Міністерство освіти і науки України
Харківський національний університет радіоелектроніки

Факультет комп'ютерних наук
(повна назва)

Кафедра програмної інженерії
(повна назва)

КВАЛІФІКАЦІЙНА РОБОТА
Пояснювальна записка

рівень вищої освіти другий (магістерський)

Дослідження методів обробки аудіо запису за допомогою ШІ для виявлення емоційного стану
(тема)

Виконав:
студент 2 курсу, групи ІПЗм-22-2

Суворов Д.С
(прізвище, ініціали)

Спеціальність 121 – Інженерія програмного забезпечення
(код і повна назва спеціальності)
Тип програми освітньо-наукова

Керівник к.т.н., доц. Афанасьєва І.В.
(посада, прізвище, ініціали)

Допускається до захисту
Зав. кафедри _____ Дудар З.В.
(підпис) (прізвище, ініціали)

2024 р.

Харківський національний університет радіоелектроніки

Факультет комп’ютерних наук
Кафедра програмної інженерії
Рівень вищої освіти другий (магістерський)
Спеціальність 121 – Інженерія програмного забезпечення
Тип програми освітньо-наукова
Освітня програма Інженерія програмного забезпечення
(шифр і назва)

ЗАТВЕРДЖУЮ:
Зав. кафедри _____
(підпис)
«___» _____ 2024 р.

ЗАВДАННЯ
НА КВАЛІФІКАЦІЙНУ РОБОТУ

студентові Суворову Даніілу Спартаковичу
(прізвище, ім’я, по батькові)

1. Тема роботи «Дослідження методів обробки аудіо запису за допомогою ШІ для виявлення емоційного стану»

Затверджена наказом по університету від 29.03.2024р. № 250 Ст

2. Термін подання студентом роботи до екзаменаційної комісії 07.06.2024

3. Вихідні дані до роботи наукові статті та публікації за темою роботи, відкриті дані у різних форматах, приклади розробки нейронних мереж на тему роботи, використання мови Python для розробки нейронних мереж

4. Перелік питань, що потрібно опрацювати в роботі

вступ, аналіз предметної галузі, аналіз стану проблеми, постановка задачі, опис можливості використання отриманих результатів у науковій і практичній діяльності, експериментальні дослідження, висновки, перелік джерел посилання, додатки

КАЛЕНДАРНИЙ ПЛАН

№	Назва етапів роботи	Термін виконання етапів роботи	Примітка
1	Аналіз робіт, пов’язаних із предметною галуззю	29.02.2024 – 05.03.2024	*виконано*
2	Аналіз предметної галузі	06.03.2024 – 26.03.2024	*виконано*
3	Аналіз та вибір методів обробки аудіо записів	27.03.2024 – 03.04.2024	*виконано*
4	Планування експериментів	04.04.2024 – 10.04.2024	*виконано*
5	Програмна реалізація методів обробки аудіо записів та побудови моделей ШІ	11.04.2024 – 20.04.2024	*виконано*
6	Експериментальні дослідження	21.04.2024 – 28.04.2024	*виконано*
7	Аналіз результатів експериментальних досліджень та розробка рекомендацій	29.04.2024 – 30.04.2024	*виконано*
8	Написання та оформлення статті	31.04.2024 – 17.05.2024	*виконано*
9	Підготовка пояснювальної записки	31.04.2024 – 20.05.2024	*виконано*
10	Підготовка презентації та доповіді	21.05.2024 – 25.05.2024	*виконано*
11	Перевірка на плагіат	30.05.2024	
12	Нормоконтроль	03.06.2024	*виконано*
12	Рецензування	04.06.2024	*виконано*
13	Занесення диплома в електронний архів	04.06.2024	*виконано*
14	Попередній захист	04.06.2024	*виконано*
15	Допуск до захисту у зав. кафедри	05.06.2024	*виконано*

Дата видачі завдання 29 лютого 2024р.

Студент _____ Суворов Д.С.
(підпис)

Керівник роботи _____ к.т.н., доц. Афанасьєва І.В.
(підпис) (посада, прізвище, ініціали)

РЕФЕРАТ / ABSTRACT

Пояснювальна записка містить 113 стор., 39 рис., 17 табл., 29 джерел, 11 додатків.

АУДІО, ЕМОЦІЇ, МАШИННЕ НАВЧАННЯ, МОВЛЕННЯ, НЕЙРОННІ МЕРЕЖІ, РОЗПІЗНАВАННЯ, ШТУЧНИЙ ІНТЕЛЕКТ, PYTHON, TENSORFLOW.

Об’єктом дослідження є аудіо записи людської мови.

Метою роботи є визначення найбільш ефективного методу обробки аудіо для визначення емоційного стану із використанням інструментів штучного інтелекту.

Методами дослідження є проведення експериментів із використанням методів обробки аудіо для розпізнавання емоційного стану людини за допомогою штучного інтелекту із подальшим порівняльним аналізом результатів.

У результаті роботи повинен бути проведений аналіз наукових досліджень за визначеною темою, визначено методи обробки аудіо записів для проведення експериментів, проаналізовано набори даних для обраної задачі, методи штучного інтелекту, що є найбільш ефективними у розпізнаванні емоційного стану за аудіо записом, методи оцінки математичних моделей побудованих на основі обраних методів, проведено експерименти на спроєктованих моделях та проаналізовані результати із подальшими висновками.

AUDIO, EMOTIONS, MACHINE LEARNING, SPEECH, NEURAL NETWORKS, RECOGNITION, ARTIFICIAL INTELLIGANCE, PYTHON, TENSORFLOW.

The object of the research is human speech recordings.

The purpose of the research is to determine the most effective audio processing method for recognition the emotional state using artificial intelligence methods.

The research methods include conducting experiments using audio processing methods to recognize a person’s emotional state using artificial intelligence, followed by a comparative analysis of the results.

As a result of the work, an analysis of scientific research on a particular topic should be carried out, methods of processing audio recordings for experiments should be determined, datasets for the selected task should be analyzed, artificial intelligence methods that are most effective in speech emotion recognition should be analyzed, methods for evaluating mathematical models built on the basis of the selected methods, experiments on the designed models should be conducted and the results should be analyzed with further conclusions.

Заява щодо самостійного виконання кваліфікаційної роботи та можливості її публікації в електронному архіві відкритого доступу ElArKhNURE.

Я, Суворов Данііл Спартакович, студент гр. ІПЗм-22-2, здобувач вищої освіти на другому (магістерському) рівні кафедри «Програмна інженерія», заявляю: моя кваліфікаційна робота на тему «Дослідження методів обробки аудіо запису за допомогою ШІ для виявлення емоційного стану», що буде представлена в екзаменаційну комісію для публічного захисту, виконана самостійно, в ній не містяться елементи плагіату і вона може бути опублікована в електронному архіві відкритого доступу ElArKhNURE. Всі запозичення з друкованих та електронних джерел мають відповідні посилання.

Я ознайомлений з діючим положенням «Про протидію академічному плагіату в ХНУРЕ», згідно з яким виявлення плагіату є підставою для відмови в допуску кваліфікаційної роботи до захисту та застосування дисциплінарних заходів.

ЗМІСТ

ВСТУП

Зростання інтересу до розпізнавання емоцій за аудіозаписами мовлення є результатом стрімкого розвитку області штучного інтелекту та його використання в повсякденному житті. Людський голос несе в собі багато інформації, яка може бути ключовою для розуміння емоційного стану. Сучасні системи розпізнавання мови та обробки сигналів мовлення надають можливість дослідження та використання цієї інформації для створення інтелектуальних систем, здатних взаємодіяти з користувачем на більш глибокому та емоційному рівні [1].

На кафедрі Програмної інженерії активно вивчається напрямок психоаналізу людини через тексти [2], звуки, зокрема мовлення. Методи штучного інтелекту, машинного навчання часто застосовуються у суміжних роботах цього напряму. Наукові дослідження на цю тему проводили та продовжують проводити Афанасьєва І.В. (яка є науковим керівником даної роботи), Смеляков К.С., Єрохін А.Л. та інші викладачі та дослідники кафедри ПІ [3].

Метою даної роботи є визначення найбільш ефективного методу обробки аудіо записів із використанням штучного інтелекту для задачі розпізнавання емоцій за мовленням.

Для досягнення поставленої мети необхідно виконати наступні завдання:

- провести аналіз існуючих методів для обробки аудіо сигналів;
- зробити висновки щодо найбільш актуальних та ефективних методів штучного інтелекту, описаних у роботах попередників;
- вивчити набори даних для даної задачі та вибрати найбільш якісні з теоретичної точки зору для поточного дослідження;
- дослідити можливості розширення наборів даних;
- визначити необхідні метрики для порівняння продуктивності методів;
- розробити математичні моделі на основі обраних методів;
- провести експерименти з метою перевірки точності моделей;
- провести аналіз отриманих результатів та розробити рекомендації для подальших досліджень.

Об’єктом дослідження є аудіо записи людської мови.

Предметом дослідження є методи обробки аудіо записів за допомогою штучного інтелекту з метою виявлення емоційного стану.

Методами дослідження є проведення експериментів із використанням методів обробки аудіо для розпізнавання емоційного стану людини за допомогою штучного інтелекту із подальшим порівняльним аналізом результатів.

Наукова новизна дослідження полягає у розробці більш досконалих підходів до розпізнавання емоційного стану людини через аудіо запис з використанням сучасних методів глибокого навчання.

Практичне значення роботи полягає у можливості застосування розроблених моделей для створення інтерактивних систем, що здатні розпізнавати емоційний стан користувача у реальному часі. Результати дослідження можуть бути використані в різних галузях, таких як медичні послуги, розумні будинки, освітні технології, системи аварійного реагування та інші системи, де важливо враховувати емоційний стан користувачів для покращення взаємодії та підвищення ефективності роботи.

Після проведеного дослідження результати було опубліковано у вигляді тез у науковому журналі «Znanstvena misel journal» (DOI: 10.5281/zenodo.11049575) [4] (див. дод. В), а також у вигляді наукової статті з виступом на конференції MoMLeT Workshop 2024 (див. дод. Г), яка індексується у Scopus, DBLP та Google Scholar.

1 АНАЛІЗ ПРЕДМЕТНОЇ ГАЛУЗІ ТА ПОСТАНОВКА ЗАДАЧІ

1.1 Аналіз наявних досліджень

Ця сфера щороку привертає увагу дослідників. Значна кількість робіт досліджує потенційні можливості застосування штучного інтелекту для розпізнавання емоцій, з особливим акцентом на аудіо. Прискорений розвиток технологій машинного навчання та глибокого навчання призвів до появи цілого ряду інструментів, які можуть бути використані для вирішення проблеми розпізнавання емоцій у більш елегантний та ефективний спосіб.

Значна кількість робіт використовує набори даних, які доволі сильно відрізняються один від одного. Найчастіше використовуються набори даних RAVDESS і TESS. Проте вже існує ряд більш репрезентативних наборів даних, які були визначені дослідниками.

Першим значним дослідженням є «Emotional Speech Recognition Using Deep Neural Networks» 2022 року [5]. У цьому дослідженні використовується великий набір даних IEMOCAP, що складається з 10 емоцій, як, щоправда, не дуже однорідні у наборі. Автори досягли точності 97,54% для моделі, що використовує GRU для чотирьох емоцій. Крім того, були протестовані моделі на основі CNN і CRNN, точність яких склала 96,96% і 97,18% відповідно. У своїй роботі дослідники використовували аугментацію даних, включаючи додавання шуму і зсув формант. Було продемонстровано позитивний вплив розширення даних таким шляхом. Методологія використовувала матричний підхід, коли дані, отримані з аудіо, подавалися на вхід моделі у вигляді матриці. В експериментах використовувалися як MFCCs (Mel-Frequency Cepstral Coefficients), так і інші характеристики, такі як спектральні.

У блозі платформи dataiku [6] наведено ілюстративний приклад обробки аудіоданих за допомогою навчання моделей, що демонструє можливості платформи dataiku. У дописі заглиблюються в специфіку методології попередньої обробки даних. Крім того, використання комбінації наборів даних CREMA-D, RAVDESS, SAVEE і TESS для завдання SER є сильною стороною, яка,

безсумнівно, підвищить об’єктивність отриманої моделі. Це пов’язано з тим, що набори даних створені абсолютно різними групами людей, які використовували різних акторів та підходи для запису голосів. Втім, у статті лише згадується можливість розширення даних, а точність розпізнавання коливається від 43% до 72% для шести класів. Автори використовували комбінацію MFCCs та Мел-спектрограм в якості характеристик аудіо.

У 2020 році автори статті «Speech Emotion Recognition with deep learning» [7] застосували Auto-Encoder і SVM на базі «Ryerson Multimedia Laboratory» набору даних для вирішення проблеми розпізнавання емоцій з аудіо. В експериментах використовувалися MFCCs, Zero Crossing Rate та інші характеристики аудіо, що дозволило досягти точності від 65% до 74%, залежно від конфігурації моделі.

Автори публікації «Speech Emotion Recognition Using Deep Learning Techniques: A Review» [8] створили оглядовий матеріал, який описує підходи до розв’язання SER та результати цих підходів станом на 2016 рік. У публікації описано особливості порядку обробки аудіо для класифікації емоцій та наведено порівняльні характеристики різних емоцій. На той час глибинне навчання продемонструвало кращі результати, що свідчить про актуальність використання нейронних мереж для цієї задачі.

Ще одне варте уваги дослідження 2023 року – «A Deep Learning Approach for Speech Emotion Recognition Optimization Using Meta-Learning» [9]. У своїй публікації автори представляють широке дослідження різних наборів даних, їх комбінацій, впливу аугментації на навчання та застосування так званого метанавчання. Метанавчання охоплює пошук оптимізатора та гіперпараметру швидкості навчання, пошук найефективнішого типу аугментації та набору характеристик аудіо.

У цій роботі використовувався векторний підхід, за допомогою якого витягнуті з аудіо характеристики перетворювалися з матриці у вектор. Цей підхід дозволив авторам досягти точності 83% і 91% для наборів даних CREMA-D і RAVDESS+TESS+SAVEE+CREMA-D відповідно. Також було помічено, що розширення даних за допомогою розтягування аудіо в часі має найбільший

позитивний вплив на точність, тоді як інші типи аугментації показують менш якісні результати. Однак деякі набори даних у цьому дослідженні демонструють 100% точність, що викликає питання про те, наскільки репрезентативними та варіативними є такі набори даних. Крім того, в цій статті проаналізовано лише модель на основі CNN та її комбінацію з декількома LSTM-шарами.

Інше дослідження, «Speech-Based Emotion Recognition», опубліковане в «International Journal for Research» [10], досліджує застосування згорткових нейронних мереж для розпізнавання емоцій. У статті використовується набір даних RAVDESS, що включає п’ять емоцій, і нейронна мережа на основі згорткових шарів (шість шарів) з одномірними згортковими шарами. Крім того, автори використовують MFCCs як характеристики аудіо, які слугують вхідними даними для моделі. Автори стверджують, що їхній підхід дає високу точність. Як метрику було використано f1-score, що дало значення 0,91. Крім того, автори припускають, що результати таких досліджень можуть бути інтегровані в алгоритм рекомендацій для маркетплейсів, що є багатообіцяючим застосуванням, яке заслуговує на подальше дослідження.

1.2 Опис використовуваних методів штучного інтелекту

Найбільш ефективними серед проаналізованих робіт методами штучного інтелекту виявилися нейронні мережі, а саме наступні:

- згорткові нейронні мережі;
- рекурентні нейронні мережі.

