УДК 004.8+004.032.26+159.942+534.4

С. О. Жуков, канд. техн. наук, доц.;О. В. Рудзевич

ОПТИМІЗАЦІЯ ГЛИБОКИХ НЕЙРОННИХ МЕРЕЖ ДЛЯ КЛАСИФІКАЦІЇ ЕМОЦІЙНОГО СТАНУ МОВЛЕННЯ З ВИКОРИСТАННЯМ ДИНАМІЧНОГО КВАНТУВАННЯ

У статті представлено результати системного дослідження з аналізу емоційної тональності аудіозаписів із використанням методів глибокого навчання. Основною метою є створення моделі, здатної ефективно класифікувати емоції мовлення в умовах обмежених обчислювальних ресурсів без суттєвих втрат точності. Обґрунтовано актуальність проблеми для різних інформаційних систем реального часу, зокрема і висвітлено огляд наявних підходів з їхніми недоліками та можливостями для покращень. Далі було обрано набір даних для проведення дослідження, яким став «Emotional Speech Dataset». Для уніфікації довжини записів застосовано модифіковану стратегію заповнення нулями, що випадково розподіляє доповнення між початком і кінцем сигналу.

У процесі передобробки видобуто як часові, так і часово-частотні ознаки аудіозаписів, що забезпечило багатше представлення властивостей вхідних даних. Після цього було виконано навчання нейронних мереж з використанням елементів з довгою короткочасною пам'яттю, з блоками багатоголової уваги, а також зі згортковими шарами. Найвищу точність показала модель зі згортковими шарами, а саме близько 95% на тестувальних даних, тоді як дві інші моделі мали точність 90 та 93 % відповідно.

Для пришвидшення інференсу та зменшення розміру моделей було застосовано динамічне квантування, внаслідок чого нейронна мережа зі згортковими шарами погіршила показник точності на тестувальних даних до 92 %, однак швидкість інференсу зменшилася вчетверо і склала 0,4 мс, а обсяг пам'яті для зберігання зменшився більш ніж удесятеро і склав 74 кБ. Схожу поведінку щодо прискорення інференсу та зменшення обсягу пам'яті за рахунок пригнічення точності демонстрували і решта нейронних мереж.

Аналіз результатів тестування на помилки показав те, що всі нейронні мережі якщо й помиляються, то найчастіше при сплутуванні радісної емоції з гнівної, а також нейтральної із сумною, що вказує на потребу подальшого розширення переліку ознак, зокрема врахування тембральних характеристик і інтонаційних закономірностей.

***Ключові слова**: емоційна тональність, розпізнавання емоцій, аудіозапис, аудіомовлення, класифікація, видобування ознак, класифікація мовлення, машинне навчання, глибоке навчання, нейронна мережа, довга короткочасна пам'ять, багатоголова увага, згортковий шар, LSTM, CNN, оптимізація нейронної мережі, динамічне квантування.*

Вступ

Сучасні інформаційні системи дедалі більше орієнтуються на врахування емоційного стану користувачів, адже це безпосередньо впливає на ефективність взаємодії і якість прийнятих рішень. Аналіз емоційної тональності аудіозаписів критично важливий у тих випадках, коли система обробляє великі обсяги голосової інформації в реальному часі або близько до нього.

Завдяки цьому можна не тільки швидше відреагувати на кризові ситуації, але й надати персоналізований сервіс, що підвищує рівень задоволеності користувачів. Розпізнавання емоційної тональності є одним із провідних компонентів у таких складних системах, як:

а) система дистанційного психологічного консультування, яка містить підсистеми обробки аудіо, збереження історій пацієнтів, взаємодії з профільними фахівцями й іншими медичними системами тощо;

б) система взаємодії з користувачами в освітній платформі, яка складається з підсистем платформи для курсів, комунікації з викладачами, аналізом прогресу, чат-ботами підтримки тощо;

в) система call-центру багаторівневої підтримки, яка має підсистеми операторів різних рівнів, автоматизованого інтерактивного голосового меню, відносин з клієнтами, аналітичних модулів тощо [1].

Таким чином, аналіз емоційної тональності аудіозаписів є невід'ємною частиною комплексних процесів в різних галузях, що забезпечують масове обслуговування та взаємодію з користувачами. Завдяки сучасним методам обробки сигналів і моделювання такі системи здатні підвищувати рівень задоволеності й довіри у взаєминах між людиною і технологіями.

Огляд останніх досліджень та публікацій

На просторах інтернету можна знайти різноманітні підходи до аналізу емоційної тональності аудіозаписів. У результаті пошуку були виявлені та досліджені такі рішення:

а) «Speech-Emotion-Classification» – репозиторій на GitHub, у якому пропонується побудова класифікаторів емоцій у аудіозаписах різного рівня деталізації за допомогою згорткових нейронних мереж та методів опорних векторів; стандартний пайплайн передбачає завантаження аудіофайлів, витягування визначеного набора ознак і збереження їх для подальшого навчання моделей машинного навчання [2];

б) «pontonkid/Audio-Sentiment-Analysis» – репозиторій на GitHub, що демонструє створення інтерактивного вебзастосунку з використанням бібліотеки Streamlit мови програмування Python; ключовим етапом є перетворення мовлення в текст засобами бібліотеки speech_recognition, а далі – аналіз сентименту за допомогою попередньо натренованої моделі[3];

в) «Sentimental analysis of audio based customer reviews without textual conversion» – стаття, у якій обґрунтовується перевага роботи з аудіоданими для поліпшення обслуговування клієнтів порівняно з аналізом текстових відгуків; автори пропонують модель нейронної мережі з використанням різних типів ознак; також описано систему, яка зберігає всі голосові відгуки у спеціальній базі даних, видобуває набір ознак для тренування моделі й тестує нові аудіозаписи на основі тих самих параметрів [4];

г) «Tilak612/Audio-Sentiment-Analysis» – репозиторій на GitHub, що пропонує реалізувати вебзастосунок із використанням Flask; тут перетворення мовлення на текст здійснюється через SpeechToTextV1 від IBM Watson, а рівень інтенсивності сентименту визначається за допомогою модуля nltk.sentiment.vader мови програмування Python задля аналізу емоцій у відгуках користувачів про сервіс задля подальшого покращення обслуговування [5].

Провівши аналіз обраних підходів, було виявлено такі недоліки:

а) у деяких реалізаціях моделей спостерігається високий ризик перенавчання;

б) відсутність будь-якого виду оптимізації вже натренованих моделей машинного навчання;

в) переважна орієнтація на використання лише текстової складової аудіозапису для визначення емоційної тональності.

Мета і завдання дослідження

Метою цієї роботи є здійснення аналізу емоційної тональності аудіозаписів із застосуванням методів машинного навчання для підвищення якості прийняття рішень шляхом оптимізації точності й швидкості інференсу як основного критерію ефективності.

Серед можливих обмежень слід врахувати недостатність обчислювальних ресурсів під час навчання моделей, спотворення вхідних аудіоданих, наявність мультиакцентного мовлення тощо. Можна також виділити такі задачі дослідження:

а) визначення репрезентативного датасету для проведення дослідження;

б) створення переліку ознак для видобування з аудіозаписів для їх багатшого представлення;

в) вибір структури моделі машинного навчання та обґрунтування підходів до її оптимізації;

г) навчання, оптимізація й тестування обраних моделей на підготовлених наборах даних;

д) виконання оцінки впливу результатів виявлення емоційної тональності на прийняття рішень та визначення оптимальної моделі машинного навчання.

Вибір даних для навчання й перевірки моделей машинного навчання

Перед виконанням аналізу емоційної тональності аудіозаписів одним з ключовихетапів є вибір правильного набору даних для навчання й перевірки моделей машинного навчання, який повинен відповідати таким вимогам:

а) достатня кількість зразків у кожному класі, що забезпечить якісне навчання та зменшить ризик перенавчання;

б) репрезентативність даних із реальним відображенням розподілу класів, щоб моделі могли працювати з різноманітними прикладами;

в) різноманітність мовців з урахуванням інтонацій, акцентів та інших характеристик, що сприятиме кращому узагальненню.

