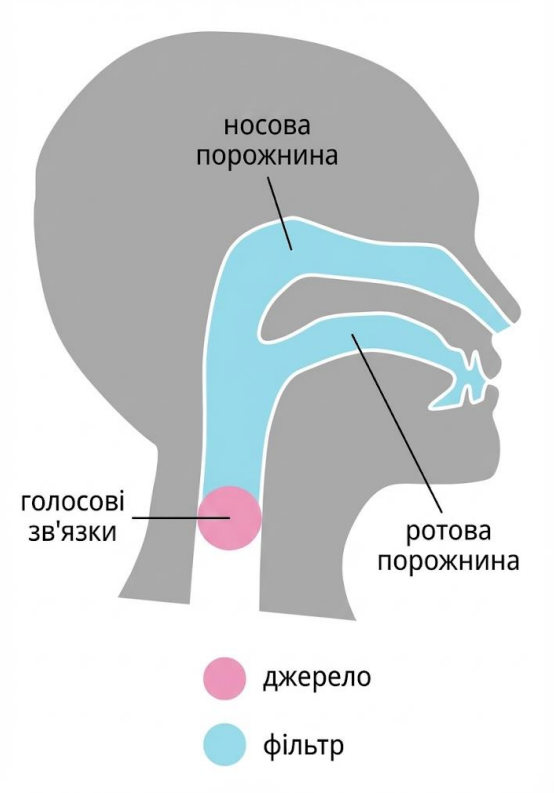


Рисунок 2.8 – Source-filter theory

На практиці зазвичай використовують вектор із перших 12–13 коефіцієнтів. Але вони дають лише статичний «зріз» тембру. Щоб захопити емоційну динаміку (наприклад, різкий зрив голосу при переляку), до базових коефіцієнтів додають їхні перші та другі похідні. Це дозволяє моделі оцінювати не тільки миттєвий стан голосу, а й швидкість та прискорення його змін, що критично важливо для виявлення різких емоційних сплесків.

2.4 Аналіз якості голосу та просодичні характеристики

Хоча спектральні та кепстральні методи дають багато інформації, вони більшою мірою фокусуються на тембрі. В той час, як психоемоційний стан людини суттєво впливає не лише на тембр, а й на стабільність голосу.

Коли людина перебуває у стані сильного стресу, страху або хвилювання, вона втрачає точний контроль над м’язами гортані. Це викликає незначні, непомітні на звичайних спектрограмах коливання в голосі, але які є важливими показчиками емоцій.

Для вимірювання цих відхилень використовують метрики якості голосу:

– частота основного тону визначає, наскільки високим чи низьким є голос. Аналіз контуру F0 дозволяє розрізняти емоції за рівнем збудження: наприклад, гнів та радість проявляється у високому та мінливому тоні, тоді як сум чи нудьга – у низькому та монотонному;
– jitter (джиттер) вимірює незначні коливання частоти голосу від одного звукового циклу до іншого. На слух високий джиттер сприймається як «тремтіння» або хрипліть. Він є одним з найточніших показників наявності проблем зі здоров’ям голосу або сильного емоційного потрясіння;
– shimmer (шиммер), подібно до джиттера, стосується амплітуди звуку. Він відображає мінливість гучності голосу. Збільшення шиммера пов’язане з задишкою або придушеного гніву.

Поєднання цих параметрів із MFCC значно підвищує надійність системи. Це допомагає уникнути плутанини між емоціями, які мають однакову інтенсивність і для звичайного спектрального аналізу виглядають майже однаково.

2.5 Висновки до розділу 2

Проведений у даному розділі аналіз методів цифрової обробки сигналів дозволив сформувати стратегію попередньої підготовки даних, яка є фундаментом для подальшого навчання нейронних мереж. Дослідження показало, що безпосереднє використання необробленого аудіосигналу є малоефективним через його високу розмірність та наявність шумів, тому було обґрунтовано необхідність побудови багатоетапного пайплайну обробки.

Визначено, що першочерговими кроками мають бути уніфікація частоти дискретизації на рівні 16–22 кГц, нормалізація амплітуди для нівелювання різниці в умовах запису та застосування алгоритмів детекції мовленнєвої активності для видалення неінформативних пауз.

Для етапу виділення ознак обрано підхід, який поєднує сильні сторони різних методів аналізу. Зокрема, використання Мел-спектрограм дозволить застосувати апарат комп'ютерного зору для пошуку візуальних патернів емоцій, тоді як розрахунок MFCC та просодичних характеристик забезпечить врахування фізичних особливостей голосоутворення. Такий комплексний підхід до формування простору ознак дозволить майбутній системі розпізнавати як явні зміни гучності, так і тонкі мікро-варіації голосу, характерні для прихованих емоційних станів.

3 ВИБІР МЕТОДІВ МАШИННОГО НАВЧАННЯ ТА НАБОРУ ДАНИХ

Розробка системи розпізнавання емоційного стану людини є складним завданням, яке вимагає не лише реалізації програмного коду, а й глибокого розуміння вхідних даних та архітектурних особливостей нейронних мереж. У даному розділі проведено аналіз існуючих наборів даних та архітектури машинного навчання.

3.1 Вибір набору даних

При виборі наборів даних критичним є не лише обсяг даних, але й акустична чистота, валідність розмітки та придатність для навчання глибоких нейронних мереж. Для практичної оцінки доцільно порівняти популярні датасети.

EMO-DB – вільнодоступний датасет німецькою мовою, створений Інститутом комунікації Технічного університету Берліна [14]. Він містить 390 висловлювань від 10 професійних акторів (5 чоловіків, 5 жінок) і покриває 7 емоцій: гнів, нудьгу, тривогу, щастя, сум, огиду та нейтральний стан.

Незважаючи на відмінну якість запису, є суттєве обмеження EMO-DB для завдань даного дослідження - мовний бар'єр. Акустичні патерни німецької мови (зокрема, твердість голосних, специфічна просодика та інтонаційні контури) кардинально відрізняються від слов'янських або англійської мов. Використання цього датасету як основного може внести в модель специфічні лінгвістичні зміщення (bias), що ускладнить подальше масштабування системи.

Датасет IEMOCAP містить записи спонтанних діалогів та імпровізацій, що робить його максимально наближеним до реальних умов спілкування. Однак для початкового навчання моделі IEMOCAP створює багато проблем: наявність фонових шумів та часті перебивання співрозмовників.

Набір даних RAVDESS створений дослідниками Університету Раєрсона та на сьогодні виступає «золотим стандартом» у сфері мультимодального аналізу емоцій [15].

У записі даних беруть участь 24 професійні актори (12 жінок і 12 чоловіків). Усі записи виконані у професійній студії. На відміну від EMO-DB, RAVDESS пропонує два рівні інтенсивності емоцій – нормальний та сильний, а також розрізняє нейтральний та спокійний стани. Кожен актор вимовляє дві лексично нейтральні фрази ("Kids are talking by the door", "Dogs are sitting by the door"). Такий підхід нівелює вплив змісту слів на сприйняття, змушуючи алгоритм аналізувати саме інтонацію.

Кожен запис у RAVDESS пройшов перевірку 247 незалежними експертами. Згідно з технічним звітом авторів датасету, середня точність розпізнавання цих записів людиною становить 62-67% [15]. Цей показник ми використовуємо як базовий орієнтир: якщо наша модель перевищить цей поріг, систему можна вважати ефективною.

