Міністерство освіти і науки України
Харківський національний університет радіоелектроніки

Факультет Центр післядипломної освіти
(повна назва)
Кафедра Штучного інтелекту
(повна назва)

КВАЛІФІКАЦІЙНА РОБОТА
Пояснювальна записка

рівень вищої освіти другий (магістерський)

Інформаційна технологія класифікації емоцій за аудіозаписами на основі нейронної мережі CNN
(тема)

Виконав:
здобувач другого року навчання,
групи СШІздм-24-1
Андрій КРОХМАЛЬ
(власне ім'я, прізвище)
Спеціальність 122 Комп'ютерні науки
(код і повна назва спеціальності)
Тип програми освітньо-наукова
(освітньо-професійна або освітньо-наукова)
Освітня програма Системи штучного інтелекту
(повна назва освітньої програми)
Керівник доц. Каріна СЕЛІВАНОВА
(посада, власне ім'я, прізвище)

Допускається до захисту

Завідувач кафедри ШІ _____ Лариса ЧАЛА
(підпис) (власне ім'я, прізвище)

2026 р.

Харківський національний університет радіоелектроніки

Факультет Центр післядипломної освіти

Кафедра Штучного інтелекту

Рівень вищої освіти другий (магістерський)

Спеціальність 122 Комп'ютерні науки
(код і повна назва)

Тип програми освітньо-наукова
(освітньо-професійна або освітньо-наукова)

Освітня програма Системи штучного інтелекту
(повна назва)

ЗАТВЕРДЖУЮ:
Зав. кафедри _____
(підпис)
«_____» _____ 20 ___ р.

ЗАВДАННЯ
НА КВАЛІФІКАЦІЙНУ РОБОТУ

здобувачеві Крохмалю Андрію Віталійовичу
(прізвище, ім'я, по батькові)

1. Тема роботи Інформаційна технологія класифікації емоцій за аудіозаписами на основі нейронної мережі CNN

затверджена наказом університету від 6 квітня 20 26 р. № 48Стз

2. Термін подання здобувачем роботи до екзаменаційної комісії 19 травня 20 26 р.

3. Вихідні дані до роботи Формат ознак: спектрограма. Набори даних емоційного мовлення: CREMA-D, RAVDESS, Surrey Audio-Visual Expressed Emotion Database, TESS, Urdu Language Speech Dataset, Berlin Database of Emotional Speech, RESD, Mexican Emotional Speech Database, Speech Noise Dataset. Архітектури ШНМ: класична CNN, MobileNet V1 та V2, DenseNet, ResNet, ShuffleNet V1 та V2, SqueezeNet, SE-Net, RepVGG, ConvMixer. Метрики оцінки моделей: accuracy, precision, recall, F1, confusion matrix. Максимальна кількість параметрів: 0,1 млн. Вимоги до часу обробки: режим реального часу.

4. Перелік питань, що потрібно опрацювати в роботі
1) Сучасні підходи до аналізу акустичних сигналів у машинному навчанні
2) Розробка вдосконаленого методу класифікації емоцій за аудіоданими
3) Реалізація інформаційної технології класифікації емоцій у форматі вбудованої системи

КАЛЕНДАРНИЙ ПЛАН

№	Назва етапів роботи	Строк / термін виконання етапів роботи	Примітка
1	Отримання завдання на кваліфікаційну роботу	06.04.2026	виконано
2	Укладання і погодження з керівником плану і	08.04.2026	виконано
	етапів виконання роботи		
3	Узагальнення даних літературних джерел	15.04.2026	виконано
4	Аналіз шляхів виконання завдання. Вибір і	17.04.2026	виконано
	погодження з керівником оптимального шляху		
	виконання завдання		
5	Реалізація практичної частини завдання	22.04.2026	виконано
6	Виконання експериментальних досліджень	01.05.2026	
7	Укладання, оформлення та погодження	13.05.2026	виконано
	пояснювальної записки з керівником		
8	Підготовка слайдів презентації та доповіді	15.05.2026	виконано
9	Надання пояснювальної записки на кафедру	19.05.2026	виконано

Дата видачі завдання 6 квітня 2026 р.

Здобувач _____
(підпис)

Керівник роботи _____ доц. Каріна СЕЛІВАНОВА
(підпис) (посада, власне ім’я, прізвище)

РЕФЕРАТ

Пояснювальна записка: 206 с., 28 рис., 5 табл., 6 дод., 58 джерел.

ВБУДОВАНІ СИСТЕМИ, ЗГОРТКОВА НЕЙРОННА МЕРЕЖА, КЛАСИФІКАЦІЯ АУДІОСИГНАЛІВ, ЛОГ-МЕЛ СПЕКТРОГРАМА, ПОСТТРЕНУВАЛЬНЕ КВАНТУВАННЯ, РОЗПІЗНАВАННЯ ЕМОЦІЙНИХ СТАНІВ, ESP32.

Об’єкт дослідження – методи цифрового кодування, виділення ознак та класифікації акустичних сигналів.

Предмет дослідження – метод та інформаційна технологія розпізнавання емоцій за голосом людини на основі глибинного навчання для Edge AI систем.

Мета роботи – вдосконалення процесу автоматичного аналізу емоційного стану людини на основі аудіоданих на пристроях з обмеженими обчислювальними ресурсами.

Методи дослідження – загальнонаукові (аналіз і синтез, порівняння, абстрагування, моделювання); конкретнонаукові (методи обробки сигналів, виділення ознак, архітектурне та алгоритмічне моделювання, методи обробки «великих даних», методи машинного навчання, експеримент), елементи теорії множин і розпізнавання образів.

В роботі проаналізовано принципи цифрового кодування аудіоданих, методи виділення ознак та класифікації для задачі розпізнавання емоційних станів за голосом людини. Розроблено удосконалений метод розпізнавання емоцій з використанням спектрограм та глибоких згорткових мереж. Сформовано експериментальний набір даних, проведено апробацію модифікованих моделей на основі популярних згорткових архітектур. Для найбільш ефективної моделі здійснено додаткову архітектурну оптимізацію та квантування з огляду на використання в Edge AI пристроях.

ABSTRACT

Master’s thesis contains: 206 pp., 28 fig., 5 tabl., 6 ann., 58 references.

AUDIO SIGNAL CLASSIFICATION, CONVOLUTIONAL NEURAL NETWORK, EMBEDDED SYSTEMS, EMOTION RECOGNITION, ESP32, LOG-MEL SPECTROGRAM, POST-TRAINING QUANTIZATION.

Object of the study – techniques for digital encoding, feature extraction, and classification of acoustic signals.

Subject of the study – a deep-learning-based method and an information technology for speech emotion recognition in Edge AI systems.

Aim of the study – to enhance automated assessment of human emotional state from audio data on resource-constrained devices.

Research methods – general: analysis and synthesis, comparison, abstraction, and modeling; domain-specific: signal processing, feature extraction, architectural and algorithmic modeling, big data processing, machine learning, experimentation, and elements of set theory and pattern recognition.

This study analyzes the principles of digital encoding of audio data, methods for acoustic feature extraction, and classification approaches for speech-based recognition of human emotional states. An improved emotion recognition method is proposed based on spectrogram representations and deep convolutional neural networks, enabling the classification of six emotional states and the detection of voice absence (non-speech). During the development of the method, an experimental dataset was compiled and modified models derived from widely used convolutional architectures were evaluated. For the most effective model, additional architectural optimization and quantization were performed regarding deployment on Edge AI devices.

ЗМІСТ

Перелік умовних позначень, символів, одиниць, скорочень і термінів 9
Вступ 11
1 Сучасні підходи до аналізу акустичних сигналів у машинному навчанні 20
1.1 Методи виділення ознак з аудіоданих 23
1.1.1 Швидкість переходу через нуль 24
1.1.2 Енергія сигналу 25
1.1.3 Енергія смуги 26
1.1.4 Спектральний потік 27
1.1.5 Спектральний спад 28
1.1.6 Спектральний центроїд 28
1.1.7 Спектрограма 29
1.1.8 Кепстральні ознаки 32
1.1.9 Коефіцієнти мел-частотного кепструму 33
1.2 Методи класифікації акустичних сигналів 34
1.2.1 Приховані Марківські моделі 36
1.2.2 Алгоритм випадкового лісу 38
1.2.3 Багатошаровий персептрон 39
1.2.4 Згорткова нейронна мережа 41
1.2.5 Рекурентні нейронні мережі 44
1.2.6 Трансформерні моделі 46
1.2.7 Гібридні моделі 48
Висновки до розділу 1 51
2 Розробка вдосконаленого методу класифікації емоцій за аудіоданими 53
2.1 Розробка методу виділення ознак 54
2.1.1 Розбиття аудіо на фрагменти 54
2.1.2 Нормалізація аудіосигналу 54
2.1.3 Пре-емпфазис 55

2.1.4 Застосування віконної функції 56
2.1.5 Обчислення спектру потужності 57
2.1.6 Застосування мел-фільтру 57
2.1.7 Перетворення у децибели 59
2.1.8 Об’єднання фрагментів спектру 60
2.2 Формування експериментального набору даних 61
2.2.1 CREMA-D 61
2.2.2 RAVDESS 62
2.2.3 SAVEE 62
2.2.4 TESS 63
2.2.5 Urdu Language Speech Dataset 64
2.2.6 EMODB 64
2.2.7 RESD 65
2.2.8 MESD 65
2.2.9 Speech Noise Dataset 66
2.3 Вибір архітектури штучної нейронної мережі 72
2.3.1 Класична CNN 73
2.3.2 MobileNet-подібні архітектури 73
2.3.3 DenseNet-подібна архітектура 74
2.3.4 ResNet-подібна архітектура 75
2.3.5 ShuffleNet-подібні моделі 76
2.3.6 SqueezeNet-подібна архітектура 77
2.3.7 SE-Net 78
2.3.8 RepVGG-подібна архітектура 79
2.3.9 Архітектура на основі ConvMixer 79
Висновки до розділу 2 87
3 Реалізація інформаційної технології класифікації емоцій у форматі вбудованої системи 89
3.1 Вибір апаратних та програмних засобів 89
3.1.1 Апаратні засоби 89

3.1.2 Програмні засоби 93
3.2 Оптимізація нейронної мережі для цільової апаратної платформи... 98
3.2.1 Оптимізація архітектури 98
3.2.2 Квантування моделі 101
3.3 Апаратно-програмна реалізація розробленої інформаційної технології 103
Висновки до розділу 3 112
Висновки 114
Перелік джерел посилання 116
Додаток А Розрахунки обчислювальної та об’ємної складності алгоритму виділення ознак для голосового сигналу 124
Додаток Б Програмний код 129
Додаток В Архітектура моделей ШНМ 176
Додаток Г Графіки скалярних метрик протягом навчання та валідації 193
Додаток Д Візуалізація матриць помилок, отриманих при тестуванні моделей 204
Додаток Е Відомість кваліфікаційної роботи 206

ПЕРЕЛІК УМОВНИХ ПОЗНАЧЕНЬ, СИМВОЛІВ, ОДИНИЦЬ, СКОРОЧЕНЬ І ТЕРМІНІВ

МК – мікроконтролер;

ОС – операційна система;

ШІ – штучний інтелект;

ШНМ – штучна нейронна мережа;

AGC – Automatic Gain Control – система автоматичного керування підсиленням;

CNN – Convolutional Neural Network – згорткова нейронна мережа;

DCT – Discrete Cosine Transform – дискретне косинусне перетворення;

DMA – Direct Memory Access – прямий доступ до пам'яті – режим обміну даними без участі центрального процесора.

DSP – Digital Signal Processing – цифрова обробка сигналів;

FFT – Fast Fourier Transform – швидке перетворення Фур’є;

FPU – Floating Point Unit – арифметичний співпроцесор – апаратний блок для прискорення операцій над числами з плаваючою комою;

GRU – Gated Recurrent Unit – керований рекурентний елемент;

HMM –Hidden Markov Model – прихована марковська модель;

ISA – Instruction Set Architecture – архітектура системи команд;

LSTM – Long Short-Term Memory – нейронна мережа з довгою короткочасною пам’яттю;

MFCC – Mel-Frequency Cepstral Coefficients – мел-частотні кепстральні коефіцієнти;

NPU – Neural Processing Unit – нейронний процесор – апаратний блок для прискорення нейромережевих обчислень;

PCM – Pulse Code Modulation – імпульсно-кодова модуляція – метод цифрового представлення аналогових сигналів;

PDM – Pulse Density Modulation – імпульсно-частотна модуляція – метод цифрового представлення аналогових сигналів;

PTQ – Post-Training Quantization – посттренувальне квантування;

QAT – Quantization-Aware Training – навчання з урахуванням квантування;

RF – Random Forest – випадковий ліс – ансамблевий метод машинного навчання;

RNN – Recurrent Neural Network – рекурентна нейронна мережа;

SER – Speech Emotion Recognition – виявлення емоційних станів за мовленням;

SIMD – Single Instruction, Multiple Data – «одна інструкція – багато даних» – тип паралельних обчислень, де одна команда застосовується одночасно до кількох елементів даних;

SoC – System-on-a-chip – система на кристалі – інтегральна мікросхема, що поєднує процесор, пам’ять, контролери та інші компоненти в одному кристалі;

SRAM – Static Random Access Memory – статичний оперативний запам’ятовуючий пристрій;

VAD – Voice Activity Detection – виявлення голосової активності;

ViT – Vision Transformer – зоровий трансформер;

ВСТУП

Розпізнавання емоційних станів людини залишається складною і недостатньо вирішеною задачею, що привертає увагу дослідників у галузях когнітивних наук, психофізіології та штучного інтелекту (ШІ). Методи автоматичного розпізнавання емоційних станів людини знаходять дедалі ширше застосування у багатьох галузях.

У медичній сфері технології емоційного аналізу використовуються насамперед у психіатрії та клінічній психології для підтримки діагностики афективних розладів, таких як депресія, генералізований тривожний розлад, біполярний розлад та посттравматичний стресовий розлад. Застосування подібних систем сприяє підвищенню об’єктивності оцінювання психічного стану пацієнтів, а також забезпечує можливість довготривалого моніторингу ефективності терапевтичних втручань. У неврологічній практиці системи розпізнавання емоцій застосовуються для раннього виявлення нейродегенеративних захворювань, зокрема хвороби Альцгеймера, хвороби Паркінсона та різних форм деменції [1]. Порушення емоційної експресії, зниження реактивності до афективних стимулів та зміни інтонаційної структури мовлення можуть виступати ранніми маркерами когнітивного зниження. Аналіз таких змін за допомогою систем розпізнавання емоційних станів дозволяє ідентифікувати патологічні процеси ще до появи виражених клінічних симптомів, що є критично важливим для своєчасного діагностування захворювань.

Важливим напрямом є застосування емоційного аналізу у реабілітаційній медицині. У пацієнтів після інсультів або черепно-мозкових травм часто спостерігаються труднощі з розпізнаванням та вираженням емоцій, що негативно впливає на їх соціальну адаптацію. Завдяки використанню автоматизованих систем оцінювання емоційного стану стає можливим відстеження динаміки відновлення когнітивних та соціально-комунікативних функцій, а також забезпечення інклюзивності

реабілітаційних програм [2]. Також подібні технології активно впроваджуються у телемедичних рішеннях для дистанційного моніторингу психоемоційного стану пацієнтів за допомогою носимих сенсорів та мобільних пристроїв.

У немедичній сфері методи емоційного аналізу відіграють значну роль у системах людино-машинної взаємодії. Інтелектуальні інтерфейси можуть адаптувати спосіб подання інформації залежно від емоційного стану користувача, змінюючи рівень складності завдань, темп комунікації або форму зворотного зв’язку. Зокрема, це може бути застосовано при побудові інтелектуальних чат-ботів з голосовим інтерфейсом для забезпечення більш природнього способу взаємодії та адаптації наданих відповідей з урахуванням емоційного стану та потреб користувача. Це особливо актуально для освітніх технологій, де автоматичне визначення рівня зацікавленості, фрустрації або когнітивного навантаження здобувача освіти дозволяє персоналізувати навчальний процес та підвищити його ефективність [3].

У сфері безпеки емоційний аналіз використовується для виявлення потенційно небезпечної поведінки у громадських місцях. Аналіз міміки, жестів і поведінкових патернів може сприяти ранньому виявленню агресії, паніки або стресу в натовпі, що дозволяє оперативно реагувати на можливі загрози. Аналогічні підходи застосовуються у транспортній сфері для моніторингу стану водія з метою виявлення сонливості, втоми або емоційного напруження, що може призвести до аварійних ситуацій.

Крім того, методи розпізнавання емоцій активно використовуються у маркетингових дослідженнях та управлінні персоналом. Аналіз емоційної реакції споживачів на надані послуги, рекламні матеріали або користувацькі інтерфейси дозволяє оптимізувати якість сервісу, дизайн продуктів і підвищити рівень задоволеності клієнтів [4]. У корпоративному середовищі такі технології можуть застосовуватися для оцінювання рівня стресу

працівників, виявлення професійного вигорання та покращення психологічного клімату в колективі.

Таким чином, сучасні методи емоційного аналізу та інтеграція сучасних методів та засобів штучного інтелекту з сенсорними системами формують основу для створення адаптивних інтелектуальних систем, здатних враховувати психоемоційний стан людини, що відкриває перспективи для покращень у сферах медицини, безпеки, освіти та людино-машинної взаємодії. Тому дослідження у галузі автоматичного емоційного аналізу на сьогоднішній день залишаються надзвичайно актуальними.

Сучасні методи та засоби емоційного аналізу пропонують різні підходи до розпізнавання емоцій, серед яких використання комплексу фізіологічних параметрів, зокрема, змін артеріального тиску, температури тіла, варіабельності серцевого ритму, електродермальної активності та електромагнітної активності мозку, що реєструється за допомогою електроенцефалографії. Використання фізіологічних сигналів у задачах емоційного аналізу забезпечує підвищену надійність результатів у порівнянні з методами, що базуються виключно на зовнішніх поведінкових ознаках, оскільки такі сигнали є менш залежними від свідомого контролю з боку досліджуваного, що є критичним для більшості медичних застосувань. Однак, більшість фізіологічних методів вимагає прямого контакту із людиною, а також досить складного та вартісного обладнання, що обмежує їх практичне застосування.

Альтернативним підходом є дистанційний аналіз емоційного стану на основі аудіальних або візуальних характеристик, зокрема за змінами у голосі людини. Такий підхід передбачає використання методів автоматичного аналізу мовленнєвого сигналу (англ. Speech Emotion Recognition, SER), що дозволяють виявляти емоційні стани на основі просодичних, спектральних та часових характеристик голосу, включаючи темп мовлення, інтонаційні особливості, варіації частоти основного тону та енергетичні параметри сигналу. У роботі [5] зазначається, що емоційні

реакції безпосередньо впливають на функціонування голосового апарату та дихальної системи, що зумовлює зміну акустичних характеристик мовлення, які можуть бути зафіксовані за допомогою мікрофона.

Очевидними перевагами даного підходу є безконтактний спосіб отримання даних, а також значно нижчі вимоги до застосованого обладнання, яке може бути виконане у форматі носимого пристрою. Метод аналізу емоцій з використанням аудіо як джерела даних не здатен забезпечити точність, характерну для мультимодальних методів, однак може бути скомбінованим з іншими джерелами інформації [6], або використаним у прикладних областях, де такі характеристики є прийнятними. До таких застосувань можна віднести покращення людино-машинних інтерфейсів при побудові голосових чат-ботів, систем автоматизованого моніторингу відгуків клієнтів call-центрів, освітні та маркетингові дослідження тощо.

Задача розпізнавання емоційних станів шляхом аналізу звукової інформації пов’язана з кількома ключовими викликами, що негативно впливають на достовірність виявлення. По-перше, аудіосигнал може бути спотворений фоновими шумами, що виникають у повсякденному середовищі: звуки техніки, голоси інших людей, акустичні викривлення приміщення, зовнішні перешкоди тощо. Це ускладнює правильне виділення ознак мовлення, що відповідають емоційному стану. Можливими рішеннями є додаткове використання методів шумозаглушення або застосування методів виявлення, стійких до впливу шумових факторів.

По-друге, емоційне забарвлення мовлення може значно змінюватися залежно від особистих характеристик мовця, його темпераменту, психофізіологічного стану та соціального контексту. Люди мають різні тембри голосу, різну манеру мовлення та інтонаційні особливості голосу, що ускладнює створення універсальної моделі розпізнавання емоцій. Для вирішення цієї проблеми необхідно формувати великі та різноманітні набори даних, а також розробляти стійкі адаптивні моделі. По-третє, рівень

емоційного вираження та його акустичні характеристики можуть суттєво відрізнятися залежно від культурного середовища та мовних традицій. Наприклад, у деяких мовах емоції передаються більш інтенсивно за рахунок тональної мови, тоді як в інших – завдяки ритмічним особливостям. Це потребує врахування мовних факторів при створенні універсальних моделей розпізнавання емоцій.

Окремо слід зазначити актуальність використання носимих інтелектуальних пристроїв, що реалізують концепцію обчислень на периферії мережі (Edge AI), для забезпечення безперервного моніторингу психофізіологічного стану людини в режимі реального часу без необхідності залучення хмарної інфраструктури. Інтеграція сучасних методів глибокого навчання у вбудовані системи дозволяє здійснювати локальний аналіз даних із мінімальними затримками передачі інформації та підвищеним рівнем захисту персональних даних. Водночас це створює обмеження щодо обчислювальних ресурсів, енергоспоживання та обсягу пам’яті носимих платформ, що зумовлює необхідність розробки ресурсоефективних моделей, методів та інформаційної технології розпізнавання емоційних станів.

Таким чином, успішна реалізація системи розпізнавання емоцій передбачає подолання низки технічних викликів. Важливим аспектом є використання вдосконалених методів обробки аудіосигналу, застосування просунутих підходів та ефективних алгоритмів машинного навчання, що забезпечують високу точність роботи системи в реальних умовах експлуатації та задовольняють апаратним обмеженням. Отже, розробка нових та вдосконалення існуючих методів розпізнавання емоційних станів за голосом людини є актуальною науково-технічною задачею.

Метою дослідження є вдосконалення процесу автоматичного аналізу емоційного стану людини на основі аудіоданих на пристроях з обмеженими обчислювальними ресурсами.

Для досягнення зазначеної мети було визначено наступні задачі дослідження:

– аналіз методів та засобів цифрового кодування звукових сигналів та їх основних характеристик;

– огляд існуючих методів виділення ознак для аудіоданих, аналіз їх придатності для використання в задачі SER;

– огляд методів класифікації акустичних сигналів, визначення особливостей їх застосування, переваг та недоліків;

– розробка методу виділення ознак для голосового акустичного сигналу;

– формування експериментального набору даних, створення навчальної та тестувальної вибірок;

– розробка архітектури штучної нейронної мережі для класифікації емоцій, оптимізація моделі для Edge AI пристрою;

– експериментальна оцінка ефективності розроблених нейромережевих моделей;

– вибір апаратних та програмних засобів для реалізації системи розпізнавання емоційних станів у форматі вбудованої системи;

– апаратно-програмна реалізація розробленої інформаційної технології.

Гіпотеза дослідження. Цифрове кодування аудіоданих із мінімальними параметрами, що гарантують розбірливість мовлення, повинно забезпечувати оптимальне використання пам’яті та обчислювальних ресурсів при збереженні інформативних характеристик голосу. Застосування ефективного методу виділення ознак для звукових сигналів та сучасних засобів глибокого навчання надасть ефективний засіб для здійснення класифікації. Якщо поєднати всі три фактори, отримаємо покращений метод розпізнавання емоційних станів за голосом людини, що має забезпечити достатню достовірність виявлення з одночасним

збереженням ефективності обчислень та придатності для використання в умовах реального часу на Edge AI пристроях.

Об’єктом дослідження є методи цифрового кодування, виділення ознак та класифікації акустичних сигналів.

Предметом дослідження є метод та інформаційна технологія розпізнавання емоцій за голосом людини на основі глибинного навчання для Edge AI систем.

Методологічна та теоретична основа дослідження. Для розробки методу розпізнавання емоційних станів використовувались теорія множин та теорія розпізнавання образів. Методи дослідження – загальнонаукові: аналіз і синтез, порівняння, абстрагування та моделювання; конкретнонаукові: загальні методи обробки сигналів, методи виділення інформативних даних, архітектурне та алгоритмічне моделювання, загальні методи обробки «великих даних» (масштабування вхідних даних, доповнення (аугментація), стратифікований поділ на вибірки), методи машинного навчання (регуляризація, оптимізація функції втрат, глибинне навчання, квантування), експеримент.

Для розробки було використано програмні засоби: платформа Python 3, фреймворки Pandas, Librosa, Scikit-Learn, PyTorch, засіб візуалізації TensorBoard. Апаратні засоби: графічний процесор з підтримкою технології Nvidia CUDA.

Наукова новизна результатів дослідження. Було отримано адаптовані для обробки спектрограм та використання на Edge AI пристроях реалізації поширених згорткових архітектур. Отримано моделі розміром до 1 мільйона параметрів: простого згорткового стеку з SE-модулями та без них, MobileNet V1 та V2, ResNet, DenseNet-BC, SqueezeNet v1.1, ShuffleNet V1 та V2, RepVGG та ConvMixer. Було виконано експериментальну оцінку ефективності модифікованих моделей в задачі класифікації емоцій за аудіоданими за кількома поширеними метриками. Запропоновано модифікацію моделі SqueezeNet, оптимізовану для використання на

пристроях класу вбудованих систем, що використовує розділені згортки та згортковий класифікатор.

Теоретична значущість дослідження полягає у дослідженні особливості визначення емоційного забарвлення мовлення на основі спектрограм та згорткових нейронних мереж. Вивчення методів виділення ознак з акустичного сигналу дозволяє забезпечити ефективне перетворення голосового сигналу у представлення, придатне для обробки засобами машинного навчання. Застосування спектрограм як вхідного простору ознак відкриває можливості для вилучення інформативних характеристик аудіосигналу з істотним зниженням розмірності даних. Це також дозволяє зменшити ризик перенавчання, підвищуючи загальну узагальнюючу здатність моделі. Теоретична цінність полягає також у дослідженні релевантних архітектур згорткових нейронних мереж, адаптованих до обмежених обчислювальних ресурсів. Зокрема, аналізуються підходи до побудови компактних нейронних моделей із зменшеною кількістю параметрів та обчислювальних операцій, що досягається шляхом оптимізації глибини мережі, використання згорток із малою розмірністю ядра, механізмів розділюваної згортки, а також заміни повнозв’язних шарів згортковими.

Практична значущість дослідження полягає у створенні ефективного інструменту для автоматизованого визначення емоційного стану людини за мовленням, що має широкий спектр застосувань у різних галузях. Розроблена інформаційна технологія може бути інтегрована в системи голосових асистентів, освітніх платформ, а також у засоби моніторингу психоемоційного стану в умовах підвищеного навантаження, зокрема, у комплексі з іншими джерелами даних. Розроблена інформаційна технологія здатна функціонувати як у середовищах з повноцінними обчислювальними ресурсами, так і на платформах з обмеженою обчислювальною потужністю, що дає можливість його використання у вбудованих системах, які включають носимі пристрої, пристрої Інтернету речей (IoT), портативну

медичну електроніку тощо. Також результати дослідження можуть бути використані для формалізації підходів до оцінювання інформативності ознак, що використовуються в задачах класифікації емоцій, і для подальшого вдосконалення моделей розпізнавання емоцій. Крім того, напрацьовані підходи до аналізу даних, виділення ознак та побудови компактних моделей можуть бути використані в інших прикладних задачах класифікації звукових сигналів.

Апробація результатів дослідження. Основні положення дослідження було висвітлено у доповідях на міжнародних наукових конференціях:

– «Інформаційні інтелектуальні системи» у рамках 30-го Міжнародного молодіжного форуму «Радіоелектроніка та молодь у XXI столітті» (22–24 квітня 2026 р., м. Харків);

– XXXIV Міжнародна науково-практична конференція «Інформаційні технології: наука, техніка, технологія, освіта, здоров’я. MicroCAD-2026» (13–16 травня 2026 р., м. Харків).

1 СУЧАСНІ ПІДХОДИ ДО АНАЛІЗУ АКУСТИЧНИХ СИГНАЛІВ У МАШИННОМУ НАВЧАННІ

Цифрове кодування аудіоданих є фундаментальним процесом у цифровій обробці звуку, що забезпечує збереження, передачу та відтворення акустичних сигналів у двійковому форматі. Даний процес передбачає перетворення аналогового звукового сигналу у цифрове представлення за допомогою аналого-цифрового перетворювача (АЦП) та забезпечує можливість його подальшого зберігання та обробки.

Залежно від способу представлення та зберігання аудіоданих розрізняють нестиснене та стиснене цифрове кодування. Методи стиснення без втрат забезпечують повне відновлення вихідного сигналу після декодування, тоді як з втратами – передбачають видалення малозначущих компонентів сигналу з метою зменшення обсягу даних.

Найбільш поширеним методом представлення нестисненого аудіосигналу є імпульсно-кодова модуляція PCM (Pulse Code Modulation), що передбачає представлення аналогового сигналу у вигляді послідовності числових значень, що отримуються шляхом дискретизації та квантування амплітуди сигналу у рівномірно розподілені моменти часу. У цьому випадку кожен семпл сигналу кодується певною кількістю бітів (наприклад, 16 або 24 біти), що забезпечує високу точність відтворення амплітудних значень. Такий підхід дозволяє безпосередньо виконувати операції цифрової обробки сигналів, включаючи фільтрацію та спектральний аналіз.

Популярність PCM зумовлена простотою реалізації, відсутністю втрат інформації під час кодування, високою точністю відтворення сигналу та можливістю ефективної інтеграції з алгоритмами цифрової обробки сигналів, що робить його базовим стандартом для зберігання та попередньої обробки аудіоданих у системах машинного навчання.