Відповідно до зазначених критеріїв, для виконання дослідження було обрано «Emotional Speech Dataset (ESD)» [6]. Цей набір містить 350 висловлювань, записаних 10 носіями англійської та 10 носіями китайської мови, та охоплює кілька емоційних категорій. Загальна тривалість записів перевищує 29 годин.

Структура датасету така: для кожного мовця створено 20 папок, а в кожній з них — підпапки, що відповідають різним емоціям, у яких зберігаються WAV-файли. Кожен мовний корпус містить по п'ять емоційних категорій: злість, радість, сум, нейтральний стан і здивування. Проте перші чотири емоції є полярними і широко представлені, водночас «здивування» може мати як позитивне, так і негативне забарвлення, що ускладнює його однозначне виявлення та здатне спотворити результати. З огляду на це в рамках цієї роботи було вирішено виключити категорію «здивування» з подальшого аналізу.

Отже, вибір датасету «EmotionalSpeechDataset» є обґрунтованим завдяки його емоційній різноманітності та структурованості.

Принципи передобробки аудіоданих та добування ознак

Оскільки вибраний датасет має чітко структуровану файлову систему, завантажувати аудіофайли в оперативну пам'ять досить просто. Водночас важливою проблемою є різна тривалість записів: кожен файл може мати різну довжину, що створює незручності під час тренування й тестування моделей, які зазвичай розраховані на вхідні дані однакової довжини.

Для вирівнювання тривалості аудіозаписів було розглянуто традиційний метод заповнення нулями: до коротших файлів додають нулі до кінця, доки їх довжина не зрівняється з найдовшим зразком. Проте такий підхід може призвести до того, що модель «звикне» до специфічного шаблону (початок завжди невкорочений, а завершення — суцільні нулі), що знижує здатність узагальнення на нових даних.

Для уникнення цього було модифіковано базову процедуру заповнення: замість додавання всіх нулів у кінець, їхню загальну кількість розділено випадково між початком і кінцем аудіозапису. Завдяки цьому заповнення нулями на обох кінцях файлу дані набувають більшої різноманітності, а навчання моделей проходитиме менш шаблонно.

Провівши огляд досліджень [7 – 9], присвячених аналізу аудіозаписів для різних прикладних задач, було здійснено комплексний підхід до вибору переліку ознак для подальших видобування й моделювання, а саме:

а) частота перетинань нуля – часова характеристика, яка показує, скільки разів сигнал змінює знак у межах певного інтервалу, та може бути знайдена за формулою (1):

$$ZCR = \frac{1}{2T}\sum_{t=1}^{T}\left|\mathrm{sgn}\big(s(t)\big) - \mathrm{sgn}\big(s(t+1)\big)\right|, \#(1)$$

де ZCR – частота перетинань нуля, T – кількість дискретних значень в досліджуваному інтервалі, $s(t)$ – функція тиску звукового сигналу;

б) середньоквадратичне значення тиску звукового сигналу, яке відображає «енергію» сигналу, визначається формулою (2) та також є часовою характеристикою:

$$RMS = \sqrt{\frac{1}{T}\sum_{t=1}^{T} s^2(t)}, \#(2)$$

де RMS – середньоквадратичне значення тиску, T – кількість дискретних значень в досліджуваному інтервалі, $\mathrm{s}(t)$ – функція тиску звукового сигналу;

в) кепстральні коефіцієнти шкали Мела (MFCC) – часово-частотна характеристика звукового сигналу, яка відображає його спектральні властивості з урахуванням сприйняття частот людським слухом та використовується для аналізу ознак тембру звукозапису;

г) висота тону – частотна характеристика звукового сигналу, яка характеризує сприйняття основної частоти звуку людиною; фізично пов'язана від частоти звукових хвиль, але не є її еквівалентом через суб'єктивність сприйняття цієї величини;

д) спектральний спад – частотна характеристика звукового сигналу, яка визначає ту частоту, нижче якої зосереджено певний відсоток (зазвичай 80 – 85 %) від загальної енергії спектру; вище цієї частоти енергія звуку вважається незначною;

е) спектральний контраст – частотна характеристика звукового сигналу, яка відображає різницю між максимальними та мінімальними амплітудами в окремих частотних діапазонах спектру; може бути знайдена за формулою (3):

$$C_1 = 10\lg\left(\frac{P_{max}}{P_{min}}\right), \#(3)$$

де C_1 – спектральний контраст, P_{max} (P_{min}) – середнє значення найбільших (найменших) амплітуд звукового сигналу;

ж) спектральний центроїд – частотна характеристика звукового сигналу, яка вказує на «центр мас» спектру, тобто частоту, навколо якої сконцентрована більша частина його «енергії»; може бути обчислена за формулою (4):

$$C_2 = \sum_{n=1}^{N} f(n)|S(n)| \div \sum_{n=1}^{N} |S(n)|, \#(4)$$

де C_2 – спектральний центроїд, $f(n)$ – функція частоти звукового сигналу, $S(n)$ – функція спектру частот на комлексній площині;

з) хрома-функція – частотна характеристика звуку, яка групує енергію сигналу за повторюваними шаблонами частот, які мають однакове співвідношення частот, але розташовані в різних діапазонах.

Обґрунтування вибору структури моделей машинного навчання

Машинне навчання – це набір методів і технологій для розв'язання прикладних завдань, у основі якого лежить «тренування» моделей на доступних даних. Воно є невід'ємною складовою «науки про дані».

Одним із різновидів машинного навчання є навчання з учителем, коли модель навчається відображати зв'язки між вхідними ознаками та заданими цільовими значеннями, щоб потім робити прогнози на нових даних. Цей підхід охоплює дві основні категорії задач: класифікацію, де потрібно вибрати правильну мітку з фіксованого набору класів, та

регресію, коли треба передбачити параметр із безперервного діапазону [10]. Для аналізу емоційної тональності аудіозаписів у цьому дослідженні застосовується саме задача класифікації.

Деякими прикладами методів машинного навчання з учителем є метод k-найближчих сусідів, лінійні моделі, дерево рішень або їх ансамбль, однак головним недоліком таких підходів є насамперед те, що вони працюють з двовимірними наборами даних, де одна вісь містить кількість прикладів, а інша – кількість ознак.

Так як з аудіоданих можна видобути корисні для класифікації тривимірні ознаки з урахуванням їхньої зміни в часі, використати вищезгадані види моделей можна за допомогою перетворення в двовимірне представлення такими способами, як, скажімо:

а) розгортання вздовж однієї з осей – перетворення частини багатовимірного масиву на одновимірний вектор уздовж обраної осі;

б) агрегація за однією з осей – обчислення агрегованих показників, наприклад, середнього значення, максимуму тощо вздовж однієї з осей.

Недоліком подібних підходів є втрата часової послідовності: при розгортанні зникає інформація про порядок подій, а при агрегації пригладжуються важливі особливості сигналу, що зменшує різноманітність ознак.

Видом машинного навчання з учителем, який дозволяє зберігати тривимірні дані в їхньому початковому вигляді, тим самим усуваючи недоліки, перелічені вище, є штучні нейронні мережі, або коротко – нейронні мережі. Це модель штучного інтелекту, що імітує принципи роботи людської нервової системи, зокрема здатність до навчання і виправлення помилок. Базовою одиницею вищеописаної структури є штучний нейрон.

Нейронна мережа складається з багатьох взаємопов'язаних нейронів, організованих у шари таких видів, як вхідний, приховані та вихідний. Під час навчання дані проходять через шари послідовно, а вагові коефіцієнти між нейронами оновлюються, щоб мінімізувати похибку прогнозів і поступово підвищувати точність моделі [11].

Для аналізу емоційної тональності аудіозаписів методами машинного навчання важливо розглядати набір ознак як часову послідовність: динаміка змін характеристик у часі може мати ключове значення для правильного визначення емоційного стану.