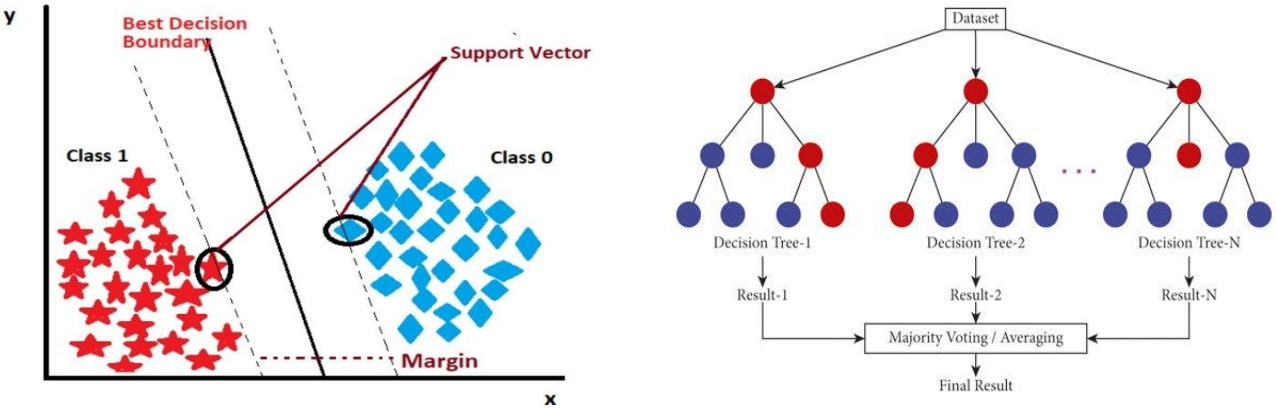
3.2 Методологія вибору архітектур машинного навчання

Обравши джерело даних, наступним кроком є вибір алгоритму класифікації. Він визначає не лише точність розпізнавання, але й вимоги до обчислювальних ресурсів, швидкість навчання та здатність системи до узагальнення. У даному розділі проведено аналіз алгоритмів трьох поколінь: класичного машинного навчання (SVM, Random forest), глибокого навчання на основі згорток (CNN) та сучасних механізмів уваги (Transformers).

3.2.1 Класичні алгоритми машинного навчання

На початкових етапах розвитку галузі розпізнавання емоцій (до 2012–2014 років) домінуючим підходом було використання класифікаторів на кшталт

Методу опорних векторів (SVM) та ансамблевих методів, таких як Random Forest (рис. 3.1).



а) метод опорних векторів б) random forest

Рисунок 3.1 – Приклади класифікаторів

Але класичні методи не вміють працювати з необробленими даними. Вони потребують, щоб людина-експерт заздалегідь перетворила звук на набір статистичних коефіцієнтів (середнє значення пітчу, дисперсія енергії, ексцес тощо). Цей процес є суб'єктивним, бо неможна точно сказати, які саме фізичні параметри відповідають за ту чи іншу емоцію.

Для подачі аудіосигналу на вхід SVM або Random Forest його зазвичай "сплющують" у вектор фіксованої довжини шляхом обчислення середніх значень по всьому файлу (Global Statistics). Це призводить до катастрофічної втрати інформації про динаміку. Наприклад, фраза з інтонацією, що різко падає (розчарування), та фраза з інтонацією, що поступово зростає (здивування), можуть мати абсолютно однакове середнє значення частоти основного тону. Для SVM ці два стани будуть нерозрізненні, тоді як емоційний зміст у них протилежний.

Тому, враховуючи, що емоція – складний процес, що змінюється та розгортається в часі а також має складну нелінійну структуру, класичні методи є недостатньо гнучкими для вирішення поставленої задачі на сучасному рівні якості.

3.2.2 Згорткові нейронні мережі

Усвідомлення обмежень класичного підходу призвело до зміни парадигми в обробці сигналів: переходу від аналізу одновимірних статистичних рядів до аналізу двовимірних частотно-часових представлень. У цьому контексті перспективним напрямком є використання згорткових нейронних мереж (CNN) [16].

Хоча ця архітектура первинно була розроблена для комп'ютерного зору, вона демонструє високу ефективність у задачах обробки аудіо, якщо представити звук візуально – у вигляді спектрограми.

Метод полягає у перетворенні акустичної задачі на задачу розпізнавання образів. Аудіосигнал візуалізується у вигляді мел-спектрограми (час на осі X, частота на осі Y, амплітуда – яскравість пікселя), що дозволяє виявити візуальні патерни, які відповідають емоційному забарвленню голосу.

Основною перевагою CNN над класичними методами є здатність до автоматичного виявлення ознак завдяки операції згортки. Мережа використовує набір невеликих фільтрів, які сканують спектрограму крок за кроком. Це забезпечує дві критичні властивості, недоступні для SVM.

– інваріантність до зсуву, завдяки якій CNN здатна розпізнати патерн емоції (наприклад, різкий сплеск енергії) незалежно від того, де саме він знаходиться – на початку, в середині чи в кінці запису [17].
– збереження локальної структури, оскільки мережа аналізує не середнє значення по всьому файлу, а взаємозв'язок між сусідніми частотними та часовими точками, що дозволяє фіксувати короткочасні акустичні події.

Проте, незважаючи на значний прогрес порівняно з класичними алгоритмами, CNN має свої обмеження. Згорткові фільтри ефективно знаходять короткочасні патерни, але мають обмежене поле сприйняття. Це означає, що мережі важко пов'язати подію на початку довгої фрази з подією в її кінці. CNN бачить окремі фрагменти емоційної картини, але може втрачати загальний контекст, що розгортається у часі. Крім того, дослідження вказують на

схильність CNN фокусуватися переважно на енергетичних характеристиках (гучності), що може призводити до помилок при розрізненні емоцій з однаковим рівнем збудження, але різною валентністю (наприклад, страх та щастя) [16].

Саме ці обмеження змушують дослідників не зупинятися на візуальному підході, а розглядати архітектури, спеціально створені для роботи з послідовностями та контекстом (рис. 3.2) [18].

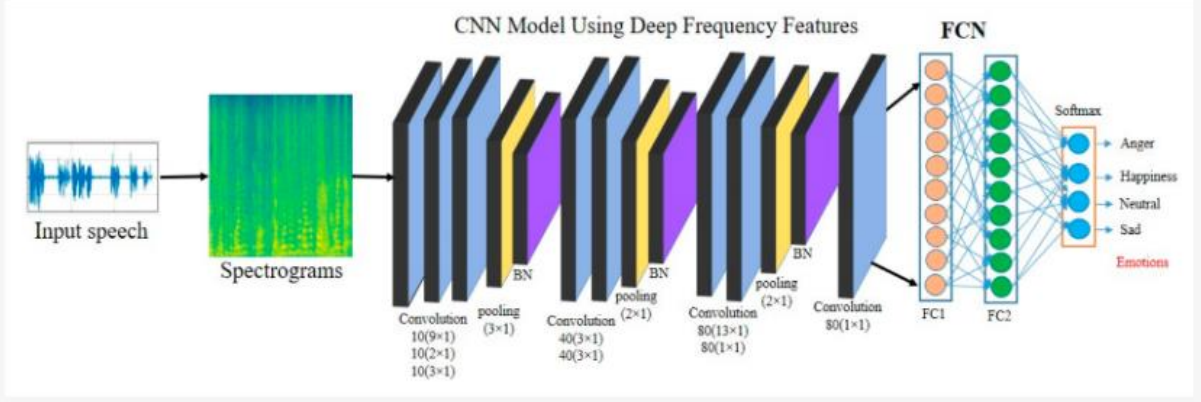


Рисунок 3.2 – Архітектура згорткової нейронної мережі

3.2.3 Трансформери

Якщо CNN намагається «побачити» емоцію на картинці спектрограми, фокусуючись на локальних патернах, то архітектура Transformer базується на спробі «зрозуміти» структуру звуку в цілому. Даний підхід сьогодні вважається найпотужнішим інструментом у сфері обробки природної мови та аудіосигналів [19].

Трансформер використовує механізм Self-Attention (рис. 3.3) [20], який дозволяє моделі динамічно оцінювати важливість кожного фрагмента аудіо у взаємозв'язку з іншими, незалежно від відстані між ними. Це дозволяє системі зрозуміти, як інтонація початку фрази впливає на емоційне забарвлення її закінчення.

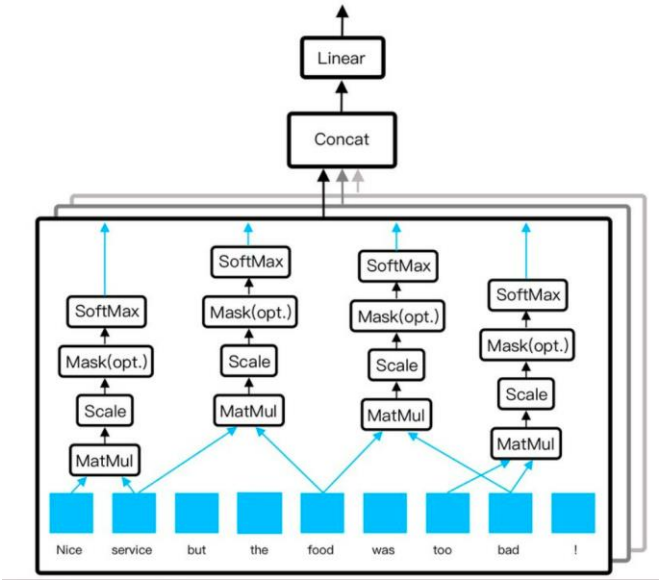


Рисунок 3.3 – Архітектура механізму Self-Attention трансформера

Однак трансформери потребують колосальних обсягів даних для навчання. Експериментальний набір RAVDESS містить лише 1440 файлів, чого критично недостатньо для навчання глибокої моделі. У такому випадку система перенавчиться, запам'ятавши лише приклади.