Розглянемо роботу кожної з мереж.

CNN (згорткова нейрона мережа) [11] – це тип глибокої нейронної мережі, спеціально розроблений для обробки та аналізу структурованих матриць даних, таких як зображення (хоча часто такі мережі використовуються для зображень, вони також є ефективні і для обробки аудіо, оскільки аудіо сигнал можна зобразити графічно, наприклад спектрограмою).

Загалом, згорткова мережа складається з кількох компонент (див. рис. 1.1):

- шару згортки, в якому використовуються певні фільтри для операції згортання;
- шару об’єднання для зменшення матриці та виокремлення найбільш значущих характеристик;
- повнозв’язних шарів (кілька шарів нейронів, що мають зв’язки кожен з кожним);
- шару активації.

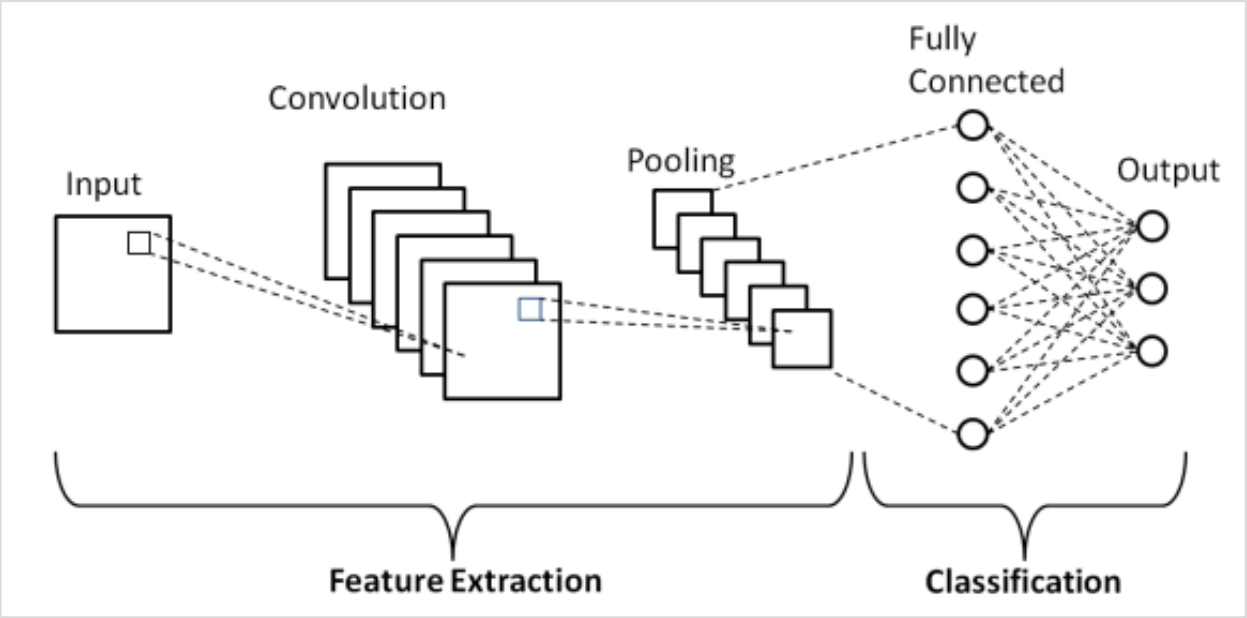

Рисунок 1.1 – Вид простої згорткової нейронної мережі (за даними [12])

RNN (рекурентна нейрона мережа) [13] – це клас нейронних мереж, призначений для роботи з послідовностями даних, де інформація передається в часі. Основна ідея полягає в тому, щоб в мережі були зв’язки, які створюють цикли, дозволяючи інформації з попередніх кроків часу впливати на поточний стан мережі.

Серед рекурентних мереж можна виділити основні типи:

- звичайна рекурентна мережа (RNN);
- рекурентна мережа з довгою короткочасною пам’яттю (LSTM);
- рекурентна мережа з вентильним рекурентним вузлом (GRU).

Звичайна RNN має зв'язки, які створюють цикли, дозволяючи інформації з попередніх кроків часу впливати на поточний стан мережі (див. рис. 1.2). Швидко

стикається із проблемою вибухаючого або зникаючого градієнта при тренуванні на довгих послідовностях, що обмежує їхню здатність до вивчання довгострокових залежностей.

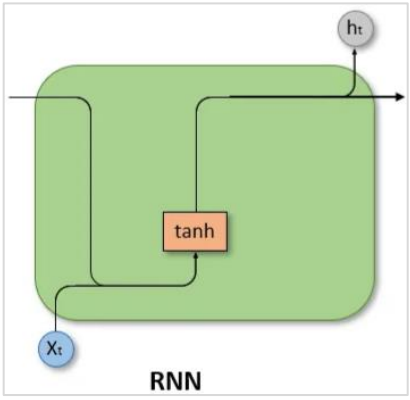

Рисунок 1.2 – Блок звичайної рекурентної мережі (за даними [14])

Довга короткострокова пам’ять (LSTM) є вдосконаленням RNN, запропонованого в 1997 році, щоб подолати обмеження короткострокової пам’яті RNN і проблеми градієнта [15].

LSTM мають «комірки» в прихованих шарах нейронної мережі, які мають три шлюзи (gates): вхідний, вихідний і шлюз забування (див. рис. 1.3). Ці шлюзи контролюють потік даних, який потрібен для прогнозування виходу рекурентного блоку [16].

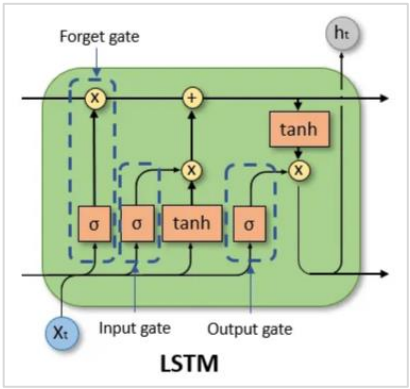

Рисунок 1.3 – Блок рекурентної мережі LSTM (за даними [14])

Рекурентна мережа на основі GRU схожа на LSTM, оскільки вона також допомагає вирішити короткочасної пам’яті рекурентних моделей.

Замість використання інформації для регулювання «стану комірки» блок GRU використовує приховані стани і має лише два шлюзи – шлюз скидання та шлюз оновлення (див. рис. 1.4) [17]. Аналогічні LSTM, шлюзи скидання та оновлення контролюють, скільки та яку інформацію зберігати.

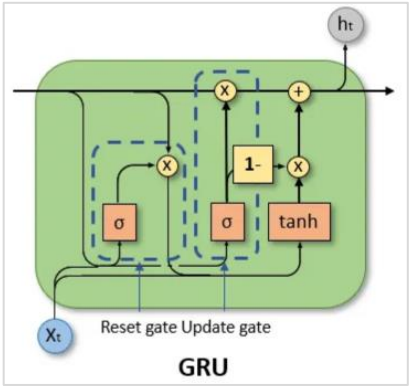

Рисунок 1.4 – Блок рекурентної мережі з GRU (за даними [14])

Також існують модифікації рекурентних мереж, так звані Bidirectional RNN. Це архітектура рекурентної нейронної мережі, яка працює з послідовностями даних в обидва напрямки (вперед та назад).

Основна ідея двоспрямованих рекурентних нейронних мереж полягає в тому, що в кожен момент часу вона обчислює не тільки прогнози на основі попередніх значень уперед, але й на основі майбутніх значень, що дозволяє працювати моделі в контексті як з минулого, так і з майбутнього (див. рис. 1.5) [18].

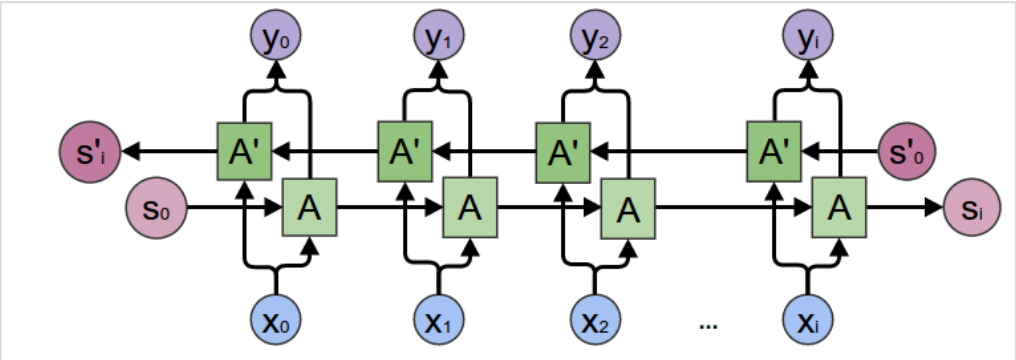

Рисунок 1.5 – Двоспрямована рекурентна мережа (за даними [19])

Таким чином, найбільш цікавими для дослідження є моделі, що базуються на наступних архітектурах:

- BiLSTM;

- GRU;
- CNN;
- CRNN (CNN + LSTM).

У цій роботі будуть розроблені та досліджені моделі на цих 4 архітектурах.

1.3 Постановка задачі

Після проведеного аналізу предметної галузі необхідно визначити задачу дослідження. Вона полягає в аналізі різних методів.

Для цього передбачається:

- в якості найбільш ефективного методу штучного інтелекту, представленого попередниками, обрати нейронні мережі, а саме архітектури BiLSTM, GRU, CNN та комбінацію CNN та LSTM –CRNN;
- визначити найбільш придатні набори даних для проведення дослідження, які включатимуть достатню варіативність емоцій людини, варіативність голосів та загалом матимуть досить велику кількість семплів;
- визначити методи для вилучення звукових характеристик, які б далі можна було використовувати як вхідні дані до нейронних мереж;
- реалізувати вилучення звукових характеристик із обраних наборів даних за допомогою інструментів Python;
- реалізувати методи розпізнавання з використанням нейронних мереж із використанням Python та Keras;
- розробити та реалізувати експериментальні дослідження з та без використання агментацій та з різними наборами даних;
- проаналізувати отримані результати та зробити висновки щодо ефективності методів обробки аудіо даних, ефективності архітектур нейронних мереж, ефективності аугментацій та їх вплив на результат, та описати подальші рекомендації.

Отже, наприкінці експериментів повинні бути отримані та проаналізовані набори метрик кожної з моделей, на основі чого зроблено висновки щодо ефективності кожного компоненту дослідження.

2 ОПИС МЕТОДІВ ДОСЛІДЖЕННЯ

2.1 Набори даних

Наразі у вільному доступі можна знайти декілька датасетів, які часто використовуються для задачі розпізнавання емоції людини. Кожен з наборів містить різну кількість прикладів, різну кількість авторів, що зачитують фрази, різний набір фраз, різну кількість емоцій і так далі. Розглянемо найбільш популярні датасети.

SAVEE (Surrey Audio-Visual Expressed Emotion) [20] – це набір даних, створений для дослідження в галузі розпізнавання емоцій за допомогою звукової та візуальної інформації. Основний акцент SAVEE спрямований на розпізнавання виражених емоцій у голосі.

Основні особливості набору даних SAVEE:

- SAVEE включає аудіозаписи, створені чотирма чоловіками, які читають речення, що виражають чотири різні емоції: радість, сум, гнів і страх. Кожен учасник читає по 15 фраз для кожної емоції;
- SAVEE фокусується на чотирьох основних емоціях, забезпечуючи баланс між позитивними (радість) та негативними (сум, гнів, страх) емоційними станами.

RAVDESS (Ryerson Audio-Visual Database of Emotional Speech and Song) [21] – це датасет, призначений для дослідження розпізнавання емоцій за допомогою звукової та візуальної інформації, який включає аудіозаписи та відеозаписи.

Основні особливості датасету RAVDESS:

- датасет включає аудіозаписи та відеозаписи професійних акторів, які виконують фрази з різними емоційними виразами. Усього датасет налічує більше 7000 файлів високої якості;
- RAVDESS включає в себе 24 емоції, включаючи радість, гнів, страх, відразу, сум і нейтральний стан. Кожен актор виконує фрази для кожної емоції;

- серед акторів 12 чоловіків та 12 жінок.

TESS (Toronto Emotional Speech Set) [22] – це датасет аудіозаписів емоційного мовлення, створений з метою дослідження та розробки систем розпізнавання емоцій за звуковою інформацією. Цей датасет складається з аудіозаписів, які передають різні емоції, виконані жінками та чоловіками.

Основні особливості датасету TESS:

- датасет включає аудіозаписи мовлення, які виконані англомовними акторами, що виражають різні емоції. Усього датасет включає більше 2800 аудіо файлів;
- TESS містить шість базових емоцій: радість, смуток, гнів, боязь, огиду і нейтральний стан. Кожен аудіозапис представляє одну з цих емоцій;
- кожен аудіозапис також має відомості про інтенсивність емоційного виразу.

CREMA-D (Crowd-sourced Emotional Mutimodal Actors Dataset) [23] – це набір даних, створений для вивчення та розробки систем визначення емоцій в музиці за допомогою аудіо- та візуальної інформації. Цей датасет є унікальним тим, що він об'єднує аудіозаписи та відеозаписи професійних акторів, які читають фрази з різними емоційними виразами.

Основні особливості датасету CREMA-D:

- датасет включає аудіозаписи та відеозаписи понад 90 професійних акторів, які читають фрази, призначені для вираження різних емоцій. Фрази спеціально розроблені для цього датасету;
- CREMA-D охоплює широкий спектр емоцій, таких як радість, сум, гнів, огиду, здивування та нейтральний стан;
- кожен актор вимовляє 12 фраз у трьох повтореннях для забезпечення різноманітності даних. Загальна кількість аудіо файлів у датасеті становить понад 7000.

Датасет IEMOCAP (Interactive Emotional Dyadic Motion Capture) [24] – це набір даних, призначений для дослідження в області розпізнавання та аналізу емоційних станів у спілкуванні на основі аудіо- та відеоінформації. Цей набір даних

унікальний тим, що містить сесії інтерактивного спілкування між акторами, що відображають різні емоційні стани.

Основні особливості датасету IEMOCAP:

- датасет включає аудіо- та відеозаписи взаємодії між акторами у спеціально підготовлених та довільних сценаріях;
- IEMOCAP охоплює широкий спектр емоцій, включаючи радість, смуток, гнів, страх, здивування, відразу, нейтральний стан тощо;
- кожна сесія включає в себе діалоги між парами акторів, які відтворюють реальні ситуації, обговорення, дискусії або просте спілкування;
- загальна кількість сесій у датасеті становить більше 1000, загальна тривалість аудіо більше 12 годин, з урахуванням різних акторів та емоційних станів.

Порівняльна характеристику відображена у таблиці (див. табл. 2.1).

Таблиця 2.1 – Порівняння датасетів (створено самостійно)

		SAVEE	**RAVDESS**	**TESS**	**CREMA-D**	**IEMOCAP**
Total samples		480	1440	2800	7442	10039
Emotions	anger	+	+	+	+	+
	happiness	+	+	+	+	+
	disgust	+	+	+	+	+
	fear	+	+	+	+	+
	sadness	+	+	+	+	+
	surprise	+	+	+		+
	neutral	+	+	+	+	+
	calmness		+			
	frustration					+
	excitation					+
Total emotions		7	8	7	6	10
Text variations		15	2	20	12	a lot of
Samples per emotion		~60	196 (96 for neutral)	400	~1270	uneven
Speakers		4 (M)	24 (12M/12F)	2F	91 (48M/43F)	10 (5M/5F)
Emotion levels		1	2	1	4	a lot of

Як видно з таблиці, найбільш цікавими є датасети CREMA-D та IEMOCAP, оскільки в них відносно велика кількість семплів, а також велика варіація тексту та акторів. Крім того CREMA-D має аж 4 рівні емоційності, а IEMOCAP має величезну кількість варіацій тексту, який не просто зачитуються, а промовляється як у реальному житті.

