МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
ХАРКІВСЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ
РАДІОЕЛЕКТРОНІКИ

МАТЕРІАЛИ XXX МІЖНАРОДНОГО МОЛОДІЖНОГО ФОРУМУ

РАДІОЕЛЕКТРОНІКА ТА МОЛОДЬ У XXI СТОЛІТТІ

Том 6

Харків 2026

МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
ХАРКІВСЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ
РАДІОЕЛЕКТРОНІКИ

МАТЕРІАЛИ XXX МІЖНАРОДНОГО
МОЛОДІЖНОГО ФОРУМУ

«РАДІОЕЛЕКТРОНІКА ТА МОЛОДЬ У XXI СТОЛІТТІ»

22 – 24 квітня 2026 р.

Том 6

КОНФЕРЕНЦІЯ «ІНФОРМАЦІЙНІ ІНТЕЛЕКТУАЛЬНІ СИСТЕМИ» INFORMATION INTELLIGENT SYSTEMS

Електронне видання

Харків 2026

УДК 004.89(06)

ХХХ Міжнародний молодіжний форум «Радіоелектроніка та молодь у ХХІ столітті». Зб. матеріалів форуму. Т. 6. / [Електронний ресурс] – Харків: ХНУРЕ. 2026. – 520 с. – pdf 14,4 Mb

ISBN 978-966-659-424-5

У збірнику представлено матеріали доповідей учасників ХХХ Міжнародного молодіжного форуму «Радіоелектроніка та молодь у ХХІ столітті».

Для науковців, викладачів, практичних працівників, здобувачів вищої освіти усіх рівней підготовки, а також широкого кола читачів, які цікавляться цією проблематикою.

Відповідальність за зміст поданого матеріалу несе його автор.

Електронне видання

Видання підготовлено факультетом комп’ютерних наук
Харківського національного університету радіоелектроніки

61166, Україна, Харків, просп. Науки, 14
тел./факс: (057) 7021397

E-mail: mref21@nure.ua

ISBN 978-966-659-424-5

УДК 004.85

АДАПТАЦІЯ АРХІТЕКТУРИ ГЛИБИННОЇ ЗГОРТКОВОЇ МОДЕЛІ ДЛЯ ЕМОЦІЙНОГО АНАЛІЗУ В EDGE AI СИСТЕМАХ

Крохмаль А.В.
e-mail: andrii.krokhmal.cpe@nure.ua
Науковий керівник – к.т.н., доц. Селіванова К.Г.
Харківський національний університет радіоелектроніки, каф. ШІ
м. Харків, Україна

This work is devoted to the selection of neural network architectures for the task of speech-based emotion recognition (SER). The objective is to enhance recognition performance on devices with constrained computational resources. For model training, publicly available datasets were employed, from which log-mel spectrograms were extracted as input representations. An experimental evaluation was carried out to investigate the effectiveness of lightweight adaptations of several widely used convolutional neural architectures.

Задача розпізнавання емоційних станів людини залишається не повністю вирішеною на сьогоднішній день та вимагає міждисциплінарних досліджень у сферах психофізіології, когнітивних наук та штучного інтелекту. Сучасні методи та засоби емоційного аналізу забезпечують високу достовірність розрізнення емоцій на основі аналізу фізіологічних параметрів (зміни тиску, температури тіла, електромагнітної активності мозку) та невербальної поведінки (рухи, міміка). Основними обмеженнями такого підходу є необхідність у постійному контакті з людиною та достатньо складному та вартісному обладнанні. Альтернативою є дистанційний моніторинг, що використовує аудіо або візуальні ознаки, зокрема зміни у голосі людини. Його очевидними перевагами є безконтактний спосіб отримання даних та значно нижчі вимоги до обладнання, яке може бути виконане у форматі носимого пристрою. Метод аналізу емоцій на основі голосу не здатен забезпечити точність, притаманну мультимодальним методам, однак може бути скомбінованим з іншими джерелами даних [1], або використаним у прикладних областях, де це є прийнятним. До них можна віднести покращення людино-машинних інтерфейсів голосових чат-ботів, систем аналізу відгуків клієнтів кол-центрів, освітні та маркетингові дослідження тощо. Отже, розробка нових та вдосконалення існуючих методів розпізнавання емоційних станів за голосом людини є актуальною науково-технічною задачею.

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

Для застосування методів машинного навчання задачу емоційного аналізу можна формалізувати як задачу багатокласової класифікації

окремих зразків. Така постановка спрощує архітектуру та алгоритм навчання моделі, оскільки зразки аналізуються поза контекстом. Однак, опис зразка однією міткою створює обмеження, пов'язані з перекриттям кількох емоцій в межах аудіозапису, що негативно впливатиме на достовірність класифікації.

Для навчання моделей та тестування їх ефективності було створено набір даних шляхом об'єднання 9 популярних датасетів емоційного мовлення у відкритому доступі: 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. Загальний обсяг склав 22808 зразків, поділених на 7 класів (6 емоційних станів та відсутність голосу). Сформовано тренувальну та тестову вибірки зі співвідношенням 0,8:0,2.