Для вирішення цієї проблеми доцільно застосувати стратегію Transfer Learning. Для реалізації даного підходу оптимальним рішенням є застосування попередньо навченої моделі Wav2Vec 2.0 XLSR-53 від компанії Meta AI [21]. Ця модель є унікальною тим, що була навчена на 56000 годин нерозміченого мовлення 53 мовами світу, включаючи українську. У процесі самонавчання (Self-Supervised Learning) вона вже сформувала глибинні репрезентації людської мови: навчилася відрізняти шум від голосу, ідентифікувати фонеми та, що найважливіше, виявляти особливості інтонації та ритму мовлення.

Але ця масивна модель не вміє ідентифікувати емоції з аудіофайлів. Для вирішення даної проблеми можна замінити її останній шар (який відповідав за розпізнавання звуків) на новий класифікаційний шар для емоцій.

Очікується, що використання Wav2Vec 2.0 дозволить не лише досягти максимальної точності на тестових даних, а й забезпечить здатність до крос-мовного використання, дозволяючи коректно працювати з українським мовленням без необхідності створення окремого великого датасету.

3.3 Висновок до розділу 3

У даному розділі сформовано методологічний фундамент дослідження. Як експериментальну базу, після порівняльного аналізу різних наборів даних, було обрано датасет RAVDESS. Його вибір обґрунтований такими перевагами: відмінною акустичною якістю, гендерною збалансованістю та наявністю двох рівнів інтенсивності емоцій. Ці фактори дозволяють ефективно мінімізувати вплив технічних артефактів на процес навчання нейронних мереж, забезпечуючи високу чистоту експерименту.

Виявлено, що традиційні підходи до машинного навчання (SVM, Random Forest) не справляються зі складними емоціями, оскільки вони ігнорують важливу часову динаміку сигналу. З огляду на це, було вирішено порівняти ефективність двох архітектур глибокого навчання: 2D згорткових нейронних мереж (CNN), що аналізують візуалізовані спектрограми, та трансформерів (Wav2Vec 2.0), які застосовують механізм перенесення знань.

Такий підхід дозволить не лише оцінити максимально досяжну точність розпізнавання, але й перевірити, чи здатні попередньо навчені мовні моделі до ефективного крос-мовного перенесення емоційних ознак, що є критично важливим для роботи з українськими даними. Отримані теоретичні висновки та обрані архітектури стануть основою для програмної реалізації системи у наступному розділі.

4 ПРОГРАМНА РЕАЛІЗАЦІЯ ТА ЕКСПЕРИМЕНТАЛЬНЕ ДОСЛІДЖЕННЯ

4.1 Реалізація та дослідження моделі на основі згорткових нейронних мереж

Першим етапом практичного експерименту стала розробка системи розпізнавання емоцій, що базується на візуальному аналізі звуку. Основною задачею цього етапу було встановити, чи призводить емоційне забарвлення мовлення до формування на спектрограмі стійких візуальних ознак, які піддаються ідентифікації за допомогою комп'ютерного зору.

Програмна реалізація виконувалася у середовищі Kaggle з використанням бібліотек librosa для цифрової обробки сигналів та TensorFlow для побудови нейронної мережі.

4.1.1 Попередня обробка даних

Критично важливим етапом для роботи згорткової мережі є уніфікація вхідних даних, оскільки архітектура CNN вимагає фіксованого розміру вхідного тензора. Вихідний набір даних RAVDESS містить аудіофайли різної тривалості, тому подача їх на вхід моделі у необробленому вигляді є неможливою. Для вирішення цієї проблеми було розроблено спеціалізований алгоритм попередньої обробки.

Першим кроком алгоритму є завантаження аудіосигналу та його ресемплінг до частоти 22 кГц. Такий вибір частоти дискретизації є стандартом у задачах обробки мовлення, оскільки він дозволяє охопити частотний діапазон до 11 кГц, де зосереджена більшість значущих акустичних характеристик голосу. Під час завантаження застосовувалося зміщення (offset) у 0.5 секунди для видалення технічної тиші на початку запису. Наступним кроком стала

нормалізація тривалості: всі файли приводилися до єдиного стандарту в 3 секунди. Якщо оригінальний запис був коротшим, він доповнювався нульовими значеннями (padding), що на спектрограмі виглядає як чорний фон і не вносить спотворень у патерни. Довші файли обрізалися до необхідного розміру.

Ключовим елементом підготовки даних стала генерація Мел-спектрограм. Параметри перетворення включали 128 частотних смуг та розмір вікна аналізу 2048 семплів із кроком 512. Для наближення до людського сприйняття гучності, амплітуда сигналу була переведена в логарифмічну шкалу (децибели) (додаток А).

У результаті роботи цього алгоритму кожен аудіофайл було перетворено на тривимірний тензор розмірністю 128×130×1 який фактично являє собою одноканальне (чорно-біле) зображення звуку. Саме ці дані слугували входом для навчання нейронної мережі.

4.1.2 Архітектура згорткової мережі

Для класифікації отриманих спектрограм було спроектовано нейронну мережу, архітектура якої натхненна моделями VGG [22], що демонструють високу ефективність у задачах розпізнавання образів. Структура мережі побудована за ієрархічним принципом: послідовне виділення ознак від простих геометричних примітивів до складних абстракцій.

Модель складається з чотирьох послідовних згорткових блоків. Кожен блок має уніфіковану структуру, що включає шар двовимірної згортки (Conv2D) з функцією активації ReLU, шар нормалізації (BatchNormalization) та шар субдискретизації (MaxPooling2D). Кількість фільтрів у згорткових шарах зростає з глибиною мережі: від 32 у першому блоці до 256 у четвертому. Це дозволяє мережі на початкових етапах детектувати прості патерни, такі як різкі вертикальні лінії (атаки звуку) або горизонтальні смуги (протяжні голосні), а на глибоких рівнях комбінувати їх у складні текстури, що відповідають емоційним станам (наприклад, хрипота або тремтіння).

Важливим є використання методів регуляризації. Оскільки датасет RAVDESS є відносно невеликим (1440 файлів), існує високий ризик перенавчання моделі. Для запобігання цьому після кожного блоку згортки та перед вихідним шаром було додано шари Dropout, які випадковим чином вимикають частину нейронів (25% у згорткових блоках та 50% у класифікаторі) під час навчання, змушуючи мережу формувати більш стійкі та розподілені ознаки. Завершується архітектура повнозв'язним шаром із функцією активації Softmax, яка формує розподіл ймовірностей для 8 класів емоцій.

4.1.3 Навчання та аналіз результатів

Навчання моделі проводилося на 80% даних, а 20% було відкладено для тестування. Для оптимізації процесу використовувався алгоритм Adam та функція втрат Categorical Crossentropy. Також було впроваджено систему зворотних викликів (callbacks): автоматичне збереження найкращої моделі (ModelCheckpoint), адаптивне зменшення швидкості навчання при виході на плато (ReduceLROnPlateau) та рання зупинка (EarlyStopping) для економії ресурсів. Динаміка процесу навчання відображена на графіках точності та функції втрат (рис. 4.1).