У таких випадках слід використовувати рекурентні нейронні мережі, які здатні враховувати контекст та динаміку змін в часових послідовностях завдяки наявності зворотніх зв'язків. Проте базові рекурентні архітектури стикаються з серйозною проблемою – «затуханням» або «вибухом» градієнтів, що унеможливлює ефективне тренування на довгих послідовностях, тому що модель фактично «забуває» важливі закономірності, які могли трапитися раніше.

Щоб вирішити цю проблему, застосовуються удосконалена версія рекурентних нейронних мереж – елементи з довгою короткочасною пам'яттю (LSTM), які мають спеціальні додаткові механізми – комірки пам'яті та керуючі вентилі, які дозволяють контролювати потік інформації, вирішуючи, яку інформацію потрібно зберігати, а яку необхідно передавати до наступних складових нейронної мережі.

Кожен елемент з довгою короткочасною пам'яттю працює з трьома основними елементами: вхідним вектором, прихованим станом та станом комірки пам'яті. Усі операції цього елемента базуються на чотирьох вентилях:

а) вентиль забування, що визначає, яку частину попередньої пам'яті (5) варто забути:

$$f_t = \sigma\left(W_f x_t + U_f h_{t-1} + b_f\right), \#(5)$$

де f_t – вектор значень для «забування», x_t – вхідний вектор, h_{t-1} – прихований стан попереднього кроку, W_f, U_f – матриці вагів для вхідного вектора та прихованого стану відповідно, C_{t-1} – стан комірки пам'яті попереднього кроку, b_t – вектор зсуву;

б) вентиль входу, що знаходить, яку нову інформацію (6) слід додати до стану пам'яті, та паралельно з яким створюється кандидатний вектор стану пам'яті (7), що представляє нову потенційну інформацію для збереження:

$$i_t = \sigma(W_i x_t + U_i h_{t-1} + b_i), \#(6)$$
$$\tilde{C}_t = \tanh(W_C x_t + U_C h_{t-1} + b_C), \#(7)$$

де i_t – вектор вентиля входу, x_t – вхідний вектор, h_{t-1} – прихований стан попереднього кроку, $\tilde{C}_t$ – кандидатний вектор стану пам'яті, W_i, U_i, b_i та W_C, U_C, b_C – матриці вагів для вхідного вектора й прихованого стану з векторами зсуву для вектору вентиля входу та кандидатного вектору стану пам'яті відповідно; після проходження через цей вентиль обчислюється оновлений стан комірки пам'яті за допомогою поєднання «забутої» інформації з попереднього кроку та нової наданої інформації за формулою (8):

$$C_t = f_t \odot C_{t-1} + i_t \odot \tilde{C}_t, \#(8)$$

де C_t – оновлений стан комірки пам'яті, f_t – вектор значень для «забування», C_{t-1} – стан комірки пам'яті попереднього кроку, i_t – вектор вентиля входу, $\tilde{C}_t$ – кандидатний вектор стану пам'яті, $\odot$ – оператор добутку Адамара [12];

в) вентиль виходу, який контролює яка частина інформації з пам'яті буде передана далі як новий прихований стан (9):

$$h_t = \sigma(W_o x_t + U_o h_{t-1} + b_o) \odot \tanh C_t, \#(9)$$

де h_t – вектор нового прихованого стану, x_t – вхідний вектор, h_{t-1} – прихований стан попереднього кроку, W_o, U_o, b_o – матриці вагів для вхідного вектора й прихованого стану з векторами зсуву, $\odot$ – оператор добутку Адамара.

Суттєвим недоліком використання будь-якого типу рекурентних нейронних мережі є поелементна обробка часових даних, яка значно сповільнює тривалість навчання моделі й ускладнює моделюванням тривалих залежностей. Використання спеціальних блоків, як-от, елементів з довгою короткочасною пам'яттю лише частково нівелює цю проблему.

У зв'язку з цим особливу роль відіграє використання механізму уваги, який дозволяє моделі визначати найбільш значущі частини вхідної послідовності на кожному кроці її обробки. За такого підходу розглядаються всі частини одразу та обчислюються ваги уваги, які визначають важливість кожної із них відносно інших.

Одним із поширених блоків з механізмом уваги, який розширює базовий функціонал, є блок з механізмом багатоголової уваги (Multi-head attention), в якому кілька незалежних «голів» уваги паралельно аналізують різні аспекти взаємозв'язків у послідовності. Це дає змогу моделі виявляти складніші закономірності у вхідних даних, підвищуючи її здатність до узагальнення.

Під час роботи кожної «голови» вхідні дані спочатку проєктуються у три окремі простори – запитів (10), ключів (11) і значень (12) за допомогою відповідних вагових матриць:

$$Q = XW^Q \#(10)$$
$$K = XW^K \#(11)$$
$$V = XW^V \#(12)$$

де X – вхідна послідовність, Q, K, V – запит, ключ і значення відповідно, W^Q, W^K, W^V – матриці вагів для запиту, ключа й значення відповідно.

Після отримання представлень у трьох просторахдалі показник уваги за цими складовими можна знайти за формулою (13):

$$A = S\left(\frac{QK^T}{\sqrt{d_k}}\right)V\#(13)$$

де A – увага однієї частини блоку багатоголової уяви, Q, K і V – запит, ключ й значення відповідно, S – нормована експоненційна функція, d_k – розмірність векторів ключів, яку можна знайти за формулою (14):

$$d_k = \frac{d_{model}}{h}\#(14)$$

де d_k – розмірність векторів ключів, d_{model} – розмірність вхідних даних, h – кількість голів у блоці багатоголової уваги.

У цій формулі $\sqrt{d_k}$ є коефіцієнтом масштабування, який стабілізує градієнти при навчанні, особливо у випадку великих розмірностей. Ігнорування цього коефіцієнта може призвести до насичення нормованої експоненційної функції, внаслідок чого модель стає менш чутливою до різниці у значеннях.

У кінці роботи блоку з механізмом багатоголової уваги поточні результати кожної складової конкатенуються, далі проходять через повнозв'язний шар, який перетворює результат в потрібну розмірність, тобто істинним є співвідношення (15), та передаються до наступних складових нейронної мережі:

$$M = (A_1 \,||\, \dots \,||\, A_n)W^O\#(15)$$

де M – результуюча матриця блоку багатоголової уваги, A_i, $i = \overline{1,n}$ – уваги частин блоку багатоголової уваги, W^O – матриця вагів повнозв'язного шару, $||$ – оператор конкатенації [13].

Ще одним прикладом побудови структури нейронної мережі для поставленої задачі є використання згорткових шарів (CNN). Це ключовий компонент однойменних нейронних мереж, основна мета яких – автоматичне виявлення таких ознак, як контури, текстури тощо. Це виявлення досягається за допомогою застосування фільтрів, які «ковзають» по вхідних даних та виконують операцію згортки для отримання карти ознак. Кожен такий фільтр навчається виявляти певний тип ознаки, а глибші шари можуть розпізнавати складніші структури.

Найчастіше згорткові шари мають двовимірні фільтри для роботи із зображеннями, проте для часових послідовностей існує відповідний одновимірний аналог. У такому випадку кожна позиція згортки відповідає часовому кроку, а фільтр виконує агрегацію значень у певному вікні часу.

Застосування згорткового шару може зумовлювати те, що розмірність вихідних даних стає меншою, аніж вхідних, що є їх важливою властивістю. За великої кількості таких шарів розмір результуючих даних в деяких задачах може бути критично низьким, тому для збереження поточної розмірності використовуються доповнення нулями таким чином, аби вихідні дані мали таку ж форму, як і вхідні.

Ще одним визначальним параметром згорткових шарів є крок згортки, який визначає, на скільки позицій фільтр зміщується під час виконання цієї операції. Чим менше це значення, тим більшим стає розмір карти ознак, тому що фільтр в такому випадку охопить більше позицій, і навпаки.