Для датасету IEMOCAP розподіл емоцій можна побачити на рисунку (див. рис. 2.1).

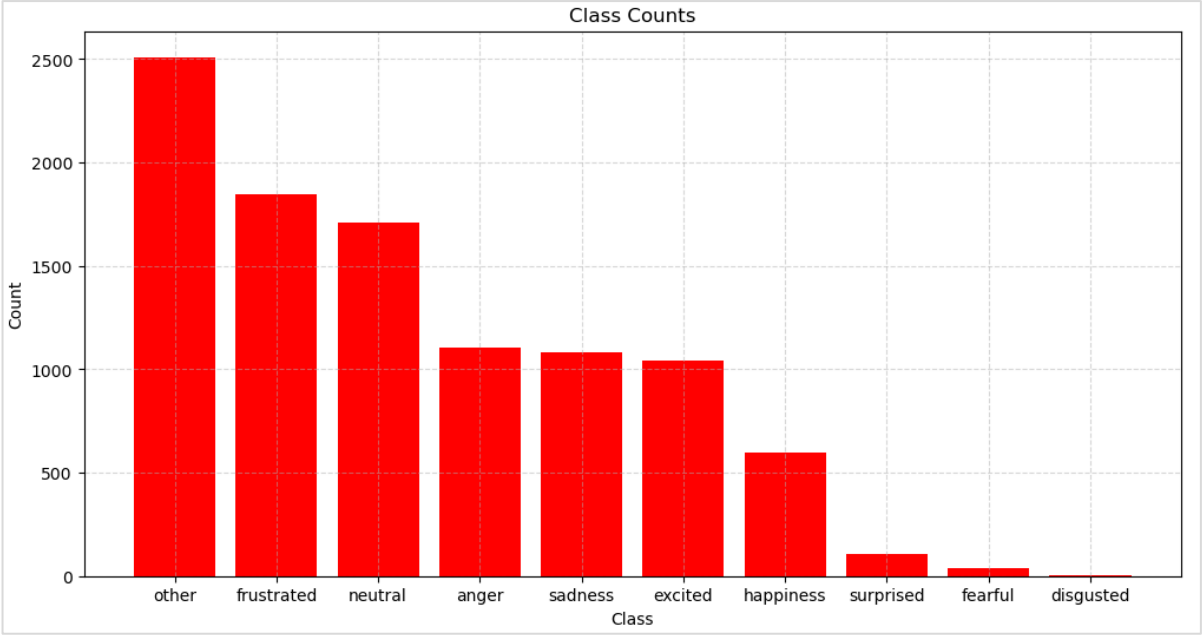

Рисунок 2.1 – Кількість аудіо за кожною емоцією у датасеті IEMOCAP (створено самостійно)

Оскільки створення власного датасету може зайняти велику кількість часу та ресурсів, для проведення експериментів обрано саме ці датасети як найбільш репрезентативні серед доступних.

2.2 Вилучення звукових характеристик

Для аналізу даних, які закодовані у аудіо, необхідно вилучити ці дані. Але перед тим, як ми матимемо доступ до цих даних (параметрів або характеристик аудіо), необхідно провести певні маніпуляції із аудіо сигналом [25].

Власне вилучення і обробка аудіо виконується так званими фреймами. Розбиття аудіо на ці частинки називається фреймінг. Розмір фреймів або frame size

(кількість семплів, які входять до фрейму) зазвичай обирається ступенем двійки (це забезпечує більш швидке застосування Fourier Transform, про який буде описано нижче).

Також, розбиття на фрейми відбувається із використанням перекриття (overlapping). Показник перекриття вказує на те, на скільки сусідні фрейми перекриваються (кількість семплів). Якщо представити, що сигнал на фрейми розбивається за допомогою зсувного вікна, то hop size – це кількість семплів, на яку це вікно зсувається. Часто використовують перекриття, щоб зберегти інформацію про зміни в сигналі, яка відбувається в межах фреймів.

Після отримання фрейму, можна обробляти (аналізувати) сигнал по частинах. Перейдемо до типів характеристик, що мають аудіо сигнали.

Під час аналізу виділяють два типи характеристик:

- часові;
- частотні;
- часово-частотні.

Часові характеристики – це характеристики сигналу, які безпосередньо пов’язані з амплітудою або силою сигналу як функції часу. Ці характеристики надають інформацію про те, як сигнал змінюється з часом без врахування його частотних компонент.

Частотні характеристики – це характеристики сигналу, які пов’язані з частотним складом сигналу. Замість того, щоб аналізувати, як сигнал змінюється з часом, ці характеристики надають інформацію про розподіл енергії по різних частотах, які присутні в сигналі.

Часово-частотні характеристики – це характеристики сигналу, що описують частотні зміни у сигналі зі зміною часу (зазвичай такі характеристики представлені у вигляді спектрограм).

Розглянемо деякі з часових характеристик, які найбільш часто використовуються при обробці аудіо методами штучного інтелекту.

Amplitude envelope.

Демонструє максимальну амплітуду серед семплів певного фрейму. Дозволяє отримати приблизне розуміння гучності сигналу, а також може допомогти визначити початок події (початок звучання ноти, початок слова і так далі).

RMS (Root Mean Square).

Середнє квадратичне значення для усіх семплів для фрейму також надає інформацію про гучність сигналу, воно менш чутливе до викидів (одномоментних стрибків гучності). В контексті машинного навчання RMS може допомогти в сегментації аудіо, тобто розрізнити фрагменти сигналу.

ZCR (Zero Crossing Rate).

Ця характеристика описує, скільки разів сигнал перетинає вісь нуля протягом певного часу. Ця характеристика також дуже часто використовується, а контексті дослідження може бути корисною для визначення, коли в мовленні відбувається перехід між фонемами або словами. Зміна ZCR може вказувати на границі або зміни в акустичних характеристиках мовлення.

Розглянемо деякі з частотних характеристик.

Важливо зазначити, що перед тим, як можна буде вилучити такі характеристики, над фреймами аудіо сигналу необхідно провести дві додаткові дії:

- застосувати windowing;
- застосування Fourier Transform.

Спершу, потрібно визначити, що Fourier Transform (або перетворення Фур’є) – це математичний інструмент, який використовується для перетворення сигналів між областями часу і частоти. Ця трансформація визначає, які частоти складають даний сигнал та якою є їх амплітуда.

Однак існує певний недолік такої трансформації через не цілу кількість періодів сигналу у фреймі. Відомо, що сигнал являє собою сукупність звукових хвиль, які мають певний період, тобто двома послідовними моментами, коли сигнал повторює свій патерн (див. рис. 2.2).

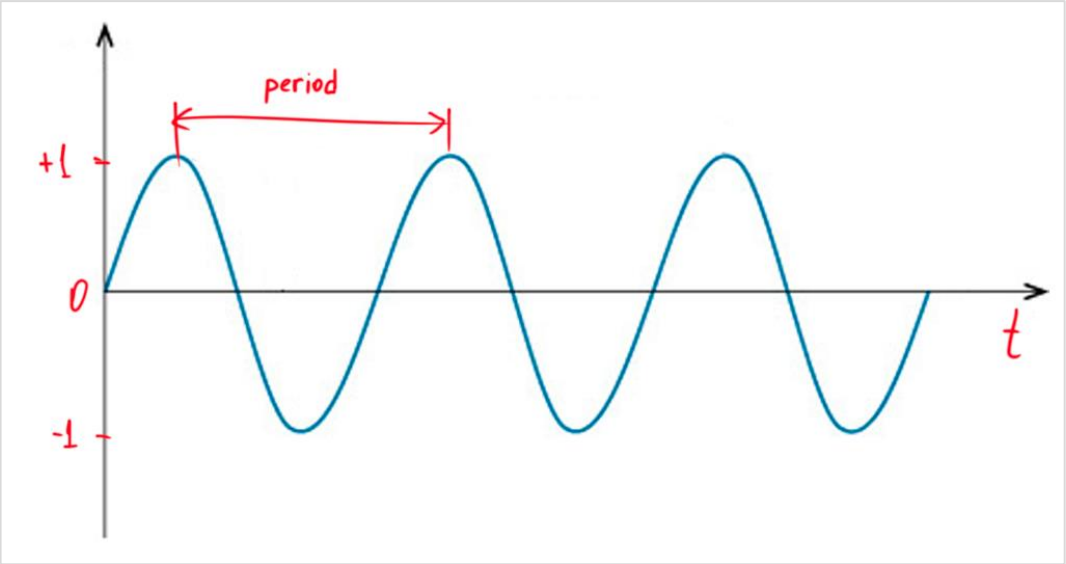

Рисунок 2.2 – Звукова хвиля (створено самостійно)

Отже, після застосування Fourier Transform, фрейм фактично множиться на вихідний сигнал. Якщо вихідний сигнал не має точно визначеної частоти, або якщо його частота не вирівнюється з періодом вікна, то при взятті вибірок відбувається змішування частот. Це призводить до появи додаткових частот у спектрі, які не існують в оригінальному сигналі (див. рис. 2.3).

Для уникнення такого ефекту якраз використовується так званий windowing.

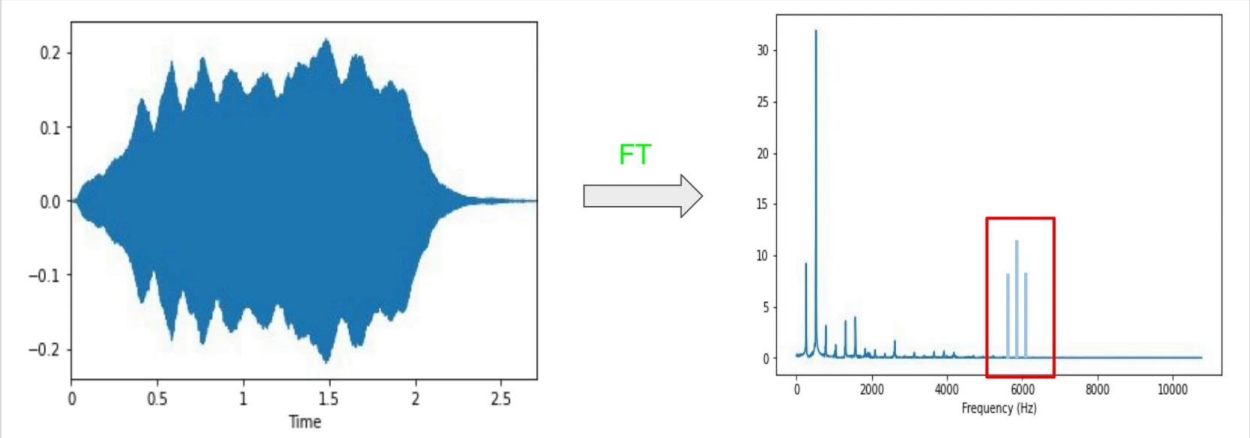

Рисунок 2.3 – Перетворення звукової хвилі через Fourier Transform (за даними [26])

Іншими словами, до сигналу (до кожного з утворених фреймів) застосовується windowing функція (наприклад, Hann window або Hamming

window). Завдяки чому семпли на обох кінцях фрейму просто зводяться до нуля, таким чином сигнал у межах фрейму стає періодичним (див. рис. 2.4).

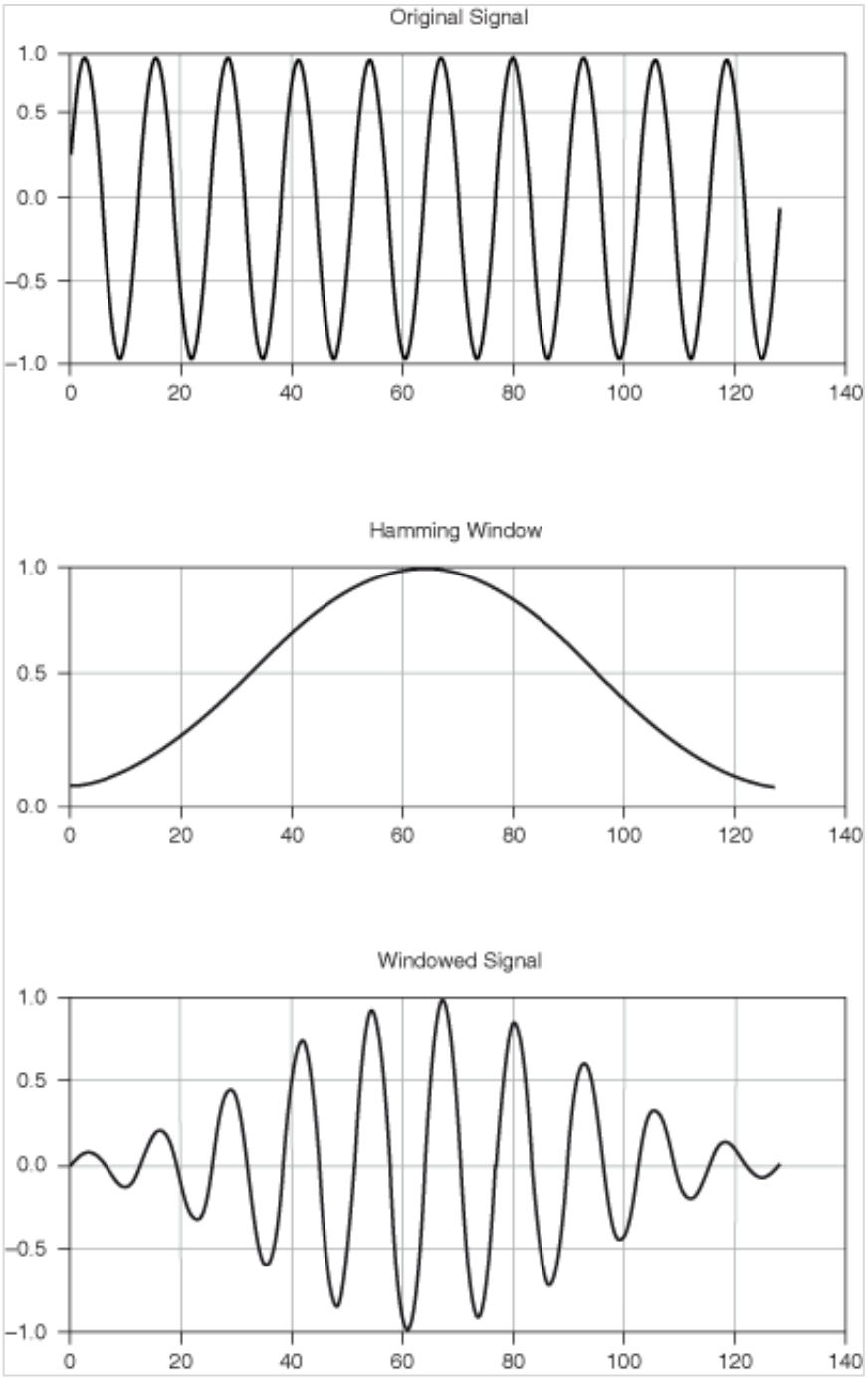

Рисунок 2.4 – Застосування функції windowing до фрейму сигналу (за даними [27])

Отже, після цих двох етапів (windowing та Fourier transform) отримується частотна характеристика сигналу (фрейму).