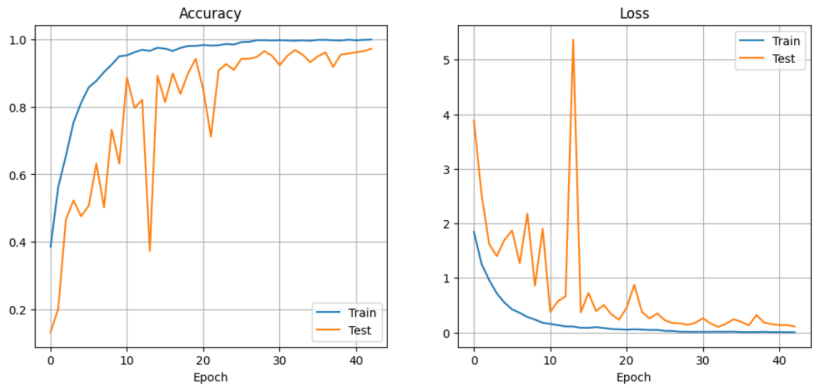


Рисунок 4.1 – Графіки зміни точності та функції втрат моделі 2D CNN

За результатами 50 епох навчання модель досягла фінальної точності на тестовій вибірці на рівні 97.2%. Детальний звіт про класифікацію наведено на рис. 4.2.

```
🏆 Фінальна точність CNN: 97.22%
📈 Графіки збережено!
18/18 ━━━━━━━━━━━━━━━━━━━━ 4s 231ms/step
🟦 Матрицю збережено!

📑 Classification Report:
              precision    recall  f1-score   support

       angry       1.00      1.00      1.00        76
        calm       0.93      1.00      0.96        77
     disgust       0.97      0.97      0.97        77
        fear       0.97      1.00      0.99        77
       happy       0.97      0.95      0.96        77
     neutral       1.00      0.84      0.91        38
         sad       1.00      1.00      1.00        77
    surprise       0.95      0.95      0.95        77

    accuracy                           0.97       576
   macro avg       0.97      0.96      0.97       576
weighted avg       0.97      0.97      0.97       576
```

Рисунок 4.2 – Звіт про метрики класифікації

Попри високий підсумковий показник, детальний аналіз матриці помилок (рис. 4.3) показав наявність систематичних похибок.

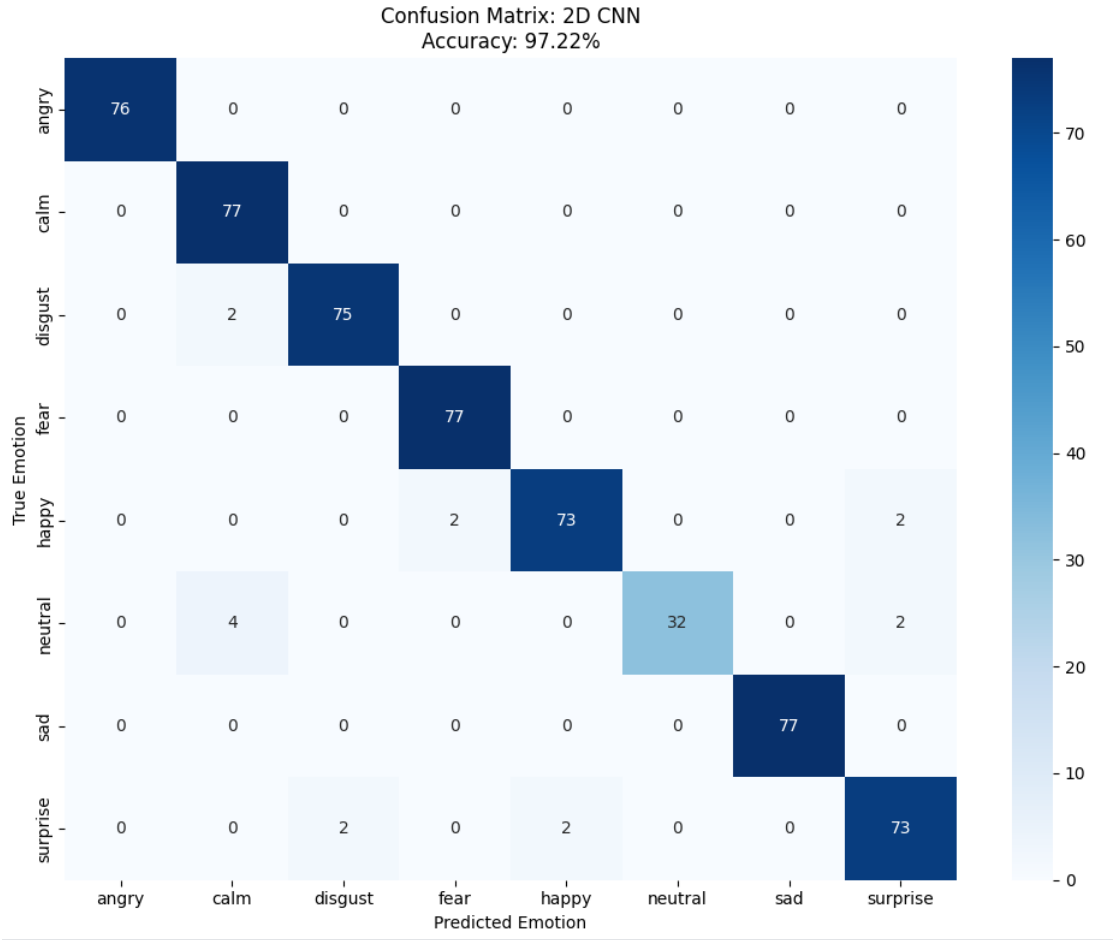


Рисунок 4.3 – Матриця плутанини моделі 2D CNN

З аналізу матриці видно, що модель ідеально розпізнає такі емоції, як спокій, страх та сум та на дуже високому рівні – злість, відразу, щастя та здивування. Найменшу кількість раз модель розпізнала нейтральну емоцію, що пов’язано з її акустичною близькістю до спокою.

Всупереч очікуванням про обмеженість візуального підходу, модель 2D CNN продемонструвала високу здатність розрізняти навіть акустично близькі емоції (як-от страх і щастя) в умовах студійного запису. Висока точність у понад 97% дозволяє стверджувати, що спектрограми містять достатньо інформативних ознак для надійної класифікації емоційного стану без залучення лінгвістичного контексту за умови високої якості вхідного аудіосигналу.

4.2 Реалізація та дослідження моделі на основі трансферного навчання

Незважаючи на високі показники точності згорткової нейронної мережі, якісний аналіз її помилок виявив суттєву залежність моделі від енергетичних характеристик звукового сигналу. Це призводило до плутанини схожих емоцій. Для подолання цих обмежень візуального підходу другим етапом дослідження стала реалізація стратегії трансферного навчання (Transfer Learning), що базується на використанні архітектури Transformer. Очікується, що завдяки попередньому навчанню на великому масиві даних, система зможе аналізувати глобальну структуру мовлення та інтонацію, а не лише миттєві характеристики звуку.

4.2.1 Підготовка даних та налаштування моделі

На відміну від попереднього методу, трансформери не потребують перетворення звуку на зображення. Вони працюють із «сирим» аудіосигналом, що дозволяє зберегти максимальну кількість інформації, включаючи мікро-паузи та інтонаційні відтінки.

Для реалізації цього підходу було використано бібліотеку Hugging Face Transformers та модель facebook/wav2vec2-large-xlsr-53 (як зазначено у розділі 3).

Реалізація здійснювалася з використанням бібліотеки Hugging Face Transformers [23]. Підготовка даних включала ресемплінг усіх записів до частоти 16 кГц, що є базовою частотою моделі Wav2Vec 2.0, та зведення їх тривалості до 3 секунд шляхом доповнення або обрізки. Фрагмент коду, що відповідає за цей процес, наведено у додатку В.

Конфігурація параметрів навчання була адаптована до обмежень апаратного забезпечення та специфіки задачі fine-tuning. Для забезпечення стабільної збіжності та запобігання руйнуванню попередньо вивчених ваг було встановлено відносно низьку швидкість навчання на рівні 1×10^{-4}. Через значний обсяг пам'яті, необхідний для роботи трансформера, розмір пакету було

обмежено 8 зразками на пристрій. Для прискорення обчислень було активовано режим змішаної точності (FP16). Загальна тривалість навчання становила 10 епох.

4.2.2 Аналіз результатів моделювання

Навчання моделі Wav2Vec пройшло успішно, про що свідчать стабільна динаміка функції втрат та швидке досягнення високої точності на валідаційних даних, як показано в логах навчання (рис. 4.4). Як видно з даних, вже після 4-ї епохи модель досягла точності понад 89%, суттєво перевершивши показники CNN на аналогічному етапі, а після 8-ї епохи вийшла на плато з показником понад 95%.