Тому, як результат, значення елемента вихідного багатовимірного масиву після проходження через згортковий шар може бути виражене формулою (16):

$$Y_{t,f} = \sum_{i=0}^{K-1} \sum_{c=0}^{C-1} W_{i,c,f} \cdot X_{t \cdot S+i-P,c} + b_f \#(16)$$

де Y (X) – вихідний (вхідний) багатовимірний масив, W – матриця вагів згорткового шару, t – індекс позиції у послідовності, f – індекс фільтра згорткового шару, b – вектор зсуву, C – кількість вхідних каналів, P – розмір доповнення вхідних даних, K – розмір фільтра, S – крок згортки.

Згорткові шари також поєднуються з шарами підвибірки для зменшення розмірності та збереження найважливішої інформації. Існує кілька видів підвибірок:

а) максимальна, яка обирає найбільше значення в кожному локальному регіоні ознак, що дозволяє зберігати їхні найяскравіші прояви;

б) середня, яка обчислює відповідне значення в кожному регіоні, згладжуючи наявну інформацію;

в) глобальна, яка застосовується до всієї карти ознак і часто використовується перед останніми шарами мережі.

Після застосування згорткового шару до вхідних ознак розмір вихідних ознак змінюється та може бути знайдений за формулою (17):

$$L_{out} = \left\lfloor \frac{L_{in} + 2P - K}{S} \right\rfloor + 1 \#(17)$$

де L_{out} – розмірність вихідних ознак, L_{in} – довжина вхідних ознак, P – розмір доповнення вхідних даних, K – розмір фільтра, S – крок згортки [14].

Опис складових структури нейронних мереж веде за собою вибір та опис функції втрат, яка визначає, наскільки добре модель машинного навчання виконує правильні передбачення. Мінімізації значення цієї функції безпосередньо впливає на рівень її роботи. Для поставленого завдання, яке містить класифікацію з кількома взаємовиключними класами, найбільш поширеною функцією втрат є перехресна ентропія.

Припустимо завдання класифікації передбачає N класів. Кожну істинну мітку $k \in \{1,2,\dots,N\}$ позначимо вектором $y = (y_1, y_2, \dots, y_N)$, елемент y_i, $i = \overline{1,N}$ якого визначений правилом (18):

$$y_i = \begin{cases} 1, & i = k \\ 0, & i \neq k \end{cases} \#(18)$$

Тоді з урахуванням вектору ймовірностей класів $p = (p_1, p_2, \dots, p_N)$ для одного вхідного прикладу, який видала модель машинного навчання, перехресна ентропія має вигляд (19):

$$L(y, p) = -\sum_{i=1}^{N} y_i \log p_i \#(19)$$

де L – розглянута функція втрат, N – кількість класів у завданні класифікації, y, p – вектори істинних значень та передбачень для одного вхідного прикладу відповідно.

Для оцінки якості роботи моделі на повному наборі даних використовується така метрика, як категоріальна точність, яка представлена відношенням кількості правильно передбачених прикладів до її загальної кількості, та виражається співвідношенням (20):

$$CA = \frac{1}{M} \sum_{i=0}^{M} 1(\arg\max_i p_i = \arg\max_i y_i) \#(20)$$

де CA – категоріальна точність, y_i, p_i,$i = \overline{1, M}$ – вектори істинних значень та передбачень для повного набору даних, M – кількість прикладів у повному наборі даних [10].

Розглянуті підходи до створення структури моделей машинного навчання для виконання системного аналізу емоційної тональності аудіозаписів спрямовані на підвищення точності класифікації шляхом збереження важливих часових закономірностей та структурних особливостей даних.

Принципи оптимізації моделей машинного навчання

Сучасні моделі машинного навчання, зокрема нейронні мережі, демонструють високу точність у різних задачах, але досягають цього часто за рахунок значних обчислювальних ресурсів. Велика кількість параметрів та складні обчислення призводять до збільшення часу навчання й інференсу, а також до зростання енергоспоживання, що особливо критично при обмежених ресурсах.

Тому в таких випадках доцільно виконувати оптимізацію моделей машинного навчання – процес покращення ефективності та продуктивності за рахунок вдосконалення їхньої структури та функціонування. Основними цілями такої оптимізації є:

а) зменшення розміру моделі – мінімізація кількості параметрів і обсягу пам'яті допомагає легше переносити модель на периферійні пристрої та економити простір для зберігання;

б) прискорення інференсу – скорочення часу обробки запиту до моделі досягається завдяки зменшенню кількості обчислень або застосуванню апаратно-оптимізованих операцій.

Зрозуміло, що оптимізована модель машинного навчання повинна передбачати з такою ж точністю, як і звичайна, або максимально близькою до неї, тому що цінність покращення одного показника буде нівелюватися значним погіршенням іншого.

Для нейронних мереж існують такі популярні методи оптимізації:

а) прунінг – видалення слабозначущих ваг або зайвих нейронів з метою зменшення надмірності; зазвичай охоплює три етапи: визначення найменш вагомих елементів, їх видалення та додаткове донавчання скороченої мережі для компенсації втрати інформації;

б) дистиляція знань – навчання компактнішої моделі «студента» на основі виходів більш складної моделі «вчителя», де перша імітує виходи другої; завдяки перенесенню знань більша модель «підказує» меншій, як поводитися, що дає змогу зберегти вагому частину продуктивності при значно меншій кількості параметрів;

в) квантування – зменшення точності числового представлення складових нейронної мережі, наприклад, за допомогою заміни 32-бітного числа з плаваючою комою на 8-бітне ціле, що значно пришвидшує обчислення та робить модель меншою за розміром; це може бути особливо корисним для сучасних апаратних платформ з графічними процесорами, які мають підтримку 8-бітних операцій; квантоване значення може бути знайдене за формулою (21):

$$Q(X, n) = \left\lfloor \frac{(X - \min X)(2^n - 1)}{\max X - \min X} + \frac{1}{2} \right\rfloor \#(21)$$

де Q – функція квантування, X – багатовимірний масив, який зазнає квантування, n – розрядність вихідних значень (зазвичай 8).

Останній вид оптимізації найкраще відіграє роль в таких трьох процедурах:

а) квантово-орієнтоване навчання, в якому під час тренування ваги й активації неодноразово квантуються, що враховує похибки самого процесу та дозволяє досягти високої точності остаточної моделі;

б) квантування звичайної моделі після завершення її навчання за допомогою перетворення вагів і активацій у меншу розрядність, використовуючи невеликий набір даних, найчастіше випадкову вибірку вихідних даних, для калібрування масштабів квантування;

в) динамічне квантування, яке полягає в тому, що ваги після тренування моделі конвертуються в нижчу розрядність одразу, а активації – лише безпосередньо під час виконання інференсу; у такому випадку довжини проміжків для квантування активацій при цьому визначаються динамічно на основі поточного діапазону значень [15].

Провівши огляд основних підходів до оптимізації моделей машинного навчання, для цієї роботи видом оптимізації було обрано динамічне квантування, оскільки цей метод, на відміну від інших, не вимагає додаткового перетренування моделі, формування окремого набору даних для калібрування, а також забезпечує узгоджену трансформацію всіх її компонентів.

Результати дослідження

Після навчання трьох нейронних мереж, а саме з елементами з довгою короткочасною пам'яттю, блоками багатоголової уваги та одновимірними згортковими шарами було виконано візуалізацію результатів, а саме графіки зміни функції втрат та точності для тренувальних та валідаційних, а також різниці між двома останніми. Візуалізацію вищеописаних показників для побудованих моделей машинного навчання представлено на рис. 1 – 3.