Розглянемо кілька з частотних характеристик, які можна використовувати для машинного навчання:

- амплітудний спектр (amplitude spectrum): спектр амплітуд показує амплітуди різних частот, які присутні в сигналі. Надає інформацію про силу або величину кожної частоти;
- спектральний центроїд (spectral centroid): представляє точку «центру мас» частотного спектра сигналу і вказує на те, де знаходиться середнє значення частоти; високий спектральний центроїд може вказувати на високочастотний контент, тоді як низький – на низькочастотний;
- спектральна ширина (spectral bandwidth): визначає ширину діапазону частот;
- спектральний контраст (spectral contrast): вимірює різницю між високо- та низькочастотними областями спектра, допомагаючи виявляти зміни яскравості між різними частотами;
- спектральна рівномірність (spectral flatness): вимірює, наскільки рівномірно розподілена енергія по всьому частотному спектру; значення близьке до 1 вказує на більш рівномірний спектр, тоді як значення близьке до 0 вказує на фокусований частотний контент;
- спектральний спад (spectral rolloff): вказує на частоту, нижчу якої розташовано певний відсоток енергії спектра (зазвичай 85%); високий спектральний спад може вказувати на те, що більшість енергії сигналу сконцентрована у вищих частотах.

Перейдемо до часово-частотних характеристик. Вони, мабуть, найбільш цікаві та репрезентативні для машинного навчання, оскільки демонструють частотну характеристику в певний момент часу.

Для отримання таких характеристик необхідно застосувати до фреймів Short Time Fourier Transform (STFT) замість Fourier Transform, який дозволяє досліджувати частотні характеристики сигналу з часовою локалізацією. Результат такого перетворення для кожного вікна утворюють спектральне представлення сигналу відносно часу. Спектральні представлення кожного вікна можуть бути

зібрані в матрицю, яку іноді називають «спектрограмою» (див. рис. 2.5). Ця матриця дозволяє візуалізувати, як частотний склад сигналу змінюється вздовж часу.

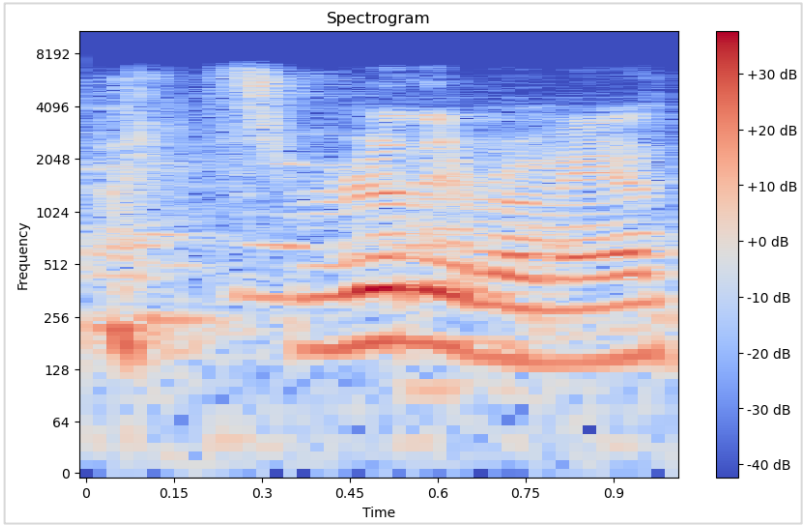

Рисунок 2.5 – Спектрограма (створено самостійно)

Одним із різновидів спектрограми є Мел-спектрограма. Це особливий вид спектрограми, який використовується в аудіо-сигнальній обробці та обробці мовлення. Вона враховує специфічні особливості сприйняття звуку людським вухом і використовує шкалу Мел для просторового представлення частот (див. рис. 2.6).

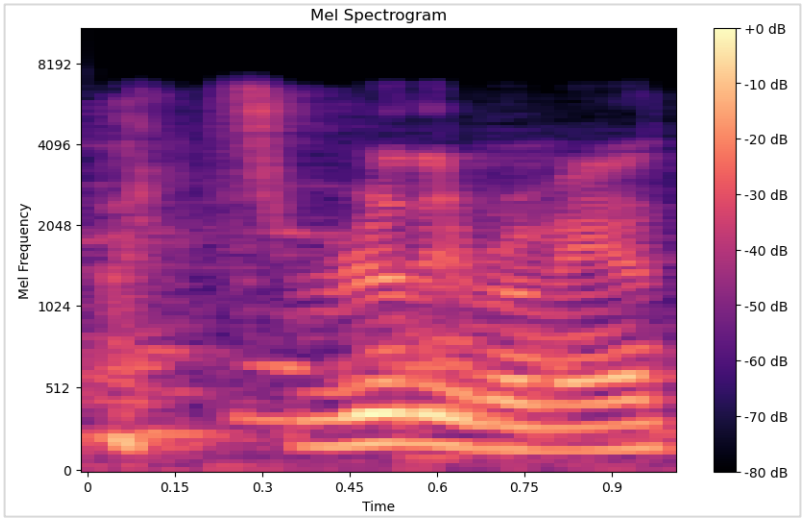

Рисунок 2.6 – Мел-спектрограма аудіо сигналу (створено самостійно)

І останнім, також широко популярним є MFCCs [28].

MFCCs (Mel-Frequency Cepstral Coefficients) – це коефіцієнти кепстрального відображення, які використовуються для опису звукових сигналів, зокрема в обробці мовлення та розпізнаванні мови. Вони виникають зі спроби моделювати сприйняття звуку людським вухом та властивості мовлення.

Для отримання цих коефіцієнтів необхідно сигнал перетворити на часово-частотну характеристику (із використанням STFT), далі застосувати Мел фільтри (результат репрезентується Мел-спектрограмою). Наступним кроком виконується обчислення логарифму для кожного фільтру. І останнім кроком застосовується кепстральне перетворення, яке включає в себе обчислення дискретного косинусного перетворення (DCT) для отримання кепстральних коефіцієнтів (див. рис. 2.7).

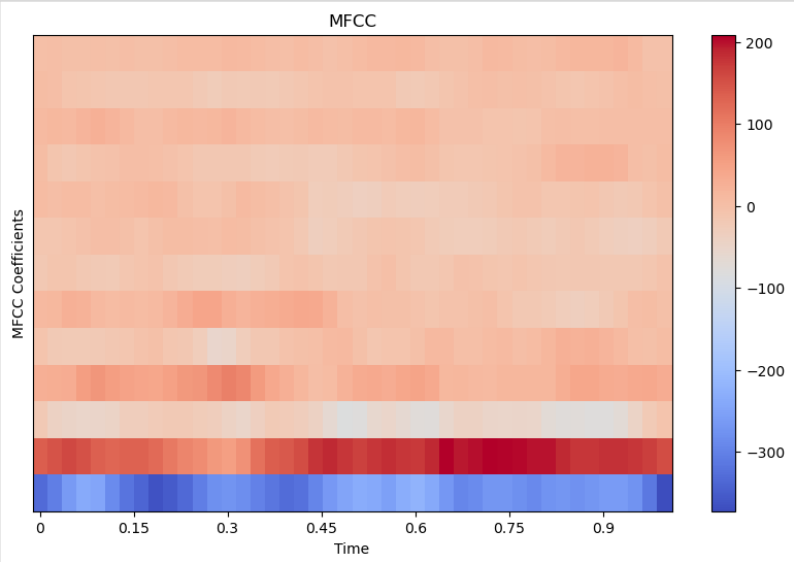

Рисунок 2.7 – Приклад MFCCs (створено самостійно)

Вибирається певна кількість коефіцієнтів для подальшого використання, оскільки не всі коефіцієнти можуть бути однаково інформативними. Зазвичай, перші коефіцієнти є найбільш інформативними (містять інформацію про форманти та інше).

Незважаючи на те, що типів характеристик дуже багато і кожна певним чином характеризує аудіо, практика показує, що найбільш вдало, повно та якісно для обробки аудіо нейронними мережами працюють саме MFCCs. Вони мають

достатньо інформації для досить точного визначення, у нашому випадку, емоційного стану.

Отже, оскільки найефективнішими характеристиками аудіо виявилися MFCCs, то алгоритм їх вилучення наступний:

- розділити кожен аудіо файл на фрейми. Frame size має бути таким, щоб кожен аудіо файл складався з 128 фреймів. Hop size знову ж таки буде вдвічі меншим за розмір кадру;
- застосувати функцію windowing до кожного фрейму;
- застосувати STFT до кожного фрейму;
- перетворити отриману спектрограму в Мел-спектрограму;
- перетворити Мел-спектрограму в MFCCs за допомогою DCT з кількістю коефіцієнтів 40 як найбільш оптимальною.

Таким чином, кожна вибірка даних буде перетворена в матрицю чисел розміром 128 на 40. Отримана матриця коефіцієнтів і буде вхідними даними до математичних моделей.

2.3 Аугментація даних

У машинному навчанні широко використовується така техніка розширення даних як аугментація. Ця техніка передбачає створення нових прикладів даних шляхом застосування різних операцій трансформації до існуючих зразків. Ця техніка застосовується з метою розширення обсягу тренувального набору та покращення загальної здатності моделі до узагальнення на нові, реальні дані.

Серед цілей аугментації можна виділити наступні:

- дозволяє моделі бачити більше різноманітності в тренувальному наборі, що може допомогти уникнути перенавчання та поліпшити узагальнювальні властивості моделі;
- додавання різних варіацій до даних допомагає моделі впоратися з різними умовами та вхідними даними;
- застосування аугментації може допомогти зробити модель менш чутливою до змін в умовах зйомки або в реальних сценаріях.

Аугментація даних є важливим етапом в побудові стійких та ефективних моделей машинного навчання.

Якщо говорити про аугментацію аудіо даних, то зазвичай можна знайти так операції з даними:

- додавання шуму до аудіо;
- розтягування в часі;
- зсув висоти тону;
- застосування різноманітних фільтрів;
- зсування форманти (зміна голосу таким чином, щоб голос чоловіків звучав більш жіночо, а голос жінок – більш чоловічо).

Таким чином, розширення набору даних із використанням зазначених операцій аугментації дозволить створити більш об’єктивну та ефективну математичну модель (модель нейронної мережі в нашому випадку).

Для даної роботи було обрано найпопулярніші типи агментації даних, а саме: додавання шуму, розтягування в часі (швидше і повільніше) та зсув висоти тону (підвищення і зниження висоти тону).

2.4 Оцінка алгоритмів

2.4.1 Метрики

Для оцінки моделей нейронних мереж широко використовуються відомі метрики точності, влучності, повноти (або чутливості) та F-міра. Загалом цих метрик достатньо для точного опису побудованої моделі. Отже, розглянемо більш детально кожну метрику.

Точність (accuracy) – стандартна метрика для оцінки математичних моделей, що описує близькість результатів вимірювання до істинних значень та виражається як відношення кількості правильно класифікованих даних до загальної кількості даних. Розраховується за формулою 2.1.

$$Accuracy = \frac{Number\ of\ Correct\ Predictions}{Total\ Number\ of\ Predictions} \quad (2.1)$$

де $Number\ of\ Correct\ Predictions$ – це кількість правильно класифікованих об'єктів вибірки,

$Total\ Number\ of\ Predictions$ – загальна кількість об'єктів вибірки.

Звісно така метрика не завжди може бути об'єктивною, зокрема, коли є дисбаланс класів у датасеті.

Влучність або прецизійність (precision) – також стандартна метрика оцінки моделей, що описує близькість замірів одне до одного та виражається як відношення правильно класифікованих вірних даних до загальної кількості вірних даних. Розраховується за формулою 2.2.

$$Precision = \frac{True\ Positives}{True\ Positives + False\ Positives} \quad (2.2)$$

де $True\ Positives$ – це кількість позитивних даних, яка була класифікована вірно,

$False\ Positives$ – це кількість позитивних даних, яка була класифікована не вірно.

Прецизійність особливо важлива, коли необхідно зменшити неправильну класифікацію вірних даних (наприклад, зменшити визначення відсутності хвороби, якщо вона дійсно є). Завжди використовується із чутливістю.

Повнота або чутливість (recall) – також стандартна метрика, що дозволяє відобразити можливість математичної моделі ідентифікувати усі вірні.

Ця метрика виражається як відношення правильно класифікованих вірних даних до загальної кількості даних, що були класифіковані як правильні.

Розраховується за формулою 2.3.

$$Recall = \frac{True\ Positives}{True\ Positives + False\ Negatives} \quad (2.3)$$

де $True\ Positives$ – це кількість позитивних даних, яка була класифікована вірно,

$False\ Negatives$ – це кількість негативних даних, яка була класифікована не вірно.

Працює разом із попередньою метрикою та має тісний зв'язок із нею, оскільки зменшення одного показника збільшує інший, тож при моделюванні моделі необхідно досягати балансу між ними.

F-міра (f1-score) – це метрика, яка використовується для оцінки якості класифікації, зокрема, у випадках, де важливо збалансувати влучність та повноту моделі. Вона представляє собою гармонічний середній між цими двома показниками і призначена для випадків, коли вибірки для класів не збалансовані, або коли точність та повнота однаково важливі. Розраховується за формулою 2.4.

$$F_1 = \frac{2 * Precisin * Recall}{Precision + Recall} \quad (2.4)$$

де $Precision$ – значення прецизійності,

$Recall$ – значення чутливості.

F-міра набуває значення від 0 до 1, де 1 вказує на ідеальну модель, яка досягла максимальної точності та повноти (на жаль, наразі таке неможливо).

Ця метрика особливо корисна в задачах, де важливо уникнути як помилок «False Positives» (коли негативний приклад помилково визначається як позитивний), так і «False Negatives» (коли позитивний приклад помилково визначається як негативний).

Усі ці метрики використовувалися у попередніх дослідження і також будуть використані у експериментах дослідження.

2.4.2 Тестування моделі

Для тестування моделі (та власне зняття метрик) у сфері штучного інтелекту застосовуються різні підходи. Загалом вони зводяться до певного розбиття датасету на дані для тренування та тестування.

Train-Test Split:

- датасет розділяється на дві частини: тренувальний та тестовий набір;
- модель тренується на тренувальному наборі, а потім тестується на тестовому для оцінки продуктивності.

K-Fold Cross-Validation:

- датасет розділяється на K підмножин;
- модель тренується та тестується K разів, кожен раз використовуючи інший підмножину як тестовий набір.

Leave-One-Out Cross-Validation:

- кожна семпл даних використовується як тестовий набір один раз;
- модель тренується K разів, де K – кількість семплів даних.

Найпопулярнішим в машинному навчанні є саме K-Fold Cross-Validation.

Цей метод включає поділ набору даних на K підмножин та ітеративне тренування та оцінювання моделі K разів. Під час кожної ітерації одна із підмножин в використовується як тестовий набір, тоді як інші K-1 підмножини використовуються для тренування моделі. На кожній ітерації записуються метрики (описані у пункті 2.4.1). Після усіх ітерацій метрики усереднюються для отримання стійкої оцінки продуктивності моделі.