Epoch	Training Loss	Validation Loss	Accuracy
1	1.989500	1.893251	0.253472
2	1.701900	1.557260	0.411458
3	1.225300	1.135345	0.586806
4	1.005700	0.667873	0.782986
5	0.470700	0.359519	0.895833
6	0.252500	0.260419	0.925347
7	0.394300	0.242263	0.920139
8	0.131900	0.159484	0.953125
9	0.155700	0.154025	0.960069
10	0.192000	0.142505	0.963542

Рисунок 4.4 – Метрики втрати та точності під час навчання трансформера

За результатами фінального тестування на відкладеній вибірці модель досягла підсумкової точності 96.35%. Деталізований звіт про класифікацію (рис. 4.5) демонструє винятково високі показники точності (precision) та повноти

(recall) для всіх класів емоцій, причому для половини з них, а саме для емоцій злості, страху, смутку та здивування, F1-міра досягла значення 0.99-1.00.

Classification Report:

	precision	recall	f1-score	support
angry	1.00	1.00	1.00	76
calm	0.99	0.99	0.99	77
disgust	0.97	0.97	0.97	77
fear	1.00	0.99	0.99	77
happy	0.97	1.00	0.99	77
neutral	0.97	0.97	0.97	38
sad	0.97	1.00	0.99	77
surprise	1.00	0.96	0.98	77
accuracy			0.99	576
macro avg	0.99	0.99	0.99	576
weighted avg	0.99	0.99	0.99	576

Рисунок 4.5 – Метрики класифікації моделі Wav2Vec

Для якісного порівняння з попереднім підходом було проаналізовано матрицю плутанини трансформера (рис. 4.6). Порівняно з матрицею CNN, спостерігаються суттєві якісні покращення. По-перше, модель значно підвищила точність розпізнавання нейтрального стану (37 правильних відповідей), мінімізувавши плутанину зі спокоєм, яка була слабким місцем згорткової мережі. По-друге, було досягнуто абсолютної точності у розпізнаванні емоцій щастя та суму (по 77 правильних відповідей), а також майже безпомилкової ідентифікації станів гніву, страху та спокою (по 76 правильних відповідей). Це свідчить про здатність механізму самоуваги Self-Attention вловлювати найтонші просодичні нюанси та контекст, які залишилися недоступними для візуального аналізу спектрограм.

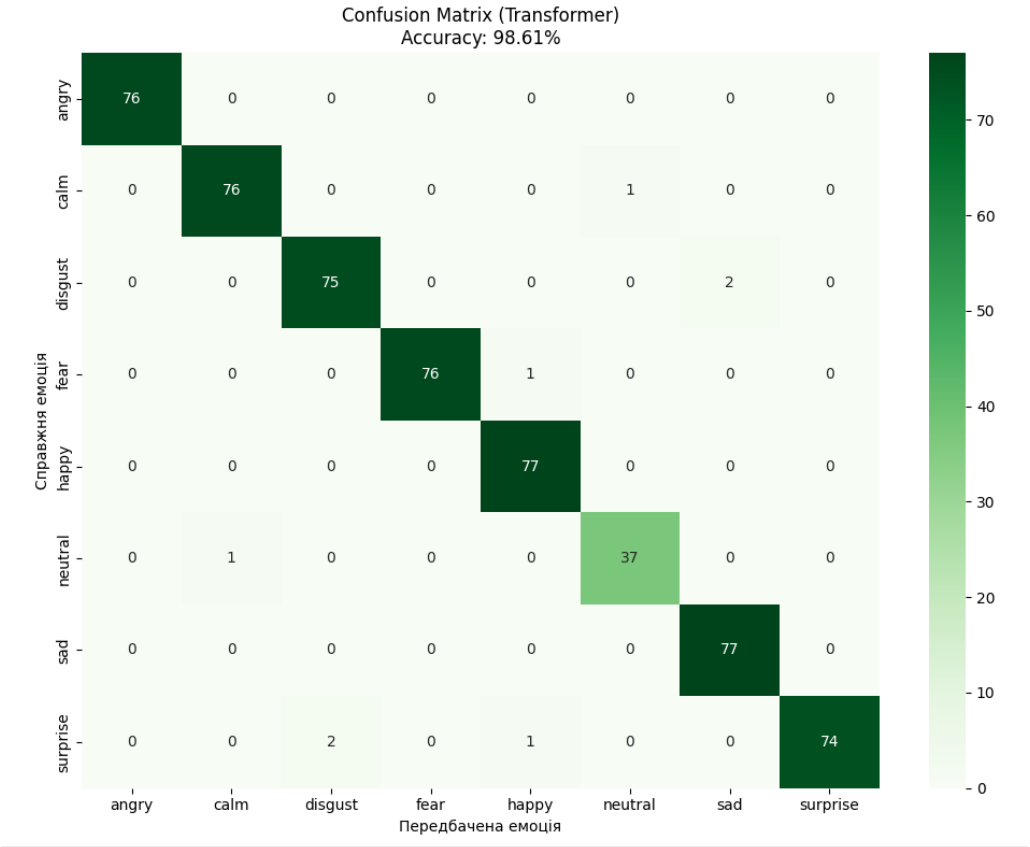


Рисунок 4.6 – Матриця плутанини моделі Wav2Vec

Отримані результати підтверджують гіпотезу про перевагу трансферного навчання для задач аудіоаналізу на малих датасетах та створюють надійну основу для проведення крос-мовних експериментів.

4.3 Формування контрольної вибірки україномовних даних

Важливим етапом експериментального дослідження була перевірка здатності розроблених систем працювати в умовах зміни мови. Оскільки моделі навчалися виключно на англомовному матеріалі (RAVDESS), необхідно було з’ясувати, чи зберігаються акустичні патерни емоцій при переході на іншу мову. Враховуючи відсутність у вільному доступі стандартизованих та валідованих наборів даних емоційного українського мовлення, було прийнято рішення про формування власної контрольної вибірки.

4.3.1 Методика відбору та структура даних

Джерелами аудіоматеріалів виступили фрагменти з відкритих медіа-ресурсів, що забезпечило різноманітність та наблизило умови тестування до реальних. Зокрема, використовувалися уривки з професійно озвучених аудіокниг та радіовистав, фрагменти емоційних висловлювань з подкастів, а також побутові записи, зроблені на непрофесійне обладнання. Такий підхід дозволив охопити різні типи запису: від студійного до побутового.

Процедура підготовки даних полягала у ручному моніторингу аудіопотоку та виокремленні сегментів із яскраво вираженими емоціями. З довгих аудіодоріжок вирізалися цільові фрагменти тривалістю від 2 до 4 секунд. Цей часовий проміжок обрано тому, що він відповідає формату, на якому навчалися моделі (3 секунди), і є достатнім для того, щоб людина встигла виразити закінчену емоційну фразу, але не настільки довгим, щоб емоція встигла змінитися на протилежну. Назва кожного аудіозапису містить англомовний ідентифікатор відповідної емоції та порядковий номер.

У результаті відбору було сформовано збалансований тестовий набір із 14 файлів, який репрезентує основні класи емоцій, використані при навчанні. Зокрема, до набору увійшли по два зразки для емоцій гніву, страху, здивування, суму та нейтрального стану. Для емоції щастя, яка є однією з найскладніших для акустичної ідентифікації, було відібрано три файли, а для стану спокою – один зразок.

Проте необхідно зазначити, що, на відміну від датасету RAVDESS, де кожен запис перевірявся сотнями асесорів, розмітка емоцій у створеній українській вибірці базується на експертній оцінці автора роботи. Однак в рамках магістерського дослідження такий підхід є допустимим для проведення якісного оцінювання здатності моделей до крос-мовного перенесення знань.

4.3.2 Попередня обробка українського контенту

Для забезпечення коректності порівняння, усі файли контрольної вибірки пройшли той самий технологічний етап попередньої обробки, що і тренувальні дані. Першим кроком стала конвертація різнорідних форматів (mp3, m4a) у формат .wav. Наступним етапом було виконано ресемплінг: для тестування згорткової мережі файли приводилися до частоти 22 050 Гц, а для трансформера – до 16 000 Гц.

Особлива увага приділялася нормалізації амплітуди. Оскільки джерела запису мали різний рівень гучності, це могло призвести до хибної класифікації, де тихі звуки автоматично відносилися б до суму чи спокою. Застосування пікової нормалізації дозволило вирівняти рівні потужності сигналів, змушуючи моделі орієнтуватися на інтонаційні характеристики, а не на абсолютну гучність. Підготовлені таким чином дані стали основою для проведення фінального порівняльного експерименту.