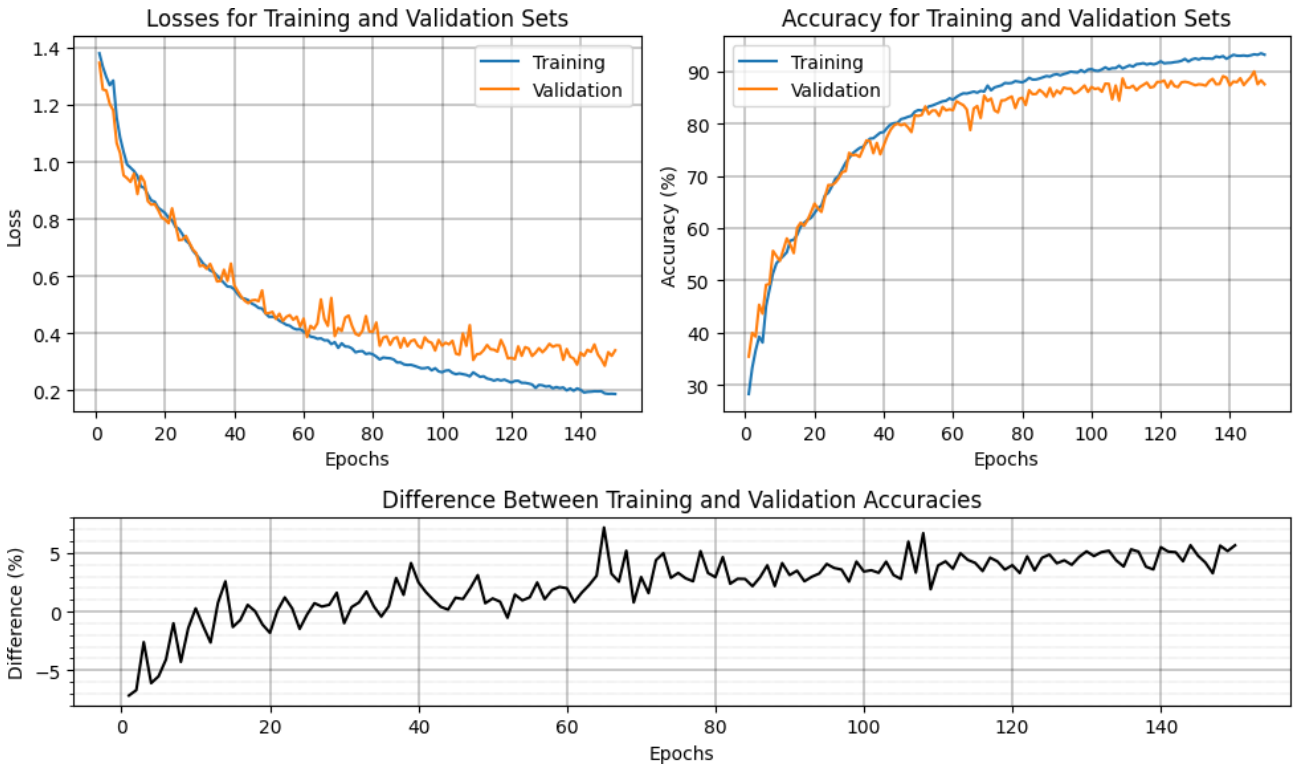

Рис. 1. Візуалізація показників тренування нейронної мережі з елементами з довгою короткочасною пам'яттю

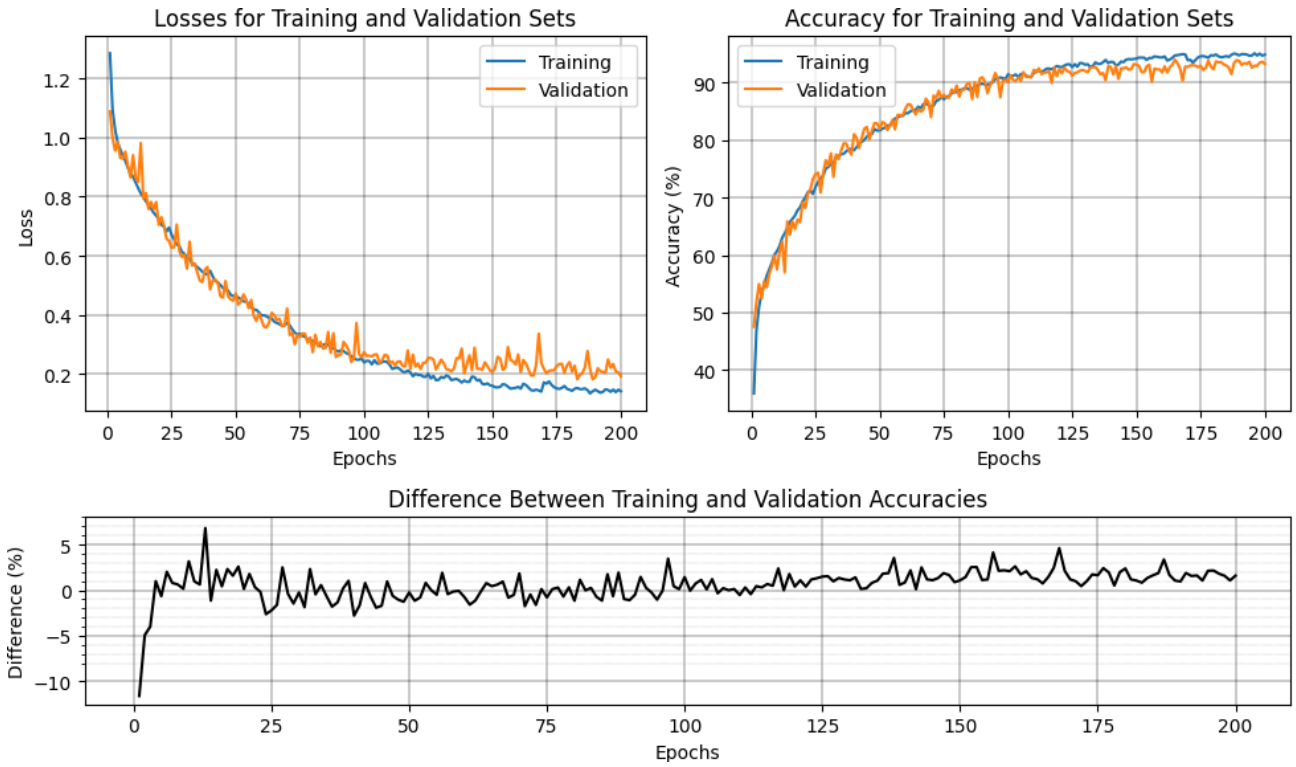

Рис. 2. Візуалізація показників тренування нейронної мережі з блоками багатоголової уваги