Цифровий аудіосигнал не є прийнятним для обробки засобами машинного навчання через високу надлишковість інформації, тому вимагає застосування методів виділення ознак. У дослідженні було використано метод лог-мел-спектрограм, що є типовим для задач обробки голосового сигналу. Він базується на масштабуванні спектрограми у частотному вимірі за мел-шкалою, а в енергетичному – за шкалою децибелів для імітації особливостей людського сприйняття звукових сигналів, яке є нелінійним. Було використано монофонічні аудіозаписи з частотою дискретизації 8 кГц. Відповідно до теореми Котельникова, це дозволяє передати ознаки з частотами до 4 кГц, що охоплює значну частку спектру людського голосу та забезпечує достатню точність в задачах SER [2]. Тривалість аудіозаписів обрана відповідно до модального значення тренувальної вибірки: 2,5 с, розмір вікна: 256 семплів, кількість мел-бінів: 78. В результаті отримано компактне представлення ознак розмірністю 78x78.

Для здійснення класифікації обрано згорткову нейронну мережу (CNN), що є однією з найефективніших архітектур для даної задачі, завдяки своїй здатності виявляти складні структурні патерни у багатовимірних представленнях, яким є спектрограма. Також CNN здатні визначати інваріантні представлення навіть за високої кореляції ознак, що знижує вимоги до методу попередньої обробки.

В результаті експериментальних досліджень було здійснено апробацію 11 найбільш використовуваних згорткових архітектур. Зважаючи на невеликий розмір вхідних ознак та необхідність запуску на пристроях класу вбудованих систем, референсну архітектуру моделей було змінено так, щоб кількість параметрів не перевищувала 0,1 М. Зокрема, «класична» CNN є тришаровим стеком з конфігурацію каналів: 1/32/64/128 зі згортками 3x3/2, BN та ReLU. Для MobileNet v1 α=0,25 та depth=0,2, а для MobileNet v2 – 0,15 і 0,2 відповідно. ResNet має згортки 3x3/1 і 3 стадії з конфігурацією каналів 16/24/48. DenseNet-BC реалізована з k=16 і θ=0.5. SqueezeNet v1.1 має і 4 Fire-модулі (squeeze=32,

expand1x1=expand3x3=64). ShuffleNet v1 має g=2 та конфігурацію 96/192/384, а ShuffleNet v2: 56/104/208. SENet є інтеграцією SE-модулів з r=16 у стек 16/32/64/128 (кожен рівень: згортка 3x3/1 та MaxPool). RepVGG в режимі інференсу є стеком 1/32/64/128 (по 1 блоку на стадію). ConvMixer має p=3, h=128, d=4 та k=5.

Для кожної моделі здійснено навчання протягом 150 епох з коефіцієнтом навчання 0,001 та оптимізатором Adam. Для запобігання перенавчанню застосовано L2-регуляризацію (λ=0,001) та згладжування міток (ε=0,01). Для аугментації даних використано метод SpecAugment [3] та адитивний гаусівський шум. Результати експериментальної оцінки ефективності розглянутих моделей за макро- та зваженими метриками: точність, прецизійність, повнота та F1-міра – наведені на рис. 1.

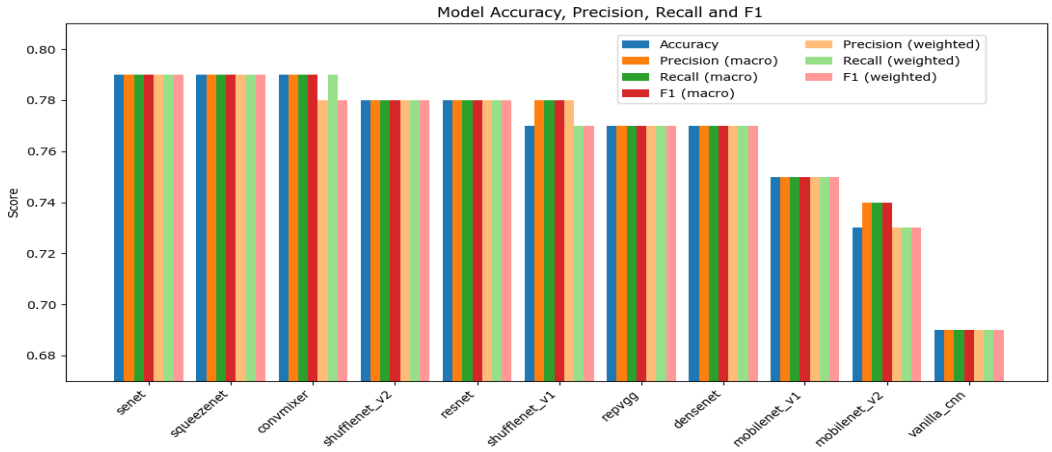

Рисунок 1 – Порівняльний аналіз ефективності обраних моделей CNN

Таким чином, результати експериментальної оцінки достовірності розпізнавання емоційних станів демострують перевагу модифікованих архітектур SENet, SqueezeNet та ConvMixer. Вони показали значення точності 79%, зберігаючи низькі вимоги до обсягів пам'яті та швидкості отримання класифікаційного рішення, тому можуть бути застосовані для виведення на Edge-пристроях в режимі реального часу.

Список використаних джерел:

1. Al-Azani S., El-Alfy ES.M. A review and critical analysis of multimodal datasets for emotional AI. Artif Intell. Rev 58, 334. 2025. DOI: 10.1007/s10462-025-11271-1.

2. Lech M. Real-Time Speech Emotion Recognition Using a Pre-trained Image Classification Network: Effects of Bandwidth Reduction and Companding / M. Lech et al. Front. Comput. Sci. 2:14. DOI: 10.3389/fcomp.2020.00014.

3. Park D. SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition / D. Park et al. Interspeech 2019. ISCA, 2019. DOI: 10.21437/interspeech.2019-2680.