4.4 Порівняльний аналіз ефективності моделей у крос-мовному середовищі

Фінальним етапом експериментального дослідження стала перевірка навчених моделей на сформованому наборі українського мовлення. Метою даного експерименту було з’ясувати, наскільки добре працюють розроблені системи, і визначити, яка з двох архітектур, CNN чи Transformer, краще ідентифікує емоції при зміні мови аудіоданих.

4.4.1 Програмна реалізація алгоритму тестування

Для проведення порівняльного аналізу було розроблено скрипт інференсу, який автоматизує процес завантаження аудіофайлів, їх попередню обробку та отримання прогнозів від обох моделей.

Алгоритм працює наступним чином: скрипт сканує директорію з тестовими файлами, зчитує справжню емоційну мітку з назви файлу та передає аудіосигнал на вхід моделям.

Важливою особливістю реалізації є те, що для кожної моделі застосовується свій специфічний метод підготовки даних, ідентичний тому, що використовувався при навчанні.

Для моделі 2D CNN аудіосигнал спочатку нормалізується за гучністю, а потім генерується спектрограма розміром 128×130 пікселів. Отриманий двовимірний масив розгортається у 4D-тензор, який очікує модель Keras. Програмна реалізація функції передбачення наведена у додатку С.

Для моделі Wav2Vec сигнал нормалізується, приводиться до частоти 16 кГц та токенізується за допомогою Wav2Vec2Processor, який перетворює звукову хвилю на тензор значень. Цей тензор подається безпосередньо на вхід моделі. Фрагмент програмного коду наведено у додатку Д.

Обидві функції повертають не лише клас емоції, а й рівень впевненості, що дозволяє оцінити сумніви моделі у складних випадках.

4.4.2 Результати тестування візуальної моделі

Результати інференсу згорткової нейронної мережі на українських даних виявилися незадовільними (рис. 4.7) та підтвердили гіпотезу про низьку стійкість візуальних методів до зміни мови. Із 14 тестових зразків модель змогла коректно класифікувати лише 2 файли (спокій та сум), що відповідає точності 14.3%.

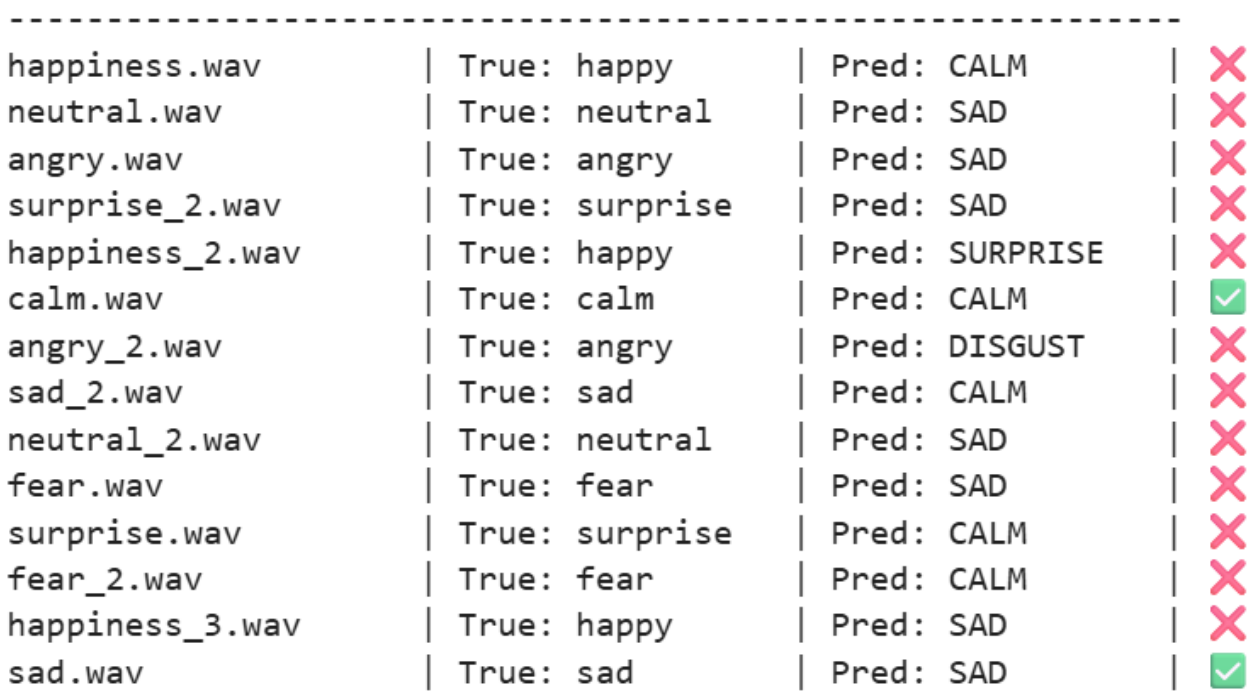


Рисунок 4.7 – Результати роботи 2D CNN на українських даних

Детальний аналіз протоколу тестування виявив ефект «колапсу передбачень». Незалежно від реальної емоції диктора, модель у переважній більшості випадків (10 з 14) класифікувала запис як сум або спокій. Наприклад, високоенергетичні емоції, такі як гнів, страх та здивування, які акустично характеризуються гучністю та різкими змінами тону, були помилково розпізнані як сум.

Для з’ясування причин низької ефективності CNN на українських даних було проведено візуальний порівняльний аналіз вхідних даних (рис. 4.8). Верхній ряд зображень демонструє спектрограми з датасету RAVDESS. Вони характеризуються чіткими горизонтальними лініями та абсолютно чорним фоном, що свідчить про відсутність шумів. Нижній ряд – записи з української вибірки. На них чітко видно спектральний шум, що виділяється фіолетовим та помаранчевим фоном замість чорного.