Основні переваги K-Fold Cross-Validation включають:

- K-Fold Cross-Validation надає більш стабільну та менш упереджену оцінку продуктивності моделі порівняно з одноразовим розбиттям на тренувальний та тестовий набори;
- гарантується, що кожна точка даних використовується для тестування рівно один раз, і модель зустрічається з різними підмножинами даних під час тренування;

- допомагає оцінити, наскільки добре модель узагальнює до різних підмножин даних, надаючи важливі відомості про її стійкість.

Вибір значення K залежить від розміру датасету та обчислювальних ресурсів. Зазвичай використовують значення K рівне 5 або 10. Загалом K-Fold Cross-Validation є цінним методом для отримання більш надійної оцінки продуктивності моделі, особливо при роботі з обмеженими обсягами даних.

Власне саме K-Fold Cross-Validation і буде використано для остаточної оцінки моделей у експериментах дослідження. Кількість підмножин для цієї валідації буде дорівнювати 5. Однак, для попередніх розробок та тестування буде використовуватися train-test split для більш швидкого отримання приблизних результатів.

2.5 Інструменти розробки

Основним інструментом розробки є мова програмування Python, яка ідеально підходить для машинного навчання, обробки даних та нейронних мереж завдяки своїй простоті та багатству спеціалізованих бібліотек. Зокрема, бібліотека librosa [29] використовується для роботи з аудіо, забезпечуючи всі необхідні функції для підготовки даних. Для побудови моделей використовується бібліотека Keras, яка є високорівневим API для TensorFlow, дозволяючи легко створювати моделі та налаштовувати параметри навчання. Допоміжними бібліотеками є numpy для роботи з масивами даних і matplotlib для візуалізації даних.

Розробка проводилася в середовищі JupyterLab, яке дозволяє об’єднувати код, результати та коментарі, сприяючи організованому процесу розробки.

Конфігурація системи, на якій здійснювалася розробка, включає:

- AMD Ryzen 3 3600X;
- 32 ГБ ОЗУ;
- NVIDIA RTX 3060 (12GB).

TensorFlow було налаштовано на використання ресурсів відеокарти для забезпечення максимальної ефективності навчання моделей.

3 ПРОВЕДЕННЯ ЕКСПЕРИМЕНТАЛЬНИХ ДОСЛІДЖЕНЬ

Практичні дослідження ставлять на меті з’ясувати, які методи та які моделі нейронних мереж продемонструють кращій результат у задачі розпізнавання людської емоції за аудіо записом.

3.1 Вилучення характеристик

Незважаючи на те, що власне даними є аудіо файли, обрані датасети CREMA-D та IEMOCAP мають трохи різну структуру. Для можливості зчитування аудіо бібліотекою librosa необхідно мати шляхи до файлів.

У випадку із CREMA-D просто необхідно зчитати усі файли підряд, оскільки назви файлів мають у собі мітку класу. Загальна назва файлів ID_SENT_EMO_LVL.wav, де:

- ID – ідентифікатор актора;
- SENT – назва речення, що промовляється;
- EMO – емоція (клас);
- LVL – рівень емоційності.

У випадку з IEMOCAP було необхідно пройти по складній вкладеній структурі по аудіо файлах та співвіднести назви файлів із назвою емоції, яка описана у іншому файлі. Код вилучення шляхів до файлів та відповідних класів зазначено у додатку Д (див. рис. Д.1).

Зчитування файлів для обох наборів даних виконувалося у циклі, що надавало змогу одразу ідентифікувавши клас, провести обробку аудіо запису у відповідності до алгоритму вилучення звукових характеристик. У кінці кожного циклу вже характеристики записувалися до відповідних полів об’єкту з даними.

Подальша обробка проходила однаково для обох наборів даних.

Схема, за якою зберігалися вилучені дані після обробки, має три поля: класи (масив назв класів), вилучені характеристики та лейбли (індекс класу, до якого відноситься файл) (див. рис. 3.1).

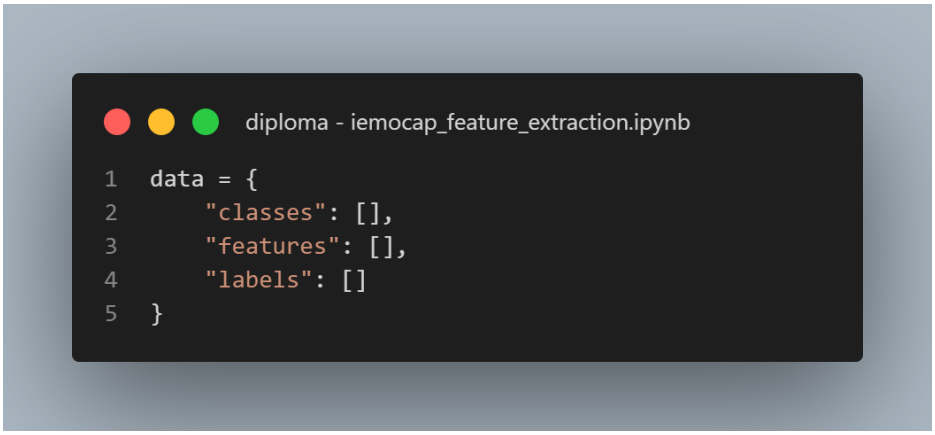

Рисунок 3.1 – Структура даних для характеристик аудіо (створено самостійно)

Також, як описувалося у розділі 2, було обрано для дослідження лише 4 емоції: anger, happiness, sadness та neutral.

Безпосередньо вилучення характеристик виглядає наступним чином:

- читання файлу з використанням бібліотеки librosa (див. рис. 3.2 рядок 1);
- збільшення кількості семплів за рахунок аугментації, якщо необхідно (описано в наступному підрозділі) (див. рис. 3.2 рядок 2);
- для кожного з елементів (оригінального та аугментованих) вилучаються характеристики (див. рис. 3.2 рядок 11-19);
- вилучені характеристики додаються до об’єкту із характеристиками усіх файлів (див. рис. 3.2 рядок 21-25).

Hop size (або hop length) розраховується за формулою 3.1.

$$hop\_length = signal\_length \,//\, frames\_amount \quad (3.1)$$

де $signal\_length$ – це кількість семплів, з яких складається сигнал

$frames\_amount$ – це бажана кількість фреймів (128 у поточному дослідженні),

$//$ – це операція цілочисельного ділення.

Це дозволяє нам рівномірно покрити всю довжину аудіозапису. Розмір власне фрейму вдвічі більше за розмір перекриття. Це забезпечує те, що кожна нова частина аудіо обробляється два рази, з різними контекстами, що покращує точність аналізу і забезпечує більш плавний перехід між фреймами (див. рис. 3.2 рядок 8).

diploma - iemocap_feature_extraction.ipynb

```
signal, sr = librosa.load(file_path, sr=sr)
signals = augment_signal(signal, sr, augs, augs, augs)

for el in signals:
    if len(el)<frames_amount:
        continue

    frame_length = (len(el)//frames_amount) * 2
    hop_length = len(el)//frames_amount

    features = get_features(signal=el,
                            sr=sr,
                            n_fft=frame_length,
                            hop_length=hop_length,
                            n_mfcc=40,
                            mfcc=True,
                            mfcc_delta=False,
                            chroma=False)
    features = features[:frames_amount]

    if not label in data["classes"]:
        data["classes"].append(label)

    data["features"].append(features.tolist())
    data["labels"].append(data["classes"].index(label))
```

Рисунок 3.2 – Код для вилучення характеристик (створено самостійно)

Розглянемо власне вилучення характеристик.

Як було зазначено раніше, найбільш ефективними себе продемонстрували MFCCs, тому саме їх ми і вилучаємо. Звісно, функція get_features має і додаткові можливості, такі як вилучення похідних MFCCs та хромограми, які в поточному дослідженні не використовувалися (див. рис. 3.3).

diploma - cremad_feature_extraction.ipynb

```
1  def get_features(signal,
2                   sr,
3                   n_fft,
4                   hop_length,
5                   n_mfcc,
6                   mfcc,
7                   mfcc_delta,
8                   chroma):
9      features = []
10     if mfcc:
11         mfccs = get_mfccs(signal, sr, n_fft, hop_length, n_mfcc)
12         features.append(mfccs.T)
13     if mfcc_delta:
14         mfccs_delta = get_mfcc_delta(signal, sr, n_fft, hop_length, n_mfcc)
15         features.append(mfccs_delta.T)
16     if chroma:
17         chroma = get_chroma(signal, sr, n_fft, hop_length)
18         features.append(chroma.T)
19
20     return np.hstack(features)
```

Рисунок 3.3 – Функція для вилучення характеристик аудіо (створено самостійно)

З цієї функції найбільш цікавим є рядки 11-12, оскільки саме там викликається функція get_mfccs, яка повертає масив коефіцієнтів (див. рис. 3.4).

diploma - cremad_feature_extraction.ipynb

```
1  def get_mfccs(signal, sr, n_fft, hop_length, n_mfcc):
2      mfccs = librosa.feature.mfcc(y=signal,
3                                   sr=sr,
4                                   n_mfcc=n_mfcc,
5                                   n_fft=n_fft,
6                                   hop_length=hop_length)
7      return mfccs
```

Рисунок 3.4 – Вилучення MFCC з сигналу (створено самостійно)

Це робиться всього однією функцією librosa.feature.mfcc з бібліотеки librosa.

Останнім робиться запис об’єкту характеристик у файл .json (див. рис. 3.5).

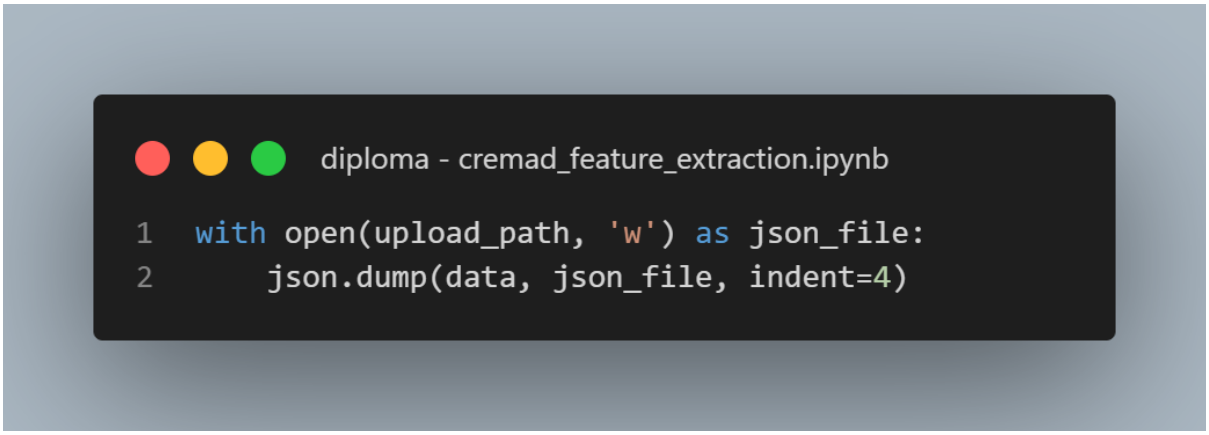

Рисунок 3.5 – Запис вилучених аудіо характеристик у файл (створено самостійно)

Таким чином, було отримано 4 файли, по 2 на кожен набір даних із аугментаціями та без. Загальна кількість семплів в утворених наборах наступна:

- CREMA-D w/o augmentations – 4 900 семплів;
- CREMA-D w augmentations – 29 400 семплів;
- IEMOCAP w/o augmentations – 4 490 семплів;
- IEMOCAP w augmentations – 26 940 семплів.

Набір IEMOCAP приблизно на 9% менший за CREMA-D.

3.2 Аугментація даних

Як вже було зазначено у попередньому розділі, з обох наборів даних були вилучені характеристики із додаванням аугментації та без.

В якості аугментацій було вирішено використовувати додавання шуму до аудіо, зміна висоти тону аудіо (пониження та підвищення) та розтягування у часі (збільшення та зменшення) (див. рис. 3.6). Це найбільш поширені типи аугментації, які демонструють у роботах попередників досить високий приріст ефективності.

Функція приймає об’єкт сигналу, sample rate та прапорці, які вказують на необхідність додавання певного типу аугментації для більш гнучкої роботи із даними.

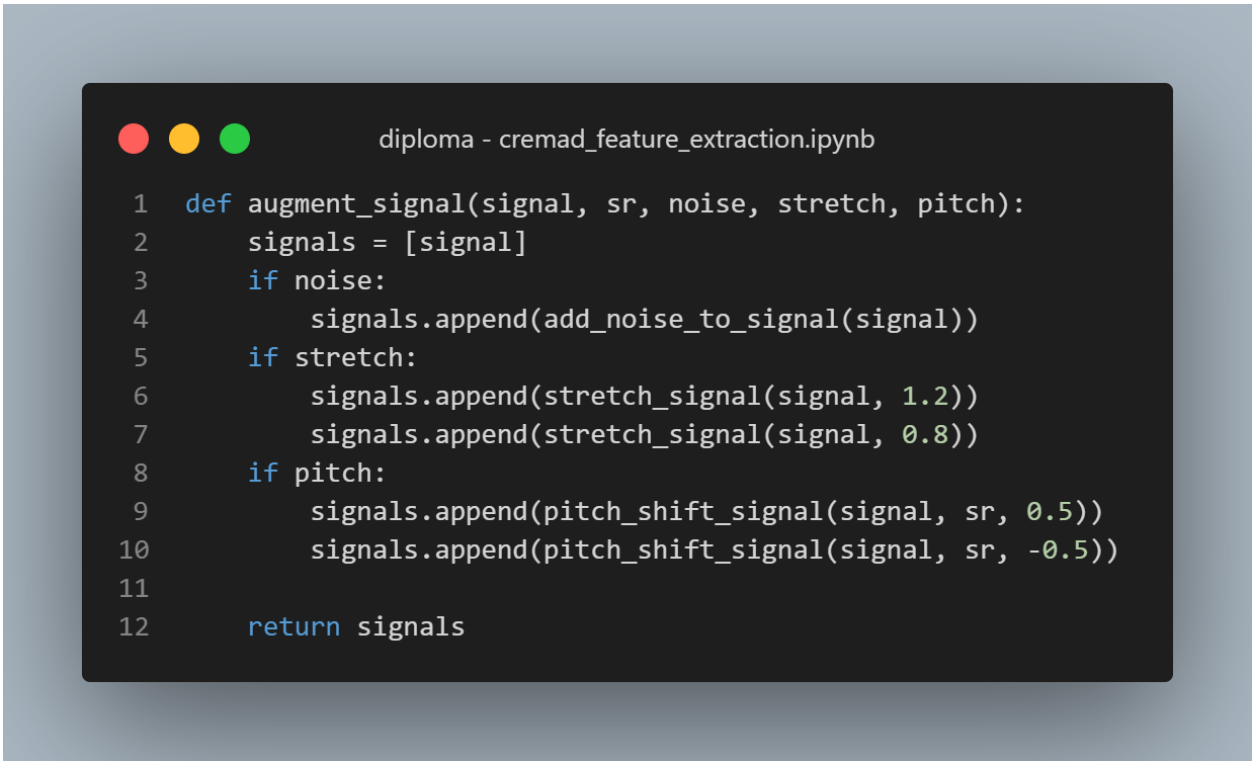

Рисунок 3.6 – Функція з додавання аугментацій до аудіо (створено самостійно)

Додавання шуму робиться через бібліотеку numpy (див. рис. 3.7).