Інший формат аудіокодування, PDM (Pulse Density Modulation), використовує однобітове представлення сигналу, у якому інформація про

амплітуду передається не абсолютним значенням семплу, а щільністю імпульсів у певному часовому інтервалі. У PDM сигнал формується як високочастотна послідовність бітів, де збільшення щільності логічних одиниць відповідає зростанню амплітуди аналогового сигналу. Такий підхід дозволяє значно спростити апаратну реалізацію аналого-цифрового перетворювача, що обумовлює його широке використання у цифрових мікрофонах. Проте для подальшої цифрової обробки сигнал PDM, як правило, потребує попереднього перетворення у формат PCM шляхом виконання децимації та цифрової фільтрації.

У імпульсно-кодовому представленні нестиснений аудіосигнал зберігається у вигляді послідовності семплів, що відображають амплітуду звукової хвилі у дискретні моменти часу (рисунок 1.1). Аудіозапис може бути монофонічним (одноканальним), стереофонічним (двоканальним) або багатоканальним, причому в останніх двох випадках кожен канал містить незалежну послідовність семплів, що відповідає окремим звуковим потокам.

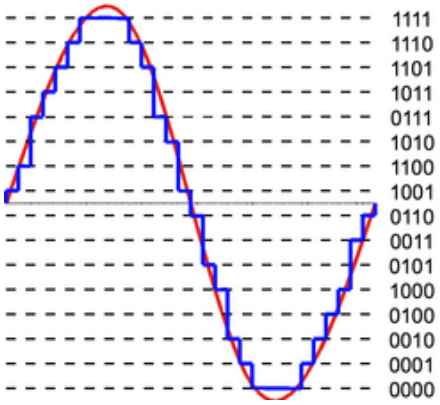


Рисунок 1.1 – Цифрове представлення аудіосигналу з використанням кодування PCM

До основних параметрів цифрового кодування належать частота дискретизації та розрядність. Розрядність визначає точність представлення амплітудних значень сигналу у цифровому форматі та задається кількістю

бітів, що використовуються для кодування одного семплу. Збільшення розрядності призводить до зростання кількості можливих рівнів квантування та зменшення похибки квантування. Зокрема, при 16-бітному представленні можливе кодування 65536 дискретних рівнів амплітуди, що забезпечує достатню динамічну роздільну здатність для більшості прикладних задач обробки аудіосигналів.

Частота дискретизації визначає кількість відліків (семплів), що реєструються за одиницю часу, і безпосередньо впливає на точність відтворення оригінального сигналу. Відповідно до теореми Котельникова, частота дискретизації повинна щонайменше вдвічі перевищувати максимальну частоту сигналу для забезпечення його коректного відновлення без виникнення ефекту накладання спектрів (аліазингу). Наприклад, для представлення звукового сигналу в діапазоні частот до 20 кГц необхідна частота дискретизації не менше 40 кГц. На практиці використовується значення 44,1 кГц, що відповідає стандарту Audio CD [7].

З огляду на реалізацію інформаційної технології розпізнавання емоцій у форматі вбудованих систем, у даному дослідженні було використано монофонічні аудіозаписи з частотою дискретизації 8 кГц. Такий формат зменшує обсяг даних, вимоги до пам’яті та обчислювальних ресурсів, що є важливим для реалізації SER у реальному часі на ресурсно-обмежених платформах. Відповідно до теореми Котельникова, частота дискретизації 8 кГц дає змогу коректно відтворювати спектральні компоненти до 4 кГц, тобто охоплювати значну частину інформативного діапазону людського мовлення, зокрема область, важливу для розбірливості та передачі емоційно значущих акустичних ознак. Такий частотний діапазон добре узгоджується з практикою використання у вузькосмуговій телефонії, де для забезпечення базової розбірливості мовлення традиційно використовується смуга близько 300–3400 Гц [8]. Критичні для розбірливості приголосні та значна частина ознак, пов’язаних із

артикуляційною чіткістю, зосереджені в області 2–4 кГц, тобто залишаються доступними при частоті дискретизації 8 кГц.

Результати дослідження [9] підтверджують практичну придатність такого підходу: зменшення частоти дискретизації з 16 до 8 кГц призвело лише до незначного зниження середньої точності SER – приблизно на 3,3%, що свідчить про збереження достатньої кількості релевантної емоційної інформації навіть у вузькосмуговому сигналі.

Отже, для вбудованих SER-систем використання такого формату вхідних даних є обґрунтованим компромісом, за якого незначна втрата точності виправдана істотним виграшем у швидкодії та енергоспоживанні.

1.1 Методи виділення ознак з аудіоданих

Цифровий аудіосигнал не є прийнятним для обробки засобами машинного навчання через високу надлишковість інформації, тому вимагає застосування методів виділення інформативних даних. Виділення ознак є важливим етапом у процесі аналізу даних, оскільки воно дозволяє виокремити найбільш значущу інформацію та встановити внутрішні зв’язки у даних. Зокрема, при роботі з необробленими аудіоданими моделі машинного навчання не можуть ефективно інтерпретувати та використовувати ці дані для аналізу або прийняття рішень. Тому застосовуються методи виділення ознак, які перетворюють складні сигнали у релевантне представлення для подальшого використання такими алгоритмами [10].

Процес виділення ознак спрямований на оптимізацію та узагальнення інформації, зберігаючи при цьому її найбільш суттєві аспекти. Правильний вибір методів перетворення сигналу впливає на якість аналізу, дозволяючи точніше визначати структуру даних та покращувати ефективність зазначених систем. У даній роботі розглянуто декілька найбільш поширених методів виділення ознак для акустичних сигналів.

Методи виділення ознак з аудіоданих можна умовно поділити на два блоки – ті, що формують ознаки часової області і частотної області. Перша група методів включає сукупність підходів, у яких характеристики мовлення обчислюються безпосередньо з часової форми сигналу, тобто з послідовності його амплітудних відліків у межах коротких вікон аналізу. Такі методи орієнтовані на виявлення часових закономірностей сигналу, зокрема його енергетичної структури, періодичності, темпу змін і переходів між різними типами звуків. Саме ці методи традиційно застосовують для сегментації мовлення, виявлення меж слів і фраз, а також для розмежування дзвінких, глухих і паузних ділянок сигналу. Перевагою методів часової області є їхня обчислювальна простота та придатність до реалізації в системах реального часу. Нижче наведено кілька поширених методів цієї групи.

1.1.1 Швидкість переходу через нуль

Швидкість переходу через нуль (англ. Zero-Crossing Rate, ZCR) є одним з фундаментальних параметрів аналізу аудіосигналів, що відображає частоту зміни знаку сигналу з позитивного на нульовий і на негативний, або навпаки – з негативного на нульовий і на позитивний. Ця характеристика відіграє важливу роль у визначенні тембральних властивостей звуку та знаходить широке застосування в різних областях обробки звукової інформації. Зокрема, швидкість переходу через нуль є одним з ключових параметрів у завданнях відокремлення мовних фрагментів від немовних [11], [12]. Крім того, даний метод активно використовується в системах пошуку музичної інформації. Наприклад, його використання є ефективним в задачі класифікації ударних звуків, оскільки такі звуки мають високий рівень зміни знаку сигналу порівняно з іншими типами аудіофрагментів.

ZCR формально визначається як:

$$ZCR = \frac{1}{T-1}\sum_{t=1}^{T-1} 1_{R<0}\left(S_t S_t - 1\right), \tag{1.1}$$

де S є сигналом довжини T;

$1_{R<0}$ є індикаторною функцією.

1.1.2 Енергія сигналу

Енергія сигналу є одним із ключових параметрів, які характеризують інтенсивність звукового сигналу в часі. Вона визначається як сума квадратів амплітудної функції сигналу x(n) протягом певного інтервалу часу і відображає загальний рівень потужності сигналу [12]. Формально енергія сигналу може бути представлена у вигляді:

$$Energy = \sum_n \left[x(n)\right]^2. \tag{1.2}$$

Цей показник є важливим для аналізу акустичних сигналів, оскільки він дозволяє оцінити їхню загальну силу та виразність, що є особливо корисним у таких задачах, як розпізнавання мовлення, класифікація аудіоданих та музичний аналіз.

Крім загальної енергії, у багатьох випадках використовується середньоквадратична енергія (англ. Root Mean Square Energy, RMSE), яка визначає усереднений рівень потужності сигналу і враховує його змінність протягом певного періоду. Середньоквадратична енергія у сигналі визначається як:

$$RMSE = \sqrt{\frac{1}{N}\sum_n^N \left[x(n)\right]^2}, \tag{1.3}$$

де N – загальна кількість відліків сигналу у заданому часовому вікні.

Блок методів виділення ознак частотної області – це сукупність підходів, у яких акустичний сигнал попередньо перетворюється із часового представлення у спектральне, після чого аналізується його частотний склад, енергетичні піки або часово-частотна динаміка. Такий підхід ґрунтується на перетворенні Фур'є, яке дозволяє подати сигнал як суму гармонічних компонент різних частот. У цифровій обробці сигналів, зокрема в аналізі мовлення, працюють не з неперервним сигналом, а з його дискретною послідовністю відліків x[n]. Для такого випадку застосовують дискретне перетворення Фур'є (англ. Discrete Fourier Transform, DFT), яке визначається формулою:

$$X[k] = \sum_{n=0}^{N-1} x[n] e^{-j\frac{2\pi}{N}kn}, \qquad k = 0,1,\ldots,N-1. \tag{1.4}$$

З математичної точки зору коефіцієнт X[k] є комплексним числом, тому для аналізу зазвичай використовують модуль спектра або енергетичний спектр (формула 1.5).

1.1.3 Енергія смуги

Енергія смуги – це загальна енергія, що міститься в даній смузі частот k кадру m представлення сигналу X[k,m] у частотно-часовому діапазоні, визначеного як квадрат його величини:

$$E[k,m] = \left|X[k,m]\right|^2. \tag{1.5}$$

У часово-частотному аналізі ця величина дозволяє оцінити, скільки енергії сигналу зосереджено в певній частотній області в конкретний момент часу. На практиці саме розподіл енергії E[k,m] по частотних смугах

і часових кадрах формує основу спектрограми та споріднених ознак, оскільки дає змогу виявляти гармонічну структуру, формантні області, локальні підсилення або ослаблення спектральних компонент та використовується для подальшого обчислення узагальнених параметрів, наприклад кепстральних коефіцієнтів [13].

1.1.4 Спектральний потік

Спектральний потік вимірює спектральну зміну між послідовними кадрами. Він визначається як квадрат різниці між нормованою величиною $\hat{X}$ двох кадрів:

$$F[m] = \sum_{k=1}^{N} \left(\hat{X}[k,m] - \hat{X}[k,m-1] \right)^2, \tag{1.6}$$

де N – кількість частотних бінів;

$\hat{X}$ [k,m] – нормована величина спектра для біну k у кадрі m.

З фізичного погляду великі значення спектрального потоку відповідають ділянкам сигналу, де відбуваються різкі спектральні перебудови, наприклад поява нових гармонік, зміна формантної структури, атака звука або перехід між різними фонетичними одиницями. Натомість малі значення цієї ознаки свідчать про відносну спектральну стабільність між сусідніми кадрами. Саме тому спектральний потік широко застосовують у задачах сегментації аудіосигналів, виявлення моментів початку звука, аналізу мовлення та музичних сигналів, а також як інформативну часово-частотну ознаку в системах автоматичної класифікації [14]. У контексті аналізу мовлення ця характеристика дозволяє кількісно оцінювати інтенсивність локальних змін у спектрі, що є корисним для опису переходів між звуками та емоційно забарвлених ділянок мовного сигналу.

1.1.5 Спектральний спад

Спектральний спад кадру вказує на асиметрію його спектральної форми та визначається як частота, нижче якої міститься заздалегідь визначена кількість (зазвичай від 85% до 99% загальної енергії) спектральної енергії. Формально ця ознака обчислюється на основі енергетичного спектра $|X[k,m]|^2$ шляхом пошуку індексу частотного біна r, для якого виконується умова накопичення заданої частки енергії:

$$\sum_{k=0}^{r}|X[k,m]|^2 \geq \alpha \sum_{k=0}^{N-1}|X[k,m]|^2, \tag{1.7}$$

де α – вибране порогове значення;

N – кількість спектральних бінів.

Таким чином, спектральний спад відображає, наскільки швидко енергія спектра концентрується в нижньочастотній області, і тому може розглядатися як ознака, що опосередковано характеризує форму спектра та співвідношення між низько- і високочастотними компонентами [15].

1.1.6 Спектральний центроїд

Спектральний центроїд є однією зі спектральних ознак, що характеризує положення «центра мас» спектра в межах окремого кадру сигналу. Інтуїтивно ця величина показує, у якій частині спектра зосереджено основну енергію або амплітуду сигналу. Саме тому спектральний центроїд тісно пов’язують із перцептивним враженням яскравості звуку: сигнали з більшою часткою високочастотних компонент зазвичай мають вищий спектральний центроїд, тоді як сигнали, енергія яких переважно зосереджена на нижчих частотах, характеризуються меншими значеннями цієї ознаки.

Він обчислюється як середнє зважене частот, присутніх у сигналі, з їх величинами як вагами:

$$Centroid = \frac{\sum_{n=0}^{N-1} f(n)x(n)}{\sum_{n=0}^{N-1} x(n)}, \tag{1.8}$$

де x(n) – величина спектра для секції з номером n;

f(n) – центральна частота цієї секції.

1.1.7 Спектрограма

Спектрограма є графічним способом представлення змін сили або гучності сигналу в часі на різних частотах, що містяться у його структурі. Вона дозволяє не лише порівнювати рівень енергії на різних частотах і аналізувати динаміку змін енергетичних рівнів у часовому просторі. У різних галузях науки спектрограми широко застосовуються для дослідження акустичних характеристик звукових хвиль, створюваних різними джерелами.

Розрахунок спектрограми виконується в такий спосіб. Необроблена форма аудіосигналу спочатку розбивається на кадри, що перекриваються. Довжина кадру зазвичай становить 15–25 мс, причому звук може сприйматися як приблизно стаціонарний. Іноді використовується тимчасовий зсув, менший за розмір вікна, зазвичай близько 10 мс, що забезпечує перекриття між кадрами. Цей зсув необхідний для зменшення втрати інформації, спричиненої поділом.

У контексті аналізу голосу людини спектрограма є одним із базових інструментів, оскільки дає можливість одночасно спостерігати гармонічну структуру, зміни основної частоти та спектральні переходи між фонетичними елементами. Саме тому спектрографічне подання (рисунок 1.3) широко використовується як для візуального аналізу

мовних сигналів, так і як вхідне представлення для алгоритмів автоматичного розпізнавання мовлення та емоцій.

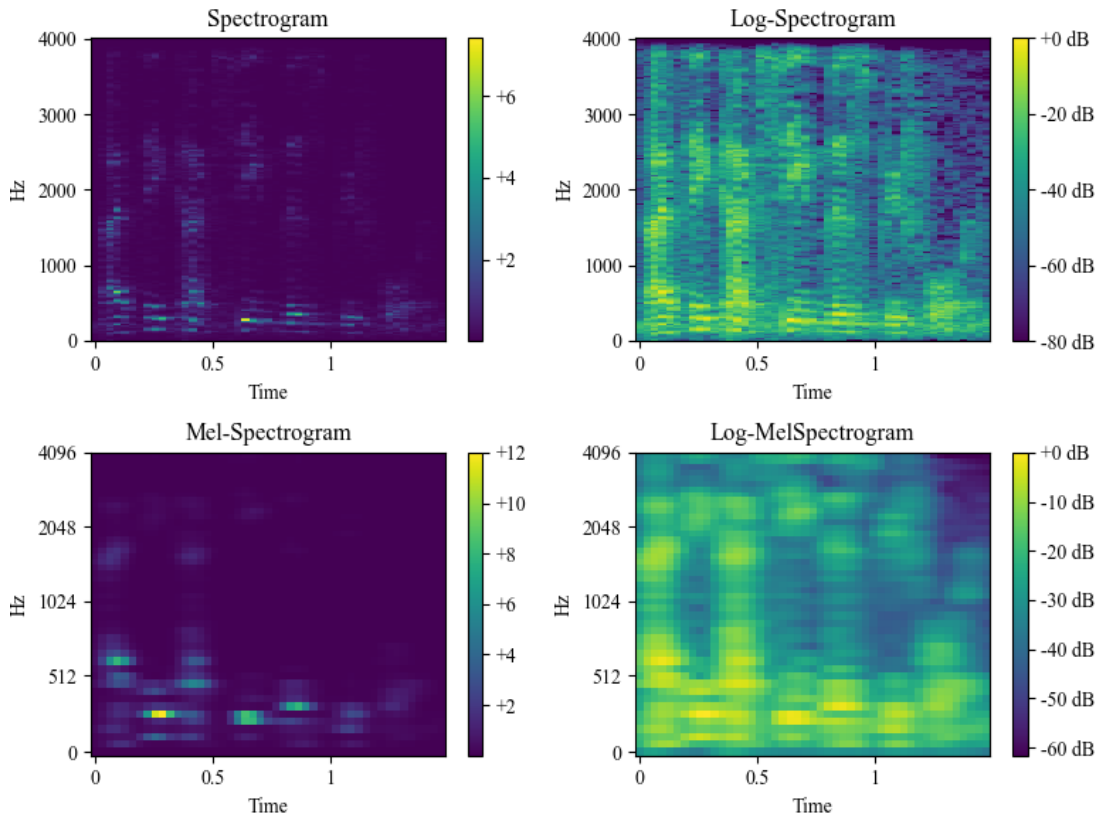


Рисунок 1.3 – Спектрограма, лог-спектрограма, мелспектрограма та лог-мелспектрограма аудіозапису для емоційного стану «злість»

Проте слухове сприйняття людини не є лінійно пропорційним частоті. З метою опису суб'єктивного сприйняття висоти тону S. S. Stevens, J. Volkmann та E. B. Newman запропонували мел-шкалу (формула 1.9), побудовану на основі психоакустичних експериментів, у яких оцінювалася психологічна величина висоти тону [16]. Графік, що демонструє мел-частотну шкалу, наведено на рисунку 1.2.

$$m = 1127 ln\left(1 + \frac{f}{700}\right), \tag{1.9}$$

де f – фактична частота;

m – мел-частота.

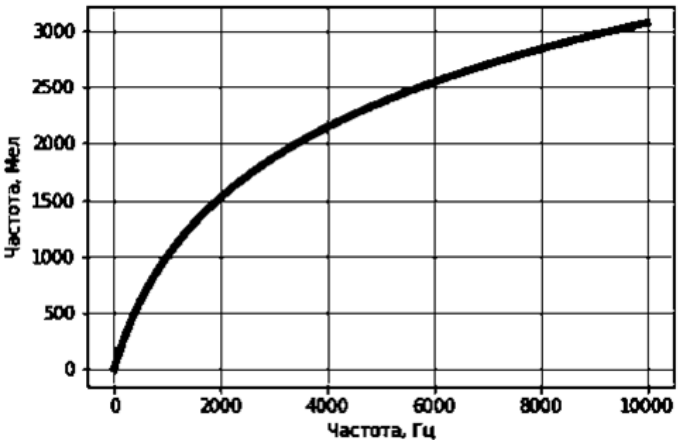


Рисунок 1.2 – Мел-шкала

Мел-спектрограма є модифікацією звичайної спектрограми, у якій лінійна частотна шкала замінюється на мел-шкалу. Для її побудови після обчислення спектра або спектрограми енергія частотних компонент проєктується на банк трикутних мел-фільтрів, центрованих за нелінійною шкалою: у низькочастотній області роздільна здатність є вищою, тоді як у високочастотній – грубішою.

Лог-спектрограма – це спектрограма, до якої застосовано логарифмічне масштабування амплітуди або енергії спектральних коефіцієнтів. Логарифмічне перетворення стискає динамічний діапазон спектральних значень, завдяки чому слабші спектральні компоненти стають краще помітними, а відносні відмінності між сильними та слабкими гармоніками зменшуються. Лог-спектрограма подається у формі:

$$S_{\log}[k,m] = ln\left|X[k,m]\right|^2. \qquad (1.10)$$

Лог-мелспектрограма є часово-частотним представленням сигналу, яке поєднує два зазначені перетворення: проєкцію спектральної енергії на

мел-шкалу та логарифмічне масштабування. Така спектрограма є більш стабільною та інформативною, ніж лінійна спектрограма, оскільки краще відповідає перцептивним властивостям слухової системи людини. Лог-мелспектрограма широко застосовується як базове вхідне представлення в сучасних системах автоматичного аналізу мовлення, включно з задачами розпізнавання мовлення, класифікації звуків та розпізнавання емоцій.

1.1.8 Кепстральні ознаки

Кепстральні ознаки є класом параметрів, що формуються в кепстральній області та використовуються для компактного опису спектральної структури звукового, зокрема мовного, сигналу. Кепстр виникає внаслідок послідовного застосування двох спектральних перетворень із нелінійною операцією логарифмування між ними: спочатку сигнал переводиться у частотну область, далі обчислюється логарифм його спектра або спектра потужності, після чого до отриманої послідовності застосовується зворотне дискретне перетворення Фур'є. У разі використання спектра потужності $|X[k]|^2$ кепстр $X_C[n]$ можна подати як:

$$X_C[n] = IDFT\left(log\left(|X[k]|^2\right)\right). \tag{1.11}$$

Перевага кепстральної області полягає в тому, що вона дозволяє відокремити джерело та фільтр(и), оскільки згортка над часом представлена як додавання в кепстральній області. Саме тому в кепстральній області повільно змінний спектр, що утворюється завдяки резонансам вокального тракту, та швидкі гармонічні коливання, пов'язані з основною частотою, можуть аналізуватися окремо. Це робить кепстральні ознаки особливо актуальними для задач автоматичного розпізнавання мовлення,

ідентифікації мовця та емоційного аналізу, оскільки вони забезпечують компактне, інформативне та порівняно стійке до варіацій представлення акустичної структури сигналу [17].

1.1.9 Коефіцієнти мел-частотного кепструму

У прикладних системах найчастіше використовують не весь кепстр, а лише його низькочастотні коефіцієнти, оскільки вони найкраще описують гладку спектральну характеристику сигналу. На цій ідеї ґрунтується, зокрема, представлення мел-частотних кепстральних коефіцієнтів.

Коефіцієнти мел-частотного кепструму (MFCC) – це представлення, визначене як реальний кепстр віконного короткочасного сигналу, отриманого за допомогою дискретного перетворення Фур'є цього сигналу. На відміну від звичайного кепстра, MFCC формуються шляхом проектування спектра на банк мел-фільтрів, логарифмування енергії окремих мел-смуг та подальшого застосування дискретного косинусного перетворення (DCT). Останнє забезпечує перехід до компактного кепстрального простору, подібного за ідеєю до класичного кепстрального аналізу, але узгодженого зі слуховими властивостями людини завдяки використанню нелінійних перетворень.

У задачах обробки мовлення MFCC набули широкого застосування завдяки тому, що вони забезпечують стійке та інформативне подання акустичної структури сигналу, придатне для розпізнавання мовлення, ідентифікації мовця та емоційного аналізу. На відміну від спектра, який містить значну кількість надлишкових або дрібномасштабних деталей, MFCC містять компактну інформацію про форму спектральної оболонки. Однак, ця властивість MFCC робить їх обмеженими для використання в задачах загальної класифікації аудіосигналів. Тому різні типи звукових подій потребують більш детального спектрального представлення.

У даному дослідженні було обрано лог-мелспектрограму в якості основного представлення сигналу для задачі розпізнавання емоцій за мовленням у вбудованих системах. Цей вибір є обґрунтованим з позицій як ефективності ознак, так і їх обчислювальної вартості. Ознаки часової області, хоча й мають дуже низьку обчислювальну складність на кадр (яка зазвичай не перевищує O(N)), демонструють недостатню інформативність для задач SER, оскільки погано відображають спектральну структуру емоційно забарвленого мовлення. Звичайна спектрограма є одним із найменш витратних способів отримання частотного представлення: при використанні алгоритму швидкого перетворення Фур'є (FFT) обчислювальна складність становить O(N log N), а потреба пам'яті – приблизно N комплексних чисел на кадр. Проте вона є надмірно чутливою до варіацій амплітуди та погано узгоджується з характеристиками людського слуху. На цьому тлі лог-мелспектрограма забезпечує компроміс між компактністю ознак та часовою складністю їх отримання: мел-фільтрація є обчислювально легкою операцією з приблизною складністю O(M), де M – кількість мел-фільтрів. Логарифмічне масштабування також додає O(M) операцій. Сукупна вартість побудови лог-мелспектрограми становить близько O(N log N + M) на кадр, що залишається прийнятним для вбудованих систем реального часу. На відміну від неї, формування кепстру і MFCC вимагає додаткового етапу IDFT або DCT зі складністю O(M log M) та збільшеними вимогами до пам'яті. Отже, лог-мелспектрограма не потребує додаткових перетворень і водночас зберігає ключову інформацію про частотні характеристики мовлення, що робить її найбільш придатним вибором для вбудованих застосувань.

1.2 Методи класифікації акустичних сигналів

Класифікація – один із розділів машинного навчання, присвячений вирішенню наступного завдання: є множина об'єктів, розділених на класи,

задано кінцеві множини об’єктів, для яких відомо клас приналежності. Початкова (вхідна) множина називається навчальною вибіркою. Класова приналежність інших об’єктів невідома. Необхідно побудувати алгоритм, здатний класифікувати довільний об’єкт із вихідної множини [10], [18].

Класифікатором називається відображення, представлене формулою 1.10:

$$\hat{c}: X \to C, C = \{C_1, C_2, \ldots, C_n\}, \tag{1.12}$$

де $\hat{c}$ – функція класифікатора;

X – множина об’єктів навчальної вибірки;

C – кінцева та зазвичай невелика множина міток класів;

$\hat{c}(x)$ – оцінка істиної, але невідомої функції *c(x)*:

$$(x, c(x)), x \in X, \tag{1.13}$$

де x – об’єкт;

c(x) – істинний клас об’єктів x.

Під навчанням класифікатора розуміється створення функції $\hat{c}$. Вона, як найкраще, наближається до вихідної функції (і не лише на навчальному наборі, але, в ідеальному випадку – на всьому просторі об’єктів, що забезпечує здатність моделі до узагальнення).

У даному випадку задача розпізнавання емоцій за мовленням може бути подана як окремий випадок задачі багатокласової класифікації. Нехай кожному мовному фрагменту $x \in X$, представленому у вигляді набору акустичних ознак, спектрального опису або часово-частотного представлення, необхідно призначити у відповідність одну мітку з множини емоційних класів:

$$C = \{C_1, C_2, \ldots, C_K\}, \tag{1.14}$$

де K – кількість емоційних категорій.

Тоді класифікатор $\hat{c}$ реалізує відображення від простору ознак мовного сигналу до множини емоційних міток, а результат його роботи $\hat{c}(x)$ інтерпретується як оцінка емоційного стану мовця для заданого аудіофрагмента. З позиції теорії статистичного навчання ця задача також може бути сформульована як оцінювання умовного розподілу ймовірностей належності об'єкта x до кожного з класів C_k, з подальшим вибором найбільш імовірного класу:

$$\hat{y} = \arg\max_{C_k \in C} P(C_k \mid x). \tag{1.15}$$

Таким чином, постановка задачі узгоджується з класичним підходом до задач класифікації в машинному навчанні та створює теоретичну основу для подальшого вибору архітектури моделі та критеріїв оцінювання її якості. Нижче наведено огляд поширених методів класифікації аудіосигналів, актуальних для вирішення задачі розпізнавання емоційних станів за голосом мовця.

1.2.1 Приховані Марківські моделі

Приховані Марківські моделі (англ. Hidden Markov model, HMM) були одними з перших підходів, застосованих у розпізнаванні емоцій з мовлення, задовго до появи поширених рекурентних нейронних мереж. Їхня робота ґрунтується на припущенні, що спостереження генеруються послідовністю прихованих станів, а переходи між цими станами підпорядковуються марківській властивості. Марківський процес – це стохастичний процес, який має властивість «відсутності пам'яті», тобто майбутній стан залежить

лише від поточного стану і не залежить від попередніх станів. Наприклад, перехід від одного стану до іншого залежить тільки від того, в якому стані система перебуває зараз. В моделі HMM є набір станів, які називаються прихованими, оскільки вони не спостерігаються безпосередньо. Натомість спостерігається послідовність сигналів, які залежать від цих прихованих станів. Ймовірності переходів визначають ймовірності переходу від одного стану до іншого. Це задається матрицею переходів A, де елемент a_{ij}, а відповідає ймовірності переходу зі стану i до стану j. Ймовірності спостережень визначають ймовірності появи спостереження (наприклад, конкретного звуку) у кожному зі станів. Це описується матрицею спостережень B, де $b_j(o_t)$ – це ймовірність того, що спостереження o_t відбувається у стані j. Початкові ймовірності станів задаються вектором π, де π_i – це ймовірність того, що початковий стан системи дорівнює i. Приховану марківську модель можна описати наступним чином:

$$\lambda = (A, B, \pi), \tag{1.16}$$

де A – матриця ймовірностей переходів між станами;

B – матриця ймовірностей спостережень;

π – початкові ймовірності станів.

У контексті вирішення задачі виявлення емоцій зазвичай навчається окрема HMM-модель для кожного класу емоцій. Під час класифікації обчислюється правдоподібність того, що певне висловлення було згенероване моделлю відповідного класу, після чого вибирається клас із максимальною ймовірністю. HMM мають суттєве обмеження: вони погано працюють із нелінійними залежностями в даних, що обмежує їх ефективність у задачах SER, де емоційні патерни часто є складними і нелінійними. Згідно результатів дослідження [19], середня точність класифікації емоцій у просторі ознак MFCC з використанням HMM склала близько 31%. Попри свою історичну значущість та досвід використання у

задачах обробки людського мовлення, станом на зараз приховані Марківські моделі є неконкурентоспроможними у порівнянні із сучасними підходами.