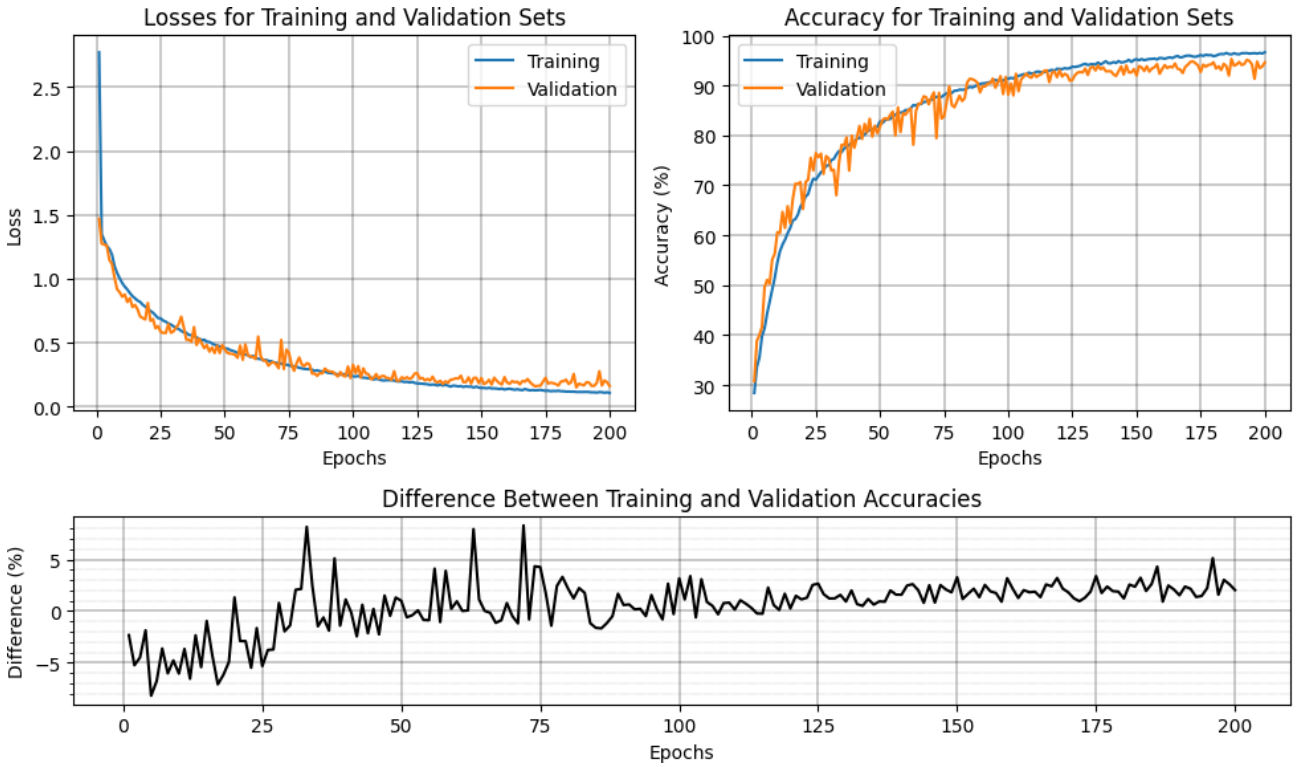

Рис. 3. Візуалізація показників тренування нейронної мережі зі згортковими шарами

Навчанням моделей показало чудові результати з найвищою точністю у моделі зі згортковими шарами (95 % на валідаційному наборі даних), трохи гірше у моделі з блоками багатоголової уваги (94 %) та ще гірше у мережі з блоками LSTM (90 %). Таку ж тенденцію можна побачити для функції втрат, де для останніх двох моделей на валідаційних даних її мінімальне значення становить близько 0,2, тоді як для першої – близько 0,3.

Також можна помітити, що орієнтовне середнє значення різниці між значеннями точностей на тренувальних та валідаційних наборах даних в моделях з блоками багатоголової уваги та зі згортковими шарами є меншою (2 – 4 %), аніж в моделі з

елементами з довгою короткочасною пам'яттю (3 – 6 %), що може бути пов'язано з тим, що останні можуть «зациклюватися» на довгострокових залежностях, порівняно з іншими, які також враховують локальні закономірності й найважливіші фрагменти.

Після тренування нейронних мереж було виконано їхнє динамічне квантування та подальше тестування з використанням візуалізації з матрицями плутанини на тестовому наборі даних для звичайної та оптимізованої моделей кожного виду, що можна побачити на рис. 4 – 6.

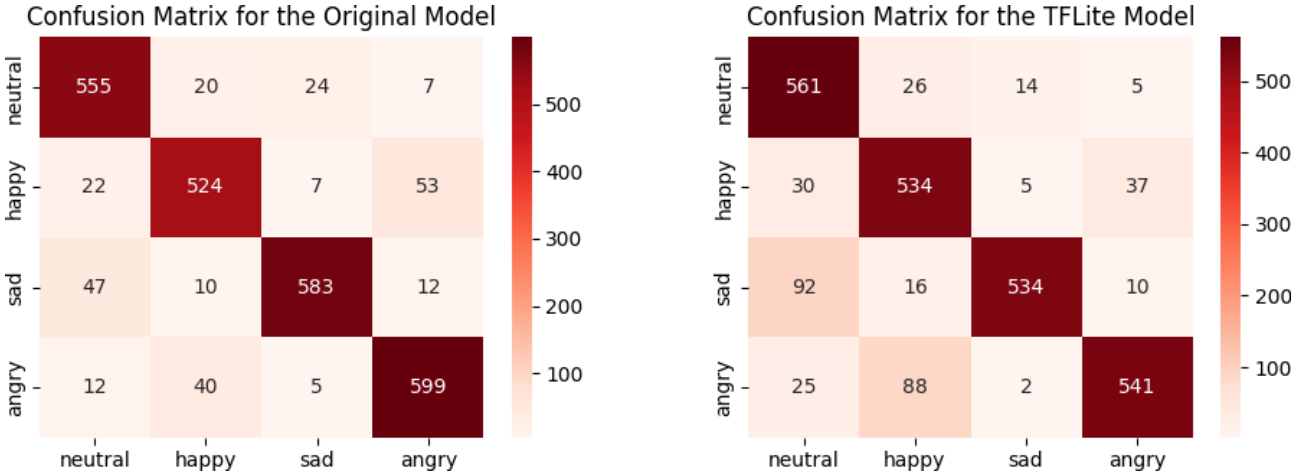

Рис. 4. Матриці плутанини для звичайної та квантованої нейронних мереж з елементами з довгою короткочасною пам'яттю

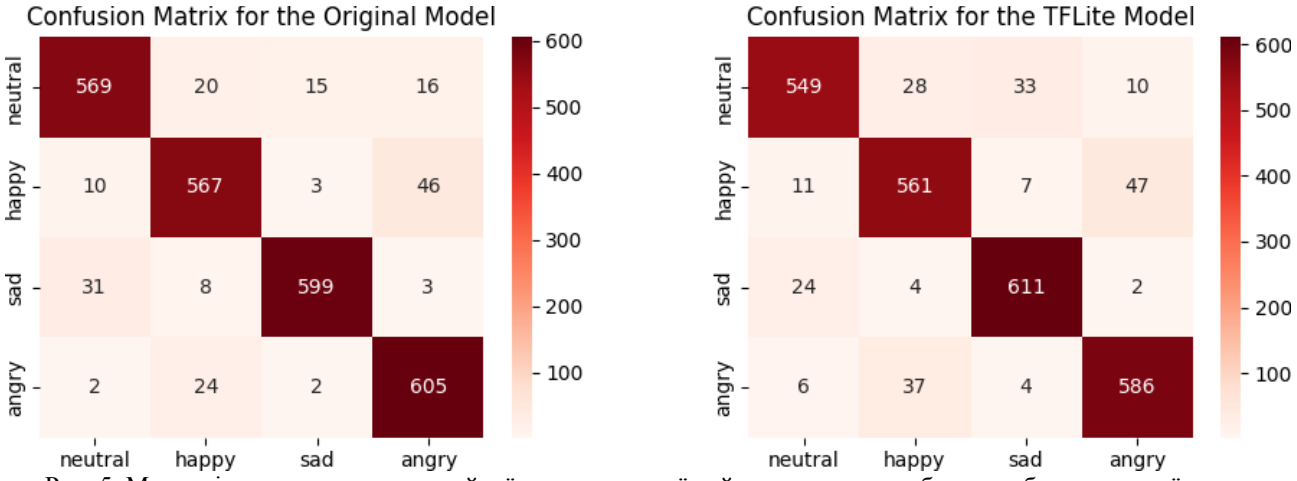

Рис. 5. Матриці плутанини для звичайної та квантованої нейронних мереж з блоками багатоголової уваги