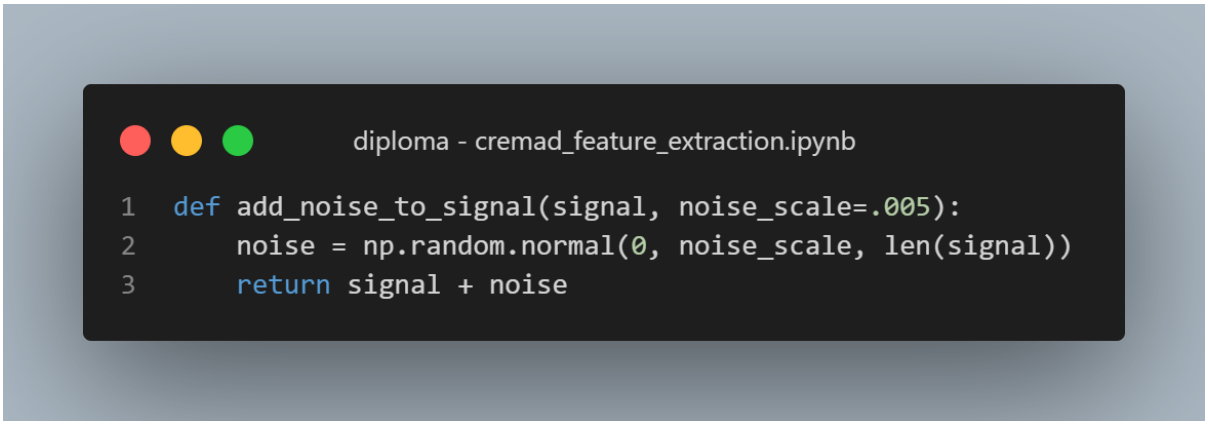

Рисунок 3.7 – Функція додавання шуму до аудіо (створено самостійно)

Розтягування у часі робиться через функцію time_stretch в бібліотеці librosa (див. рис. 3.8). Вона приймає сигнал та значення rate (на скільки змінити довжину аудіо; 1 = оригінальна довжина).

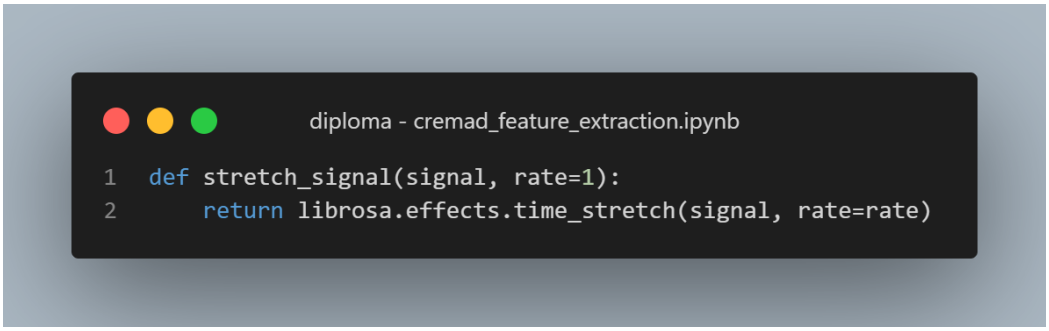

Рисунок 3.8 – Функція розтягування аудіо у часі (створено самостійно)

Зміна тону робиться через функцію pitch_shift бібліотеки librosa (див. рис. 3.9). В якості значення для зміни вона приймає параметр n_steps – кількість тонів на які підняти або підвищити аудіо.

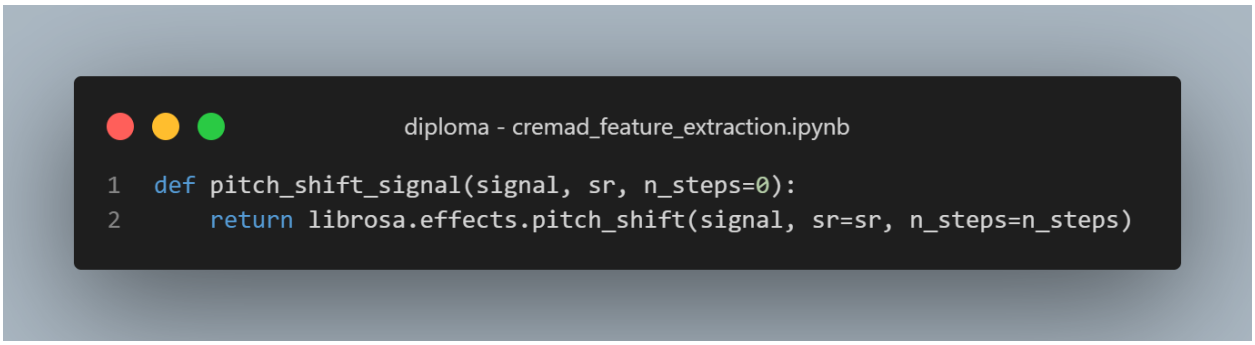

Рисунок 3.9 – Функція зміни тону аудіо (створено самостійно)

Значення noise_scale, rate та n_steps у відповідних функціях обирались таким чином, щоб людський слух чув відміну від оригіналу, але все ще мав змогу розпізнати емоцію. Приклади порівняння оригінальної звукової хвилі (синя) із аугментованою (червона) наведено у додатку Е. Зображено фрагмент аудіо класу «happiness».

3.3 Тренування моделей на наборах без аугментації

У цьому підрозділі будуть описані розроблені моделі із використанням обраних для дослідження архітектур та результати навчання нейронних мереж цих моделей. Результати будуть представлені через метрики, які були отримані в ході K-Fold Cross-Validation (K = 5). Усі тренування проводилися 150 епох, а batch_size

дорівнював 32. Останній шар кожної мережі це повнозв’язний шар із 4 нейронами та функцією активації softmax для формування вірогідності для кожного з 4 класів.

В усіх моделях зустрічаються шари Batch Normalization та Dropout. Перший покликаний прискорювати навчання моделі, другий слугує для зменшення перенавчання моделі та усюди має значення 0.2 (частка нейронів/блоків, які будуть вимкнені)

Усі інші гіперпараметри описані індивідуально.

3.3.1 BiLSTM

В ході тестування вдалося розробити найбільш прийнятну архітектуру двоспрямованої мережі із використанням LSTM блоків. Схема архітектури зображена на рисунку (див. рис. 3.10).

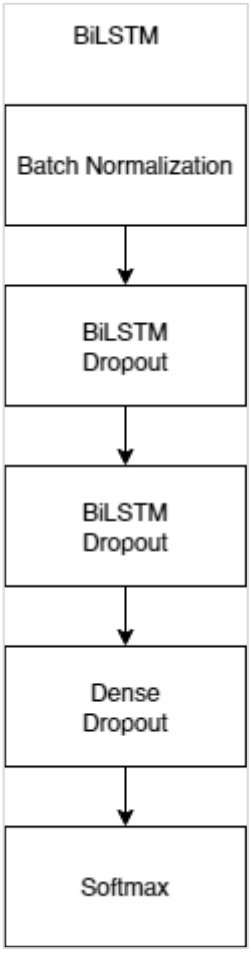

Рисунок 3.10 – Схема архітектури моделі із використанням BiLSTM (створено самостійно)

Перший шар BiLSTM має 128 блоків, другий 256 (однак через те, що шари двоспрямовані, кількість блоків кожного шару вдвічі більша, тобто 256 та 512). Ці шари використовують функцію активації tanh, яка йде за замовчуванням. Повнозв'язний шар мережі складається зі 128 нейронів, після чого йде шар із softmax. Загалом у цій моделі кількість параметрів для навчання складає приблизно 1 290 000.

Більше інформації про шари можна подивитися у додатку Ж на рисунку Ж.1. Код побудови моделі представлено у додатку Д (див. рис. Д.2).

Розглянемо графік навчання для обох наборів даних без аугментацій. Як можна побачити на рисунках нижче для наборів CREMA-D (див. рис. 3.11) та IEMOCAP (див. рис. 3.12) тенденція досить схожа.

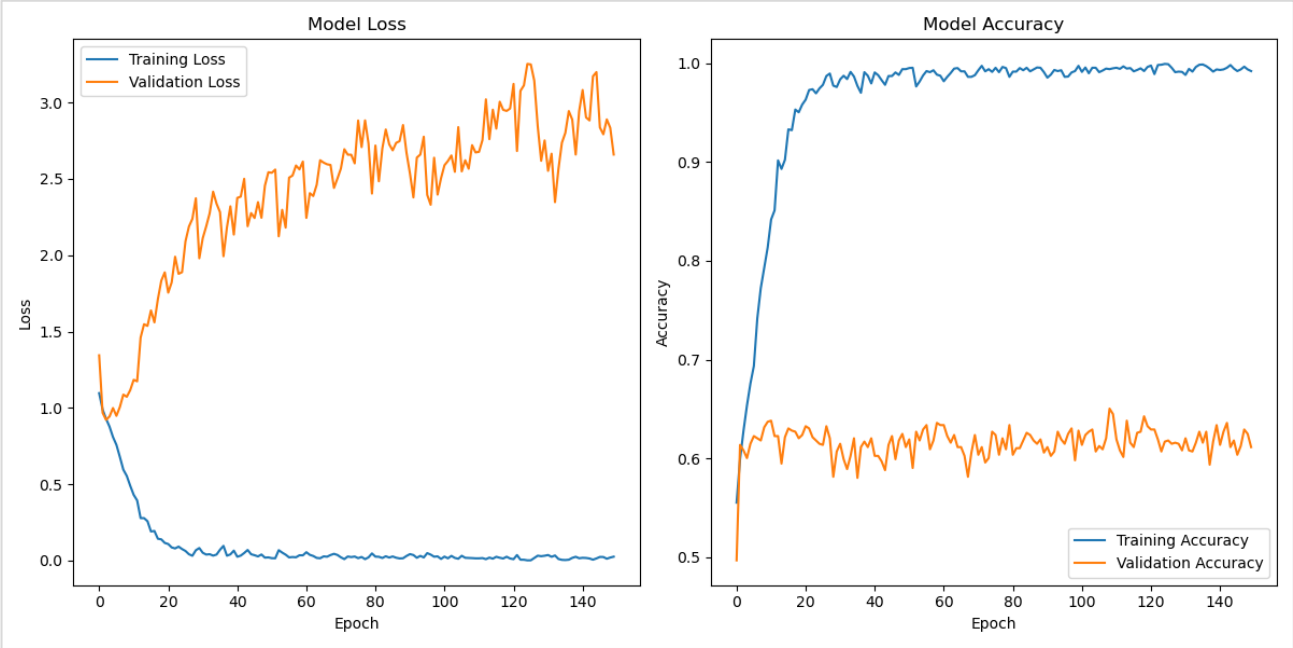

Рисунок 3.11 – Графік навчання моделі із BiLSTM на наборі CREMA-D без аугментацій (створено самостійно)

Вже на 30 епосі модель досягає свого максимуму. Нажаль майже одразу далі іде зростання функції втрат, що погано для моделі.

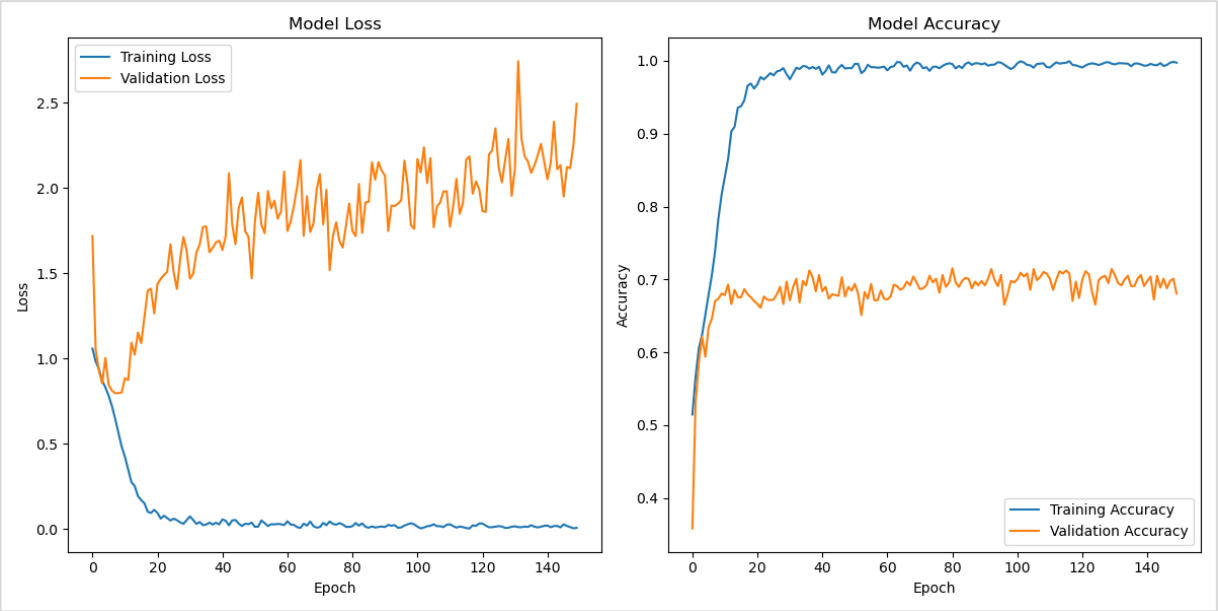

Рисунок 3.12 – Графік навчання моделі із BiLSTM на наборі IEMOCAP без аугментацій (створено самостійно)

Для більш наглядної оцінки результатів, поглянемо на метрики. У таблицях нижче зображені результати навчання наборів CREMA-D (див. табл. 3.1) та IEMOCAP (див. табл. 3.2) відповідно.

Таблиця 3.1 – Результати навчання моделі BiLSTM на наборі CREMA-D без аугментацій (створено самостійно)

	precision	**recall**	**f1-score**	**support**
anger	0,749	0,802	0,772	254
happiness	0,686	0,592	0,633	254
neutral	0,620	0,596	0,607	217
sadness	0,700	0,762	0,729	254
accuracy			0,691	
macro avg	0,689	0,688	0,685	980
weighted avg	0,691	0,691	0,688	980

З таблиці 3.1 видно, що модель найкраще розпізнає емоцію злості, потім сум, а далі досить погано розпізнаються емоції щастя та нейтральна. Це дає змогу зрозуміти, як чітко BiLSTM може розрізнити емоцію на доволі невеликому наборі даних.

Таблиця 3.2 – Результати навчання моделі BiLSTM на наборі IEMOCAP без аугментацій (створено самостійно)

	precision	**recall**	**f1-score**	**support**
anger	0,701	0,725	0,712	221
happiness	0,346	0,227	0,272	119
neutral	0,637	0,655	0,645	342
sadness	0,632	0,693	0,661	217
accuracy			0,624	
macro avg	0,579	0,575	0,572	898
weighted avg	0,613	0,624	0,616	898

Хочеться виділити, що у наборі IEMOCAP досить не рівномірний розподіл класів. Тож, як видно, для емоції «happiness», яка у наборі тестування представлена у меншості значення f1-score (найбільш об’єктивна метрика з погляду точності) й найнижчим і складає приблизно 0,27, коли для інших це значення вище за 0,6. Набір CREMA-D не має такого ефекту.

Загалом, вже можна побачити, що більш рівномірний розподіл класів позитивно впливає на якість моделі. І, як видно, якість моделі на CREMA-D вища на ~7%.

Щодо аналізу BiLSTM із даними без аугментації, то обидва демонструють досить низькі результати до 70% точності, а графік навчання досить неприємний.