1.2.2 Алгоритм випадкового лісу

Алгоритм випадкового лісу (англ. Random Forest, RF) – це контрольований алгоритм машинного навчання, який складається з ансамблю дерев рішень. Основою методу є ансамблеве навчання, яке поєднує велику кількість незалежних класифікаторів для покращення узагальнювальної здатності моделі та підвищення точності результатів.

Побудова випадкового лісу ґрунтується на множині дерев рішень (рисунок 1.4), кожне з яких тренується на випадковій підвибірці даних та ознак. Зазвичай використовується техніка пакетування, яка зменшує варіативність моделі та запобігає перенавчанню. Алгоритм формує кінцеве рішення на основі агрегації прогнозів окремих дерев, зазвичай шляхом голосування (для класифікації) або усереднення (для регресії). Збільшення кількості дерев у лісі, як правило, позитивно впливає на точність, хоча й підвищує обчислювальні витрати.

Дерева рішень є базовими структурними елементами випадкового лісу. Кожне дерево – це модель, що послідовно розділяє навчальні дані відповідно до певних критеріїв, формуючи деревоподібну структуру, яка складається з кореневого вузла, вузлів рішень і листових вузлів. Перевагами дерева рішень є простота інтерпретації, швидкість навчання та здатність працювати з нелінійними залежностями, тоді як його основними недоліками є схильність до перенавчання та чутливість до шумових даних. У випадковому лісі кожне дерево будується із використанням випадкових підмножин ознак та об’єктів, що забезпечує різноманітність дерев і підвищує стабільність результатів.

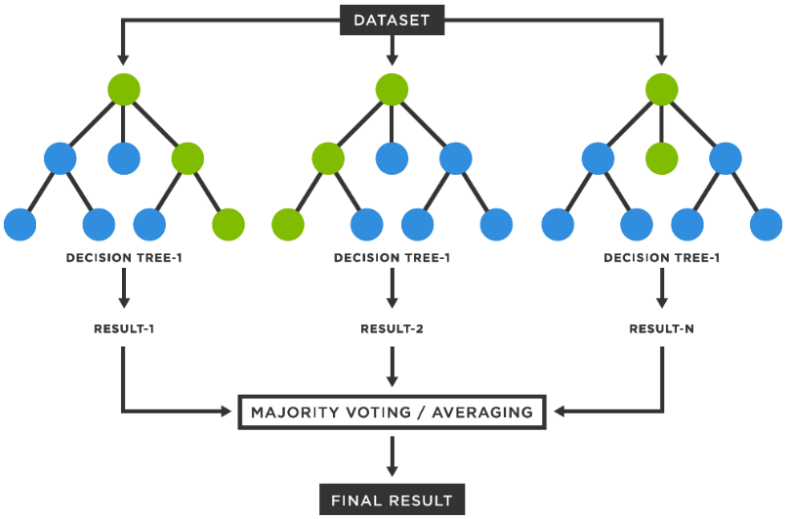
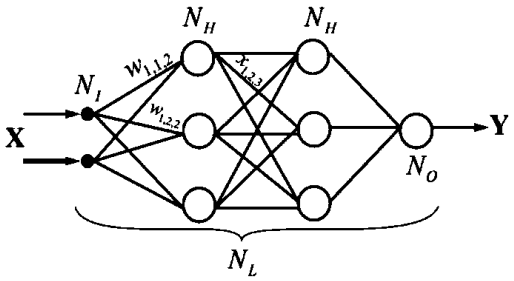


Рисунок 1.4 – Ансамбль дерев рішень

У задачі розпізнавання емоцій з мовних сигналів алгоритм випадкового лісу традиційно застосовується як легко інтерпретований і обчислювально ефективний класифікатор, який доцільно застосовувати при обмежених обсягах даних та їх високій варіативності. Дослідження [12] показало, що RF може демонструвати конкурентну, а інколи й вищу точність порівняно з глибокими нейронними мережами, якщо модель працює на невеликих датасетах та компактних представленнях ознак (таких як MFCC). Середня точність класифікації за допомогою Random Forest склала 69%. Зокрема, автори підкреслюють, що алгоритм випадкового лісу забезпечує високу ефективність у класифікації емоцій за рахунок високої узагальнювальної здатності та низьких вимогах до обсягу навчальних даних, що підтверджує релевантність його використання в задачах SER.

1.2.3 Багатошаровий персептрон

Багатошаровий персептрон – це одна із перших архітектур штучних нейронних мереж (ШНМ), що складається з розташованих підряд кількох шарів нейронів: вхідний; один або декілька прихованих шарів; вихідний шар (рисунок 1.5).



N_I – кількість нейронів вхідного шару;

N_H – кількість нейронів у кожному схованому шарі;

N_O – кількість вихідних нейронів;

x – вектор вхідних сигналів мережі;

y – вектор вихідних сигналів;

N_L – повна кількість шарів у мережі, враховуючи вхідний.

Рисунок 1.5 – Багатошаровий персептрон

Робота багатошарового персептрона описується формулами 1.17–1.19:

$$NET_{jl} = \sum_i w_{ijl} x_{ijl} \tag{1.17}$$

$$OUT_{jl} = F\left(NET_{jl} - \Theta_{jl}\right) \tag{1.18}$$

$$x_{ij(l+1)} = OUT_{il}\,, \tag{1.19}$$

де i – номер входу;

j – номер нейрона в шарі;

l – номер шару;

x_{ijl} – i-й вхідний сигнал j-го нейрона в шарі l;

w_{ijl} – ваговий коефіцієнт i-го входу нейрона номер j у шарі l;

NET_{jl} – сигнал j-го нейрона в шарі l;

OUT_{jl} – вихідний сигнал нейрона;

θ_{jl} – порогове значення (зміщення) j-го нейрона шару l;

F – функція активації штучного нейрона.

До основних переваг багатошарового персептрона належить його широке поширення у науковій спільноті та надійність у розв'язанні багатьох задач. Водночас до недоліків належить обмежена здатність до роботи з динамічними процесами, а також потреба у великій навчальній вибірці для забезпечення адекватного узагальнення знань. Також недоліком є неефективність у випадках високорозмірних та сильно корельованих ознак, оскільки така структура вхідних даних ускладнює оптимізацію ваг.

У системах SER багатошаровий персептрон зазвичай використовувався як класифікатор для компактного представлення ознак, таких як MFCC, спектральні та просодичні параметри. Його застосування у таких системах дозволяє забезпечити кращу узагальнювальну здатність порівняно з традиційними алгоритмами машинного навчання. Експериментальні дослідження демонструють, що MLP може досягати конкурентної точності у задачі класифікації емоцій порівняно з іншими сучасними методами. У дослідженні, присвяченому порівнянню CNN, k-NN, Random Forest та MLP [20], останній показав найвищу середню точність понад 90% на емоційному датасеті Berlin Database of Emotional Speech. Це підтверджує його ефективність для задач SER за умови правильного вибору ознак та архітектури мережі.

1.2.4 Згорткова нейронна мережа

Згорткова нейронна мережа (англ. Convolutional neural network, CNN) – спеціалізована архітектура ШНМ, націлена на ефективне розпізнавання, і відноситься до технологій глибокого навчання (англ. Deep learning). CNN будуються на трьох базових концепціях: локальне

рецептивне сприйняття, колективні ваги (ядро згортки), субдискретизація (пулінг).

У підходах глибокого навчання вхідними даними зазвичай є багатовимірні масиви даних, а ядро згортки також подається як багатовимірний масив параметрів, які оптимізуються в процесі навчання. Оскільки лише скінченна кількість елементів масиву містить ненульові значення, згорткова операція реалізується як підсумовування по обмеженій області, що робить її обчислювально ефективною.

Згортки часто використовуються над більш ніж однією віссю одночасно. У випадку, якщо використовується двовимірний об'єкт I як вхідні дані, також потрібно використовувати двовимірне ядро K:

$$S(i,j) = (I * K)(i,j) = \sum_m \sum_n I(m,n) K(i-m, j-n). \qquad (1.20)$$

У двовимірній згортці вихід визначається лише для тих позицій, де ядро повністю накладається на вхідну матрицю, що відповідає поняттю «дійсної» згортки. На рисунку 1.6 схематично показано процес згортки, де проілюстровано формування верхнього лівого елемента вихідного тензора шляхом застосування ядра до відповідної ділянки вхідного масиву.

Традиційні шари нейронних мереж виконують множення повної матриці ваг на вхідний вектор, тобто кожен вихідний елемент взаємодіє з кожним вхідним елементом. Натомість CNN використовують розріджену взаємодію (розріджену зв'язність), оскільки ядро охоплює лише невелику локальну область вхідних даних. Це досягається шляхом зменшення розміру ядра порівняно зі входом, що зменшує кількість параметрів, потребу в пам'яті та підвищує статистичну ефективність моделі.

Це також означає, що обчислення результату вимагає значно менше операцій. Якщо традиційний повнозв'язний шар із m входами та n виходами потребує m×n параметрів і має часову складність O(m×n), то розріджено підключений згортковий шар із k з'єднаннями на один вихід включає

лише k×n параметрів і виконується за O(k×n), що забезпечує суттєве прискорення обчислень [21]. У глибоких CNN нейрони внутрішніх шарів завдяки послідовному застосуванню згорток та пулінгу формують розширені рецептивні поля, що дозволяє їм опосередковано взаємодіяти з більшістю вхідних даних і охоплювати широкі контекстні залежності.

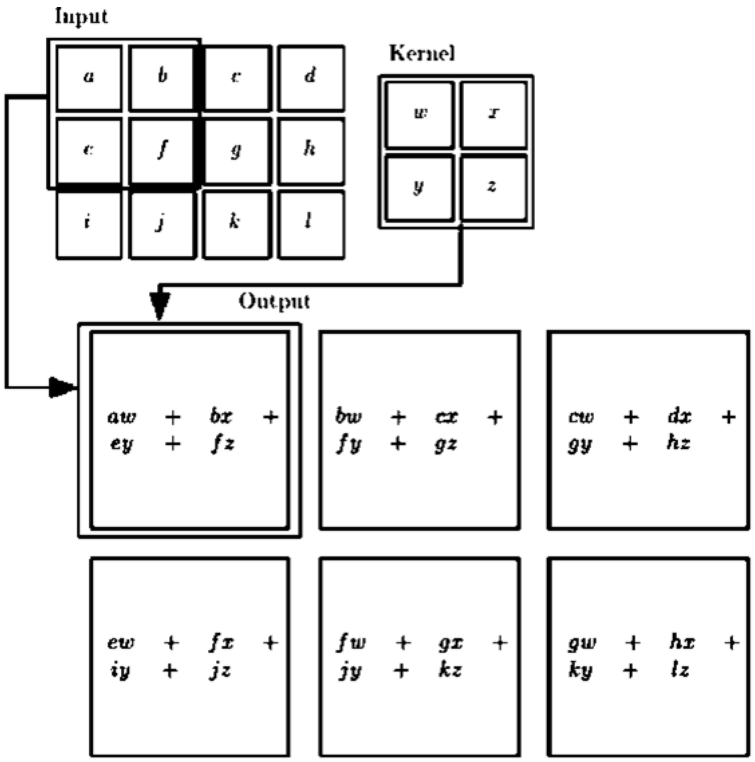


Рисунок 1.6 – Схема виконання операції згортки

Згорткові моделі посідають важливе місце серед сучасних методів розпізнавання емоцій з мовлення, оскільки здатні ефективно опрацьовувати спектрограми та інші двовимірні представлення аудіосигналу, автоматично виділяючи релевантні просторово-частотні ознаки. Завдяки локальним рецептивним полям та можливості формувати глибоку ієрархію представлень, вони демонструють високу узагальнювальну здатність навіть за умов варіативності мовленнєвих характеристик. Сучасні дослідження підтверджують ефективність застосування CNN у задачах SER. Зокрема, у роботі [22] запропоновано 1D-CNN на основі гібридних

ознак MFCCT (MFCC + часові характеристики), яка досягла точності 97% на Emo-DB, 93% на SAVEE та 92% на RAVDESS. Подібні висновки отримано у інших роботах: 1D-CNN у дослідженні [23] забезпечила точність 94.0% на MELD і 91.9% на RAVDESS. Водночас результати свідчать, що продуктивність таких моделей істотно залежить від способу подання вхідного сигналу та характеристик вибірки. 2D-CNN на сегментованому комбінованому наборі даних та спектрограмах показала точність майже 84% [24]. Тобто приріст точності визначається не ускладненням архітектури, а якістю попередньої обробки, обраним представленням ознак і стратегіями аугментації даних. Тому коректне порівняння результатів між окремими роботами потребує врахування відмінностей у датасетах, ознакових просторах і протоколах оцінювання.

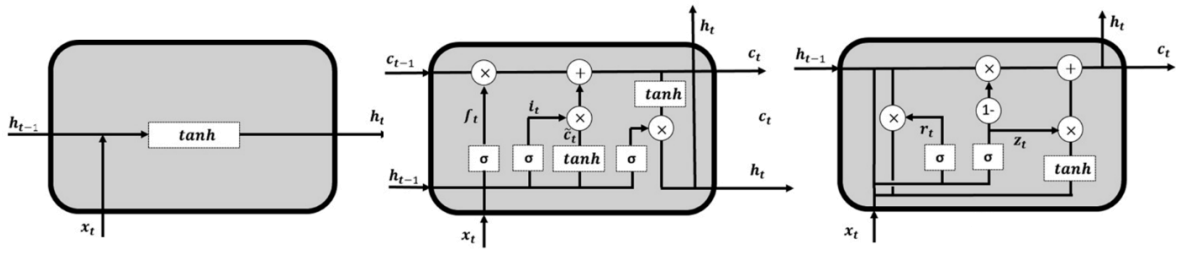
1.2.5 Рекурентні нейронні мережі

Рекурентні нейронні мережі (англ. Recurrent neural network, RNN) належать до класу моделей, спеціалізованих для роботи з послідовностями даних, де поточний вихідний стан залежить не лише від поточного входу, а й від попередніх станів. Ключова особливість полягає у введенні прихованого стану, який оновлюється через нелінійну функцію активації на основі поточного входу та попереднього стану (рисунок 1.7 (а)), що дозволяє мережі накопичувати інформацію про контекст. Водночас класичні RNN стикаються з проблемою зникання або вибуху градієнтів під час навчання на довгих послідовностях.

Моделі типу Long Short-Term Memory (LSTM) були запропоновані для подолання проблеми зникання градієнта у стандартних RNN під час навчання на довгих послідовностях. Її ключова відмінність полягає у введенні стану пам’яті та керувальних брам, які забезпечують стабільніший перенос інформації між часовими кроками. У комірці функціонують три основні брами: forget gate, input gate та output gate (рисунок 1.7 (б)), які

відповідно визначають, яку інформацію слід зберегти, оновити або передати далі.

Gated Recurrent Unit (GRU) є спрощеним різновидом керованих рекурентних мереж, у якому механізм пам’яті реалізовано компактніше, ніж у LSTM. Використовується лише дві брами: update gate та reset gate (рисунок 1.7 (в)), завдяки чому процес оновлення прихованого стану стає структурно простішим, але зберігає здатність моделювати довші часові залежності.



а) б) в)

Рисунок 1.7 – Рекурентні нейронні мережі: а) RNN, б) LSTM, в) GRU

Для задачі розпізнавання емоцій за аудіоданими використання рекурентних моделей є особливо актуальним, оскільки емоційно значущі ознаки мовлення часто є довгими та нерівномірно розподіленими в часі. Показовим є дослідження [25], де безпосередньо порівнювалися RNN, GRU і LSTM для роботи з ознаками MFCC. Результати показали, що LSTM найкраще працює на однорідному датасеті TESS, досягаючи точності 99,6%, тоді як GRU продемонструвала кращу точність на складніших і менших наборах, зокрема на SAVEE. Отже, вибір між трьома основними типами рекурентних мереж не є очевидним та істотно залежить від структури корпусу та зразків аудіозаписів у ньому. У іншому дослідженні [26] було порівняно глибокі моделі в задачі SER, серед яких рекурентні (LSTM) та згорткові ШНМ. Також використано різні формати вхідного представлення, зокрема необроблений аудіосигнал без етапу виділення ознак. У такій

постановці задачі LSTM продемонструвала найкращий результат лише на однорідному корпусі TESS, досягнувши точності 99,5%. Однак на більшості інших наборів даних, незалежно від представлення ознак, перевагу отримала згорткова ШНМ. Для «сирого» сигналу її точність склала понад 90% на EMO-DB та RAVDESS, майже 86% на SAVEE, майже 70% на CREMA. Це свідчить, що застосування рекурентних моделей у SER не є універсальним. LSTM показує кращі результати на чистіших і менш варіативних даних, тоді як CNN демонструє вищу стабільність на різнорідних корпусах.

1.2.6 Трансформерні моделі

Трансформери є класом глибоких нейромережевих архітектур, у яких моделювання залежностей між елементами послідовності або фрагментами вхідного сигналу здійснюється через механізм уваги, що дає можливість ефективно враховувати глобальний контекст і довготривалі зв’язки. У класичному варіанті трансформер має структуру «енкодер-декодер», де перший формує контекстне подання вхідної послідовності, а другий на основі цього подання генерує вихідну послідовність. Базовими компонентами архітектури є масштабована увага на основі скалярного добутку, багатоголова увага, блоки прямого поширення, залишкові з’єднання, нормалізація шару та позиційне кодування. Останнє вводиться тому, що механізм уваги не містить вбудованого уявлення про порядок елементів. На рівні механізму самоуваги кожен елемент послідовності проєктується у три подання: query, key і value. Ваги уваги обчислюються через подібність між query та key, після чого формується контекстне подання як зважена комбінація value-векторів. Використання кількох голів уваги дає змогу моделі одночасно враховувати різні типи залежностей у різних підпросторах ознак: локальні, глобальні, синтаксичні або семантичні тощо. Завдяки відсутності рекурентної обробки трансформери дозволяють

краще використовувати паралельні обчислення, ніж RNN-моделі, а також здатні ефективно обробляти довгі послідовності. Це стало однією з ключових причин їхнього домінування в сучасному глибинному навчанні.

Зорові трансформери (англ. Vision Transformers, ViT) є адаптацією трансформерної моделі до двовимірних вхідних даних, насамперед зображень. У базовій архітектурі ViT (рисунок 1.8) зображення спочатку розбивається на патчі фіксованого розміру, кожен з яких розгортається у вектор. Після цього до них застосовують лінійну проєкцію в простір вкладеного подання та інтерпретують як послідовність токенів, аналогічно словам. До цієї послідовності додаються позиційні вкладання, а також спеціальний [CLS]-токен, який акумулює інформацію з усіх патчів і надалі використовується для класифікації. Наступним кроком послідовність подається до стека стандартних блоків трансформерного кодувальника. Отже, ViT не потребує згорткових шарів для виділення ознак. Порівняно з CNN, такі моделі мають слабші властивості обробки локальних патернів та трансляційної інваріантності, але краще захоплюють глобальні просторові взаємозв'язки між далекими ділянками зображення. Водночас їхня ефективність, як правило, сильніше залежить від обсягу даних для навчання.

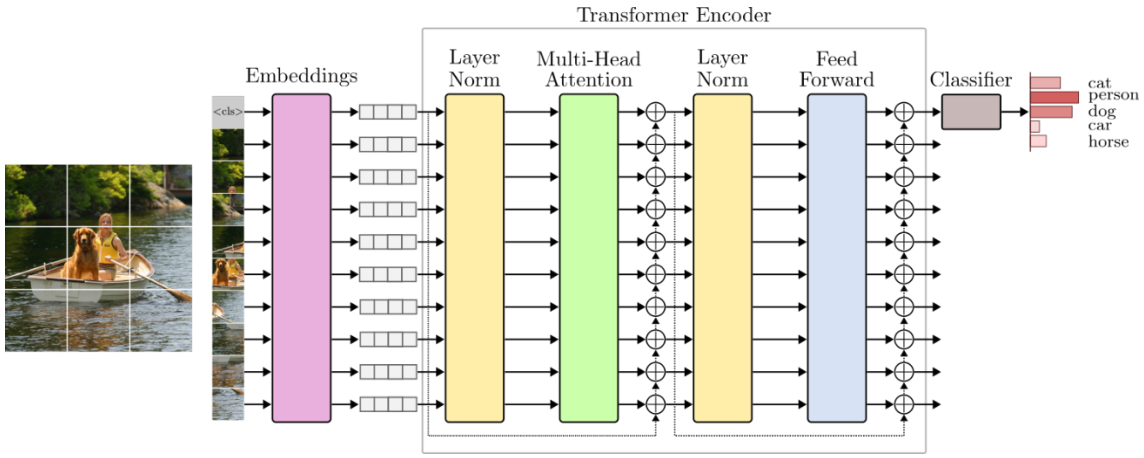


Рисунок 1.8 – Vision-трансформер

Аналіз результатів нещодавніх досліджень свідчить, що трансформерні підходи демонструють конкурентоспроможні результати в задачі SER, хоча їхня ефективність істотно залежить від типу представлення ознак та характеристик тренувального корпусу. Зокрема, у роботі [27] запропоновано легковагову ViT-модель, яка на мел-спектрограмах досягла значення точності 98% на TESS, 91% на EMODB, що підтверджує здатність зорових трансформерів виявляти емоційно значущі просторові залежності у спектрограмних представленнях мовлення. У роботі [28] представлено MelTrans – чисту трансформерну архітектуру для аналізу мел-спектрограм, яка забезпечила 92,5% на EMODB та 76,5% на IEMOCAP. Дані результати вказують, що трансформери є ефективними не лише на відносно простих і контрольованих наборах даних, а й на складніших корпусах, де важливим є моделювання довгих часових послідовностей. Отже, моделі Transformer/ViT становлять самостійний методологічний напрям у вирішенні задач SER та є одним з найбільш сучасних та перспективних підходів у даній галузі.

1.2.7 Гібридні моделі

У сучасних підходах до розпізнавання емоцій за аудіоданими слід окремо відзначити гібридні архітектури, що поєднують згорткові блоки і рекурентні моделі або трансформерні енкодери. Завдяки цьому вони демонструють здатність ефективно вилучати як локальні спектрально-часові ознаки, так і глобальні контекстні залежності.

У роботі [29] було запропоновано гібридну глибинну архітектуру CNN+BiLSTM, застосовану до лог-мелспектральних ознак для задачі SER. Згорткові шари виконували роль екстрактора просторової інформації, ефективно вилучаючи локальні спектральні патерни з аудіосигналу. Отримані представлення подавалися у двонаправлену модель LSTM, яка моделює часову динаміку в емоційно значущих змінах

мови. На датасетах TESS і SAVEE ця гібридна модель досягла понад 96% точності, що перевищило результати чисто згорткових або рекурентних архітектур.

У нещодавньому дослідженні [30] представлено архітектуру CNN-n-GRU, що поєднує CNN для вилучення ієрархічних локальних ознак із GRU, що моделює часову структуру мовного сигналу без використання явного виділення ознак. Такий підхід забезпечив точність 99% на TESS, 81% на IEMOCAP та 86% на RAVDESS. При цьому модель мала компактну архітектуру та продемонструвала здатність ефективно обробляти необроблений сигнал без попереднього перетворення у компактне представлення.

У роботі [31] представлено модель CTENet, в якій дві паралельні згорткові мережі обробляють просторове представлення (лог-мелспектрограму або MFCC), а трансформер обробляє часову динаміку. На корпусі RAVDESS модель досягла 82% точності, а на IEMOCAP – 79%. Отже, даний гібридний підхід демонструє перевагу перед чистими CNN та трансформерами, завдяки синергетичному поєднанню локальних та глобальних спектрально-часових ознак.

У даній роботі в якості методу класифікації ознак обрано чисту згорткову архітектуру ШНМ, яка є однією з найефективніших моделей для класифікації аудіосигналів і зазвичай перевершує класичні методи на кшталт RF та HMM. Такі моделі здатні виявляти складні структурні патерни у багатовимірних представленнях звуку, яким є лог-мелспектрограма. Згорткові нейронні мережі здатні автоматично визначати значущі ознаки, зменшуючи вимоги до методу виділення ознак, які можуть бути висококорельованими, варіабельними та містити шуми. Основною перевагою згорткової архітектури у задачі класифікації аудіосигналу є її здатність ефективно обробляти просторово-залежні структури даних. Використання згорткових шарів дозволяє моделі навчатися локальним особливостям сигналу, таким як текстура спектру, зміни частоти та часова

еволюція амплітуди. Крім того, застосування операцій згортки та субдискретизації сприяє зменшенню обчислювальної складності, зберігаючи ключову інформацію для класифікації.

На відміну від багатошарових персептронів, обчислювальна складність яких масштабується як O(n×m) (де n та m – розміри повнозв'язних шарів), згорткові мережі перетворюють обробку сигналу на сукупність згорткових операцій із набагато меншою кількістю параметрів. Це забезпечує час виконання приблизно O(K×H×W) (де K – розмір ядра, H×W – розміри карти активацій), що набагато ефективніше за повнозв'язні шари подібного розміру.

Порівняно з рекурентними мережами, часові витрати яких зростають лінійно зі збільшенням довжини послідовності: $O(T \times d^2)$ (де T – кількість часових кроків, а d – розмірність стану), CNN не потребують покрокової обробки і здатні виконувати інференс в паралельному режимі. Це особливо важливо для задач реального часу на Edge-пристроях.

У порівнянні з трансформерами та ViT, згорткові моделі є ще вигіднішими з точки зору продуктивності на вбудованих системах. Шари уваги у трансформерах мають часову складність $O(T^2)$, оскільки кожен токен повинен взаємодіяти з усіма іншими, що у випадку довгих фрагментів мовлення призводить до непридатності для Edge AI пристроїв. Зоровий трансформер додатково має квадратичну залежність складності від кількості патчів. До цього додається і те, що апаратна та програмна підтримка механізмів уваги на вбудованих платформах усе ще обмежена: більшість фреймворків та апаратних прискорювачів (DSP, NPU) мають оптимізовані інструкції для 2D-згорток, а не для роботи з операціями масштабованої уваги. Через це CNN сьогодні мають значно кращу апаратну оптимізацію, тоді як реалізація повноцінних трансформерних моделей на Edge-пристроях часто вимагає складної програмної оптимізації.

Висновки до розділу 1

Даний розділ кваліфікаційної роботи присвячено огляду різних аспектів предметної галузі – розпізнавання емоцій за аудіоданими. Розглянуто процес цифрового кодування аудіосигналу, методи та моделі, що застосовуються у цьому процесі. Визначено основні характеристики цифрових аудіоданих та їх співвідношення з фізичними характеристиками акустичного сигналу. У дослідженні було обрано частоту дискретизації 8 кГц, оскільки вона забезпечує баланс між інформативністю та ресурсними обмеженнями вбудованих систем, охоплюючи критичний діапазон для розбірливості мовлення та передачі емоційних ознак.

Здійснено огляд існуючих методів виділення ознак для аудіосигналу, що можуть бути застосовані у задачах класифікації емоційного мовлення. Зокрема, розглянуто такі поширені характеристики, як ZCR, енергія сигналу, спектрограми та кепстральні ознаки. У даному дослідженні лог-мелспектрограма була обрана як основне представлення сигналу, оскільки вона поєднує низьку обчислювальну складність перетворення із високою інформативністю. На відміну від ознак часової області, вона адекватно відображає частотно-часову структуру мовлення, а порівняно з кепстральними ознаками не потребує додаткових перетворень. У порівнянні зі звичайною спектрограмою, лог-мелспектрограма адаптована до особливостей слухового сприйняття акустичних сигналів людиною. Це робить її адекватним вибором для задач розпізнавання емоцій у ресурсно-обмежених системах.

Крім того, було розглянуто основні підходи до класифікації ознак отриманих з аудіосигналів, проаналізовано досвід їх використання в задачі SER. Серед розглянутих підходів були класичні методи машинного навчання (ансамблевий метод випадкового лісу), імовірнісні підходи на основі HMM, нейромережеві архітектури: MLP, CNN, RNN, а також сучасні

трансформерні та гібридні моделі. У задачах SER найбільш ефективними вважаються чотири групи моделей:

– згорткові, які забезпечують високу якість та обчислювальну ефективність завдяки аналізу локальних патернів;

– рекурентні (LSTM/GRU), здатні моделювати довготривалі часові залежності;

– трансформерні та ViT, що найбільш ефективно враховують глобальний контекст і міжфрагментні характеристики сигналу;

– гібридні моделі CNN+RNN та CNN+Transformer, що поєднують локальне виділення ознак і глобальне охоплення контексту та часто забезпечують найвищу точність на складних корпусах.

У даній роботі для виконання класифікації ознак було обрано згорткову ШНМ, оскільки вона дозволяє ефективно класифікувати двовимірні представлення аудіосигналів (яким є спектрограма), перевершуючи класичні методи машинного навчання. CNN оптимально обробляють просторово-залежні структури даних, зменшують обчислювальну складність завдяки операціям згортки і субдискретизації та забезпечують паралельний інференс, що є критичним для застосувань реального часу на Edge-пристроях. Крім того, вони мають кращу апаратну оптимізацію порівняно з рекурентними та трансформерними моделями, що робить їх найбільш придатним вибором для вбудованих систем.

2 РОЗРОБКА ВДОСКОНАЛЕНОГО МЕТОДУ КЛАСИФІКАЦІЇ ЕМОЦІЙ ЗА АУДІОДАНИМИ

Для досягнення мети роботи було запропоновано комбінований метод розпізнавання емоцій через обробку аудіоданих. Він складається з двох основних стадій: виділення ознак акустичного сигналу та їх класифікацію. Блок-схема алгоритмічної реалізації методу наведена на рисунку 2.1.