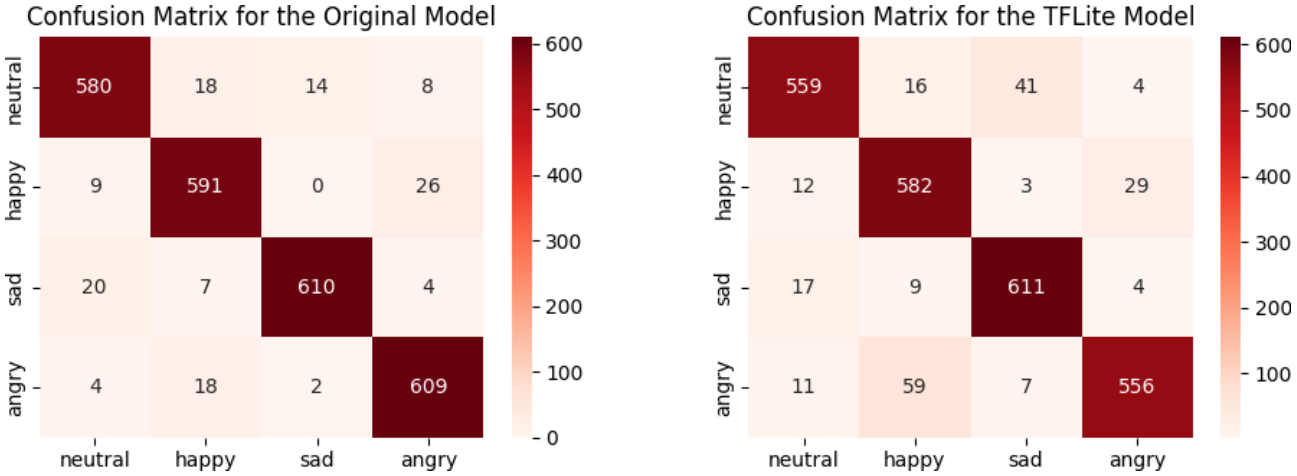

Рис. 6. Матриці плутанини для звичайної та квантованої нейронних мереж з елементами зі згортковими шарами

Аналіз помилок згідно з матрицями плутанини показав, що як і звичайні, так і квантовані нейронні мережі здатні відносно точно ідентифікувати правильний клас емоційної тональності, хоч і в оптимізованих моделей рівень точності трохи знизився.Слід відмітити, що найчастіше моделі плутають категорії «happy» і «angry», а також «neutral» і «sad».

Наявність першої пари плутанин може бути зумовлена тим, що обраний набір ознак у переважно орієнтується на рівень емоційного збудження голосу, тож радісний та розлючений тон можуть мати схожі закономірності в інтенсивності.

У другому випадку нейтральна інтонація в окремих мовців може звучати глибоко або меланхолійно, через що модель інтерпретує її як сум. Така плутанина свідчить про необхідність розширення або переоцінки списку ознак, зокрема додавання параметрів, що краще відрізняють емоційні стани з подібним рівнем інтенсивності, але різним сентиментом.

Після тестування звичайних та квантованих нейронних мереж було виконано побудову гістограм розподілу швидкості інференсу для кожного виду, що зображено на рис. 7 – 9.

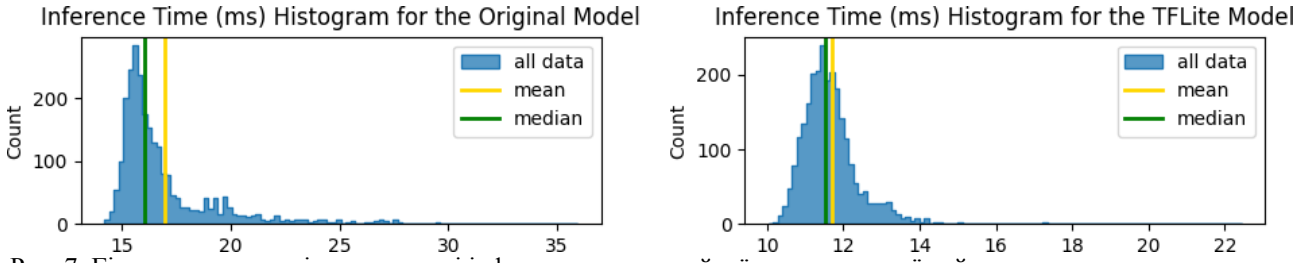

Рис. 7. Гістограми розподілу швидкості інференсу для звичайної та квантованої нейронних мереж з елементами з довгою короткочасною пам'яттю

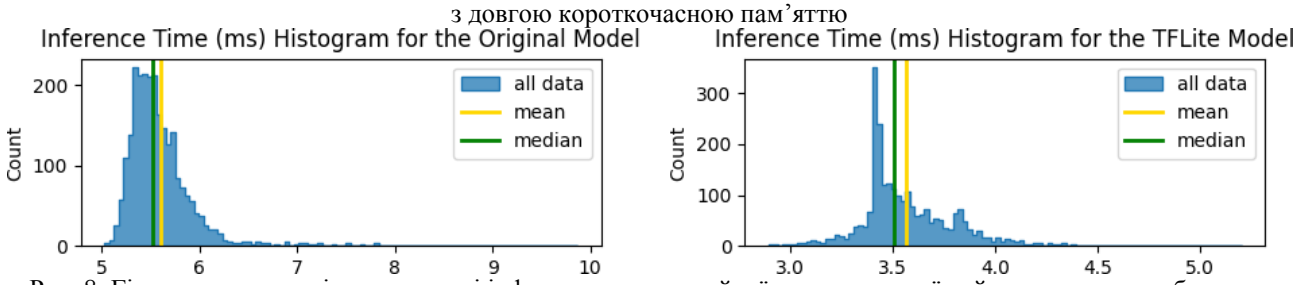

Рис. 8. Гістограми розподілу швидкості інференсу для звичайної та квантованої нейронних мереж з блоками багатоголової уваги

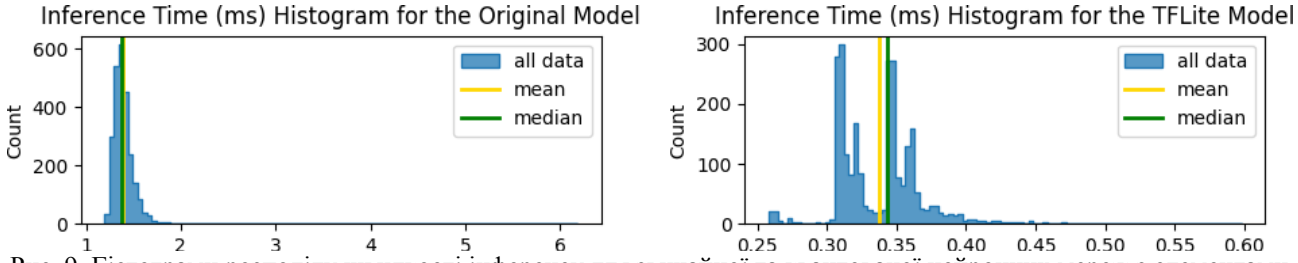

Рис. 9. Гістограми розподілу швидкості інференсу для звичайної та квантованої нейронних мереж з елементами зі згортковими шарами

Побудова гістограм допомагає пересвідчитися в тому, що динамічне квантування нейронних мереж допомогло прискорити швидкість інференсу в 1,5 – 4 рази. Найкращий результат прискорення є в нейронної мережі зі згортковими шарами, що, ймовірно, пов'язано з тим, що їхня структура є більш простою порівняно з елементами з довгою короткочасною пам'яттю чи блоками багатоголової уваги.

Наостанок результати дослідження з урахуванням рівня точності на тренувальних, валідаційних та тестувальних даних, точності оптимізованої моделі, швидкості інференсу та обсягу пам'яті для зберігання було занесено в таблицю 1 [16, 17].

Таблиця 1

Порівняльна таблиця з характеристиками натренованих нейронних мереж

№	Вид моделі	Рівень точності (%)				Швидкість інференсу (мс)		Обсяг пам'яті для зберігання (кБ)	
		звич. моделі			квант. моделі	звич. моделі	квант. моделі	звич. моделі	квант. моделі
		на тренув. даних	на валід. даних	на тест. даних					
1	З елементами з довгою короткочасною пам'яттю	93,4	90	89,7	86,1	16	11,7	795,9	99,63
2	Із блоками багатоголової уваги	95,5	94	92,9	91,5	5,6	3,57	1995,14	197,97
3	Зі згортковими шарами	96,7	95,3	94,8	91,6	1,4	0,34	779,95	73,79

У таблиці 1 також видно, що після квантування нейронної мережі обсяг пам'яті для зберігання було зменшено в 8 – 10,5 разів. З урахуванням усіх результатів тестування можна зробити висновок, що нейронна мереж зі згортковими шарами є оптимальною як і за показником точності що звичайної, що квантованої моделей, так і за швидкістю інференсу моделі. Також варто відмітити те, що найменшу втрату в точності після квантування має нейронна мережа з використанням блоків з мультиголовою увагою, однак вона вимагає для збереження навіть більше пам'яті, ніж модель з елементами з довгою короткочасною пам'яттю.