3.3.2 GRU

Архітектура нейронної мережі із використанням блоків GRU аналогічна до мережі із BiLSTM (див. рис. 3.13).

Шар нормалізації, два шари із блоками GRU з додатковим шаром Dropout для уникнення перенавчання моделі, повнозв’язний шар та шар із softmax для власне класифікації.

Оскільки внутрішньо склад блоку GRU простіше за блок BiLSTM, то і процес навчання повинен бути швидшим. Власне, під час проведення експериментів так і було: навчання моделі з GRU було на порядок швидше. Це досить актуальна перевага під час дослідження.

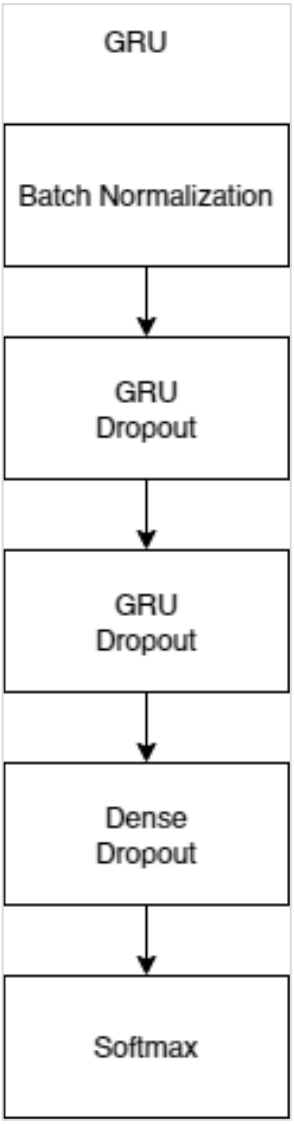

Рисунок 3.13 – Схема архітектури моделі із використанням GRU (створено самостійно)

Більше інформації про шари можна подивитися у додатку Ж (див. рис. Ж.2). Загальна кількість параметрів навчання складає приблизно 1 478 000 (трохи більше за BiLSTM). Код побудови моделі представлено у додатку Д (див. рис. Д.3).

Кількість блоків GRU така ж сама, як і у випадку з BiLSTM: 256 та 512 для першого та другого шару відповідно. Інші параметри такі самі.

Тобто відмінності лише у структурі рекурентного блоку.

Подивимось на графік навчання на наборі CREMA-D без аугментації (див. рис. 3.14). Для набору IEMOCAP графік майже ідентичний.

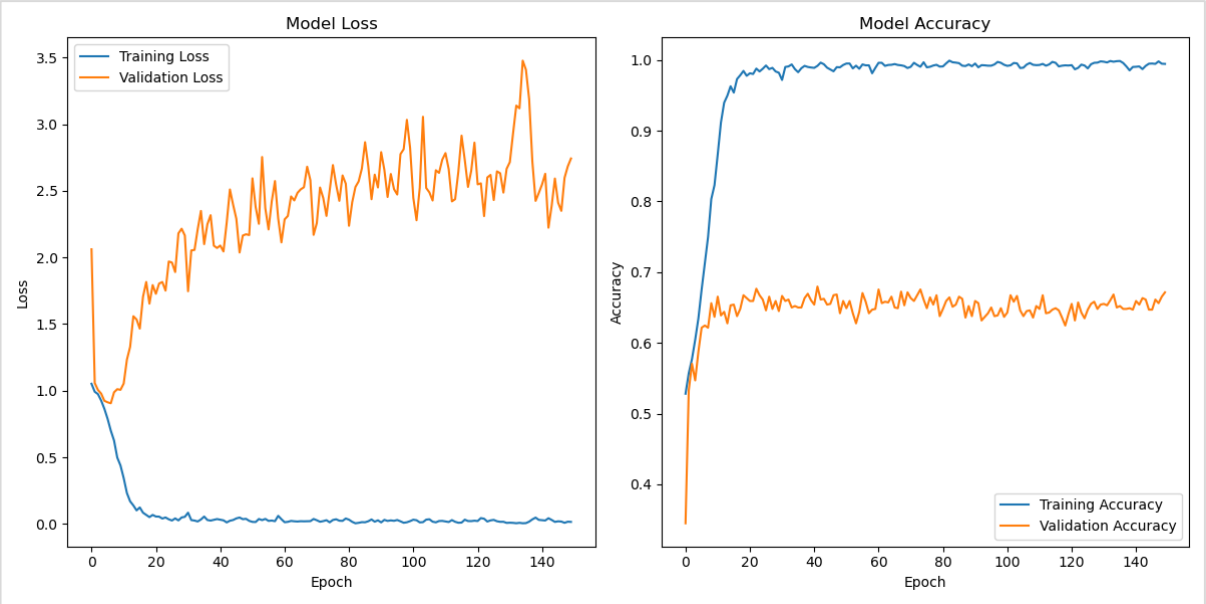

Рисунок 3.14 – Графік навчання моделі із GRU на наборі CRERMA-D без аугментацій (створено самостійно)

Загалом, можна побачити, що тенденція навчання така сама як і з блоками BiLSTM: точність не висока, збільшення витрат із часом.

Подивимось на метрики у таблиці нижче для CREMA-D (див. табл. 3.3) та IEMOCAP (див. табл. 3.4) відповідно.

Таблиця 3.3 – Результати навчання моделі GRU на наборі CREMA-D без аугментацій (створено самостійно)

	precision	**recall**	**f1-score**	**support**
anger	0,766	0,752	0,757	254
happiness	0,621	0,603	0,608	254
neutral	0,480	0,522	0,498	217
sadness	0,655	0,625	0,638	254
accuracy			0,630	
macro avg	0,630	0,626	0,625	980
weighted avg	0,636	0,630	0,630	980

Знову можна побачити, що нерівномірність класів у IEMOCAP знижує точність цього набору, але вже не на стільки сильно.

Таблиця 3.4 – Результати навчання моделі GRU на наборі IEMOCAP без аугментацій (створено самостійно)

	precision	**recall**	**f1-score**	**support**
anger	0,663	0,693	0,674	221
happiness	0,275	0,180	0,215	119
neutral	0,608	0,625	0,613	342
sadness	0,606	0,650	0,627	217
accuracy			0,589	
macro avg	0,538	0,537	0,532	898
weighted avg	0,577	0,589	0,579	898

Із використанням GRU для обох наборів без аугментації точність виявляється нижче на 3-6%. Але знову тенденція лідерства набору CREMA-D залишається і її точність складає близько 63%.

3.3.3 CNN

Ця архітектура повинна демонструвати більш якісні результати завдяки використанню більш глибокої структури і ефективних методів регуляції. Фінальна архітектура мережі із використанням згорткових шарів представлена на рисунку (див. рис. 3.15).

Модель містить 6 згорткових блоків, кожен з яких складається з:

- conv 2D (розмір ядра 5 на 5; функція активації ReLU; однакові відступи);
- batch normalization;
- max pooling 2D (розмір 4 на 4);
- dropout (значення відключення 0.2).

Після проходження через 6 згорткових блоків дані передаються до проміжного шару Flatten, який перетворює вихідну багатовимірну матрицю в одномірний вектор. Це необхідно для підготовки даних до входу в повнозв'язні шари. Після чого ідентичні до попередніх архітектур повнозвязний шар із 128 нейронами та шаром Dropout та Softmax для класифікації.

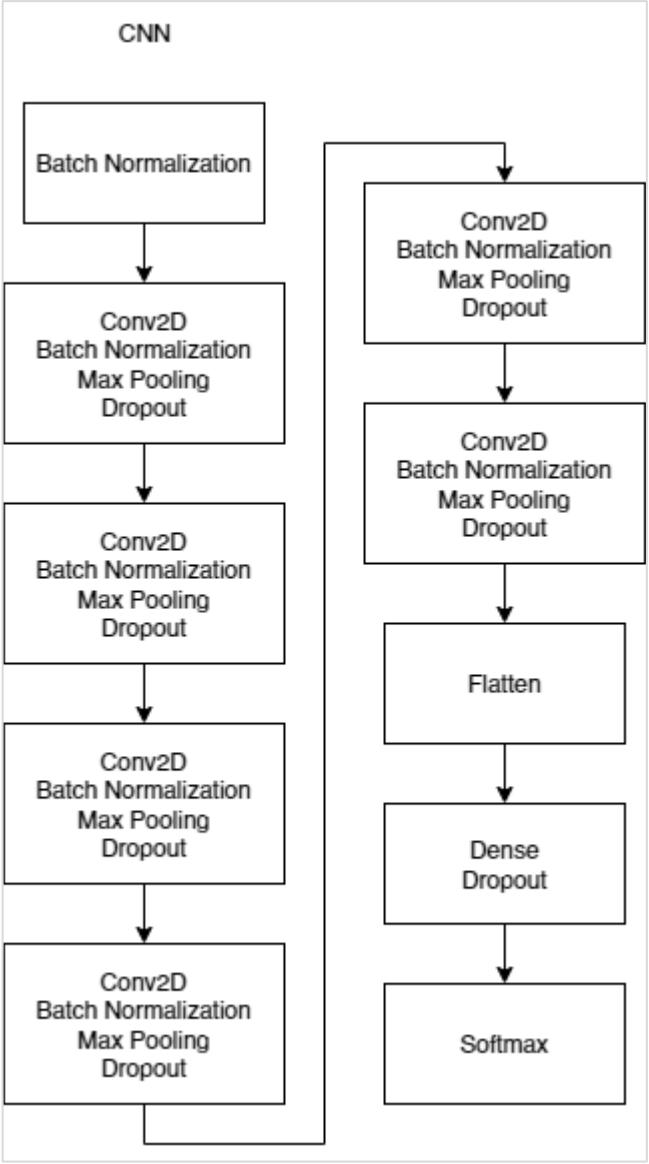

Рисунок 3.15 – Схема архітектури моделі із використанням CNN (створено самостійно)

Більше інформації про архітектуру можна знайти ну додатку Ж (див. рис. Ж.3), а код створення моделі у додатку Д (див. рис. Д.4). Загальна кількість параметрів для навчання у цій моделі складає приблизно 2 млн.

Подивимось на графіки навчання цієї моделі на наборах CREMA-D (див. рис. 3.16) та IEMOCAP (див. рис. 3.17) відповідно. Видно, що графіки виглядають значно краще у порівнянні з рекурентними мережами, хоча точність на приблизно такому ж рівні, а стабільність навчання досить низька.

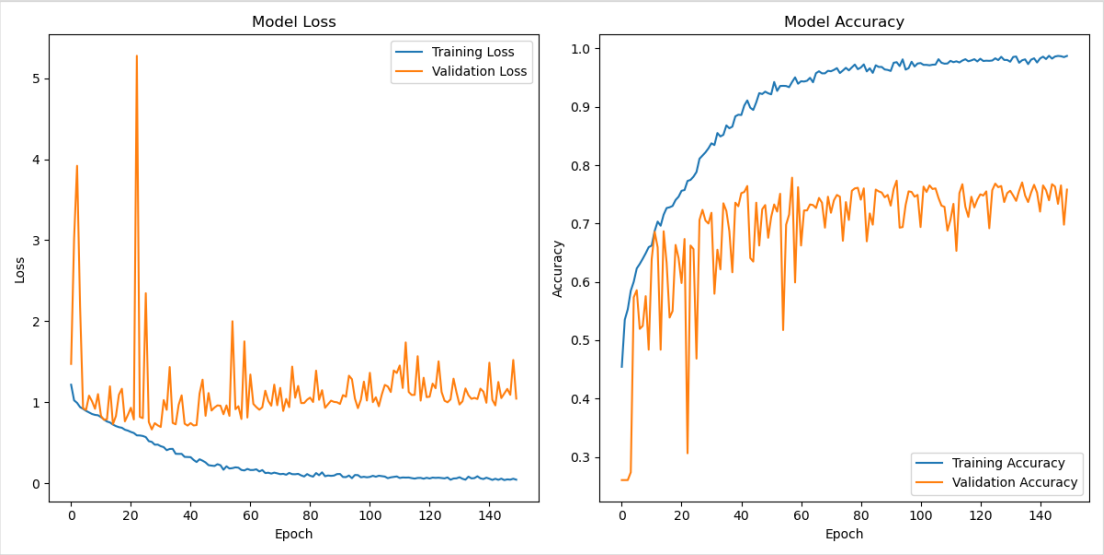

Рисунок 3.16 – Графік навчання моделі із CNN на наборі CRERMA-D без аугментацій (створено самостійно)

Загалом також видно, що функція втрат показує кращий результат на наборі з більш збалансованим розподілом класів, тобто CREMA-D.

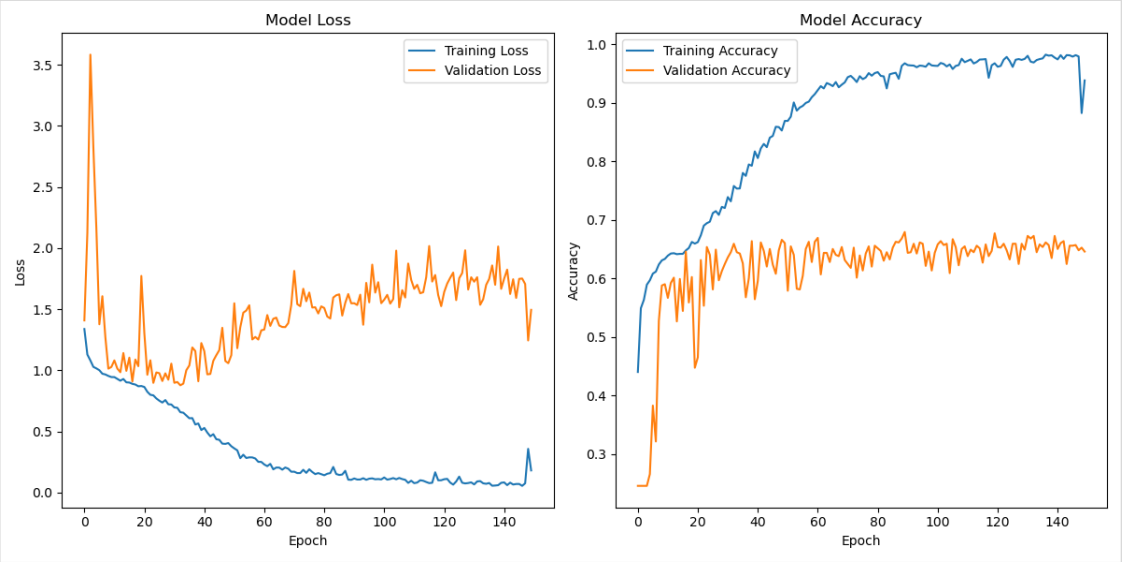

Рисунок 3.17 – Графік навчання моделі із CNN на наборі IEMOCAP без аугментацій (створено самостійно)

Подивимось на метрики CREMA-D (див. табл. 3.5) та IEMOCAP (див. табл. 3.6).

Знову ж таки можна бачити тенденцію лідерства якості моделі із використанням набору CREMA-D. Крім того, також можна зробити висновок, що

модель із використанням згорткових шарів краща за моделі із рекурентними шарами приблизно на 3-5% (у порівнянні із кращою BiLSTM серед рекурентних).

Таблиця 3.5 – Результати навчання моделі CNN на наборі CREMA-D без аугментацій (створено самостійно)

	precision	**recall**	**f1-score**	**support**
anger	0,828	0,789	0,803	254
happiness	0,727	0,710	0,718	254
neutral	0,678	0,732	0,698	217
sadness	0,799	0,762	0,771	254
accuracy			0,749	
macro avg	0,758	0,748	0,747	980
weighted avg	0,761	0,749	0,749	980

Доволі непогані значення згорткової мережі із точністю у ~75%.