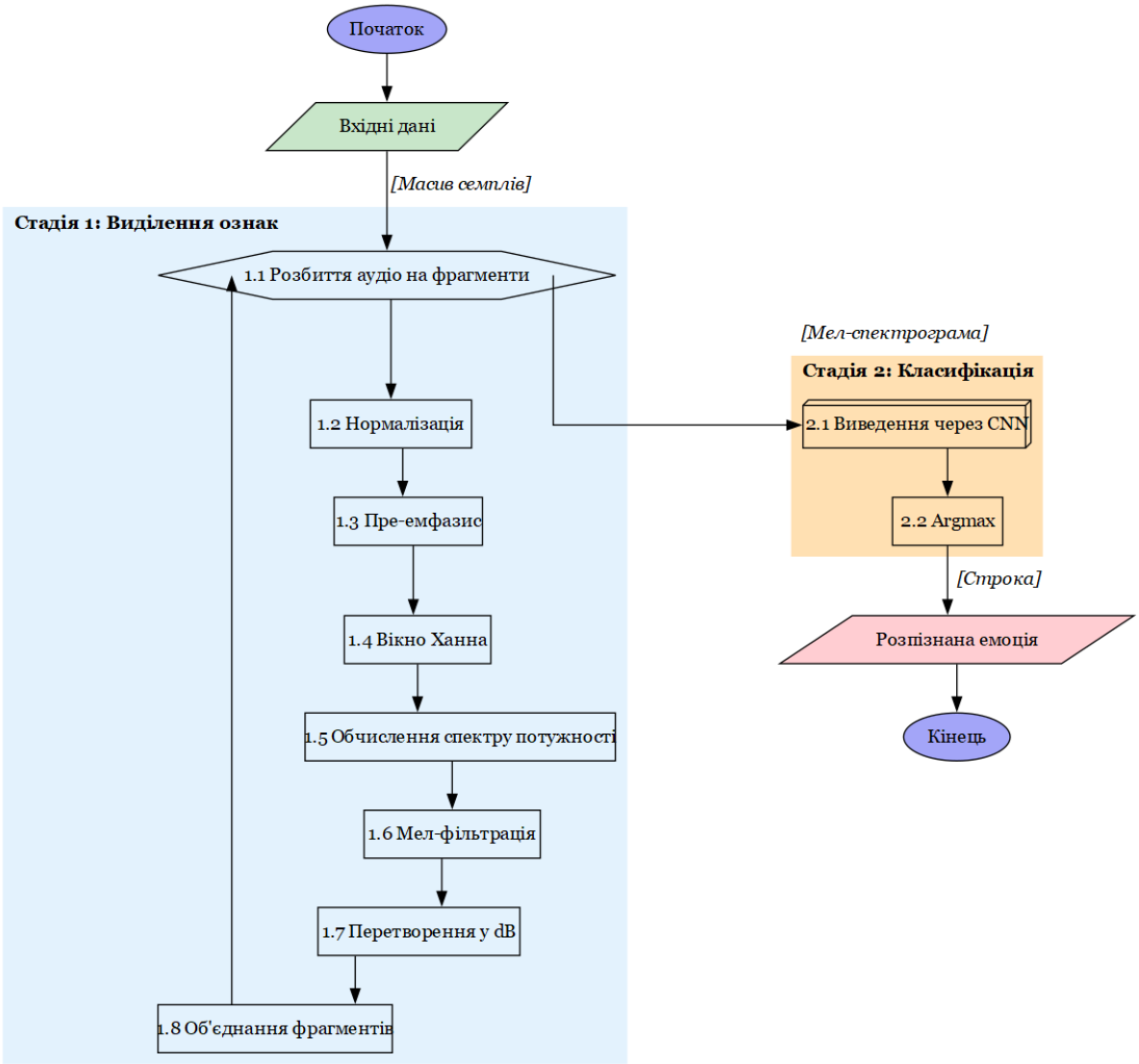


Рисунок 2.1 – Блок-схема алгоритму, що реалізує запропонований метод класифікації емоцій в акустичному сигналі

2.1 Розробка методу виділення ознак

Перша стадія методу розпізнавання емоційних станів має на меті попередню обробку вхідного аудіосигналу для перетворення його у більш ефективне представлення звукової інформації, зручне для аналізу згортковою нейронною мережею, через обчислення лог-мелспектрограми сигналу. Запропонований метод виділення ознак відбувається у вісім етапів.

2.1.1 Розбиття аудіо на фрагменти

Вхідний сигнал представляє собою необроблений масив семплів, у монофонічному форматі, тип даних – signed short (16-бітне цілочисельне значення зі знаком). Довжина аудіозапису становить 2,5 с, при частоті дискретизації 8 кГц, довжина масиву становить 20000 семплів (вхідна розмірність – 1×20000). Обробка сигналу виконується вікном розміром 3 мс, що становить 256 семплів. Таким чином вхідний масив поділяється на 78 повних фрагменти. Отже, реальна довжина аудіозапису, що використовується в обробці, становить 78·256/8000 = 2,496 с. Розмірність даних да цьому етапі – 256×78 значень. Подальша обробка (нормалізація, пре-емпфазис, вікнування, обчислення спектрограми, мел-фільтрація та перетворення у децибели) виконується циклічно над кожним рядком отриманої матриці (розмірністю 1×256).

2.1.2 Нормалізація аудіосигналу

Відповідно до типу вхідних даних signed short (цілочисельне 16-розрядне число), діапазон значень становить [-32768, 32767]. Метою нормалізації є приведення діапазону до [-1, 1] та перетворення типу даних у формат з плаваючою комою float32.

2.1.3 Пре-емпфазис

Пре-емпфазис аудіосигналу – це попередній етап обробки, що полягає у підсиленні високочастотних компонентів сигналу з метою компенсації спектрального спадання, характерного для мовлення, та сприяє покращенню співвідношення сигнал/шум для високих частот. Крім того, пре-емфазис дозволяє частково компенсувати вплив фізіологічних та акустичних факторів, які зумовлюють переважання низькочастотних компонент у мовному сигналі [32]. Вплив даного перетворення на вигляд спектрограми сигналу продемонстровано на рисунку 2.2.

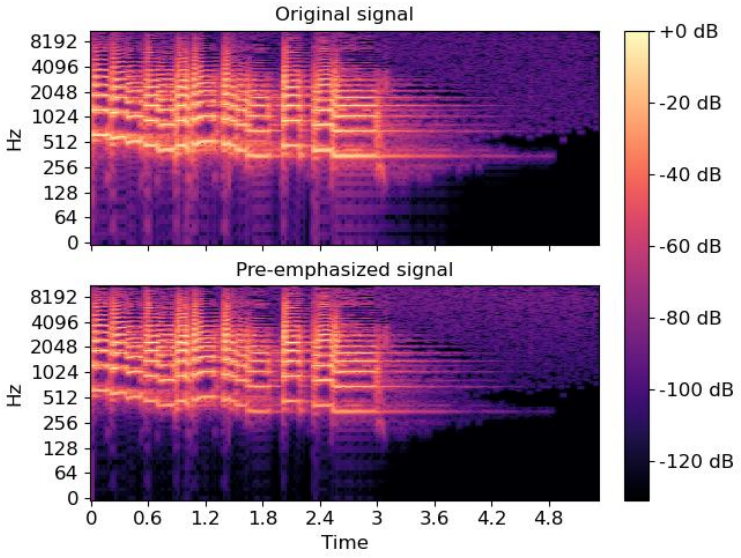


Рисунок 2.2 – Лог-спектрограма сигналу до та після застосування пре-емпфазису

Такий фільтр, як правило, реалізується як лінійна операція першого порядку:

$$y_n = x_n - \alpha x_{n-1}, \tag{2.1}$$

де α=0,97.

Отже, застосування пре-емпфазису підвищує ефективність подальших етапів спектрального аналізу у завданнях розпізнавання мовлення та емоцій.

2.1.4 Застосування віконної функції

На даному етапі до фрагменту сигналу застосовується вікно Ханна (рисунок 2.3) – віконна функція згладжування, яка застосовується операцією множення перед обчисленням спектральних характеристик, зокрема при перетворенні Фур'є. Його мета – зменшення ефектів витікання спектра (англ. spectral leakage), що виникають при дискретному аналізі скінчених за довжиною сигналів [33]. Вікно визначається як:

$$\omega_n = 0.5\left[1-\cos\left(\frac{2\pi n}{N-1}\right)\right], \tag{2.2}$$

де N – довжина вікна.

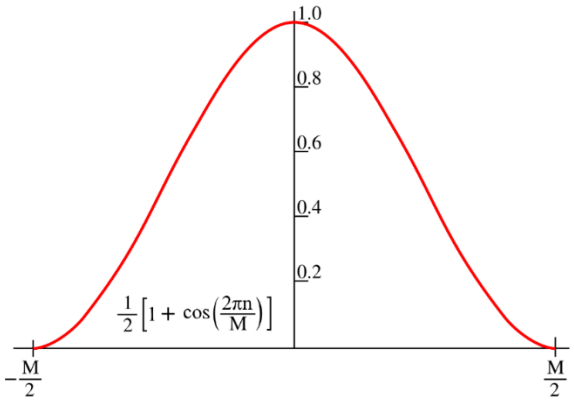


Рисунок 2.3 – Віконна функція Ханна

Під час дискретного аналізу сигнал фактично обрізається в часовій області, що еквівалентно його множенню на прямокутне вікно. Такий підхід призводить до появи різких розривів на межах аналізованого фрагмента, а

отже, до небажаного розширення спектральних компонент у частотній області. Використання вікна Ханна дає змогу плавно зменшити амплітуду сигналу до нуля на краях інтервалу спостереження, що суттєво знижує рівень бічних пелюсток у спектрі та покращує достовірність оцінювання частотного складу сигналу. Дана операція є стандартною процедурою у задачах цифрової обробки сигналів, зокрема акустичних.

2.1.5 Обчислення спектру потужності

Даний етап дозволяє визначити розподіл енергії сигналу за частотами. У рамках запропонованого підходу, до кожного фрейму сигналу довжиною 256 семплів застосовується нульове доповнення до 512 відліків, що забезпечує кращу частотну дискретизацію та дозволяє виконати перетворення у частотну область за допомогою алгоритму швидкого перетворення Фур'є відповідного розміру (формула 2.3). Після отримання комплексного спектру шляхом FFT обчислюється спектр потужності як квадрат модуля спектральних компонент (формула 2.4).

$$Y = FFT_{512}(X), \tag{2.3}$$

$$P_k = |Y_k|^2, k = 0,1,...,257, \tag{2.4}$$

де X – доповнений сигнал;

Y – результат обчислення FFT;

P – спектр потужності сигналу.

2.1.6 Застосування мел-фільтру

Застосування мел-фільтру є етапом перетворення спектру потужності аудіосигналу у форму, узгоджену з особливостями людського сприйняття

звуку. У даному дослідженні використовується набір з 78 мел-фільтрів, що представляє собою банк трикутних фільтрів, розташованих рівномірно у мел-просторі, але з нелінійною прив'язкою до частот у герцах (рисунок 2.4). Спектр потужності P_K зважується кожним фільтром H_{ik}, і результатом є мел-спектр M_i:

$$M_i = \sum_{k=0}^{N-1} P_k \cdot H_{ik}, i = 1,2,...,78. \quad (2.5)$$

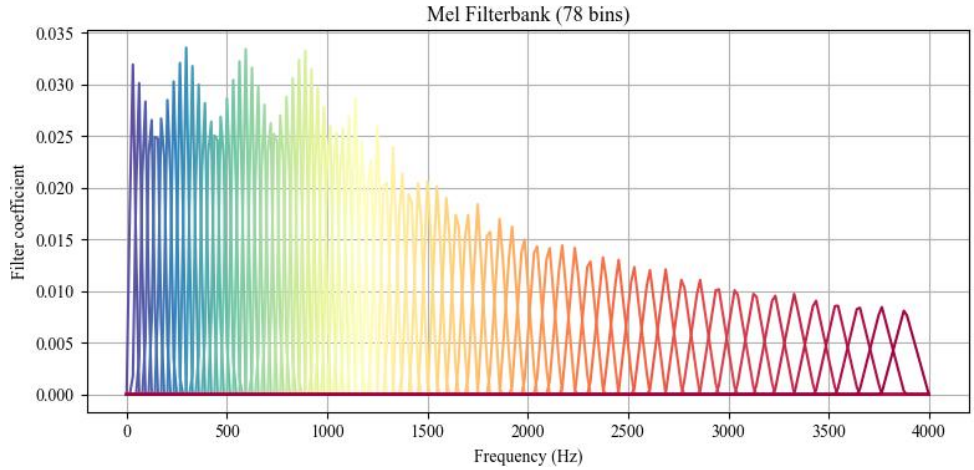


Рисунок 2.4 – Набір мел-фільтрів для 78 частотних бінів

Вибір кількості мел-фільтрів, рівної 78, не є типовим для класичних систем аналізу мовлення, оскільки у традиційних підходах до побудови набору мел-фільтрів найчастіше використовують кількість близько 40, що вважається достатнім для репрезентації мовного сигналу за помірної обчислювальної складності. З цієї точки зору використання 78 мел-бінів можна розглядати як потенційно надмірну деталізацію частотної осі. Водночас у даному дослідженні вибір такого значення зумовлений не лише спектральними міркуваннями, а й архітектурними особливостями згорткових нейронних мереж. Для більшості типових CNN принципово важливою є узгодженість розмірності вхідних тензорів, а на практиці візуальні представлення даних часто приводять до фіксованого, переважно

квадратного формату, оскільки це спрощує побудову мережі, пакетну обробку даних та подальше застосування стандартних операцій згортки і підвибірки. Саме тому в межах запропонованого підходу доцільним виявилося узгодити кількість мел-бінів із часовою розмірністю спектрограми, сформувавши квадратне вхідне представлення розмірністю 78×78. Альтернативним рішенням могло бути застосування нульового доповнення (zero padding), однак такий підхід вносить до вхідного зображення штучні, неінформативні області, а також може змінювати граничну статистику даних.

2.1.7 Перетворення у децибели

На даному етапі здійснюється перетворення енергії сигналу у логарифмічну шкалу децибелів, що дозволяє наблизити машинне представлення звукової інформації до особливостей людського слуху. Оскільки сприйняття гучності звуку людиною є логарифмічним, перетворення спектру потужності у децибели є необхідним кроком для підвищення інформативності ознак у задачах класифікації мовлення. Ця процедура дозволяє зменшити вплив неінформативних або шумових компонентів, зберігаючи найбільш значущі спектральні характеристики. Обчислення логарифмічного спектра L_i реалізовано за формулами 2.6–2.7:

$$L_i = 10 \cdot \lg\left(\frac{\max\left(P_i, a_{min}\right)}{\text{ref}}\right), \tag{2.6}$$

де P_i – значення спектру потужності після мел-фільтрації;

$a_{\min}$ – мінімальний поріг для уникнення логарифму нуля, $a_{\min} = 10^{-10}$;

ref – опорне значення, ref=1.

Після логарифмування виконується обмеження динамічного діапазону до заданого максимуму:

$$L_i = \max\left(L_i, L_{max} - \text{topdb}\right), \tag{2.7}$$

де topdb – максимальний допустимий динамічний діапазон, topdb=80 дБ;

L_{max} – максимальне значення логарифмічного спектру.

2.1.8 Об’єднання фрагментів спектру

У процесі аналізу повного аудіосигналу, перетворених фрагментів накопичується певна кількість (78), яка визначається довжиною сигналу, розміром вікна та кроком сегментації. У результаті всі вектори-стовпці об’єднуються по часовій осі, утворюючи вихідну матрицю ознак розмірності 78×78. Таке представлення слугує вхідним тензором для згорткових нейронних мереж, дозволяючи моделі ефективно враховувати як частотну, так і часову структуру звукового сигналу.

Таким чином, розмірність вхідних даних становитиме 1×20000, розмірність вихідного вектору ознак – 78×78, що означає зменшення інформаційного простору в майже 3,3 рази.

Для виділення ознак з цифрового аудіосигналу довжиною 19968 семплів (майже 2,5 с) алгоритм потребує виконання 5471076 арифметичних операцій (додавання, віднімання, множення, ділення, логарифмування). В пам’яті необхідно одночасно зберігати: коефіцієнти вікна Ханна, розріджене представлення коефіцієнтів мел-фільтрів, значення twiddle-коефіцієнтів, пошукову таблицю бітових перестановок, а також робочі буфери, включаючи вихідну матрицю ознак. При статичному виділенні пам’яті для всіх робочих буферів, необхідний об’єм пам’яті становить близько 538 КіБ. Однак, за умови динамічного керування виділенням пам’яті та видаленні даних, що не використовуються на кожному етапі, необхідний обсяг пам’яті зменшується до 319 КіБ. Проміжні розрахунки наведено в додатку А. Таким чином, навіть за

обмежених апаратних ресурсів алгоритм залишається придатним для практичної реалізації у системах реального часу.

2.2 Формування експериментального набору даних

У межах даного дослідження для навчання та тестування ефективності глибинної моделі було створено набір даних шляхом об’єднання кількох відкритих датасетів емоційного мовлення, що забезпечують достатню кількість якісних аудіозаписів. Вибір цих наборів даних ґрунтується на репрезентативності, різноманітності мовців, наявності необхідних емоційних класів та достовірності емоційного вираження. Нижче наведено перелік використаних наборів даних та їхню характеристику.

2.2.1 CREMA-D

Набір даних CREMA-D (Crowd-sourced Emotional Multimodal Actors Dataset) містить 7442 аудіо- та відеокліпи від 91 актора (48 чоловіків, 43 жінки) віком від 20 до 74 років. Усі записи виконані англійською мовою та включають як мовні, так і візуальні емоційні прояви, що робить цей датасет придатним для аналізу емоцій за допомогою аудіальних, візуальних або мультимодальних підходів.

Висловлювання представлені у шести базових емоційних категоріях: злість (angry), огида (disgust), страх (fear), радість (happy), смуток (sad), нейтральний стан (neutral). Кожна емоція (за винятком нейтральної) подана у чотирьох рівнях емоційної інтенсивності, що дозволяє моделювати не лише тип емоції, а й ступінь її вираженості. Анотація емоцій здійснювалася шляхом краудсорсингу, де кожен запис оцінювався кількома незалежними слухачами, що підвищує надійність розмітки [34]. Завдяки збалансованому розподілу емоцій, чіткій структурі анотацій та високій якості

записів, RAVDESS є поширеним набором даних для попереднього тестування моделей розпізнавання емоцій у мовленні. Розподіл зразків за емоційними класами є майже збалансованим: angry – 1271, disgust – 1271, fear – 1271, happy – 1271, sad – 1271, neutral – 1087.

2.2.2 RAVDESS

RAVDESS (Ryerson Audio-Visual Database of Emotional Speech and Song) – стандартизований аудіо-візуальний набір даних, призначений для досліджень у галузі розпізнавання емоцій у мовленні. Датасет містить 1440 аудіофайлів, записаних за участю 24 професійних акторів (12 жінок та 12 чоловіків), що створює гендерно-збалансовану вибірку.

Записи охоплюють сім емоційних категорій: спокій (calm), радість (happy), смуток (sad), злість (angry), страх (fear), здивування (surprise), огида (disgust). Крім того, до набору включено окремий клас нейтрального стану (neutral). Для всіх емоцій, за винятком нейтральної, передбачено два рівні інтенсивності. Усі записи виконані в контрольованих студійних умовах, що гарантує високу якість аудіосигналу та мінімальний рівень шуму [35]. Розподіл зразків за емоційними класами є майже збалансованим: calm – 384, happy – 384, sad – 384, angry – 384, fear – 384, disgust – 384, surprise – 384, neutral – 192. Завдяки чіткій розмітці зразків, збалансованому розподілу емоцій та високій якості записів, RAVDESS широко використовується для розробки моделей розпізнавання емоцій у мовленні.

2.2.3 SAVEE

Датасет SAVEE (Surrey Audio-Visual Expressed Emotion Database) являє собою аудіо-візуальний набір даних, розроблений для досліджень у галузі автоматичного розпізнавання емоцій у мовленні. Він містить 480

емоційно забарвлених висловлювань, записаних за участю чотирьох професійних британських акторів-чоловіків, що виконували фіксовані фрази англійською мовою.

Висловлювання охоплюють сім категорій: злість (anger), огида (disgust), страх (fear), радість (happiness), смуток (sadness), здивування (surprise) та нейтральний стан (neutral). Розмітка зразків була отримана на основі оцінювання 10 незалежними учасниками, які сприймали записи в трьох різних умовах: аудіо, відео та аудіовізуальній, що дозволило підвищити надійність анотацій [36]. Розподіл зразків за емоційними класами є майже збалансованим: neutral – 120, angry – 60, disgust – 60, fear – 60, happy – 60, sad – 60, surprise – 60. SAVEE характеризується порівняно невеликим обсягом даних та обмеженою кількістю акторів, однак широко використовується для дослідження підходів до розпізнавання емоцій.

2.2.4 TESS

TESS (Toronto Emotional Speech Set) – це повністю аудіальний набір даних, створений для досліджень у сфері розпізнавання емоцій у мовленні. Датасет ідеально збалансований, включає 2800 аудіофайлів, записаних за участю двох професійних акторок віком 26 та 64 роки. Під час запису вони вимовляли 200 окремих англійських слів, кожне з яких було озвучене у семи емоційних станах: злість (angry), огида (disgust), страх (fear), радість (happy), приємне здивування (pleasant), смуток (sad), нейтральний стан (neutral). Записи створювалися в умовах студії, що дозволило забезпечити високу якість аудіозапису та мінімальний рівень шуму [37]. Завдяки цьому, даний набір даних також широко використовується для розробки та порівняння класифікаційних моделей в задачі розпізнавання емоційних станів. Розподіл зразків за емоційними класами є повністю збалансованим: angry – 800, disgust – 800, fear – 800, happy – 800, neutral – 800, pleasant – 800, sad – 800.

2.2.5 Urdu Language Speech Dataset

Даний аудіальний набір даних створено для досліджень розпізнавання емоцій за голосом мовою урду. Датасет містить 400 емоційно забарвлених висловлювань, отриманих від 38 мовців (27 чоловіків та 11 жінок), що забезпечує гендерну різноманітність мовців [38]. Записи охоплюють чотири базові емоційні стани: злість (angry), радість (happy), смуток (sad), нейтральний стан (neutral).

Аудіозаписи були зібрані з ток-шоу мовою урду, що робить їх наближеними до природного, спонтанного мовлення порівняно зі студійними акторськими записами. Це підвищує цінність датасету для досліджень проявів емоцій у реальних умовах, але водночас може ускладнювати задачу класифікації через фоновий шум і високу варіативність. Розподіл зразків за емоційними класами є повністю збалансованим: angry – 100, happy – 100, neutral – 100, sad – 100.

2.2.6 EMODB

EMODB (Berlin Database of Emotional Speech) – це класичний аудіальний набір даних, широко використовуваний у задачах SER. Датасет містить 535 емоційних висловлювань, записаних за участю 10 професійних німецькомовних акторів (5 чоловіків та 5 жінок), які вимовляли заздалегідь підготовлені фрази в контрольованих студійних умовах. У наборі представлені сім емоційних категорій: злість (angry), нудьга (boredom), тривога/страх (fear), радість (happy), смуток (sad), огида (disgust), нейтральний стан (neutral). Записи створювалися з чітким контролем артикуляції, інтонації та емоційної виразності, а емоційна розмітка була додатково перевірена шляхом перцептивної оцінки, що дало можливість забезпечити високу якість анотацій [39]. Розподіл зразків за емоційними класами: angry – 127, boredom – 81, neutral – 79, happy – 71, fear – 69, sad – 62,

disgust – 46. EMODB характеризується відносно невеликим обсягом даних і нерівномірним розподілом емоцій, однак через високу якість записів та чітке відокремлення класів він залишається одним із базових еталонних наборів даних для порівняльної оцінки методів розпізнавання емоцій у мовленні.

2.2.7 RESD

RESD (Russian Emotional Speech Dialogs) – це відкритий мультимодальний набір даних емоційних діалогів, призначений для розв'язання задачі SER. Датасет містить приблизно 3,5 години живої діалогічної мови, записаної за участю акторів, які відтворювали емоційно забарвлені діалоги тривалістю близько 3 хвилин кожен [40]. У наборі представлені сім емоційних станів: злість (anger), огида (disgust), страх (fear), ентузіазм (enthusiasm), радість (happiness), нейтральний стан (neutral), смуток (sadness). Особливістю RESD є мультимодальність, а саме наявність як аудіозаписів мовлення, так і відповідних текстових транскриптів діалогів. Це дозволяє застосовувати як аудіальні, так і комбіновані (аудіально-текстові) підходи до аналізу емоцій. Діалоговий формат записів забезпечує більшу наближеність до реальних умов спілкування порівняно з ізольованими фразами чи словами. Розподіл зразків за емоційними класами: fear – 223, angry – 219, happy – 218, enthusiasm – 198, neutral – 191, disgust – 185, sad – 162.

2.2.8 MESD

MESD (Mexican Emotional Speech Database) – це аудіальний набір даних емоційного мовлення, призначений для досліджень в галузі SER для іспаномовних мовців із мексиканським мовним та фонетичним забарвленням. Датасет складається з однословних емоційно забарвлених

висловлювань, що робить його зручним для аналізу акустичних характеристик емоцій без впливу контексту довгих фраз. У наборі представлені шість емоційних станів: злість (angry), огида (disgust), страх (fear), радість (happy), смуток (sad), нейтральний стан (neutral). Записи були виконані як дорослими мовцями (3 жінки та 2 чоловіки), так і дітьми (6 мовців), що дозволяє створювати гендерно- та віково-інваріантні моделі. Усі аудіофайли записані у контрольованих умовах із чіткою артикуляцією та вираженою емоційною інтонацією [41]. Розподіл зразків за емоційними класами: disgust – 144, fear – 144, happy – 144, sad – 144, angry – 143, neutral – 143. Розподіл емоцій є майже повністю збалансованим, що також є важливою перевагою даного датасету.

2.2.9 Speech Noise Dataset

Speech Noise Dataset (Dataset for Speech and Noise Classification) – це аудіальний набір даних, призначений для вирішення задач класифікації мовлення та шуму, а також для досліджень у сфері покращення розбірливості голосу і попередньої обробки аудіосигналів. Він не відноситься до датасетів емоційного мовлення, однак у даному дослідженні використовується для отримання зразків шуму (зразків, що позначені як такі, що не містять мовлення). Датасет містить аудіофрагменти, що представляють як мовлення, так і різні типи шуму. Усі аудіофайли поділені на три класи: clean_speech (чисте мовлення без фонових завад), noisy_speech (мовлення з накладеним шумом), noise_only – виключно шум без мовленнєвого сигналу. Загальний обсяг датасету становить приблизно 3380 аудіофайлів [42]. Розподіл зразків: clean_speech – 1217, noisy_speech – 1217, noise_only – 946.

Вибір зазначених наборів даних можна обґрунтувати наступним чином. Першочергово було обрано CREMA-D, RAVDESS, SAVEE та TESS, оскільки вони містять велику кількість аудіофайлів із чітко визначеними

емоціями, записаними професійними акторами. Дані набори широко використовуються в дослідженнях у галузі розпізнавання емоційних станів через їхню якість та надійність. Набори EMODB, RESD, MESD, а також Urdu Language Speech Dataset були додані до вибірки задля збільшення мовного та культурного різноманіття, оскільки вони містять записи німецькою, російською, іспанською та мовою урду відповідно. Це сприяє врахуванню особливостей емоційного вираження у різних мовних спільнотах, а також зменшенню упередженості моделі до англомовних записів, забезпечуючи здатність працювати незалежно від мовного контексту.

Перші вісім наборів даних були об’єднані в один за допомогою програмного коду, наведеного в лістингу 1 додатку Б. Було перевірено та встановлено, що жоден з датасетів не містить пустих значень та некоректних міток класів. Стосовно останнього датасету, з нього вилучено лише зразки класу noise_only та додано до об’єднаного набору даних, призначивши мітку «noise». Розподіл класів у об’єднаному наборі даних наведено в таблиці 2.1.

Для подальшої оптимізації набору даних було виконано попередню обробку, спрямовану на зменшення кількості емоційних класів. З цією метою, було залишено 6 найбільш чисельних емоційних класів (відповідно до таблиці 2.1) та клас, що позначає відсутність мовлення. Емоційні категорії «calm», «enthusiasm» та «boredom» були рекласифіковані до класу «neutral», що дало можливість вирівняти баланс між нейтральними та емоційно забарвленими висловлюваннями у вибірці. Також з набору даних вилучено записи, що містять емоції «pleasant» та «surprise», для того щоб обмежити задачу до розпізнавання базових емоційних станах, які мають виражені акустичні особливості. На цьому етапі отримано корпус емоційного мовлення, що містить сумарно понад 19 тисяч зразків та охоплює 6 базових емоційних категорій: злість, радість, смуток, страх, огида, нейтральний стан та додатковий клас «noise». Розподіл класів в

очищеному датасеті наведено в таблиці 2.2 (стовпець 2). Було проведено аналіз тривалості аудіозаписів у отриманому наборі даних (рисунок 2.5). Модальне значення становить близько 2,5 с, що було обрано як цільова тривалість аудіозапису для виконання емоційного аналізу.