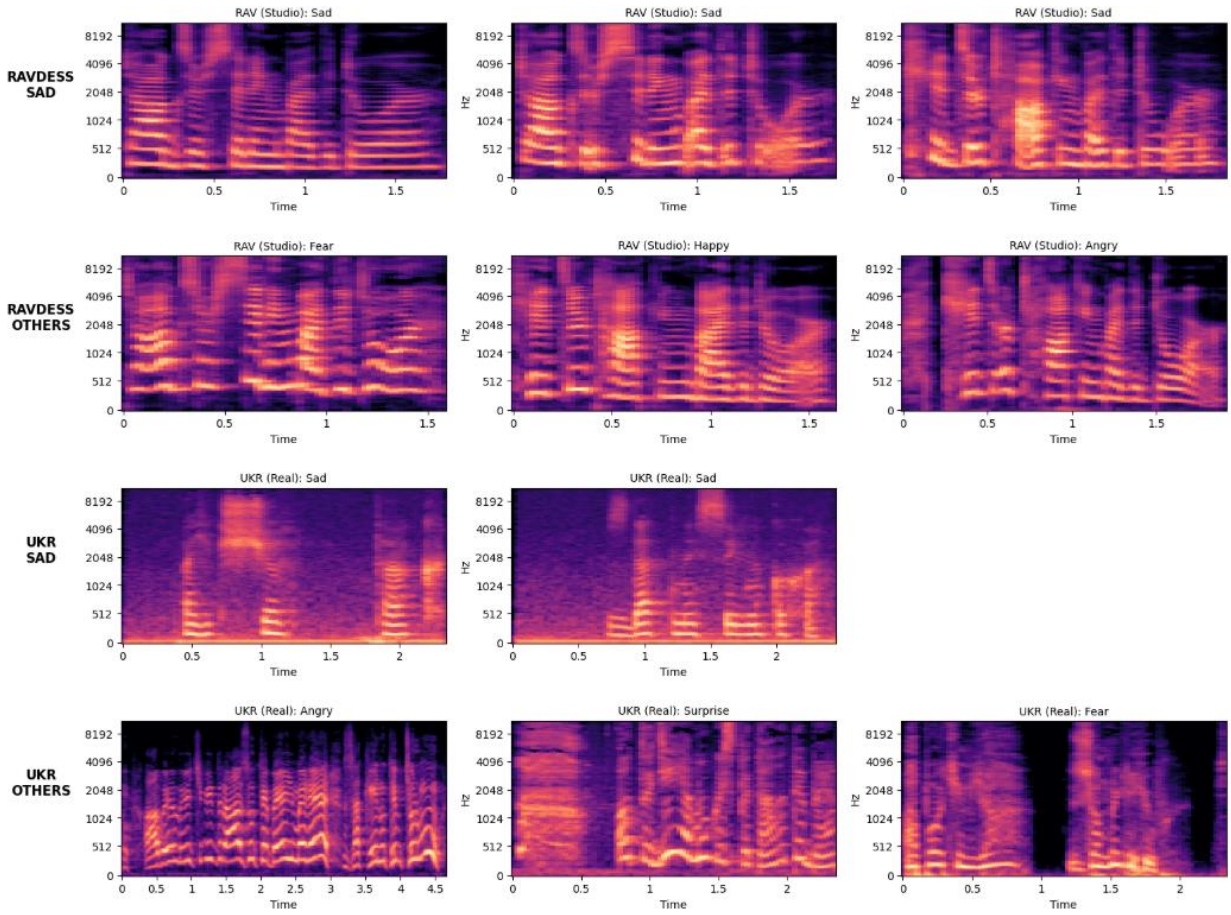


Рисунок 4.8 – Візуальне порівняння спектрограм

Модель, навчена на ідеально чистих студійних записах RAVDESS, інтерпретувала цей фоновий шум як корисний сигнал. Оскільки візуально суцільний шум нагадує текстуру емоцій низької інтенсивності, модель помилково класифікувала більшість файлів як сум.

Експеримент довів, що 2D CNN не здатна абстрагуватися від технічної якості запису і не підходить для використання у реальних умовах без попередньої адаптації до мікрофону користувача.

4.4.3 Результати тестування трансформерної моделі

Модель на базі архітектури Transformer досягла точності 50%, коректно класифікувавши 7 із 14 файлів. Враховуючи, що модель ніколи не навчалася на

українській мові, цей результат є значущим підтвердженням ефективності трансферного навчання для задач крос-мовного аналізу.

Детальний аналіз отриманих результатів (рис. 4.9) дозволяє виявити чітку закономірність у роботі алгоритму. Зокрема, для класів страх та здивування було досягнуто 100% точності розпізнавання. Показники для цих передбачень сягали 99%, що свідчить про те, що механізм самоуваги успішно виділяє специфічні просодичні маркери, такі як характерне тремтіння голосу (jitter) при страху або різкі інтонаційні стрибки при здивуванні незалежно від мови, якою говорить диктор. Також модель успішно ідентифікувала агресію у голосі, класифікувавши його вірно з упевненістю понад 99%.

ФАЙЛ	СПРАВЖНЯ	ПРОГНОЗ	ВПЕВНЕНІСТЬ	СТАТУС
happiness.wav	happy	CALM	87.2%	❌
neutral.wav	neutral	HAPPY	97.6%	❌
angry.wav	angry	HAPPY	89.9%	❌
surprise_2.wav	surprise	SURPRISE	92.8%	✅
happiness_2.wav	happy	HAPPY	64.7%	✅
calm.wav	calm	CALM	67.8%	✅
angry_2.wav	angry	ANGRY	99.2%	✅
sad_2.wav	sad	CALM	69.9%	❌
neutral_2.wav	neutral	ANGRY	99.1%	❌
fear.wav	fear	FEAR	99.4%	✅
surprise.wav	surprise	SURPRISE	79.9%	✅
fear_2.wav	fear	FEAR	99.3%	✅
happiness_3.wav	happy	SAD	89.0%	❌
sad.wav	sad	CALM	99.4%	❌

Рисунок 4.9 – Результати тестування моделі Wav2Vec на українських даних

Водночас, експеримент виявив і слабкі сторони моделі, пов'язані з відсутністю розуміння лінгвістичного контексту. Значна частина помилок зосереджена у площині розрізнення акустично схожих низькоенергетичних станів. Наприклад, обидва тестові зразки емоції сум були помилково класифіковані як спокій. Це є цілком очікуваним результатом, оскільки в українській мові ці стани характеризуються схожим темпоритмом (тихий, повільний голос), а смислове навантаження слів, яке могло б допомогти розрізнити сумну історію від спокійної розповіді, неможливо розрізнити за 3 секунди.

Окрім того, зафіксовано помилки у визначенні валентності емоцій для нейтрального стану. В умовах наявності фонових шумів модель схильна плутати нейтральні висловлювання з активними емоціями, такими як радість або гнів, орієнтуючись, ймовірно, на загальну енергетику запису, а не на інтонаційні нюанси. Тим не менш, досягнутий результат у 50% на незнайомій мові є набагато вищим за випадкове вгадування і свідчить про високий потенціал трансформерів для створення мультимовних інтерфейсів.

4.5 Висновки до розділу 4

У даному розділі виконано програмну реалізацію та проведено експериментальне дослідження двох систем розпізнавання емоційного стану, побудованих на принципово відмінних архітектурах: згортковій нейронній мережі, що здійснює візуальний аналіз спектрограм, та моделі Wav2Vec 2.0, яка базується на технології трансферного навчання. Основний фокус дослідження був зосереджений на порівнянні ефективності цих підходів як у контрольованих умовах навчання, так і в умовах зміни мовного та акустичного середовища.

Результати тестування на валідаційній вибірці англомовного датасету RAVDESS засвідчили високу ефективність обох моделей в умовах студійного запису. Згорткова мережа досягла точності 97.22%, а трансформер – 98.61%, що підтверджує можливість використання візуального підходу для задач, де акустичні умови є стабільними та передбачуваними. Проте перехід до крос-мовного експерименту на створеному наборі українського мовлення виявив критичну вразливість простих архітектур. Модель 2D CNN продемонструвала повну нездатність до узагальнення нових даних, знизивши точність до рівня випадкового вгадування у 14.3%. Аналіз помилок показав, що мережа інтерпретувала фоновий шум побутових записів та відмінності української фонетики як ознаки низькоенергетичних емоцій, що призвело до систематичної класифікації більшості зразків як сум або спокій.

Натомість використання попередньо навченої моделі Wav2Vec 2.0 підтвердило гіпотезу про ефективність трансферного навчання для подолання бар'єрів між різними мовними доменами. Трансформер досяг точності 50.0% на незнайомій мові, продемонструвавши здатність ідентифікувати універсальні просодичні маркери, такі як тремтіння голосу при страху або різкі інтонаційні зміни при здивуванні. Хоча модель допустила певні помилки у розрізненні валентності активних емоцій через відсутність розуміння контексту, її стійкість до шумів та здатність виділяти ознаки роблять цю архітектуру однозначним вибором для побудови сучасних мультимовних інтерфейсів.

ВИСНОВКИ

У даній роботі було вирішено завдання створення системи машинного слуху, здатної розпізнавати психоемоційний стан людини. Головною метою дослідження було не просто отримати високі цифри точності в лабораторних умовах, а з’ясувати, яка технологія краще підходить для реального життя, особливо коли йдеться про перехід на іншу мову.

На теоретичному етапі було визначено, що для комп’ютерного аналізу найкраще підходить модель базових емоцій. Однак головною проблемою стала природа звуку: необроблені аудіозаписи містять забагато шуму, тому для успішного навчання нейромереж їх необхідно ретельно обробляти: вирівнювати гучність та перетворювати на спектрограми або вектори ознак.

Основним етапом роботи став експеримент, у якому було порівняно два принципово різні підходи. Перший – це спроба навчити комп’ютер «бачити» емоцію на спектрограмі за допомогою згорткової мережі (CNN). Другий – використання великої мовної моделі (Wav2Vec 2.0), яка вже попередньо прослухала тисячі годин людського мовлення.

Результати виявилися показовими. На ідеально чистих, студійних записах англійською мовою обидві системи спрацювали блискуче, досягнувши точності понад 97%. Це довело, що в стерильних умовах навіть проста модель здатна чудово розпізнавати емоції. Проте справжній тест відбувся на створеному наборі українського мовлення, записаного в побутових умовах. Тут модель CNN, не впізнавши знайомих картинок звуку через шуми та іншу фонетику, почала плутати всі емоції з сумом, знизивши свою точність до рівня випадкового вгадування (14%).

Натомість модель на базі трансформера підтвердила свою перевагу. Навіть не знаючи української мови, вона змогла вловити універсальні інтонації страху та здивування, досягнувши 50% точності без жодного додаткового навчання. Отримані результати доводять, що подальший розвиток систем машинного

слуху має базуватися на технологіях трансферного навчання, які забезпечують формування стійких, мовно-незалежних ознак.