Висновки

У роботі системно проаналізовано підходи до розпізнавання емоційної тональності аудіозаписів із використанням різних архітектур глибокого навчання – нейронна мережа з довгою короткочасною пам'яттю (LSTM), нейронна мережа з блоками багатоголової уваги (Multi-head attention network) та нейронна мережа зі згортковими шарами (CNN). Для забезпечення уніфікації вхідних даних при збереженні динамічної структури сигналу запропоновано модифіковану стратегію доповнення нулями з випадковим розподілом на початку та в кінці аудіозапису. Виділено репрезентативний набір ознак – часових, частотних та часово-частотних – зокрема MFCC, спектральний спад, контраст, висоту тону та інші характеристики, що забезпечило глибше відображення властивостей мовлення.

У результаті проведеного експериментального моделювання встановлено, що модель зі згортковими шарами демонструє найвищу точність на тестових даних (94,8 %) а після динамічного квантування забезпечила найменший розмір (74 кБ) і найвищу швидкість інференсу (0,34 мс). Такі характеристики роблять її найбільш придатною до інтеграції в системи з обмеженими обчислювальними ресурсами.

Для зниження обчислювальних витрат і ресурсоємності моделей було застосовано динамічне квантування – без потреби в повторному навчанні або калібруванні. У результаті обсяг пам'яті моделей зменшився в 8 – 10,5 разів, а час інференсу – до 4 разів. Зокрема, модель зі згортковими шарами зменшила розмір із 780 кБ до 74 кБ і забезпечила найвищу швидкість роботи, зменшивши її з 1,4 мс до 0,34 мс.

Аналіз матриць плутанини виявив обмеження наявного набору ознак, зокрема часті помилки між категоріями з подібним рівнем збудження (радість та злість). Це свідчить про необхідність подальшого розширення ознакового простору – зокрема, за рахунок тембральних та інтонаційних параметрів. Також перспективним є залучення мультимодальних даних (аудіо та текст) для кращої диференціації емоцій.

Отримані результати демонструють ефективність поєднання глибокого навчання з методами оптимізації моделей, що відкриває можливості для впровадження розроблених

рішень у реальні інформаційні системи з обмеженими ресурсами – зокрема в медичних, освітніх та сервісних середовищах. Запропоновані моделі є придатними для інтеграції у пристрої з низьким енергоспоживанням або мобільні платформи, що робить дослідження актуальним у контексті розвитку edge computing та голосових інтерфейсів нового покоління. Запропонований підхід має високу прикладну цінність для інформаційних систем реального часу: call-центрів, освітніх платформ, медичних застосунків.

СПИСОК ЛІТЕРАТУРИ

1. Feng Y., Devillers L. End-to-End Continuous Speech Emotion Recognition in Real-life Customer Service Call Center Conversations. 2023. arXiv preprint. URL: https://arxiv.org/abs/2310.02281 (дата звернення: 24.05.2025).

2. Speech Emotion Classification. URL: https://github.com/Jason-Oleana/speech-emotion-classification/tree/main (дата звернення: 25.05.2025).

3. Audio Sentiment Analysis. URL: https://github.com/pontonkid/Audio-Sentiment-Analysis-/blob/main (дата звернення: 25.05.2025).

4. Maradithaya S., Katti A. Sentimental analysis of audio-based customer reviews without textual conversion. *International Journal of Electrical and Computer Engineering (IJECE).* 2024. №14 (1): 653. P. 653–661. DOI:10.11591/ijece.v14i1. URL: https://www.researchgate.net/publication/377878724_Sentimental_analysis_of_audio_based_customer_reviews_without_textual_conversion (дата звернення: 25.05.2025).

5. Audio Sentiment Analysis. URL: https://github.com/Tilak612/Audio-Sentiment-Analysis (дата звернення: 25.05.2025).

6. Emotional Speech Dataset (ESD). Kaggle. URL: https://www.kaggle.com/datasets/nguyenthanhlim/emotional-speech-dataset-esd (дата звернення: 26.05.2025).

7. Apple Machine Learning Research Team. Personalized Hey Siri: An On-Device DNN-HMM Voice Trigger System. Machine Learning Research at Apple.2023. URL: https://machinelearning.apple.com/research/voice-trigger (дата звернення: 27.05.2025).

8. Van Lieshout P., Pouplier M., Chartier J. Speech Sound Disorders in Children: An Articulatory Phonology Perspective. *Frontiers in Psychology*. 2020. URL: https://doi.org/10.3389/fpsyg.2019.02998 (дата звернення: 27.05.2025.

9. Gondohanindijo J., Muljono E., Noersasongko E., Pujiono, Setiadi D. R. M. Multi-Features Audio Extraction for Speech Emotion Recognition Based on Deep Learning. *International Journal of Advanced Computer Science and Applications (IJACSA).* 2023. №6. P. 23–29. URL: https://doi.org/10.14569/IJACSA.2023.0140623 (дата звернення: 27.05.2025).

10. Мокін В. Б., Дратований М. В. Наука про дані: машинне навчання та інтелектуальний аналіз даних : електронний навчальний посібник комбінованого (локального та мережевого) використання. Вінниця : ВНТУ, 2024. 258 с.

11. Мілян Н. Аналіз методів машинного навчання з вчителем. *Міжнародна студентська науково-технічна конференція "Природничі та гуманітарні науки. Актуальні питання": матеріали конф*., ТНТУ ім. Івана Пулюя. URL: https://elartu.tntu.edu.ua/bitstream/lib/25035/2/MSNK_2018v1_Milian_N-Analysis_of_supervised_machine_51-52.pdf (дата звернення: 28.05.2025).

12. Лосенко А. В., Козачко О. М., Варчук І. В. Нейромережевий ансамбль для прогнозування часових рядів на основі Prophet та LSTM. *Наукові праці Вінницького національного технічного університету*. 2024. №4. URL: https://doi.org/10.31649/2307-5376-2024-4-49-57 (дата звернення: 28.05.2025).

13. Chen Y., Pu H., Qu Y. An analysis of attention mechanisms and its variance in transformer. *Applied and Computational Engineering*. 2024. №47. P. 164–176. URL: https://doi.org/10.54254/2755-2721/47/20241291 (дата звернення: 29.05.2025).

14. Bhatt M., Sharma A., Singh A. A review of convolutional neural networks in computer vision. *Artificial Intelligence Review*. 2024. №57. URL: https://doi.org/10.1007/s10462-024-10721-6 (дата звернення: 29.05.2025).

15. Dantas P. V., Silva Jr. W. S., Cordeiro L. C., Carvalho C. B. A comprehensive review of model compression techniques in machine learning. *Applied Intelligence*. 2024. V. 54. P. 11804–11844. URL: https://doi.org/10.1007/s10489-024-05747-w (дата звернення: 30.05.2025).

16. SECEDC. Kaggle. URL: https://www.kaggle.com/code/olesatthewheel/sec-edc.

17. SEC EDC FINAL. Kaggle. URL: https://www.kaggle.com/code/olesatthewheel/sec-edc-final.

Стаття надійшла до редакції 05.06.2025.

Стаття пройшла рецензування 10.06.2025.

Жуков Сергій Олександрович – канд. техн. наук, доцент, доцент кафедри системного аналізу та інформаційних технологій, e-mail: sazhukov@gmail.com.

Рудзевич Олександр Володимирович – студент групи СА-21б, факультет інтелектуальних інформаційних технологій та автоматизації, e-mail: sasha.rrrr.1@gmail.com.

Вінницький національний технічний університет.