Таблиця 2.1 – Розподіл зразків за класами об’єднаної вибірки

Клас	К-ть зразків
angry	3104
happy	3048
sad	2983
fear	2951
disgust	2890
neutral	2712
noise	947
pleasant	800
surprise	444
calm	384
enthusiasm	198
boredom	81

Таблиця 2.2 – Розподіл зразків за класами

Клас	К-ть зразків			
	Після об’єднання	Після очищення	Після стратифікованого розподілення	
			Тренувальна	Тестувальна
neutral	3375	3666	3020	742
angry	3104	3329	3064	744
happy	3048	3201	2989	804
sad	2983	2824	2995	742
fear	2951	2999	3099	784
disgust	2890	3089	3060	761
noise	947	7574	3123	755

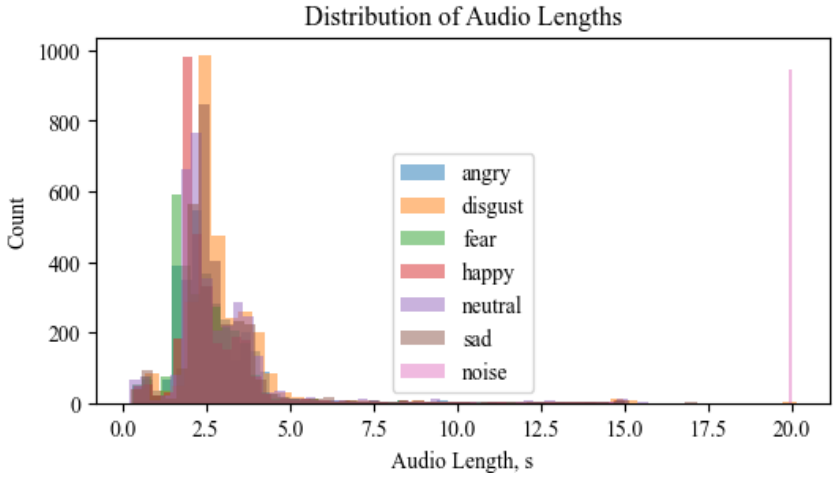


Рисунок 2.5 – Розподіл тривалості аудіозаписів у датасеті

Наступним етапом обробки набору даних є підготовка фрагментів аудіосигналу фіксованої тривалості. Попереднє прослуховування та перегляд графіків сигналу у зразках датасету показав, що вони містять значну частку немовленнєвих фрагментів, таких як тиша та шум навколишнього середовища. Тому під час поділу довгих аудіозаписів на фрагменти є ризик отримання неінформативних зразків, які знижуватимуть якість набору даних, що призведе до втрати точності під час навчання моделі. Для очищення даних було виконано аналіз кожного фрагменту з використанням методу виявлення голосової активності (VAD), після чого фрагменти, що не містять голосу, тривалістю понад 200 мс було видалено, тоді як мовленнєві фрагменти довжиною не менш, ніж 300 мс, було залишено. Отримані аудіозаписи тривалістю понад 2,5 с поділяються на фрагменти, записи меншої тривалості доповнюються нульовими значеннями до цільової тривалості. Фрагменти тривалістю менше 1 с ігноруються. Програмна реалізація даного алгоритму наведена в лістингу 2 додатку Б. Для виявлення голосової активності використано реалізацію VAD проєкту WebRTC [43]. В результаті отримано дещо незбалансовану вибірку, зі значною перевагою класу «noise» (таблиця 2.2, стовпець 3).

Наступним етапом підготовки даних є виділення ознак, що дозволяє отримати інформативні характеристики аудіозаписів для подальшого аналізу та навчання моделі. Для прикладу було переглянуто отриманий результат для кожного класу: зроблено візуалізацію форми вхідного аудіосигналу в часовому домені та перетвореної лог-мелспектрограми (рисунок 2.6).

Для підвищення стійкості моделі до варіативності мовлення та розширення обсягу навчального набору даних було застосовано аугментацію аудіоданих за методом SpecAugment [44] та додаванням випадкового гаусівського шуму. Даний метод включає зміну темпу мовлення, а також часове та частотне маскування спектрограм, що імітує природні коливання швидкості вимови, пропуски фрагментів сигналу та варіацію спектральних характеристик, зумовлену тембром голосу, акустичними умовами й апаратурою звукозапису. Додавання шуму, у свою чергу, моделює фонові завади реальних середовищ і підвищує стійкість моделі до шумових спотворень. У сукупності ці підходи сприяють запобіганню перенавчання та зменшують залежність моделі від локальних ознак, покращуючи її робастність та здатність до узагальнення. Програмна реалізація наведена у лістингу 3 додатку Б. Аугментація реалізована засобами пакету TorchAudio фреймворку PyTorch.

Наступним етапом є формування навчальної та тестової вибірок. Розбиття датасету виконано у співвідношенні 80%:20%, використовуючи стратифіковане розподілення. Також виконано балансування шляхом формування вибірки за допомогою зваженого випадкового семплера (лістинг 4 додатку Б). Вагові коефіцієнти, що визначають ймовірність вибору зразка для заданого класу, задано обернено пропорційними кількості зразків у цьому класі. Таким чином було згенеровано збалансований набір даних, що забезпечує виконання 150 навчальних епох. Розмір валідаційної вибірки склав 5332 зразки, навчальної – 21350 (стовпці 4 і 5).

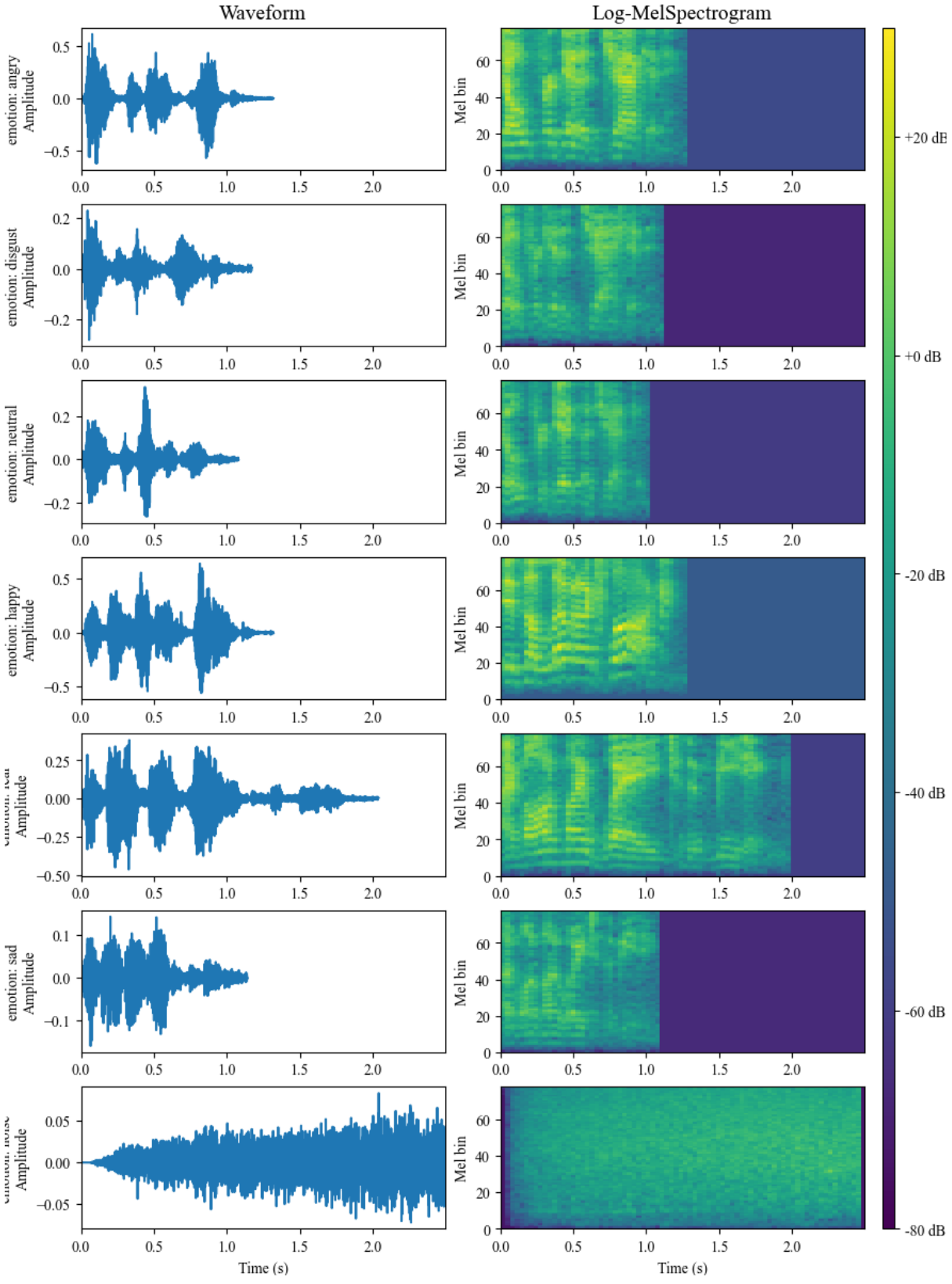


Рисунок 2.6 – Візуалізація необробленого сигналу (зліва) та лог-мелспектрограми (справа) для кожного класу вибірки

2.3 Вибір архітектури штучної нейронної мережі

Було проведено апробацію одинадцяти найбільш поширених згорткових нейромережевих архітектур з метою визначення найбільш ефективної в контексті розв’язання поставленої задачі розпізнавання емоційних станів. Зокрема, в якості базових архітектур було розглянуто:

– класичну CNN («Vanilla CNN»);

– MobileNet V1 та V2;

– DenseNet;

– ResNet;

– ShuffleNet V1 та V2;

– SqueezeNet;

– SE-Net (VGG-подібна CNN з SE-модулями);

– RepVGG;

– ConvMixer.

З урахуванням специфіки вхідних даних: невеликого розміру простору ознак, а також вимог до подальшого розгортання моделі на пристроях класу вбудованих систем – базові (референсні) архітектури було відповідним чином модифіковано. Основною метою такої адаптації було зменшення обчислювальної складності моделей та забезпечення їхньої ресурсної ефективності без значної втрати якості. Зокрема, архітектурні зміни здійснювалися таким чином, щоб загальна кількість параметрів кожної моделі не перевищувала 0,1 млн. Це відповідає обмеженням, характерним для систем із низькою обчислювальною потужністю та обмеженим обсягом пам’яті, якими є більшість сучасних мікроконтролерів.

Програмна реалізація кожної моделі побудована із забезпеченням можливості гнучкої зміни ширини та глибини, а також інших властивостей, специфічних для конкретної архітектури. Таким чином, було введено гіперпараметри, що визначають архітектуру моделі, які було підібрано емпіричним шляхом для досягнення фіксованої кількості параметрів. Нижче

розглянуто кожну модифіковану архітектуру з її відмінностями від базової реалізації.

2.3.1 Класична CNN

Ця модель являє собою класичну згорткову нейронну мережу, використану як базову архітектуру для подальшого порівняння з іншими спеціалізованими моделями.

Її структура побудована за послідовним принципом і включає стек згорткових блоків, кількість яких визначається параметром глибини мережі (depth=3). Кожен блок містить двовимірний згортковий шар із ядром розміру 3×3, за яким, залежно від конфігурації, може слідувати шар пакетної нормалізації та нелінійна активація ReLU. Просторове зменшення розмірності виконується або за допомогою операції максимального пулінгу, або шляхом використання згорток з кроком stride=2. Кількість каналів у мережі поступово зростає від початкового значення base_channels=32 із масштабуванням через параметр width_mult=1, що дає можливість реалізувати гнучке налаштування моделі відповідно до ресурсних обмежень. На завершальному етапі застосовується глобальний адаптивний середній пулінг, який агрегує просторові ознаки до компактного векторного представлення, після чого виконується класифікація.

Подібна архітектура часто використовується як референтна (baseline) модель для оцінювання ефективності альтернативних підходів. Програмна реалізація наведена в лістингу 5 додатку Б. Схематичне зображення архітектури моделі наведено на рисунку 1 додатку В.

2.3.2 MobileNet-подібні архітектури

У межах дослідження також було апробовано моделі MobileNet V1 та MobileNet V2, адаптовані до задачі обробки спектрограм. Перша з них

базується на використанні розділюваних згорток, у межах яких стандартна згортка декомпонується на просторову поканальну згортку та подальшу лінійну міжканальну проєкцію [45] Такий підхід істотно зменшує кількість параметрів і обчислювальну складність порівняно з класичними згортковими ШНМ, що робить модель ефективною для використання на ресурсно-обмежених платформах.

Архітектура MobileNet V2 [46] реалізує більш сучасний підхід на основі інвертованих залишкових блоків (inverted residual blocks), а також вузьких лінійних представлень, де спочатку виконується розширення простору ознак.

На відміну від референсних реалізацій MobileNet, у даному дослідженні обидві архітектури були адаптовані до одноканального вхідного представлення аудіосигналів. Також застосовано глобальний адаптивний усереднюючий пулінг, що забезпечує незалежність класифікатора від фіксованого просторового розміру карти ознак, а завершальний класифікатор реалізовано у вигляді одного повнозв’язного шару. Для першої версії MobileNet взято значення параметрів ширини та глибини: 0,25 та 0,2, для другої – 0,15 та 0,2 відповідно. Програмна реалізація наведена в лістингах 6–7 додатку Б. Схематичне зображення архітектури моделі наведено на рисунках 2–3 додатку В.

2.3.3 DenseNet-подібна архітектура

Модифікована модель DenseNet* реалізує DenseNet-подібну архітектуру, введену як одну з альтернатив для порівняння з базовою CNN. Її основою є принцип щільних міжшарових з’єднань, відповідно до якого кожний новий шар усередині щільного блоку отримує на вхід конкатенацію ознак, сформованих усіма попередніми шарами цього блоку. Це сприяє повторному використанню ознак та підвищує параметричну ефективність моделі.

Кожний щільний шар побудований за послідовністю: пакетна нормалізація, активація ReLU, згортка 1×1, пакетна нормалізація, активація ReLU, згортка 3×3. Тобто використовується структура з вузьким місцем для попереднього зменшення проміжної розмірності ознак. Приріст числа каналів контролюється параметром growth_rate=16. Між щільними блоками розміщено шари, які виконують стискання (зниження кількості каналів) за допомогою згортки 1×1 з коефіцієнтом compression=0,5 та подальше зменшення просторової розмірності через усереднюючий пулінг. На відміну від референсного виконання DenseNet [47], модифікована архітектура використовує компактний початковий згортковий шар 3×3 для одноканального вхідного представлення. Також зменшено кількість блоків і шарів у них (конфігурація шарів: 2, 3, 2), знижено початкову ширину мережі, а класифікатор реалізовано у вигляді глобального адаптивного усереднюючого пулінгу та одного повнозв'язного шару. Дана модифікація зберігає концепцію щільного поширення ознак, але має спрощену структуру та пристосована до роботи з малорозмірними вхідними представленнями. Програмна реалізація наведена в лістингу 8 додатку Б, схематичне зображення – на рисунку 4 додатку В.

2.3.4 ResNet-подібна архітектура

Дана модифікована модель являє собою компактну ResNet-подібну архітектуру, адаптовану для задачі класифікації спектрограм. Її основою є послідовність залишкових блоків (ResidualBlock), кожен з яких містить дві згортки з ядром 3×3 з пакетною нормалізацією, а також shortcut-з'єднання, яке забезпечує додавання вхідного сигналу до виходу блоку.

На відміну від канонічної реалізації ResNet [48], у запропонованій моделі використано одноканальний вхід, компактний початковий згортковий шар 3×3 без початкового max-пулінгу, зменшену кількість залишкових рівнів (3 рівні з кількістю залишкових блоків 1, 2, 2), а також

кероване масштабування ширини через параметр width_mult=0,75. Голова моделі складається з глобального адаптивного усереднюючого пулінгу та одного повнозв’язного шару класифікації. Програмна реалізація моделі наведена в лістингу 9 додатку Б, схематичне зображення – на рисунку 5 додатку В.

2.3.5 ShuffleNet-подібні моделі

Модель ShuffleNet V1* реалізує полегшену архітектуру, що базується на поєднанні pointwise group convolution, depthwise convolution та операції channel shuffle, що забезпечує міжканальний обмін інформацією за низької обчислювальної складності. У модифікованій реалізації кожен базовий блок містить групову згортку 1×1, операцію перестановки каналів, поканальну згортку 3×3 та завершальну групову згортку 1×1, а при зменшенні просторової розмірності використовується shortcut-гілка з усереднюючим пулінгом та конкатенацією ознак. Також реалізація відрізняється від референсної конфігурації [49], тим, що вона адаптована до одноканального вхідного представлення, не містить початкового max-пулінгу після першої згортки, та використовує адаптивний глобальний середній пулінг перед класифікатором. Також було введено параметр width для масштабування ширини стадій, а кількість блоків у стадіях винесено в параметр depths. Загалом модель має 3 стадії шириною 96, 192, 384 і кількістю ShuffleNet-блоків 2, 4 і 2 відповідно.

Модель ShuffleNet V2* використовує розщеплення каналів (channel split), обробку лише частини ознак у нелінійній гілці з конкатенацію та операції channel shuffle, що відповідає практичним рекомендаціям щодо побудови оригінальної архітектури [50]. У запропонованій модифікації блок із кроком 1 розділяє вхідні канали на дві частини, одна з яких передається без змін, тоді як інша проходить послідовність згорток: 1×1, 3×3-depthwise, 1×1, після чого виконується об’єднання

результатів та міжканальне перемішування. Порівняно з референсною архітектурою ShuffleNetV2, модифікована версія також адаптована до одноканального входу, не використовує початковий max-пулінг, не містить фінального згорткового шару 1×1 перед глобальним пулінгом, а застосовує адаптивний двовимірний усереднюючий пулінг та один повнозв'язний класифікаційний шар. Аналогічно до попередньої моделі, дана реалізація має параметри ширини: 50, 100, 200 та глибини: 2, 4, 2. Програмна реалізація міститься в лістингах 10–11 додатку Б, схематичні зображення – на рисунках 6–7 додатку В.

2.3.6 SqueezeNet-подібна архітектура

Модель SqueezeNet* реалізує SqueezeNet-подібну архітектуру, для якої характерне попереднє стискання ознак за допомогою згортки 1×1 та їх подальше розширення паралельними гілками зі згортками 1×1 і 3×3 з наступною конкатенацією. На відміну від оригінальної архітектури SqueezeNet, описаної в [51], запропонована реалізація є істотно спрощеною. Зокрема, було використано одноканальний вхід, початкову згортку 3×3 і один шар max-пулінгу. Кількість Fire-блоків визначається параметром depth=4, що дозволяє гнучко змінювати глибину мережі. Крім того застосовано параметри width_mult=2, base_channels=64 і squeeze_ratio=0,25, що забезпечують масштабування ширини моделі та ступеня стискання. Також замість фінальної згортки-класифікатора, використано адаптивний глобальний усереднюючий пулінг та повнозв'язний шар, що спрощує роботу моделі з довільними розмірами вхідних ознак. Програмна реалізація наведена в лістингу 12 додатку Б, схематичне зображення – на рисунку 8 додатку В.

2.3.7 SE-Net

Розроблена модель SE-Net побудована як послідовна згорткова архітектура, у якій вхідне одноканальне представлення послідовно проходить через стек із SE-Conv-блоків, кожен з яких містить згортку 3×3, пакетну нормалізацію, модуль Squeeze-and-Excitation (SE) та активацію ReLU, після чого виконується зменшення просторової розмірності за допомогою двовимірного max-пулінгу. Кількість каналів у мережі зростає каскадно від значення base_channels=16 з подвоєнням після кожного блоку, а завершальна частина моделі складається з глобального адаптивного середнього пулінгу та одного повнозв'язного шару класифікації, що робить архітектуру простою, масштабованою та придатною для використання в умовах обмежених обчислювальних ресурсів.

Власне SE-блок повторює структуру оригінального модуля SENet [52]: на етапі squeeze виконується глобальне усереднення по просторових координатах, а на етапі excitation – нелінійне перетворення з коефіцієнтом редукції r за допомогою двох повнозв'язних шарів, після чого одержані значення використовуються для поканального масштабування карт ознак. SE-модуль не є самостійною архітектурою, а розглядається як надбудова над вже існуючими архітектурними блоками (ResNet, ResNeXt, Inception тощо). У даній модифікації SE-блок застосовується у простішому послідовному SE-Conv-блоці. Крім того, дана архітектура адаптована до одноканального входу, містить відносно невелику кількість блоків (4) і використовує двовимірний max-пулінг після кожного SE-Conv-блоку. Також варто зазначити, що в референсній реалізації SE-Net використано низку оптимізацій, зокрема спеціальний шар для поєднання операцій channel-wise scaling та element-wise summation у мережах зі shortcut-з'єднаннями, тоді як у запропонованій модифікації вони відсутні.

Програмна реалізація наведена в лістингу 13 додатку Б, схематичне зображення – на рисунку 9 додатку В.

2.3.8 RepVGG-подібна архітектура

Дана модель реалізує RepVGG-подібну архітектуру, у якій використано ідею структурної репараметризації. Під час навчання кожен RepVGG-блок має багатогілкову топологію, що складається з основної гілки (згортка 3×3), паралельної гілки (згортка 1×1) та, за умови збігу розмірностей, shortcut-гілки, усі з пакетною нормалізацією. Після завершення навчання ці гілки згортаються в одну еквівалентну згортку 3×3 зі зсувом, що в результаті утворює більш просту та швидку VGG-подібну модель. Порівняно з референсною архітектурою [53], запропонована модель є суттєво спрощеною та адаптованою до одноканального вхідного тензору ознак і складається з початкового шару stage0 та трьох наступних стадій із RepVGG-блоків (проти 5 в оригінальній реалізації). Також модифікована версія має параметри глибини (1) і base_channels (64, 128, 256 для кожної з трьох стадій), що забезпечує гнучке масштабування моделі. Класифікаційна частина також була спрощена: після послідовного проходження трьох стадій застосовується двовимірний усереднюючий пулінг та один повнозв’язний шар. Програмна реалізація моделі наведена в лістингу 14 додатку Б, схематичне зображення – на рисунку 10 додатку В.

2.3.9 Архітектура на основі ConvMixer

Запропонована модифікована модель реалізує ConvMixer-подібну архітектуру, основна ідея якої – робота з patch-embedding представленням, розділення просторового змішування та канального змішування, а також підтримання сталої просторової роздільності ознак упродовж усього стеку основних блоків. У запропонованій модифікації початкове одноканальне

представлення перетворюється у простір ознак за допомогою згортки з ядром і кроком, рівними patch_size, після чого до нього застосовується послідовність блоків ConvMixer. Кожен такий блок містить depthwise convolution для просторового змішування із залишковим з'єднанням, а також pointwise convolution 1×1 для канального змішування. Як і в оригінальній імплементації [54], в обох підблоках використовуються пакетна нормалізація та нелінійна активація GELU. Однак, модифікована модель істотно відрізняється шириною і глибиною. Зменшено кількість блоків (depth=4) та значення patch_size=3 і kernel_size=5, що робить модель придатною для ресурсно-обмежених застосувань. Крім того, на відміну від референсної реалізації, де в шарі patch-embedding використовується схема згортка, активація, пакетна нормалізація та GELU, у адаптованій моделі застосовано лише пакетну нормалізацію та активацію ReLU. Класифікаційна частина моделі також була модифікована та складається з адаптивного глобального усереднюючого пулінгу і одного повнозв'язного шару. Програмна реалізація наведена в лістингу 15 додатку Б, схематичне зображення – на рисунку 11 додатку В.

У межах дослідження було виконано навчання з валідацією вищезгаданих 11 згорткових моделей, для того щоб провести їх комплексне порівняння за єдиних експериментальних умов. Для оцінювання якості класифікації було використано набір метрик, які відображають різні аспекти роботи моделей:

– accuracy (точність), макро- та зважена;

– precision (прецизійність), макро- та зважена;

– recall (повнота), макро- та зважена;

– F1-score (F1-міра), макро- та зважена;

– confusion matrix (матриця помилок).

Точність тут застосована як базова інтегральна характеристика загальної частки правильно класифікованих зразків. Макро- та зважені варіанти точності і F1-міри дозволяють врахувати дисбаланс між класами та

оцінити якість розпізнавання з точки зору рівномірності роботи моделі на всіх класах. Обчислення прецизійності та повноти забезпечило більш детальний аналіз здатності моделі коректно ідентифікувати об’єкти кожного класу та мінімізувати хибні позитивні та хибні негативні спрацювання. Матриця помилок була використана для поглибленого аналізу структури помилок і виявлення пар класів, між якими спостерігається найбільша кількість хибних спрацювань. Такий набір метрик забезпечує не лише узагальнену оцінку точності моделей, але й дозволяє всебічно охарактеризувати їхню поведінку в задачі багатокласової класифікації.

Моделі було навчено протягом максимум 150 епох з розміром пакету 128, коефіцієнтом навчання 0,001 та використанням оптимізатору Adam. Для оптимізації використано кросентропічну функцію втрат. Також застосовано планувальник, що зменшував швидкість навчання вдвічі на плато (в якості цільової метрики взято значення функції втрат при валідації). Додатково використано механізм ранньої зупинки, який завершував процес тренування при відсутності покращення метрики макро-F1 на етапі валідації. Саме ця метрика була обрана, оскільки вона однаковою мірою враховує внесок кожного класу незалежно від його представленості у вибірці та є більш інформативною в умовах багатокласової класифікації та потенційного дисбалансу класів. Для запобігання перенавчанню застосовано L2-регуляризацію з параметром 0,0001 та згладжування міток з параметром 0,01. Програмна реалізація тренувального циклу наведена в лістингу 16 додатку Б. Графіки метрик під час навчання та валідації наведені в додатку Г. Матриці помилок наведено в додатку Д. Для програмної реалізації навчального циклу використано фреймворк Pytorch, модуль TorchMetrics для обчислення метрик, а також програмний засіб TensorBoard для візуалізації метрик у реальному часі.

Додатково проведено тестування моделей, під час якого було розраховано кількість параметрів моделі, а також часову та об'ємну складність інференсу. Програмна реалізація наведена в лістингу 17 додатку Б. Для виконання обчислень використано вбудовані засоби фреймворку Pytorch. Діаграма результатів обчислення метрик для кожної з моделей наведено на рисунку 2.7. Діаграми порівняння моделей за кількістю параметрів, об'ємною та обчислювальною складністю наведено на рисунках 2.8–2.10.

Найвищу точність (accuracy=0,79) продемонстрували моделі на основі ConvMixer, SqueezeNet та SENet. Вони також показують збалансовані значення макро- та зваженої F1-міри, що свідчить про стабільну роботу як на збалансованих, так і на потенційно незбалансованих класах. Класична CNN має найнижчі показники (близько 0,69 за всіма метриками), що підтверджує важливість сучасних архітектурних оптимізацій. Отже, складніші вдосконалені архітектури забезпечують приріст точності у межах 8–10% порівняно з базовим згортковим стеком.