ПЕРЕЛІК ПОСИЛАНЬ

1. T. M. Wani A Comprehensive Review of Speech Emotion Recognition Systems. *IEEE Access*. 2021. Vol. 9. P. 47795–47814. URL : https://doi.org/10.1109/access.2021.3068045 (дата звернення : 20.07.2025).

2. Al-Dujaili M. J., Ebrahimi-Moghadam A. Speech Emotion Recognition : A Comprehensive Survey. *Wireless Personal Communications*. 2023. URL : https://doi.org/10.1007/s11277-023-10244-3 (date of access: 20.07.2025).

3. Yoon T.-J. How Much Does the Dynamic F0 Curve Affect the Expression of Emotion in Utterances? *Applied Sciences*. 2024. Vol. 14, no. 23. P. 10972. URL : https://doi.org/10.3390/app142310972 (date of access : 20.07.2025).

4. Vidisha R. Y. V. P. P. G. Next-Generation Virtual Assistants. *A Secure and Creative AI System*. 2025. URL: https://doi.org/10.5281/zenodo.15390420 (дата звернення: 22.07.2025).

5. Affectiva – Humanizing Technology with Emotion AI. Affectiva. URL: https://www.affectiva.com/ (дата звернення: 01.08.2025).

6. Introducing Hume’s Empathic Voice Interface (EVI) API, the first conversational voice AI with emotional intelligence. Hume AI. Home. URL : https://www.hume.ai/blog/introducing-hume-evi-api (date of access : 10.12.2025).

7. Müller M. Fundamentals of Music Processing : Audio, Analysis, Algorithms, Applications. Springer, 2016. 516 p.

8. Proakis J. G., Manolakis D. K. Digital Signal Processing. Pearson Education, Limited, Fourth Edition, 2014. 1019 p.

9. Y. Li Speech emotion recognition based on Graph-LSTM neural network. *EURASIP Journal on Audio, Speech, and Music Processing*. 2023. Vol. 2023, no. 1. URL : https://doi.org/10.1186/s13636-023-00303-9 (date of access : 11.08.2025).

10. Y. Li Speech emotion recognition based on Graph-LSTM neural network. *EURASIP Journal on Audio, Speech, and Music Processing*. 2023. Vol. 2023, no. 1. URL : https://doi.org/10.1186/s13636-023-00303-9 (дата звернення : 12.08.2025).

11. Abimbola J., Kostrzewa D., Kasprowski P. Time Signature Detection: A Survey. *Sensors*. 2021. Vol. 21, no. 19. P. 6494. URL : https://doi.org/10.3390/s21196494 (дата звернення : 13.08.2025).

12. Sarkar, S., & Ghosh, P. K. (2022). Unsupervised Voice Activity Detection by Modeling Source and System Information using Zero Frequency Filtering. *Interspeech*, 2022, 3218–3222. Doi : 10.21437/Interspeech.2022-10903.

13. D. G. Prathibha Speech Emotion Recognition Using Deep Learning. *Interantional journal of scientific research in engineering and management*. 2024. Vol. 08, no. 07. P. 1–13. URL : https://doi.org/10.55041/ijsrem36262 (дата звернення : 20.08.2025).

14. F. Burkhardt A database of German emotional speech. *Interspeech 2005*. ISCA, 2005. URL : https://doi.org/10.21437/interspeech.2005-446 (дата звернення : 25.08.2025).

15. Livingstone S. R., Russo F. A. The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS). *A dynamic, multimodal set of facial and vocal expressions in North American English. PLOS ONE*. 2018. Vol. 13, no. 5. P. e0196391. URL : https://doi.org/10.1371/journal.pone.0196391 (дата звернення : 02.09.2025).

16. Rusk N. Deep learning. *Nature Methods*. 2015. Vol. 13, no. 1. P. 35. URL : https://doi.org/10.1038/nmeth.3707 (дата звернення : 04.09.2025).

17. O. Abdel-Hamid Convolutional Neural Networks for Speech Recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*. 2014. Vol. 22, no. 10. P. 1533–1545. URL : https://doi.org/10.1109/taslp.2014.2339736 (дата звернення : 04.09.2025).

18. Anvarjon T., Mustaqeem, Kwon S. Deep-Net : A Lightweight CNN-Based Speech Emotion Recognition System Using Deep Frequency Features. *Sensors*. 2020. Vol. 20, no. 18. P. 5212. URL : https://doi.org/10.3390/s20185212 (дата звернення : 04.09.2025).

19. Zhihong Zeng A Survey of Affect Recognition Methods: Audio, Visual, and Spontaneous Expressions. *IEEE Transactions on Pattern Analysis and Machine*

Intelligence. 2009. Vol. 31, no. 1. P. 39–58. URL : https://doi.org/10.1109/tpami.2008.52 (дата звернення : 08.09.2025).

20. M. Jiang Transformer Based Memory Network for Sentiment Analysis of Web Comments. *IEEE Access*. 2019. Vol. 7. P. 179942–179953. URL : https://doi.org/10.1109/access.2019.2957192 (дата звернення : 08.09.2025).

21. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 5998–6008.

22. A Self-Supervised Model for Speech Representation. *Interspeech 2021*. *ISCA*, 2021. URL : https://doi.org/10.21437/interspeech.2021-717 (дата звернення : 15.09.2025).

23. Shi Zhong, Li K., Rui Feng. Deep convolutional hamming ranking network for large scale image retrieval. *2014 11th World Congress on Intelligent Control and Automation (WCICA), Shenyang, China*, 29 June – 4 July 2014. 2014. URL : https://doi.org/10.1109/wcica.2014.7052856 (дата звернення : 20.09.2025).

24. T. Wolf Transformers : State-of-the-Art Natural Language Processing. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing : System Demonstrations, Online*. Stroudsburg, PA, USA, 2020. URL : https://doi.org/10.18653/v1/2020.emnlp-demos.6 (дата звернення : 10.10.2025).

ДОДАТОК А

Програмний код генерації Мел-спектрограм для згорткової нейронної мережі

```
def generate_spectrogram(path):
    y, sr = librosa.load(path, sr=22050, duration=3.0, offset=0.5)
    target_len = int(22050 * 3.0)

    if len(y) < target_len:
        y = np.pad(y, (0, target_len - len(y)), mode='constant')
    else:
        y = y[:target_len]

    mel_spec = librosa.feature.melspectrogram(
        y=y,
        sr=sr,
        n_mels=128,
        hop_length=512
    )

    mel_spec_db = librosa.power_to_db(mel_spec, ref=np.max)
    return mel_spec_db[..., np.newaxis]
```

ДОДАТОК Б

Програмний код попередньої обробки та токенізації даних для моделі Wav2Vec 2.0

```
def preprocess_function(examples):
    audio_arrays = []
    for path in examples['audio']:
        speech, _ = librosa.load(path, sr=16000, duration=3.0)

        target_len = int(16000 * 3.0)
        if len(speech) < target_len:
            speech = np.pad(speech, (0, target_len - len(speech)), 'constant')
        else:
            speech = speech[:target_len]
        audio_arrays.append(speech)

    return feature_extractor(
        audio_arrays,
        sampling_rate=16000,
        padding=True,
        max_length=target_len,
        truncation=True
    )
```

ДОДАТОК В

Програмна реалізація алгоритму класифікації емоцій моделлю 2D CNN

```
def predict_emotion_cnn(path, model):
    input_tensor = preprocess_cnn_inference(path)

    if input_tensor is not None:
        probs = model.predict(input_tensor, verbose=0)[0]

        pred_idx = np.argmax(probs)
        pred_label = CLASSES[pred_idx]
        confidence = probs[pred_idx]

        return pred_label, confidence
    return "error", 0.0
```

ДОДАТОК Г

Програмна реалізація алгоритму класифікації емоцій моделлю на основі архітектури Transformer

```
def predict_emotion_transformer(path, model, processor):
    speech, _ = librosa.load(path, sr=16000)
    speech = librosa.util.normalize(speech)

    inputs = processor(speech, sampling_rate=16000, return_tensors="pt", padding=True)

    inputs = {k: v.to(device) for k, v in inputs.items()}

    with torch.no_grad():
        logits = model(**inputs).logits

    scores = torch.nn.functional.softmax(logits, dim=1).cpu().numpy()[0]
    predicted_id = np.argmax(scores)

    return model.config.id2label[predicted_id], scores[predicted_id]
```