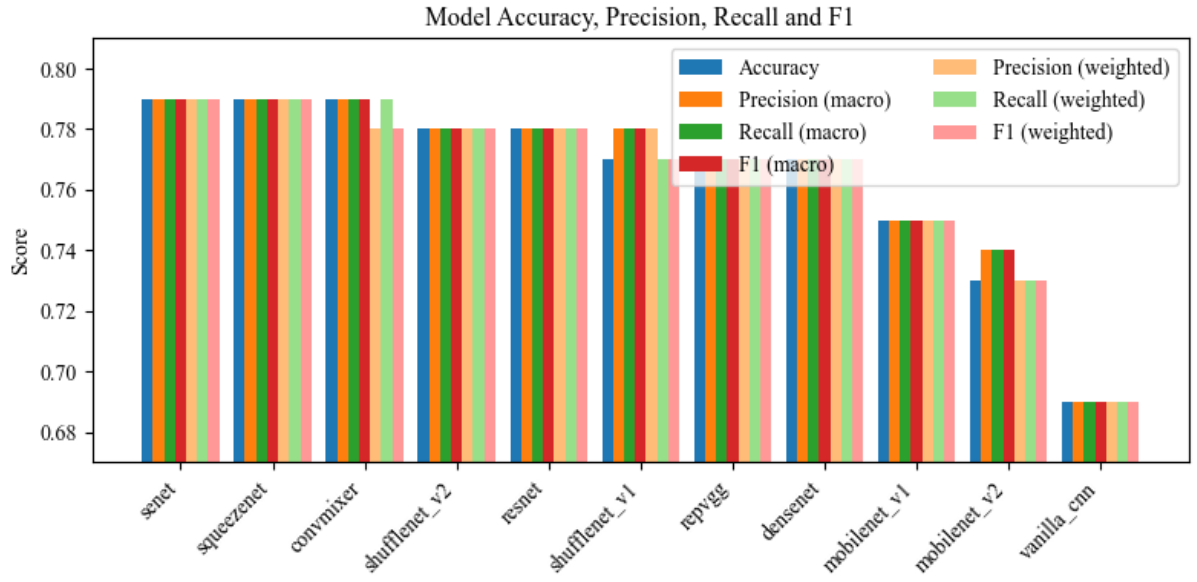


Рисунок 2.7 – Порівняння ефективності модифікованих моделей

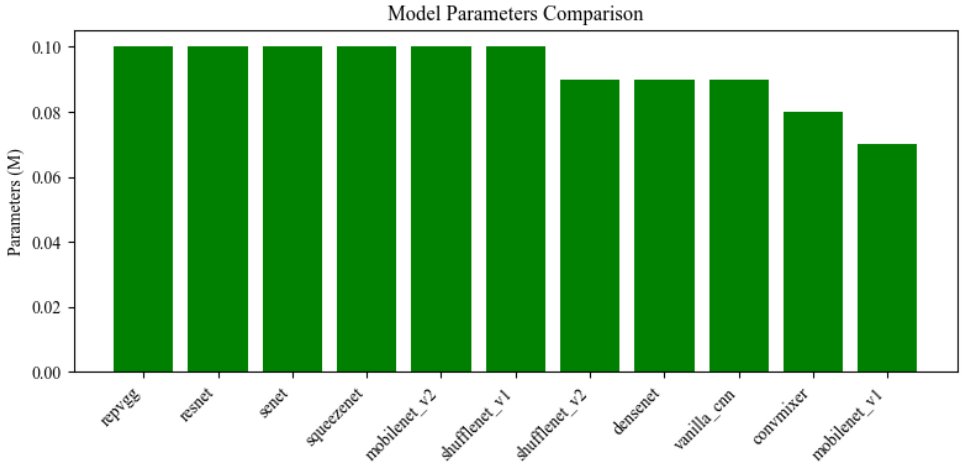


Рисунок 2.8 – Порівняння моделей за кількістю параметрів

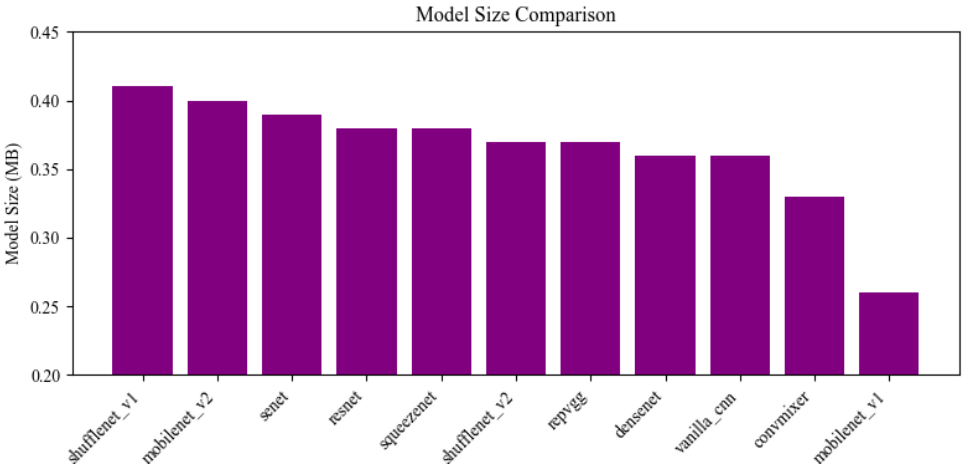


Рисунок 2.9 – Порівняння моделей за об'ємною складністю

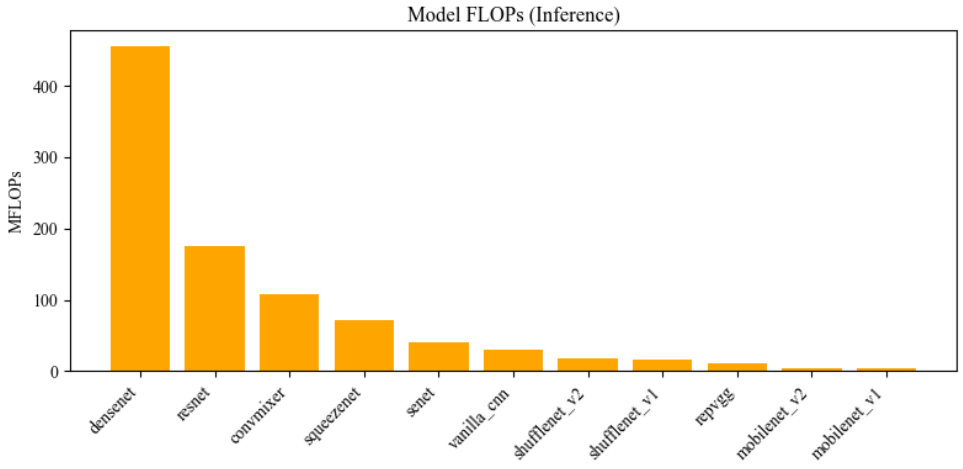


Рисунок 2.10 – Порівняння моделей за обчислювальною складністю

Найбільш обчислювально важкою згортковою архітектурою є DenseNet-подібна модель (455 MFLOPs), проте її точність є порівняно низькою (accuracy=0,77). Моделі на основі ResNet і ConvMixer також мають високу обчислювальну складність (176 та 107 відповідно), при цьому ConvMixer показує вищу точність. Найефективнішими моделями в контексті складності обчислень є модифікації MobileNet V1 (менше 4 MFLOPs), MobileNet V2 (майже 5 MFLOPs), RepVGG (12 MFLOPs), ShuffleNet V1/V2 (приблизно 17 MFLOPs). Отже, висока часова складність не гарантує кращої точності, тобто деякі архітектури мають вище співвідношення «якість–складність». Усі моделі знаходяться у вузькому діапазоні кількості параметрів (0,07–0,10 млн), однак необхідний обсяг пам’яті варіює у межах від 0,26 до 0,41 МіБ. Модифікована версія MobileNet V1 має найменшу об’ємну складність (0,26 МіБ) з цілком прийнятною точністю (accuracy=0,75). Також, ConvMixer-подібна модель має один із найменших обсягів (0,33 МіБ) при високій максимальній точності (accuracy=0,79). Підсумкове порівняння одинадцяти моделей за чотирма характеристиками у вигляді spider-діаграм наведено на рисунку 2.11.

Квантування моделей ШНМ є ефективним засобом зменшення обчислювальної складності та збереження пам’яті шляхом зменшення розрядності параметрів і активацій [55]. Найпоширенішими методами квантування є:

– посттренувальне квантування (англ. Post-Training Quantization, PTQ): проводиться після завершення навчання моделі без додаткового донавчання, характеризується простотою та швидкістю впровадження;

– квантування під час навчання (англ. Quantization-Aware Training, QAT): моделює вплив низької розрядності під час тренування, що дозволяє досягти вищої точності, але потребує повторного навчання.

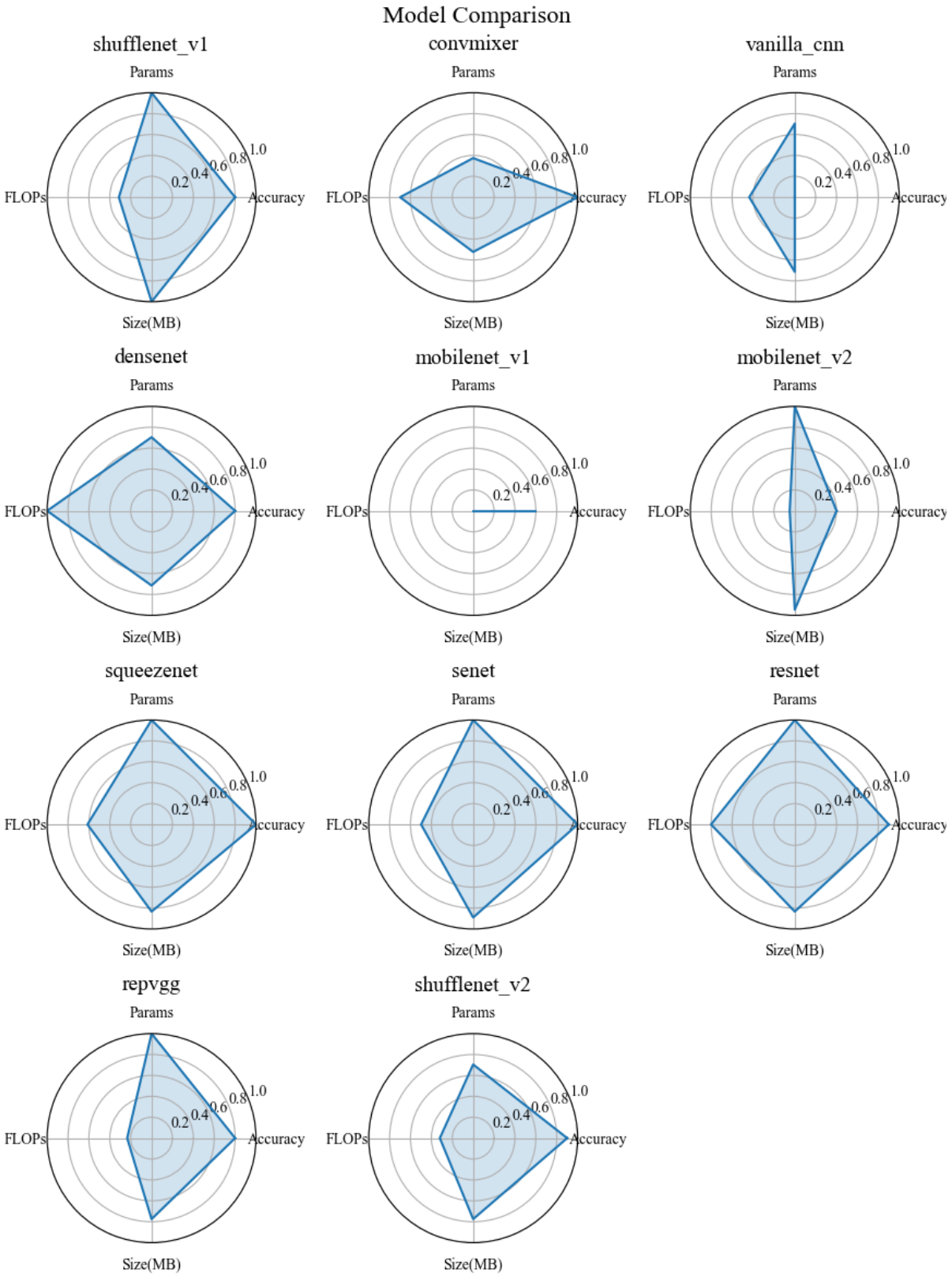


Рисунок 2.11 – Порівняння ефективності моделей за Spider-діаграмами

Статичне посттренувальне квантування вагів і активацій виконується на основі калібрувальної вибірки, що дозволяє заздалегідь оцінити діапазони активацій. Для цього часто застосовуються методи підбору порогів, зокрема метод KL-divergence, що сприяє зменшенню втрат інформації. Динамічне PTQ-квантування передбачає квантування ваг наперед, тоді як статистики активацій обчислюються безпосередньо під час інференсу, завдяки чому відпадає необхідність у калібрувальних даних, але зазвичай характеризується меншою точністю. Також застосовується квантування зі змішаною точністю (англ. Mixed-precision Quantization), що поєднує квантування різної розрядності в різних шарах мережі для забезпечення балансу швидкодії та точності.

З урахуванням апаратних обмежень обрано статичне пост-тренувальне квантування до цілочисельного восьмирозрядного типу даних int8 з використанням методу KL-divergence. Цей метод дозволяє ефективно зменшити розмір моделі до 4 разів, прискорити інференс завдяки цілочисельним 8-бітним операціям, а також не вимагає повторного тренування. У даному дослідженні апробовані моделі було піддано пост-тренувальному квантуванню засобами фреймворку PPQ після завершення навчання. Квантування не вдалось виконати для моделей на основі ShuffleNet та SE-Net через обмеження використаного фреймворку щодо можливостей роботи з операціями, застосованими у цих архітектурах.

Результати оцінки достовірності класифікації квантованих моделей за різними метриками наведено на рисунку 2.14. Отже, посттренувальне int8-квантування моделей призвело до помірного зниження якості класифікації, при цьому втрата точності становить близько 1–7% і залежить від архітектури моделі. Найбільш стійкими до зменшення розрядності виявилися моделі на основі MobileNet, ResNet, SqueezeNet та RepVGG, тоді як складніші архітектури на основі ConvMixer і DenseNet продемонстрували більшу деградацію метрик. Близькі значення макро- та зважених метрик свідчать про збереження класової збалансованості після квантування. В

цілому, найкращою з точки зору точності (0,78), і досить ефективною в контексті кількості параметрів (0,10 млн.) і обсягу необхідної пам’яті (0,38 МіБ) є модель на основі архітектури SqueezeNet.

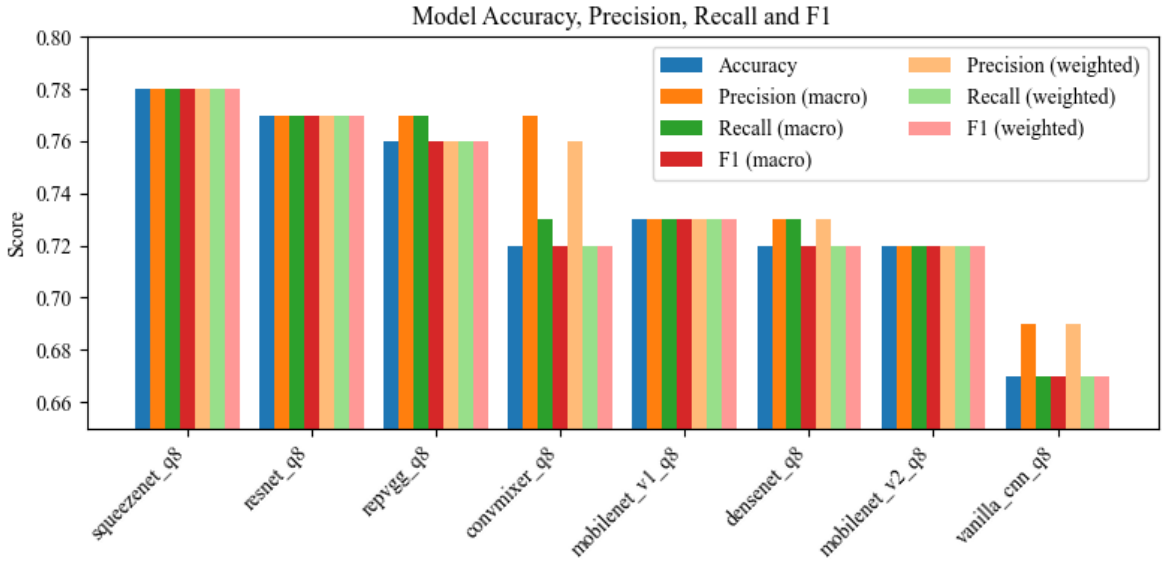


Рисунок 2.14 – Порівняння ефективності квантованих моделей

Висновки до розділу 2

Даний розділ кваліфікаційної роботи присвячено розробці методу розпізнавання емоційних станів людини за аудіоданими на основі лог-мелспектрограми та глибокої згорткової ШНМ. Обрано ефективний метод виділення інформативних ознак мовлення через обчислення лог-мелспектрограми, що має низьку обчислювальну та об’ємну складність та може бути застосований на пристроях класу вбудованих систем. Попередня обробка сигналу здійснюється через етапи пре-емфазису, застосування вікна Ханна, розрахунку спектру потужності, застосування мел-фільтрів та переведення в логарифмічну шкалу.

Було розглянуто дев’ять датасетів емоційного мовлення, що знаходяться у відкритому доступі, серед яких RAVDESS, CREMA-D, TESS

та інші, здійснено їх аналіз за параметрами якості, балансу емоцій, форматів аудіо та умов запису. Сформовано узагальнений набір аудіозаписів із достатнім рівнем різноманітності та інформативності шляхом об’єднання кількох датасетів. Підготовлено вибірки для навчання та тестування ШНМ за принципом стратифікованої розбивки. Виконано аугментацію даних шляхом додавання гаусівського шуму та застосування методу SpecAugment. Останній дозволяє варіювати темп мовлення та застосовувати часове та частотне маскування спектрограми. Загалом створено набір даних, одна епоха якого містить понад 21 тис. зразків для навчальної та понад 5 тис. зразків для тестувальної вибірок.

Запропоновано одинадцять варіантів архітектури глибинних згорткових ШНМ, адаптованих під апаратні обмеження обчислювальної платформи. Серед них: класична CNN та моделі на основі MobileNet, DenseNet, ResNet, ShuffleNet, SqueezeNet, SE-Net, RepVGG та ConvMixer. В результаті апробації модифікованих моделей шляхом тренування та тестування було отримано порівняльну характеристику за кількома метриками достовірності класифікації, кількістю параметрів, обчислювальною та об’ємною складністю. Найкращі результати продемонстрували моделі SqueezeNet, SE-Net та ConvMixer, досягши точності приблизно 0,79 на тестовій вибірці. Досліджено вплив посттренувального квантування моделей на якість класифікації. Застосовано статичне int8-квантування з метою оптимізації часової складності та зменшення обсягу пам’яті для виконання інференсу. За результатами експериментального дослідження було обрано модифіковану модель на основі SqueezeNet, що забезпечує прийнятний компроміс між якістю класифікації та складністю реалізації.

3 РЕАЛІЗАЦІЯ ІНФОРМАЦІЙНОЇ ТЕХНОЛОГІЇ КЛАСИФІКАЦІЇ ЕМОЦІЙ У ФОРМАТІ ВБУДОВАНОЇ СИСТЕМИ

3.1 Вибір апаратних та програмних засобів

3.1.1 Апаратні засоби

Наступним етапом є вибір апаратних засобів для реалізації та розгортання розробленого методу розпізнавання емоційних станів. У сучасних вбудованих системах застосування мікроконтролерів (МК) із різним рівнем продуктивності визначається балансом між обчислювальними можливостями, енергоспоживанням та наявністю периферійних інтерфейсів. У таблиці 3.1 наведено деякі з популярних мікроконтролерів даної категорії швидкодії.

Представники серії STM32L4 на основі ядра Cortex-M4 орієнтовані, в першу чергу, на енергоефективність. Вони підтримують базові DSP-інструкції та мають обмежені обсяги оперативної пам’яті (64–320 КіБ), що робить їх придатними для реалізації простих алгоритмів обробки сигналів та невеликих квантованих моделей нейронних мереж. Серія STM32F4, що має ядро Cortex-M4 із тактовою частотою до 180 МГц та підтримкою DSP-інструкцій, забезпечує вищу продуктивність порівняно з L-серією. Завдяки більшому обсягу пам’яті (128–256 КіБ) можливе виконання виведення за допомогою оптимізованих нейронних мереж у реальному часі. Однак відсутність спеціалізованих апаратних засобів прискорення векторних обчислень залишається основним обмеженням. Серія STM32H7, побудована на ядрі Cortex-M7 (у деякій підсерії – у двоядерній конфігурації Cortex-M7+Cortex-M4), забезпечує суттєво вищу обчислювальну продуктивність завдяки високій тактовій частоті до 480–600 МГц, підтримці DSP-інструкцій, наявності L1-кешу та

наявності FPU. Також дана серія має значно більший обсяг оперативної пам’яті – від 564 КіБ до 1.4 МіБ залежно від підсерії. Завдяки TCM-пам’яті, кільком DMA-контролерам і швидким інтерфейсам зовнішньої пам’яті, мікроконтролери STM32H7 краще підходять для складніших алгоритмів цифрової обробки сигналів, задач керування в реальному часі та запуску оптимізованих моделей глибинного навчання.

Мікроконтролери на ядрі Cortex-M0+ (наприклад, ATSAMD21) характеризуються низькою тактовою частотою (до 48 МГц) та порівняно малим обсягом пам’яті. Відсутність SIMD-інструкцій обмежує їх застосування у складних задачах, зокрема виведення для глибоких ШНМ, однак можуть використовуватися для отримання та базової попередньої обробки даних та керування периферією.

Мікроконтролери nRF52840 на основі ядра Cortex-M4 із тактовою частотою 64 МГц поєднують базові DSP-інструкції з інтегрованим Bluetooth Low Energy 5.0. Це робить платформу придатною для енергоефективних застосувань, де результати обробки даних передаються бездротовим каналом.

Мікроконтролери Raspberry Pi Pico (RP2040: має два ядра Cortex-M0+ (133 МГц) із 264 КіБ SRAM та RP2350: два ядра Cortex-M33 з 520 КіБ SRAM) вирізняються гнучкою системою програмованих інтерфейсів (PIO). Хоча відсутність SIMD-інструкцій у RP2040 обмежує швидкість обробки, даний мікроконтролер може використовується для прототипування та реалізації мінімальних моделей машинного навчання з невисокими вимогами щодо обсягу ОЗП.

Мікроконтролери Ambiq Apollo3 характеризуються архітектурою Cortex-M4e, яка оптимізована для наднизького енергоспоживання. При тактовій частоті до 96 МГц та обсязі пам’яті близько 384 КіБ ці мікроконтролери орієнтовані на автономні IoT-системи, де ресурсоємні обчислення виконуються періодично з мінімальними витратами енергії.

Таблиця 3.1 – Порівняльна характеристика мікроконтролерів

Платформа	Ядро, тактова частота	Векторні та DSP інструкції	RAM (вбудована)	Інтерфейси передачі даних	Орієнтовне енергоспоживання	Підтримка фреймворків DL	Орієнтовна вартість
1	2	3	4	5	6	7	8
STM32 (L4 / F4 / H7)	L4: Cortex-M4 80 МГц; F4: Cortex-M4 180 МГц; H7: Cortex-M7 480 МГц (у H7R/S –600 МГц).	L4, F4: DSP + FPU у деяких підсеріях; H7: DSP + single/double-precision FPU на M7.	L4: до 320 КіБ SRAM; F4: 192 КіБ у F407; H7: до 1 МІБ	L4: SPI, I2C, UART, USB, Quad-SPI; F4: SPI, I2C, UART, USB, Ethernet, DCMI, ext. memory; H7: USB OTG, SPI, I2C, UART, Ethernet, SDMMC, QSPI/FMC.	L4: run 30 мкА/МГц, stop 1 мкА; F4: run вищий, ніж у L4, standby 2–3 мкА; H7: run вищий, ніж у L4 і F4, standby: 3 мкА.	STM32Cube.AI, ST Edge AI, CMSIS-NN, TFLM; найкраща підтримка AI – H7.	STM32L476RET6: $6–8; STM32F407VGT6: $7–10; STM32H743VIT6: $12–16.
ATSAMD21	Cortex-M0+ до 48 МГц.	Немає SIMD/DSP	32 КіБ (ATSAMD 21G18)	I2C, SPI, UART, USB.	Run: 6–12 мА; Standby: 1,5 мкА; Idle: 2–10 мкА	Не підтримується	$5–7
nRF52840	Cortex-M4, 64 МГц.	DSP, без SIMD.	256 КіБ	BLE 5 + 802.15.4, USB, SPI, UART, I2C, I2S.	Run: 5–7 мА; RX: 4,6 мА, TX: 4,8 мА; idle: 2–3 мкА; OFF: 0,3–0,5 мкА	TFLM, CMSIS-NN, Edge Impulse	$7
RP2040 / RP2350	RP2040: 2 Cortex-M0+ 133 МГц; RP2350: 2 Cortex-M33 або 2 Hazard3 RISC-V, до 150 МГц	RP2040: без DSP/SIMD; RP2350: DSP + FPU	RP2040: 264 КіБ; RP2350: 520 КіБ	USB 1.1, SPI, I2C, UART, PIO	Run при 168–180 МГц: 70–100 мА; Stop: 10–50 мкА; Standby: 1–5 мкА	TFLM	RP2040: $1; RP2350: $1,5
Ambiq Apollo3	Cortex-M4F до 96 МГц.	Є DSP і SIMD	384 КіБ	BLE 5, SPI, I2C, UART.	Run при 48 МГц: 1–1,5 мА; 96 МГц: 2–3 мА; Deep sleep: 1–5 мкА	neuralSPOT, TFLM	$3,5–6,5
ESP32 (базовий / C3 / S3)	ESP32: 1 або 2 Xtensa LX6 до 240 МГц; ESP32-C3: RISC-V до 160 МГц; ESP32-S3: 2 Xtensa LX7 до 240 МГц	ESP32 base: без SIMD; C3: без SIMD; S3: має SIMD, vector-friendly розширення	ESP32: 536 КіБ; C3: 400 КіБ; S3: 512 КіБ, до 8 МІБ SPI PSRAM у корпусі модуля	ESP32: Wi-Fi + BT 4.2, SPI, I2C,UART,SDIO; C3: Wi-Fi + BLE 5, GPIO, I2C, I2S, SPI, UART, USB; S3: Wi-Fi + BLE 5, USB	ESP32: RX 112 мА, TX 239 мА; C3: RX 84 мА, TX 345 мА, light-sleep 130 мкА, deep-sleep 5 мкА; S3-N8R8: RX 100 мА, TX 330 мА	ESP-DL, ESP-NN, TFLM; найкраща підтримка AI – ESP32-S3	ESP32: $4,5–5,5; C3: $3,5–4; S3 (N8R8): $5,0–6,5
Kendryte K210	2 64-bit RISC-V, 400 МГц	KPU (hardware CNN accelerator), FFT, APU	8 МІБ	GPIO, UART, SPI, I2C, I2S, DVP, DMAC, FPIOA	Run >300 мА	MaixPy, MAIX ecosystem, kmodel-based flow	MAIX-I: $9

Kendryte K210 – це енергоефективна система на кристалі для задач Edge AI, побудована на архітектурі RISC-V та оснащена двоядерним 64-бітним процесором з FPU. Вона містить апаратний прискорювач KPU для згорткових нейронних мереж, що робить платформу придатною для задач комп'ютерного зору та аудіосигналів. K210 також має 8 MiB вбудованої SRAM, підтримує роботу з камерою через DVP та широкий набір периферії. Типове енергоспоживання самого чипа оцінюється на рівні менше 300 мВт, що робить його привабливим для компактних вбудованих IoT-пристроїв.

Мікроконтролер ESP32 у базовій версії використовує одне або два ядра Xtensa LX6 з тактовою частотою до 240 МГц, підтримує DSP-інструкції та апаратні MAC-операції, а також має близько 520 КіБ SRAM. Поєднання вбудованих модулів Wi-Fi і Bluetooth/BLE робить його придатним для реалізації задач бездротового IoT, локальної обробки даних і виконання нескладних алгоритмів цифрової обробки сигналів. ESP32-C3 є більш енергоефективним і бюджетним представником сімейства, побудованим на одноядерному 32-бітному RISC-V процесорі з частотою до 160 МГц. Він підтримує 2.4 GHz Wi-Fi та Bluetooth 5 LE, має близько 400 КіБ внутрішньої RAM і орієнтований насамперед на захищені підключені пристрої, де важливими є низьке енергоспоживання, бездротова взаємодія та помірні обчислювальні вимоги. Мікроконтролер ESP32-S3 має ядро Xtensa LX7 (до 240 МГц) доповнене SIMD-інструкціями та апаратною підтримкою MAC-операцій. Наявність зовнішньої PSRAM та інтегрованих модулів Wi-Fi/BLE робить платформу придатною для реалізації засобів збору та обробки даних з можливістю бездротової передачі результатів.

Враховуючи розглянуті варіанти, у рамках даного дослідження в якості апаратної платформи було обрано мікроконтролер ESP32-S3 від Espressif Systems Co., Ltd., що є високопродуктивним і водночас економічно доступним рішенням для реалізації систем з елементами штучного інтелекту на вбудованих пристроях. Детальні характеристики

мікроконтролера [56] наведені в таблиці 3.2, зовнішній вигляд – на рисунку 3.1.



а) б)

а) МК у корпусі QFN56, б) модуль з МК та обв’язкою

Рисунок 3.1 – Зовнішній вигляд SoC ESP32-S3

Таблиця 3.2 – Основні характеристики мікроконтролера ESP32-S3

Характеристика	Значення
Ядро ЦП	П’ятиступеневий конвеєр, 128-бітна шина даних, L1 cache, FPU одинарної точності
ОЗП	Вбудована 512 КіБ PSRAM + до 8 МіБ PSRAM, 16 КіБ SRAM у RTC-домені
Flash	До 16 МІБ (зовнішня), 384 КіБ ROM
Wi-Fi	2.4 GHz, IEEE 802.11 b/g/n, 20/40 МГц канал, швидкість до 150 Mbps, 4 віртуальні Wi Fi інтерфейси, antenna diversity, 802.11mc FTM
Bluetooth	5.0 LE + Mesh, Coded PHY (Long Range), 2 Mbps PHY, advertising extensions
GPIO	45 програмованих GPIO
Апаратна безпека	AES, SHA, RSA, HMAC, Secure Boot, Flash Encryption, цифровий підпис.
Енергозбереження	Є ULP-співпроцесор та кілька режимів низького енергоспоживання.
Додаткові інтерфейси	LCD/Camera controller, SD/MMC host, TWAI (CAN-сумісний), USB Serial/JTAG, MCPWM, RMT, PCNT, GDMA
Ізольоване виконання	Є World Controller
Орієнтовна ціна	5,0–6,5 USD

Він поєднує високу обчислювальну потужність з підтримкою сучасних периферійних пристроїв, інтерфейсів та протоколів. Важливою перевагою ESP32-S3 є наявність ефективної вбудованої SRAM до 8 МіБ, що значно розширює можливості розгортання глибинних нейронних мереж

порівняно з попередніми поколіннями цієї платформи. Ядро також підтримує інструкції векторної обробки, прискорюючи виконання матричних обчислень, характерних для виведення ШНМ.

3.1.2 Програмні засоби

Для програмної реалізації ШНМ для зазначеної апаратної платформи було обрано програмний пакет ESP-DL (Espressif Deep Learning) – спеціалізований фреймворк для розгортання та виконання моделей глибинного навчання на вбудованих платформах сімейства ESP32. Він функціонує на базі середовища ESP-IDF і використовується для мікроконтролерів ESP32-S3 та ESP32-P4, оскільки вони мають відповідні апаратні засоби прискорення обчислень, завдяки чому досягається вища ефективність. Підтримка класичного ESP32 також присутня, але зі зниженою швидкодією. Фреймворк призначений для локального виведення для глибинних моделей на Edge-пристроях і забезпечує повний цикл їх використання: від завантаження та запуску до тестування, профілювання й інтеграції в прикладні сценарії комп’ютерного зору та обробки даних. Важливою особливістю ESP-DL є застосування власного формату моделі .espdl, тому перед розгортанням ШНМ на пристрої необхідно виконати її перетворення у сумісне представлення. Для цього використовується інструмент ESP-PPQ, що базується на PPQ, підтримує квантування моделей ONNX і PyTorch, містить спеціалізовані засоби експорту для безпосереднього завантаження моделі у вбудовану систему. Структурно екосистема ESP-DL охоплює модулі dl, vision, audio та fbs_loader, що містять реалізацію операцій глибинного навчання, комп’ютерного зору, обробки аудіосигналів та завантаження моделей.

Для обробки аудіосигналу в даному фреймворку відповідає бібліотека ESP-DSP [57], що містить як ANSI C, так і асемблерні реалізації FFT. Конвеєр обробки сигналу складається з ініціалізації таблиць

коефіцієнтів, застосування віконної функції, виклику FFT, bit-reverse перестановки та перетворення спектра до форми, придатної для подальшого аналізу. Програмна реалізація даних перетворень включає декілька інженерних оптимізацій для цільової апаратної платформи. Таблиці twiddle-коефіцієнтів формуються шляхом динамічного виділення вирівняної пам’яті через memalign(16), що є важливим для систем реального часу. Віконна обробка винесена в окремі модулі зі специфічними оптимізованими реалізаціями для кожної архітектури. Функція Hann-вікна задається аналітичною формулою, розраховується при ініціалізації та може використовуватись багаторазово під час обробки. Обчислення FFT для вікна здійснюється послідовністю викликів: dsps_fft2r_fc32, dsps_bit_rev_fc32, dsps_cplx2reC_fc32 для переходу від комплексного внутрішнього представлення до спектральної форми (лістинг 3.1). Особливо важливо, що перша з цих функцій реалізована на асемблері для ESP32-S3 та імплементує butterfly-алгоритм, аналогічний до C-коду у коментарі (лістинг 3.2). Додаткове прискорення досягається і на етапі bit-reverse перестановки, де в використовуються 64-бітні завантаження та збереження (лістинг 3.3). Отже, навіть допоміжні етапи FFT-конвеєра реалізовані з урахуванням апаратних можливостей цільової архітектури, що підтверджує ефективність програмної реалізації перетворення на спектрограму засобами бібліотек фреймворку ESP-DL.

Лістинг 3.1 – Реалізація обчислення FFT для вікна

```
dsps_fft2r_fc32(data, length);
dsps_bit_rev_fc32(data, length);
dsps_cplx2reC_fc32(data, length);
for (int i = 0; i < length / 2; i++) {
    data[i] = 10 * log10f(
        (data[i * 2 + 0] * data[i * 2 + 0] +
         data[i * 2 + 1] * data[i * 2 + 1]) / N);
}
```

Лістинг 3.2 – реалізація функції dsps_fft2r_fc32

```
; for (int N2 = N/2; N2 > 0; N2 >>= 1) {
;     ia = 0;
;     for (int j = 0; j < ie; j++) {
;         c = w[2 * j];
;         s = w[2 * j + 1];
;         for (int i = 0; i < N2; i++) {
;             m = ia + N2;
srli a6, a3, 1
movi.n a7, 1
.fft2r_l1:
movi.n a8, 0
movi.n a11, 0
.fft2r_l2:
addx8 a10, a8, a4
lsi f0, a
```

Лістинг 3.3 – Фрагмент реалізації dsps_bit_rev_fc32

```
EE.LDF.64.IP f0, f2, a5, 0
EE.LDF.64.IP f1, f3, a6, 0
EE.STF.64.IP f0, f2, a6, 0
EE.STF.64.IP f1, f3, a5
```

Бібліотека ESP-DL [58] використовується для виведення ШНМ на мікроконтролерах ESP32 і підтримує оптимізовані оператори, статичне планування пам’яті та dual-core scheduling для шарів Conv2D і DepthwiseConv2D. Паралелізація в ESP-DL реалізується не на прикладному рівні (користувацькою програмою), а є вбудованою безпосередньо в механізм оператора згортки. Під час обчислення згортки створюється дві задачі FreeRTOS, прив’язані до різних ядер через xTaskCreatePinnedToCore, а синхронізація виконується лічильним семафором (лістинг 3.4). Арифметичне ядро обчислення згорток спирається на оптимізовані на рівні архітектури (ISA) SIMD-реалізації для ESP32-S3

і ESP32-P4, а саме – імплементація dot-product і depthwise-згортки. Лістинг 3.5 демонструє характерні векторні інструкції, що лежать в основі оптимізованих обчислень згортки: у фрагменті спостерігається завантаження 128-бітних векторів у регістри q0…q3 та застосування векторних multiply-accumulate інструкцій ee.vmulas.s8.accx..., які дозволяють обробляти одразу групи елементів. Отже, прискорення в ESP-DL реалізується не лише на рівні C-коду, а через спеціалізовані асемблерні мікроядра, адаптовані до конкретної архітектури. Це обґрунтовує вибір даного фреймворку для найбільш ефективної програмної реалізації запропонованої інформаційної технології.

Лістинг 3.4 – Фрагмент коду, що містить реалізацію dual-core scheduling для згорткових шарів

```
    SemaphoreHandle_t semaphore = xSemaphoreCreateCounting(2,
0);
    xTaskCreatePinnedToCore(
        module_forward_task, NULL, 2048, &task_data1,
        current_priority, &xHandleTask1, (current_core_id + 1)
% 2);
    xTaskCreatePinnedToCore(
        module_forward_task, NULL, 2048, &task_data2,
        current_priority, &xHandleTask2, current_core_id);
    xSemaphoreTake(semaphore, portMAX_DELAY);
    xSemaphoreTake(semaphore, portMAX_DELAY);
```

Лістинг 3.5 – Фрагмент коду, що реалізує операцію dot-product

```
    ee.vld.128.ip q0, a3, 16
    ee.vld.128.ip q1, a4, 16
    loopnez a7, loop_end_dl_tie728_dotprod_i8k8o16
    ee.vld.128.ip q2, a3, 16
    ee.vmulas.s8.accx.ld.ip q3, a4, 16, q0, q1
    ee.vld.128.ip q0, a3, 16
    ee.vmulas.s8.accx.ld.ip q1, a4, 16, q2, q3
```

3.2 Оптимізація нейронної мережі для цільової апаратної платформи

3.2.1 Оптимізація архітектури

На основі попередніх експериментальних досліджень було обрано SqueezeNet-подібну модель для виконання класифікації емоцій у рамках запропонованого підходу. Для ефективного виконання виведення на мікроконтролері доцільною є додаткова оптимізація архітектури моделі шляхом зменшення споживання обчислювальних ресурсів.

У запропонованій модифікації архітектури (рисунок 3.2) було збережено базову архітктуру Fire-модуля (операції squeeze та expand), проте внесено низку змін, спрямованих на підвищення обчислювальної ефективності. Було поєднано базові переваги архітектур SqueezeNet та MobileNet: гілку розширення 3x3 замінено на глибинно-роздільну згортку, що суттєво зменшує кількість операцій множення-накопичення без істотної втрати репрезентативної здатності. Операцію конкатенації каналів у Fire-модулі замінено на поелементне додавання, що дозволяє знизити пікове використання оперативної пам’яті та спростити обчислювальний граф. Додатково введено залишковий зв’язок всередині Fire-блока для покращення поширення градієнтів і стабілізації процесу навчання. Для забезпечення коректного навчання використано шари пакетної нормалізації після згорткових операцій, які компенсують варіації розподілів активацій, що особливо критично при використанні depthwise-згорток. На етапі розгортання (виведення) передбачено злиття шарів Conv і BatchNorm у єдині згорткові оператори з попередньо обчисленими параметрами, а також заміну функції активації ReLU на ReLU6, яка є більш придатною для низькорозрядного квантування. Сукупність цих модифікацій забезпечує компроміс між точністю моделі, швидкодією та обмеженнями пам’яті цільової вбудованої платформи.

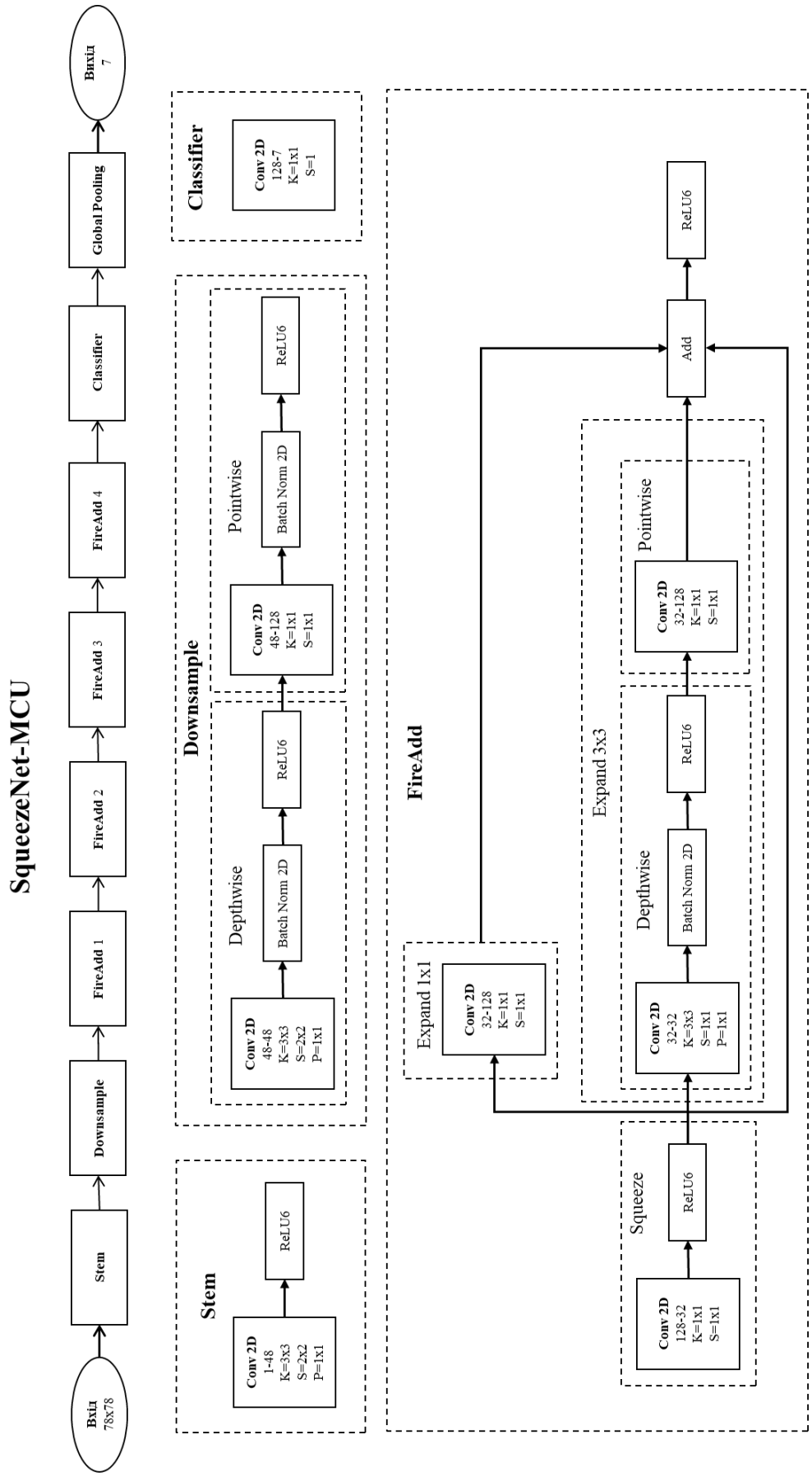


Рисунок 3.2 – Архітектура модифікованої ШНМ «SqueezeNet-MCU»

У результаті навчання моделі протягом 150 епох за аналогічної до попередніх дослідів незмінної конфігурації тренувального середовища було отримано наступні значення показників якості (таблиця 3.3). Точність класифікації: 0,7793, точність передбачення позитивного класу: 0,7797, повнота (Recall): 0,7805 та F1-міра: 0,7798.

Отримані результати свідчать про збалансованість та узгодженість з показниками Precision і Recall, що підтверджується значенням F1-міри. Це вказує на стабільність процесу навчання та ефективність запропонованої архітектури і гіперпараметрів тренування. Візуалізація метрик впродовж навчання наведена на рисунку 3.3. Хоча показники достовірності класифікації є дещо нижчими, ніж для попереднього варіанту модифікованої моделі, значною мірою покращилась обчислювальна ефективність нової моделі. Обчислювальна складність моделі на етапі інференсу становить 47,53 MFLOPs, кількість параметрів – 0,06 млн, а статичний розмір моделі – 0,23 МіБ, що характеризує її як компактну та обчислювально ефективну для практичного використання.

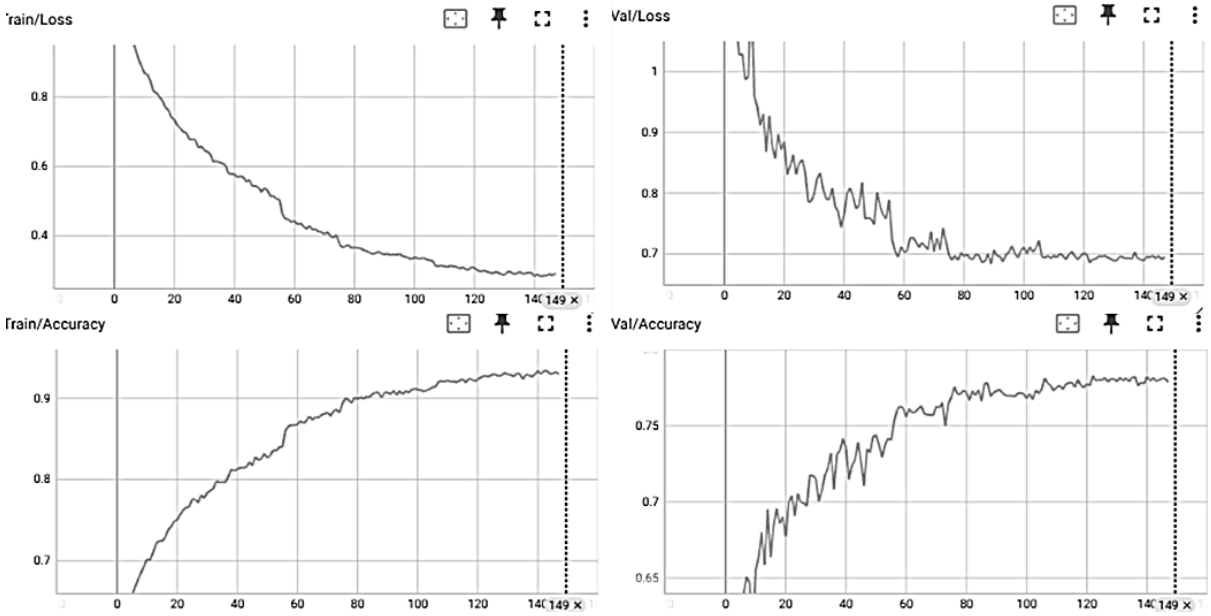


Рисунок 3.3 – Графіки функції втрат та точності під час навчання

Таблиця 3.3 – Значення метрик при різних типах квантування

Метрика	Цільовий тип даних		
	FP32	INT8	INT8+INT16
Accuracy	0,7793	0,7667	0,7697
Precision	0,7797	0,7686	0,7712
Recall	0,7805	0,7681	0,7708
F1	0,7798	0,7669	0,7703

3.2.2 Квантування моделі

Наступним етапом є квантування моделі з урахуванням можливостей апаратної платформи ESP32-S3. Найкращу продуктивність матричних обчислень даний мікроконтролер забезпечує при цілочисельному типі даних розрядністю 8 біт (INT8). Навчену модель було квантовано за допомогою програмного засобу ESP-PPQ. Зниження достовірності класифікації знизилось на майже 1,2% порівняно з початковою моделлю (таблиця 3.3) при зниженні вимог до споживання пам’яті у 4 рази. Аналіз впливу шуму квантування на результати обчислень на кожному шарі мережі (рисунок 3.4) продемонстрував максимальне відхилення (понад 0,2%) для першої згортки стему, причому вплив похибки на інших шарах є значно меншим. Це обґрунтовує доцільність застосування методу квантування зі змішаною точністю з обчисленням даної згортки при вищій розрядності типу даних. Фреймворк ESP-DL та апаратна платформа ESP32-S3 підтримують квантування зі змішаною точністю INT8+INT16, що забезпечує баланс між точністю, обчислювальною ефективністю та необхідним обсягом пам’яті. При застосуванні такого квантування (16 розрядів для першої згортки, 8 – для решти шарів) метрики якості класифікації показали в середньому зниження на приблизно 0,9% у порівнянні з FP32-моделлю (таблиця 3.3). Така конфігурація ШНМ була обрана для апаратно-програмної реалізації запропонованої інформаційної технології.

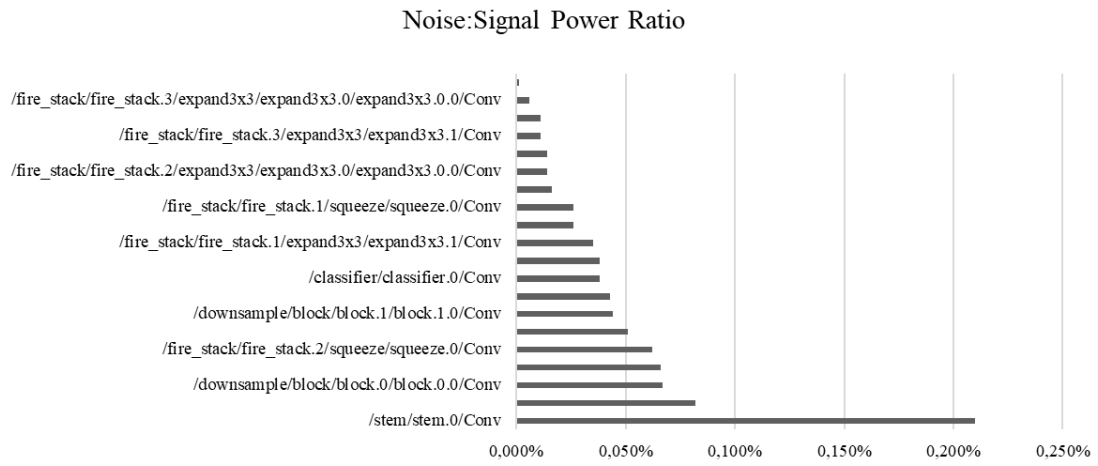


Рисунок 3.4 – Значення сигнал/шум при квантуванні для кожного шару

Матриця помилок (рисунок 3.5) свідчить про здатність моделі коректно розрізняти більшість емоційних класів, що підтверджує домінування значень на головній діагоналі. Найвищу точність класифікації модель демонструє для класу «noise», що вказує на його чітку відокремленість у просторі ознак. Високі показники правильних класифікацій також спостерігаються для класів «angry», «fear», «neutral» та «sad». Разом із тим помітна наявність взаємних помилок між класами, зокрема між «happy», «angry», «fear» та «neutral», що проявляється підвищеною кількістю хибних класифікацій поза діагоналлю. Це може свідчити про часткове перекриття їхніх ознакових представлень. Зокрема, найсуттєвіша взаємна плутанина зафіксована між класами happy та angry, а також між «happy» та «fear». Помітна кількість хибних передбачень також спостерігається між класами «disgust» і «sad», а також між «neutral» і «sad», що може бути зумовлено близькістю емоційних проявів або слабкою вираженістю відповідних сигналів. Взаємні помилки між «angry» і «fear», та «fear» і «sad» також показують складність розмежування емоцій зі схожими акустичними проявами. Загалом матриця помилок вказує на помірно збалансовану роботу моделі та адекватну узгодженість між класами. Водночас вона висвітлює напрями для подальшого покращення розпізнавання подібних емоцій.

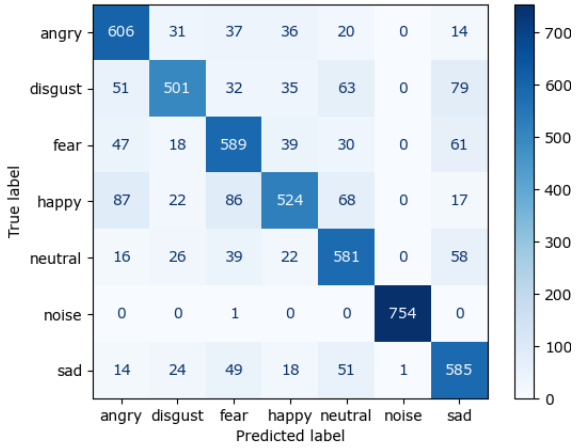


Рисунок 3.5 – Матриця помилок для моделі зі змішаним квантуванням

3.3 Апаратно-програмна реалізація розробленої інформаційної технології

У рамках виконаного дослідження було розроблено пристрій для автоматизованого визначення емоційного стану людини за голосовими ознаками. Загальне призначення пристрою полягає в аналізі акустичних характеристик мовлення з подальшою класифікацією емоцій у режимі, наближеному до реального часу, та в автономному режимі обробки аудіоданих. Розроблене рішення орієнтоване на вбудовані системи з обмеженими обчислювальними ресурсами та може застосовуватися в системах людино-машинної взаємодії, мобільних та носимих пристроях, та навчально-дослідницьких стендах.

Функціонально пристрій підтримує два режими роботи. Перший передбачає розпізнавання емоцій у реальному часі (онлайн) з використанням мікрофона як джерела вхідного сигналу. Аналоговий аудіосигнал реєструється, обробляється в цифровому представленні, після чого виконується класифікація, і результати відображаються на графічному дисплеї. Даний режим забезпечує мінімальні затримки між появою мовного сигналу та виведенням результатів аналізу, що є критично важливим для застосувань реального часу. Другий режим роботи реалізує

тестовий (офлайн) сценарій обробки, у якому аудіодані попередньо зберігаються на карті пам’яті формату SD-Card. У цьому режимі здійснюється завантаження аудіозаписів із зовнішнього носія, їх відтворення за допомогою гучномовця з одночасним виконанням емоційного аналізу. Результати розпізнавання виводяться на дисплей у процесі відтворення сигналу, що дозволяє оцінювати коректність роботи пристрою на тестових даних у контрольованих умовах та проводити його експериментальну верифікацію.

Апаратна частина пристрою (рисунок 3.6) побудована на базі мікроконтролера ESP32-S3, який виконує функції керування периферією та реалізації обчислювальних алгоритмів обробки сигналів і виведення ШНМ.

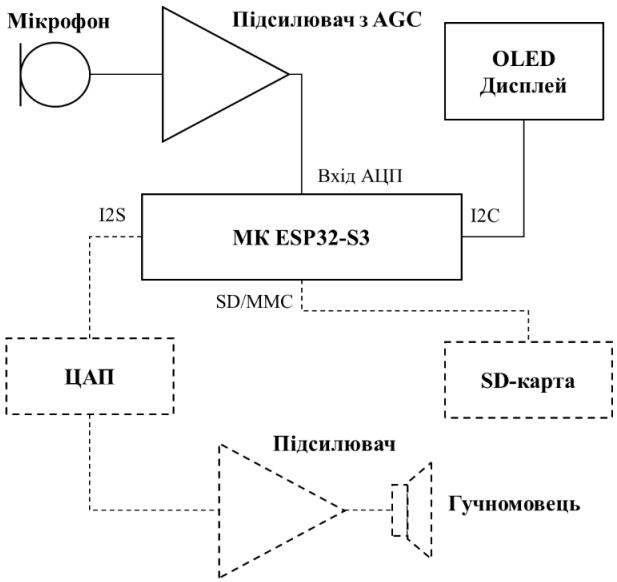


Рисунок 3.6 – Структурна схема пристрою

Для захоплення звукового сигналу використовується електретний мікрофонний модуль та малошумлячий підсилювач, оснащений вбудованою системою автоматичного регулювання підсилення (AGC), що забезпечує стабільний рівень сигналу незалежно від гучності мовлення. Підключення мікрофона здійснюється до входу аналого-цифрового перетворювача (АЦП) мікроконтролера. Для візуалізації результатів аналізу

застосовано графічний OLED-дисплей з інтерфейсом I^2C та роздільною здатністю 128×32 пікселі (на базі контролеру ssd1306), який забезпечує компактне відображення інформації. У другому режимі роботи апаратна конфігурація доповнюється цифро-аналоговим перетворювачем з інтерфейсом I^2S, підсилювачем низької частоти, гучномовцем, а також модулем карти пам’яті SD/MMC (позначено пунктиром на структурній схемі).

Запропоновану інформаційну технологію емоційного аналізу було реалізовано у вигляді вбудованої програми пристрою мовою C++ з використанням офіційного програмного пакету інструментів для розробки ESP-IDF v5.5.2 та бібліотеки ESP-DL v 3.2.4 однойменного фреймворку. Вбудована програма реалізує два режими роботи згідно алгоритмів, наведених на рисунку 3.7. Багатозадачне виконання реалізоване за допомогою засобів ОС FreeRTOS. Потокова передача аудіоданих через інтерфейс I2S використовує DMA для зменшення навантаження на ядро. Також застосовано гнучке динамічне виділення пам’яті з урахуванням її апаратних властивостей для оптимізації використання ресурсів вбудованої платформи. У програмній реалізації використано стандартні бібліотеки фреймворку ESP-IDF: sdmmc_cmd.h для роботи з SD/MMC-носіями, ssd1306.h для керування OLED-дисплеєм, esp-bsp.h для абстрагування апаратної конфігурації плати, adc_mic.h для захоплення сигналу з аналогового мікрофона через АЦП, esp_codec_dev.h для уніфікованої взаємодії з аудіотрактом, esp_timer.h для високоточного вимірювання відліків, iot_button.h і button_gpio.h для обробки кнопкового вводу, а також driver/i2s_std.h для конфігурування стандартного цифрового аудіоінтерфейсу. Також використано бібліотеки ESP-DL: dl_model_base.hpp для завантаження та виконання нейромережевої моделі, dl_audio_wav.hpp для декодування WAV-аудіо у PCM-представлення, dl_fbank.hpp для формування мел-спектрограм, а dl_image.hpp – для представлення та попередньої обробки зображень.

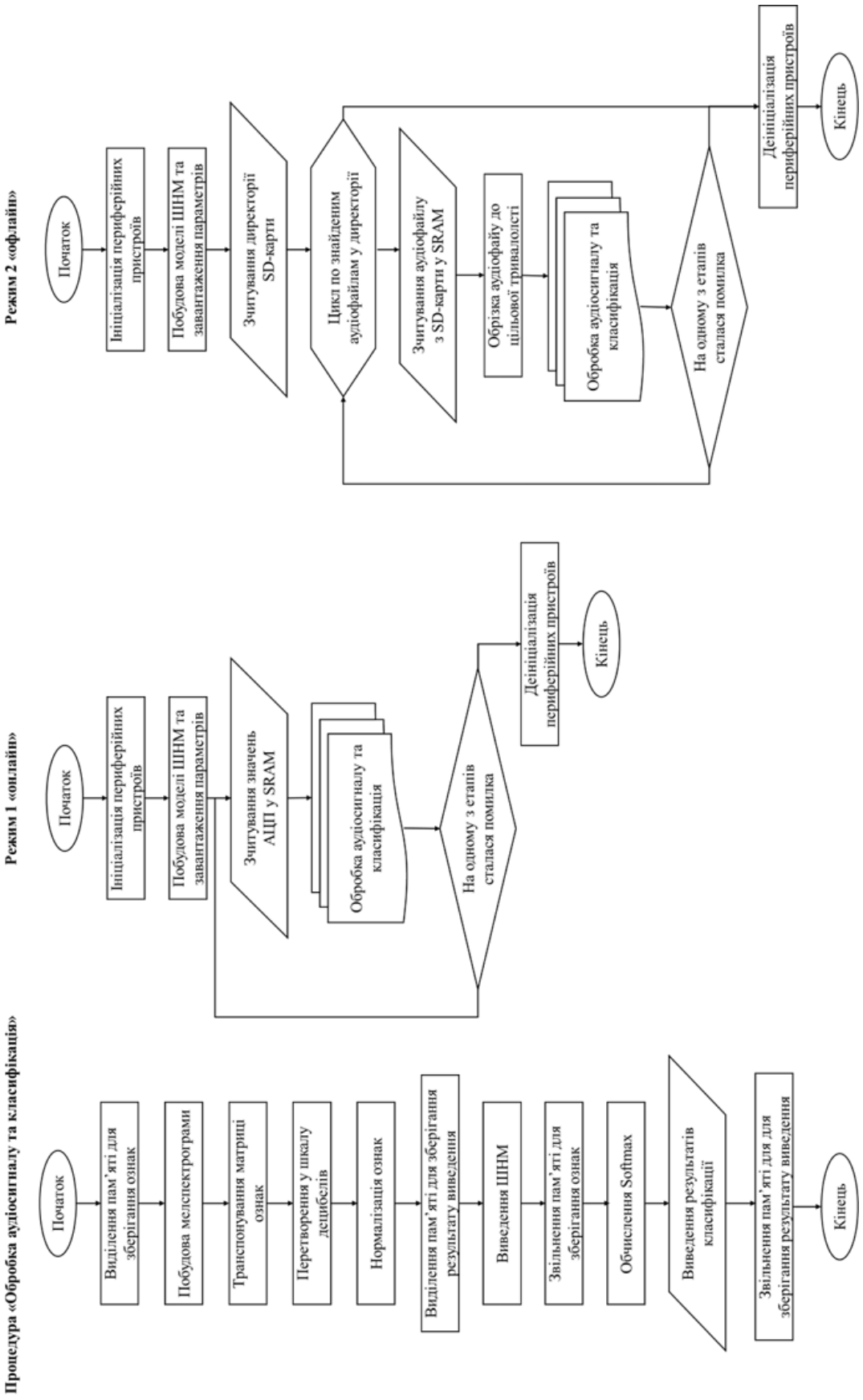


Рисунок 3.7 – Блок-схема алгоритмів, реалізованих у вбудованій програмі пристрою

Завантаження параметрів моделі можливе як із вбудованої Flash-пам’яті (більш ефективно через вищу пропускну здатність інтерфейсу), так і з SD-карти (забезпечує швидшу заміну моделі, краще для тестування системи), що наведено в лістингу 3.6. Основна логіка реалізована у вигляді послідовних викликів функцій, що виконують специфічні завдання обробки даних (лістинг 3.7). Зокрема, виділення ознак реалізовано як окрему функцію (лістинг 3.8), що включає етапи перетворення на мелспектрограму, транспонування, перетворення у шкалу децибелів та нормалізацію. Виконання виведення здійснюється викликом функції, наведеної у лістингу 3.9, що виконує квантування ознак для цільового типу даних INT8 або INT16, запускає інференс, виконує деквантування результатів та застосовує softmax для їх перетворення у ймовірнісне подання.

Лістинг 3.6 – Програмна реалізація функції ініціалізації моделі

```
void init_dl_model(){
#if CONFIG_MODEL_LOCATION_SD
    model = new dl::Model(MODEL_PATH, fbs::MODEL_LOCATION_IN_SDCARD);
#else
    dl::Model *model = new dl::Model((const char *)model_espdl, fbs::MODEL_LOCATION_IN_FLASH_RODATA);
#endif
    if (model == nullptr) {
        ESP_LOGE(TAG, "Failed to create model");
        return;
    }
    std::map<std::string, dl::TensorBase *> model_inputs = model->get_inputs();
    model_input = model_inputs.begin()->second;
    std::map<std::string, dl::TensorBase *> model_outputs = model->get_outputs();
    model_output = model_outputs.begin()->second;
}
```

Лістинг 3.7 – Реалізація конвеєра обробки аудіосигналу

```
uint32_t timestamp = esp_timer_get_time();
        float *features_buffer = (float *)malloc(sizeof(float) * features_len);
        extract_features(buffer, target_samples, features_buffer, features_len);
        uint32_t fe_time = esp_timer_get_time() - timestamp;
        float *probabilities = (float *)malloc(sizeof(float) * model_output->size);
        predict(features_buffer, probabilities);
        free(features_buffer);
        uint32_t infer_time = esp_timer_get_time() - timestamp - fe_time;
        ESP_LOGI(TAG, "Extracted features in %d ms", fe_time / 1000);
        ESP_LOGI(TAG, "Model inference completed in %d ms", infer_time / 1000);
        int label = argmax(probabilities, model_output->size);
        display_probabilities(probabilities, model_output->size);
        free(probabilities);
        char label_buf[32];
        snprintf(label_buf, sizeof(label_buf), "FE:%ld,INF:%ld", fe_time / 1000, infer_time / 1000);
        ssd1306_clear_display_page(display_handle, 0, false);
        ssd1306_display_text(display_handle, 0, label_buf, true);
```

Лістинг 3.8 – Реалізація функції для виконання виділення ознак

```
void extract_features(int16_t *audio_buffer, int audio_len, float *features_buffer, int features_len) {
```

Продовження лістингу 3.8

```
                fbank->process(audio_buffer,      audio_len,
features_buffer);
                    transpose(features_buffer,        fbank-
>config().num_mel_bins,
                features_len / fbank->config().num_mel_bins);
         power_to_db(features_buffer, features_len, 1.0f, 1e-
10f, 80.0f);
        for (int i = 0; i < features_len; i++)
            features_buffer[i] = (features_buffer[i] + 100.0f)
/ 120.0f; // normalize to [0, 1]
    }
```

Лістинг 3.9 – Реалізація функції для виконання виведення

```
    void predict(float *features, float *probabilities) {
        for (int i = 0; i < model_input->size; i++) {
            if          (model_input->get_dtype()          ==
dl::DATA_TYPE_INT8) {
                int8_t*               data_ptr               =
static_cast<int8_t*>(model_input->data);
                data_ptr[i]                                  =
dl::quantize<int8_t>(features[i],       DL_RESCALE(model_input-
>exponent));
            }    else    if    (model_input->get_dtype()    ==
dl::DATA_TYPE_INT16) {
                int16_t*              data_ptr               =
static_cast<int16_t*>(model_input->data);
                data_ptr[i]                                  =
dl::quantize<int16_t>(features[i],      DL_RESCALE(model_input-
>exponent));
            }    else    if    (model_input->get_dtype()    ==
dl::DATA_TYPE_FLOAT) {
                float*                data_ptr               =
static_cast<float*>(model_input->data);
                data_ptr[i] = features[i];
```

Продовження лістингу 3.9

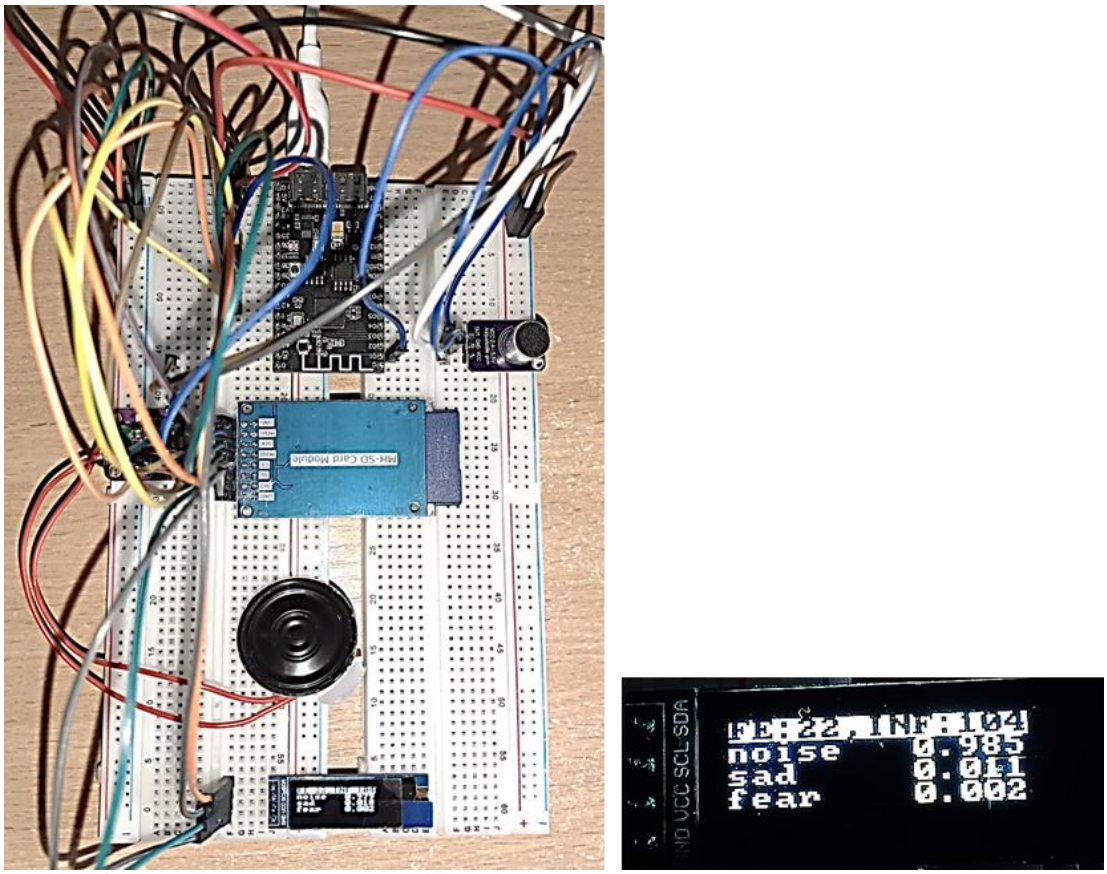
```
            } else {
                ESP_LOGE(TAG, "Unsupported model input data type");
                return;
            }
        }
        model->run(CONFIG_MODEL_RUN_MULTICORE ? dl::RUNTIME_MODE_MULTI_CORE : dl::RUNTIME_MODE_SINGLE_CORE);
        int8_t *output_ptr = static_cast<int8_t*>(model_output->data);
        for (int i = 0; i < model_output->size; i++)
            probabilities[i] = dl::dequantize(output_ptr[i], DL_SCALE(model_output->exponent));
        softmax(probabilities, model_output->size);
    }
```

Для перевірки функціоналу розробленої системи було зібрано тестовий стенд, зовнішній вигляд якого наведено на рисунку 3.8 (а). Приклад вихідних даних, що виводяться на OLED-дисплеї (час виконання виділення ознак та інференсу, а також мітки трьох класів з найбільшою ймовірністю), наведено на рисунку 3.8 (б).

Експериментальна перевірка показала, що час отримання класифікаційного рішення склав 126 мс. З них 22 мс припадає на етап виділення ознак і 104 мс – на виведення (інференс). Отже, система здатна працювати в режимі реального часу, що є однією з ключових вимог до практичного застосування в носимих пристроях та системах дистанційного моніторингу.

Перспективними напрямами подальшого вдосконалення розробленого пристрою є реалізація функціоналу журналювання результатів, накопичення статистики, засобів автоматизованої обробки та інтерпретації отриманих даних, реалізація бездротової передачі

інформації (через BLE) на мобільний пристрій, покращення користувацького інтерфейсу, забезпечення захисту даних (локальне шифрування, контроль доступу), підвищення автономності роботи системи (використання режимів low-power), а також інтеграція з іншими сенсорними каналами для мультимодального аналізу.



а) б)

Рисунок 3.8 – Зовнішній вигляд тестового стенду (а) та приклад відображення вихідних даних

Таким чином, розроблений пристрій є інтегрованим рішенням, яке поєднує засоби реєстрації аудіосигналу, візуалізації та інтелектуальної обробки, забезпечує гнучкість експлуатації та придатність до подальшого розширення функціональних можливостей.

Висновки до розділу 3

У даному розділі кваліфікаційної роботи було описано апаратно-програмну реалізацію розробленої інформаційної технології емоційного аналізу у форматі вбудованої системи. Зважаючи на поставлені вимоги до інформаційної технології: можливість роботи в режимі, наближеному до реального часу, та обмежень апаратної платформи постає потреба в подальшій оптимізації запропонованих підходів.

Перш за все було обрано цільову апаратну платформу для реалізації інформаційної технології. Розглянуто найбільш поширені мікроконтролери, що використовуються у системах Edge AI, серед яких: STM32 (серій L4, F4, H7), ATSAMD21, nRF52840, RP2040, RP2350, Ambiq Apollo3, ESP32 (базовий, C3, S3) та Kendryte K210. Проаналізовано їх ключові характеристики, здійснено порівняння з акцентом на баланс між обчислювальною потужністю, енергоспоживанням та вартістю. В результаті було обрано мікроконтролер ESP32-S3 завдяки наявності спеціалізованих апаратних засобів прискорення векторних обчислень, достатньому об’єму оперативної пам’яті, низькій вартості та наявності широкого різноманіття інтерфейсів, зокрема Wi-Fi та BLE.

Для програмної реалізації було обрано фреймворк ESP-DL, що забезпечує повний цикл обробки сигналу: від обчислення спектрограм до виведення глибоких згорткових мереж. Він безпосередньо інтегрується з ESP-IDF та забезпечує оптимізовану реалізацію базових нейромережевих операторів, статичне планування пам’яті, двоядерне планування обчислень і засоби квантування моделей.

Наступним етапом було здійснено додаткову оптимізацію архітектури SqueezeNet для ресурсно-обмеженої мікроконтролерної платформи через використання глибинно-розділених згорток, операцій додавання замість конкатенації, згорткового класифікатору замість повнозв’язного та інших модифікацій. Це дозволило знизити обчислювальну складність інференсу на

третину, а розмір моделі – на 40% при втраті точності 0,8%. Для розробленої моделі було здійснено посттренувальне квантування для цілочисельних значень зі змішаною точністю. Це дозволило додатково знизити обсяг займаної пам’яті майже вчетверо при незначній втраті точності, яка склала майже 0,77.

Було розроблено прототип пристрою для виконання емоційного аналізу за голосом людини, зібрано тестовий стенд для експериментальної верифікації його роботи. Пристрій підтримує два режими роботи: онлайн-розпізнавання емоцій у реальному часі з використанням мікрофона та офлайн-обробка тестових аудіоданих, збережених на SD-карті. Результати у вигляді трьох найбільш імовірних емоційних станів виводяться на графічний дисплей. При тестуванні пристрою час отримання класифікаційного рішення склав 126 мс, що підтвердило можливість його роботи в режимі наближеному до реального часу.

ВИСНОВКИ

В ході проведеного дослідження у рамках випускної кваліфікаційної роботи було розроблено інформаційну технологію розпізнавання емоційних станів людини за аудіоданими, що використовує спектрографічне представлення ознак та класифікатор на основі глибинної згорткової мережі. Проведені експериментальні дослідження підтвердили прийнятну точність і узгодженість моделі при збереженні низьких вимог до обчислювальних ресурсів, що робить запропоноване рішення перспективним для практичного застосування в автономних та енергоефективних пристроях. Таким чином, основна мета кваліфікаційної роботи – досягнута. Всі поставлені задачі були вирішені. В результаті проведеного дослідження було отримано наступні основні результати:

– було розглянуто методи та апаратно-програмні засоби цифрового кодування акустичних сигналів, визначено основні характеристики цифрового аудіосигналу та їх співвідношення з фізичними характеристиками акустичного сигналу;

– розглянуто найбільш поширені методи виділення ознак часової та частотної області для звукових сигналів, що застосовуються в задачах обробки голосових аудіозаписів, досліджено їх особливості та специфіку застосування; обґрунтовано використання лог-мелспектрограми для представлення ознак, як персептивно-орієнтованого та ресурсно-ефективного підходу;

– здійснено огляд сучасних підходів до класифікації аудіоданих, зокрема класичних методів машинного навчання, імовірнісних підходів та глибинних нейромережевих моделей, досліджено їх застосування в задачах емоційного аналізу; в результаті було обрано CNN як найбільш ефективну архітектуру, що поєднує достатню точність класифікації з невисокою обчислювальною складністю та простотою реалізації;

– розроблено вдосконалений метод розпізнавання емоцій за голосом, що забезпечує низьку обчислювальну та об’ємну складність і складається з двох етапів: виділення ознак та здійснення їх класифікації; він забезпечує розрізнення п’яти емоційних класів, нейтрального стану та відсутності голосу;

– для формування експериментального набору даних було обрано дев’ять відкритих датасетів емоційного мовлення, що містять зразки, записані у різній обстановці, акторами та непрофесіоналами різного віку та статі, різними мовами; це забезпечує стійкість до звукового фону, мовну та гендерну інваріантність класифікатора;

– запропоновано модифікації одинадцяти популярних згорткових архітектур, що враховують апаратні обмеження обчислювальної платформи. Здійснено експериментальну апробацію модифікованих моделей за кількома метриками; досліджено вплив квантування на ефективність моделей. Найкращі результати продемонструвала SqueezeNet-подібна архітектура, досягши точності 78%;

– здійснено аналіз апаратних засобів для реалізації запропонованої інформаційної технології. В результаті було обрано мікроконтролер ESP32-S3 з позицій компромісу між обчислювальними можливостями, енергоспоживанням та наявністю периферійних інтерфейсів. Відповідно до обраної апаратної платформи було визначено програмні засоби для реалізації вбудованої програми;

– проведено додаткову оптимізацію обраної згорткової архітектури через застосування найкращих практик, виявлених в результаті проведеного експериментального дослідження, що дозволило суттєво знизити вимоги до часової та об’ємної складності при незначній втраті точності до 77%;

– Розроблену інформаційну технологію було розгорнуто на мікроконтролерній платформі ESP32-S3, де було успішно протестовано її працездатність; час отримання класифікаційного рішення склав 126 мс, що підтверджує можливість роботи у режимі реального часу.

ПЕРЕЛІК ДЖЕРЕЛ ПОСИЛАННЯ

1. AI-Based Facial Emotion Analysis for Early and Differential Diagnosis of Dementia / L. Bergamasco et al. *Bioengineering*. 2025. Vol. 12, no. 10. P. 1082. URL: https://doi.org/10.3390/bioengineering12101082 (date of access: 25.04.2026).

2. Detection and monitoring of stress using wearables: a systematic review / A. Pinge et al. *Frontiers in Computer Science*. 2024. Vol. 6. URL: https://doi.org/10.3389/fcomp.2024.1478851 (date of access: 25.04.2026).

3. Emotion recognition for enhanced learning: using AI to detect students' emotions and adjust teaching methods / S. A. Salloum et al. *Smart Learning Environments*. 2025. Vol. 12, no. 1. URL: https://doi.org/10.1186/s40561-025-00374-5 (date of access: 25.04.2026).

4. Bhardwaj S., Thapa S. B., Gandhi A. Advances in neuromarketing and improved understanding of consumer behaviour: analysing tool possibilities and research trends. *Cogent Business & Management*. 2024. Vol. 11, no. 1. URL: https://doi.org/10.1080/23311975.2024.2376773 (date of access: 25.04.2026).

5. Emotion recognition for enhanced learning: using AI to detect students' emotions and adjust teaching methods / S. A. Salloum et al. *Smart Learning Environments*. 2025. Vol. 12, no. 1. URL: https://doi.org/10.1186/s40561-025-00374-5 (date of access: 25.04.2026).

6. Al-Azani S., El-Alfy E.-S. M. A review and critical analysis of multimodal datasets for emotional AI. *Artificial Intelligence Review*. 2025. Vol. 58, no. 10. URL: https://doi.org/10.1007/s10462-025-11271-1 (date of access: 25.04.2026).

7. Цифровий звукозапис – Вікіпедія. *Вікіпедія [Електронний ресурс]*. URL: https://uk.wikipedia.org/wiki/Цифровий_звукозапис (дата звернення: 25.04.2026).

8. International Telecommunication Union. Recommendation ITU-T P.342 (05/00): Transmission characteristics for telephone band (300–3400 Hz) digital loudspeaking and hands-free telephony terminals. 2000. URL: https://www.itu.int/rec/T-REC-P.342/en (дата звернення: 25.04.2026).

9. Real-Time Speech Emotion Recognition Using a Pre-trained Image Classification Network: Effects of Bandwidth Reduction and Companding / M. Lech et al. *Frontiers in Computer Science*. 2020. Vol. 2. URL: https://doi.org/10.3389/fcomp.2020.00014 (date of access: 25.04.2026).

10. Zakhozhai O. I., Krohmal A. V. An extensional approach to the recognition of raster images based on their symbolic transformations. *Scientific news of Dahl university*. 2023. URL: https://doi.org/10.33216/2222-3428-2023-25-2 (date of access: 25.04.2026).

11. Panagiotakis C., Tziritas G. A speech/music discriminator based on RMS and zero-crossings. *IEEE Transactions on Multimedia*. 2005. Vol. 7, no. 1. P. 155–166. URL: https://doi.org/10.1109/tmm.2004.840604 (date of access: 25.04.2026).

12. Rezapour Mashhadi M. M., Osei-Bonsu K. Speech emotion recognition using machine learning techniques: Feature extraction and comparison of convolutional neural network and random forest. *PLOS ONE*. 2023. Vol. 18, no. 11. P. e0291500. URL: https://doi.org/10.1371/journal.pone.0291500 (date of access: 25.04.2026).

13. Adavanne S. Sound event detection in multichannel audio using spatial and harmonic features / Adavanne S. et al. *Scenes and Events 2016 Workshop (DCASE2016)*. 2016. P. 6. URL: https://doi.org/10.48550/arXiv.1706.02293 (date of access: 25.04.2026).

14. SpectralFlux - Spectral flux for audio signals and auditory spectrograms - MATLAB. *MathWorks - Maker of MATLAB and Simulink*. URL: http://www.mathworks.com/help/audio/ref/spectralflux.html (date of access: 25.04.2026).

15. Feature extraction algorithms | OpenAE. *OpenAE*. URL: https://openae.io/standards/features/ (date of access: 25.04.2026).

16. Stevens S. S., Volkmann J. The Relation of Pitch to Frequency: A Revised Scale. *The American Journal of Psychology*. 1940. Vol. 53, no. 3. P. 329. URL: https://doi.org/10.2307/1417526 (date of access: 25.04.2026).

17. Bhalke D. G., Rao C. B. R., Bormane D. S. Automatic musical instrument classification using fractional fourier transform based- MFCC features and counter propagation neural network. *Journal of Intelligent Information Systems*. 2015. Vol. 46, no. 3. P. 425–446. URL: https://doi.org/10.1007/s10844-015-0360-9 (date of access: 25.04.2026)

18. Захожай О. І. Критерії визначення інформативності та ранжування образів при прийнятті рішень в багатопараметричних комбінованих системах розпізнавання. Електротехнічні та комп'ютерні системи *О.:ОНПУ*. 2018. № 27(103). с. 196-204.

19. Venkataramanan K., Rajamohan H. R. Emotion Recognition from Speech. *arXiv:1912.10458*. 2019. URL: https://arxiv.org/abs/1912.10458 (date of access: 25.04.2026).

20. Kaur J., Kumar A. Speech Emotion Recognition Using CNN, k-NN, MLP and Random Forest. *Computer Networks and Inventive Communication Technologies*. Singapore, 2021. P. 499–509. URL: https://doi.org/10.1007/978-981-15-9647-6_39 (date of access: 25.04.2026).

21. Boureau Y., Ponce J., Lecun. A Theoretical Analysis of Feature Pooling in Visual Recognition. *ICML 2010 – Proceedings of the 27th International Conference on Machine Learning*. 2010. P. 111–118.

22. Speech Emotion Recognition through Hybrid Features and Convolutional Neural Network / A. S. Alluhaidan et al. *Applied Sciences*. 2023. Vol. 13, no. 8. P. 4750. URL: https://doi.org/10.3390/app13084750 (date of access: 25.04.2026).

23. Waleed G. T., Shaker S. H. Speech Emotion Recognition on MELD and RAVDESS Datasets Using CNN. *Information*. 2025. Vol. 16, no. 7. P. 518. URL: https://doi.org/10.3390/info16070518 (date of access: 25.04.2026).

24. Reggiswarashari F., Sihwi S. W. Speech emotion recognition using 2D-convolutional neural network. *International Journal of Electrical and Computer Engineering (IJECE)*. 2022. Vol. 12, no. 6. P. 6594. URL: https://doi.org/10.11591/ijece.v12i6.pp6594-6601 (date of access: 25.04.2026).

25. Benzirar A., Hamidi M., Filali Bouami M. Building a speech emotion recognition system using RNN, GRU and LSTM. *International Journal of Speech Technology*. 2025. URL: https://doi.org/10.1007/s10772-025-10214-z (date of access: 25.04.2026).

26. Kilimci Z. H., Bayraktar Ü., Küçükmanisa A. Evaluating raw waveforms with deep learning frameworks for speech emotion recognition. *Multimedia Tools and Applications*. 2025. URL: https://doi.org/10.1007/s11042-025-20930-y (date of access: 25.04.2026).

27. Akinpelu S., Viriri S., Adegun A. An enhanced speech emotion recognition using vision transformer. *Scientific Reports*. 2024. Vol. 14, no. 1. URL: https://doi.org/10.1038/s41598-024-63776-4 (date of access: 25.04.2026).

28. MelTrans: Mel-Spectrogram Relationship-Learning for Speech Emotion Recognition via Transformers / H. Li et al. *Sensors*. 2024. Vol. 24, no. 17. P. 5506. URL: https://doi.org/10.3390/s24175506 (date of access: 25.04.2026).

29. Speech emotion recognition and classification using hybrid deep CNN and BiLSTM model / S. Mishra et al. *Multimedia Tools and Applications*. 2023. URL: https://doi.org/10.1007/s11042-023-16849-x (date of access: 25.04.2026).

30. From unaltered raw waveform to emotion: Synergizing convolutional and gated recurrent networks for holistic speech emotion analysis / A. Nfissi et al. *Applied Intelligence*. 2025. Vol. 55, no. 8. URL: https://doi.org/10.1007/s10489-025-06620-0 (date of access: 25.04.2026).

31. Speech Emotion Recognition Using Convolution Neural Networks and Multi-Head Convolutional Transformer / R. Ullah et al. *Sensors*. 2023. Vol. 23, no. 13. P. 6212. URL: https://doi.org/10.3390/s23136212 (date of access: 25.04.2026).

32. Rabiner L. R. Digital processing of speech signals. Englewood Cliffs, N.J : Prentice-Hall, 1978. 512 p.

33. Harris F. J. On the use of windows for harmonic analysis with the discrete Fourier transform. *Proceedings of the IEEE*. 1978. Vol. 66, no. 1. P. 51–83. URL: https://doi.org/10.1109/proc.1978.10837 (date of access: 25.04.2026).

34. CREMA-D: Crowd Sourced Emotional Multimodal Actors Dataset. *Kaggle [Електронний ресурс]*. URL: https://www.kaggle.com/datasets/ejlok1/cremad (дата звернення: 25.04.2026).

35. RAVDESS: Emotional speech audio dataset [Електронний ресурс] // Kaggle. URL: https://www.kaggle.com/datasets/uwrfkaggler/ravdess-emotional-speech-audio (дата звернення: 25.04.2026).

36. SAVEE Database: Speech emotion annotated data for emotion recognition systems. *Kaggle [Електронний ресурс]*. URL: https://www.kaggle.com/datasets/barelydedicated/savee-database (дата звернення: 25.04.2026).

37. Toronto Emotional Speech Set (TESS). *Kaggle [Електронний ресурс]*. URL: https://www.kaggle.com/datasets/ejlok1/toronto-emotional-speech-set-tess (дата звернення: 25.04.2026).

38. Urdu Language Speech Dataset: Urdu Language Speech Emotional Corpus. *Kaggle [Електронний ресурс]*. URL: https://www.kaggle.com/datasets/bitlord/urdu-language-speech-dataset (дата звернення: 25.04.2026).

39. EmoDB Dataset: Emotional Speech database for classification problem. *Kaggle [Електронний ресурс]*. URL: https://www.kaggle.com/datasets/piyushagni5/berlin-database-of-emotional-speech-emodb (дата звернення: 25.04.2026).

40. Russian Emotional Speech Dialogs (RESD): Open dataset for solving SER task with two modalities (voice & text). *Kaggle [Електронний ресурс]*. URL: https://www.kaggle.com/datasets/ar4ikov/resd-dataset (дата звернення: 25.04.2026).

41. Mexican Emotional Speech Database (MESD): Dataset of single-word emotional utterances of adult and child voices. *Kaggle [Електронний ресурс]*. URL: https://www.kaggle.com/datasets/saurabhshahane/mexican-emotional-speech-database-mesd (дата звернення: 25.04.2026).

42. Speech Noise Dataset. *Kaggle [Електронний ресурс]*. URL: https://www.kaggle.com/datasets/abdullahhaydarkadolu/speech-noise-dataset (дата звернення: 25.04.2026).

43. VAD. *webrtc Git repositories - Git at Google [Електронний ресурс]*. URL: https://webrtc.googlesource.com/src/+/refs/heads/main/common_audio/vad/ (date of access: 25.04.2026).

44. SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition / D. S. Park et al. *Interspeech 2019*. ISCA, 2019. URL: https://doi.org/10.21437/interspeech.2019-2680 (date of access: 25.04.2026).

45. MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications / A. Howard et al. 2017. https://doi.org/10.48550/arXiv.1704.04861 (date of access: 25.04.2026).

46. MobileNetV2: Inverted Residuals and Linear Bottlenecks / M. Sandler et al. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Salt Lake City, UT, 18–23 June 2018. 2018. URL: https://doi.org/10.1109/cvpr.2018.00474 (date of access: 25.04.2026).

47. Densely Connected Convolutional Networks / G. Huang et al. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, 21–26 July 2017. 2017. URL: https://doi.org/10.1109/cvpr.2017.243 (date of access: 25.04.2026).

48. Deep Residual Learning for Image Recognition / K. He et al. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, 27–30 June 2016. 2016. URL: https://doi.org/10.1109/cvpr.2016.90 (date of access: 25.04.2026).

49. ShuffleNet: An Extremely Efficient Convolutional Neural Network for Mobile Devices / X. Zhang et al. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Salt Lake City, UT, 18–23 June 2018. 2018. URL: https://doi.org/10.1109/cvpr.2018.00716 (date of access: 25.04.2026).

50. ShuffleNet V2: Practical Guidelines for Efficient CNN Architecture Design / N. Ma et al. *Computer Vision – ECCV 2018*. Cham, 2018. P. 122–138. URL: https://doi.org/10.1007/978-3-030-01264-9_8 (date of access: 25.04.2026).

51. SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <0.5MB model size / Iandola F. et al. 2016. URL: https://doi.org/10.48550/arXiv.1602.07360 (date of access: 25.04.2026).

52. Squeeze-and-Excitation Networks / Hu J. et al. 2017. URL: https://doi.org/10.48550/arXiv.1709.01507 (date of access: 25.04.2026).

53. RepVGG: Making VGG-style ConvNets Great Again / X. Ding et al. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Nashville, TN, USA, 20–25 June 2021. 2021. URL: https://doi.org/10.1109/cvpr46437.2021.01352 (date of access: 25.04.2026).

54. Trockman A., Kolter J. Patches Are All You Need? *Trans. Mach. Learn. Res.* 2023. URL: https://doi.org/10.48550/arXiv.2201.09792 (date of access: 25.04.2026).

55. Contemporary Advances in Neural Network Quantization: A Survey / M. Li et al. *2024 International Joint Conference on Neural Networks (IJCNN)*, Yokohama, Japan, 30 June – 5 July 2024. 2024. P. 1–10. URL: https://doi.org/10.1109/ijcnn60899.2024.10650109 (date of access: 25.04.2026).

56. ESP32-S3 Series Datasheet. Version 2. *Espressif Systems [Електронний ресурс]* URL: https://www.espressif.com/sites/default/files/documentation/esp32-s3_datasheet_en.pdf (дата звернення: 25.04.2026).

57. GitHub - espressif/esp-dsp: DSP library for ESP-IDF. *GitHub*. URL: https://github.com/espressif/esp-dsp (date of access: 25.04.2026).

58. GitHub - espressif/esp-dl: Espressif deep-learning library for AIoT applications. *GitHub*. URL: https://github.com/espressif/esp-dl (date of access: 25.04.2026).