Таблиця 3.6 – Результати навчання моделі CNN на наборі IEMOCAP без аугментацій (створено самостійно)

	precision	**recall**	**f1-score**	**support**
anger	0,732	0,754	0,742	221
happiness	0,414	0,190	0,259	119
neutral	0,632	0,714	0,669	342
sadness	0,686	0,704	0,684	217
accuracy			0,652	
macro avg	0,616	0,590	0,589	898
weighted avg	0,641	0,652	0,636	898

Однак, результати навіть із CNN без аугментації досить низькі, щоб можна було використовувати створену модель.

3.3.4 CRNN

Розглянемо останню архітектуру, які використовую як згорткові шари, так і шари з рекурентними блоками.

Подивимось на структуру моделі CRNN (див. рис. 3.18).

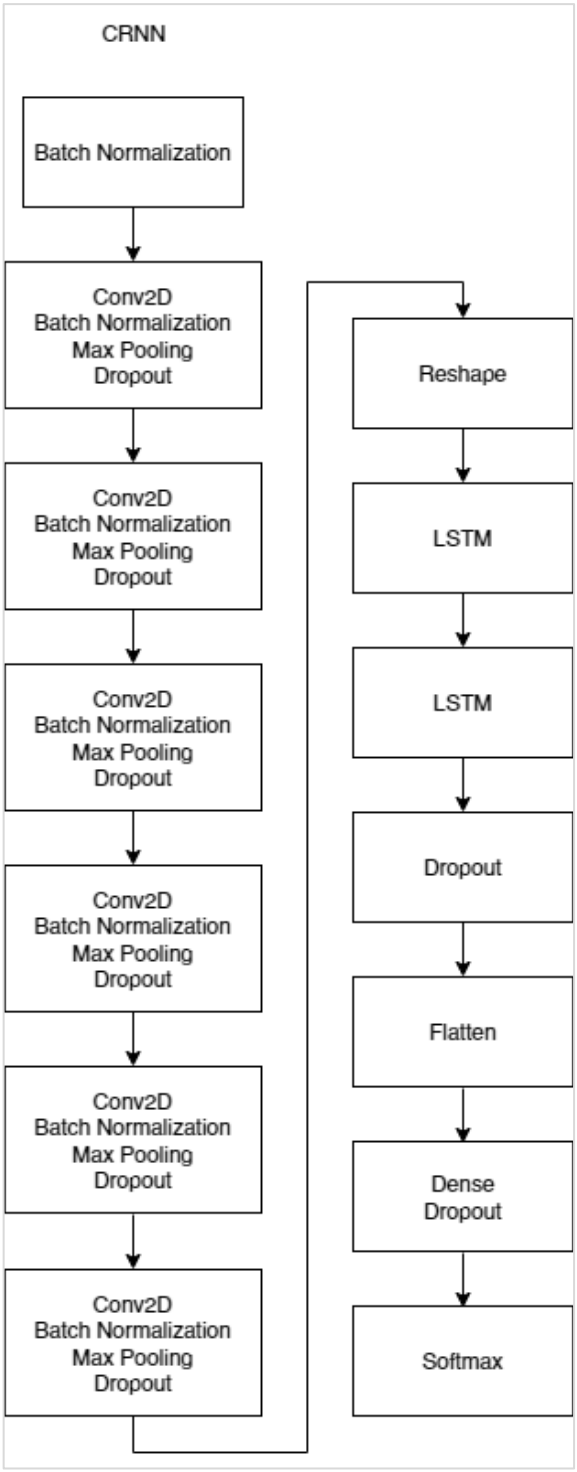

Рисунок 3.18 – Схема архітектури моделі із використанням CNN та LSTM (створено самостійно)

Перша частина мережі складається із тих самих 6 згорткових блоків, що використовуються й у звичайній CNN. Після них йде проміжний шар Reshape для того, щоб перетворити багатовимірні дані після згортання у відповідно форму для згорткових шарів.

Два згорткових шари мають по 256 та 512 блоків LSTM відповідно. Після них йде Dropout зі значенням 0.2, шар Flatten для отриманя одномірної структури, повнозвязний шар із 128 нейронами та шаром Dropout та Softmax для класифікації.

Більше інформації щодо моделі можна переглянути у додатку Ж (див. рис. Ж.4), а код побудови моделі знаходиться у додатку Д (див. рис. Д.5). Загальна кількість параметрів для навчання у цій моделі складає приблизно 4 млн.

Одразу подивимось на графік навчання на наборі CREMA-D (див. рис. 3.19).

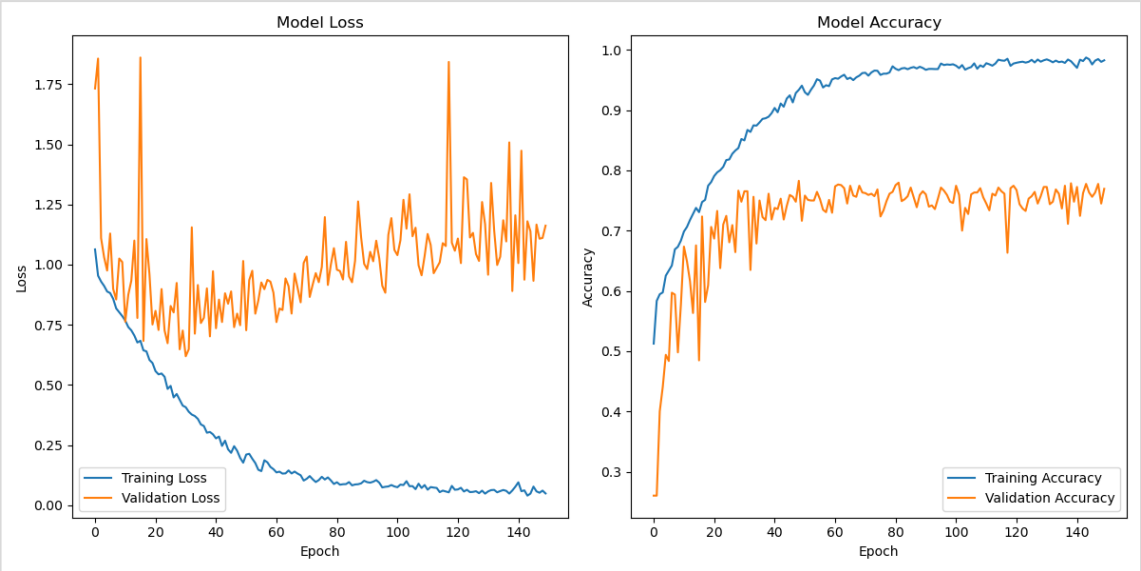

Рисунок 3.19 – Графік навчання моделі CRNN на наборі CREMA-D без аугментацій (створено самостійно)

Загалом графік такий як і для CNN, але менш стабільний для функції втрат. Подивимось на метрики CREMA-D (див. табл. 3.7) та IEMOCAP (див. табл. 3.8).

Таблиця 3.7 – Результати навчання моделі CRNN на наборі CREMA-D без аугментацій (створено самостійно)

	precision	**recall**	**f1-score**	**support**
anger	0,780	0,826	0,790	254
happiness	0,724	0,666	0,689	254
neutral	0,668	0,790	0,720	217
sadness	0,856	0,691	0,758	254
accuracy			0,741	
macro avg	0,757	0,743	0,739	980
weighted avg	0,760	0,741	0,740	980

Якщо порівнювати із моделлю звичайної CNN, то результати гірше у межах 1% для обох наборів.

Таблиця 3.8 – Результати навчання моделі CRNN на наборі IEMOCAP без аугментацій (створено самостійно)

	precision	**recall**	**f1-score**	**support**
anger	0,781	0,704	0,733	221
happiness	0,402	0,276	0,317	119
neutral	0,636	0,690	0,659	342
sadness	0,660	0,713	0,681	217
accuracy			0,644	
macro avg	0,620	0,596	0,598	898
weighted avg	0,646	0,644	0,637	898

Для такої невеликої кількості даних модель CRNN дещо гірша за CNN. Але знову ж таки, ні одна з моделей не надала якісних результатів. Розглянемо результати навчання на даних із аугментаціями.

3.4 Тренування моделей на наборах з аугментаціями

Як вже було зазначено, для в якості аугментації було використано додавання шуму, розтягування за часом (довше та швидше) та змінення тону (пониження та підвищення). Тобто розмір початкового набору даних збільшився у 5 разів. Звісно, це не звичайне збільшення, оскільки нові семпли відрізняються від оригінальних.

Усі особливості моделей залишились такими ж, зокрема набори шарів та інші гіперпараметри. Окрім даних нічого не змінилося.

3.4.1 BiLSTM

Графіки навчання можна побачити на рисунках нижче для CREMA-D (див. рис. 3.20) та IEMOCAP (див. рис. 3.21). У порівнянні із набором без аугментації, навчання проходило значно краще і більш стабільно. Значення функції витрат та власне точності дуже приємні.

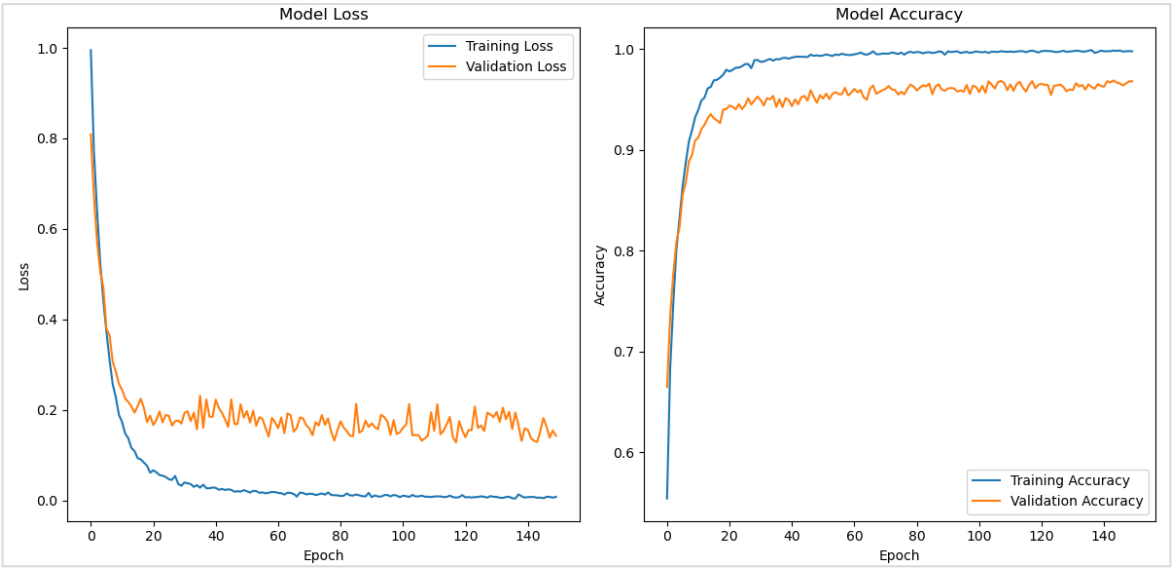

Рисунок 3.20 – Графік навчання моделі BiLSTM на наборі CREMA-D з аугментаціями (створено самостійно)

Але навіть з аугментаціями видно, що функція витрат має більші (отже, гірші) значення для набору IEMOCAP. До того ж значення точності для CREMA-D також вище.

Також можна констатувати, що створена модель BiLSTM є не ідеальною, оскільки графік точності тренування та валідації вже після 20 епохи розбігаються, тобто модель зазнає перенавчання у даній конфігурації.

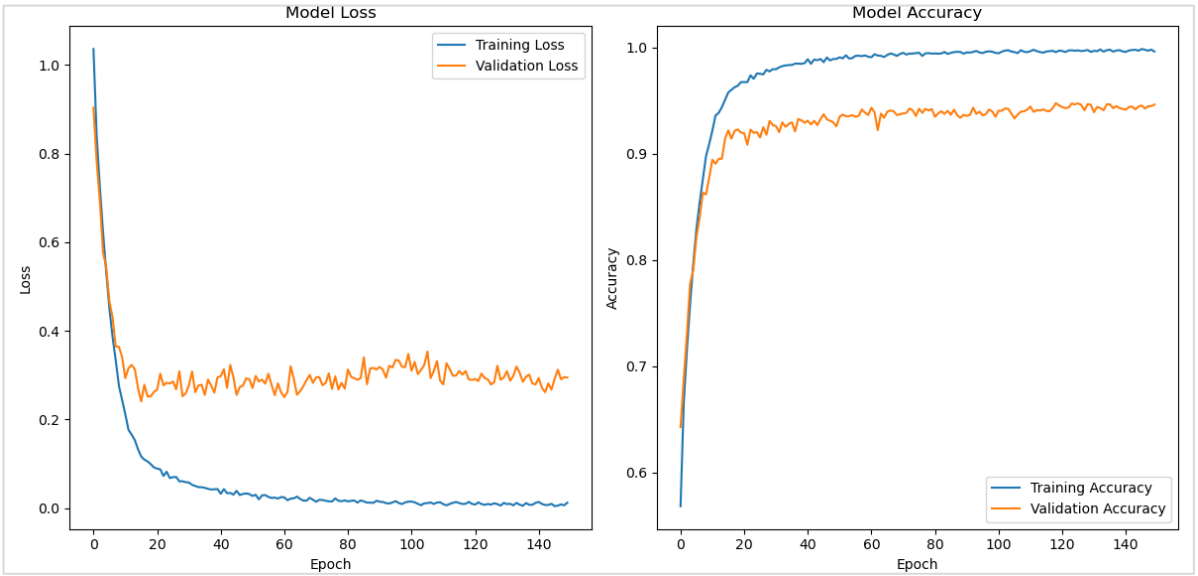

Рисунок 3.21 – Графік навчання моделі BiLSTM на наборі IEMOCAP з аугментаціями (створено самостійно)

Розглянемо отримані метрики для CREMA-D (див. табл. 3.9) та IEMOCAP (див. табл. 3.10).

Таблиця 3.9 – Результати навчання моделі BiLSTM на наборі CREMA-D з аугментаціями (створено самостійно)

	precision	**recall**	**f1-score**	**support**
anger	0,978	0,979	0,979	1 525
happiness	0,969	0,960	0,965	1 525
neutral	0,944	0,963	0,953	1 304
sadness	0,969	0,958	0,963	1 525
accuracy			0,965	
macro avg	0,965	0,965	0,965	5 880
weighted avg	0,966	0,965	0,965	5 880

Архітектура BiLSTM демонструє досить високі результати навчання.

Таблиця 3.10 – Результати навчання моделі BiLSTM на наборі IEMOCAP з аугментаціями (створено самостійно)

	precision	**recall**	**f1-score**	**support**
anger	0,969	0,962	0,965	1 324
happiness	0,945	0,884	0,913	714
neutral	0,936	0,947	0,942	2 050
sadness	0,927	0,950	0,938	1 301
accuracy			0,943	
macro avg	0,944	0,936	0,940	5 388
weighted avg	0,943	0,943	0,943	5 388

З таблиць видно, що обидва набори дозволяють отримати точність на моделі із використанням блоків BiLSTM більше 90%. Однак, знову ж таки, на наборі CREMA-D модель показує результати на ~2-2,5% краще. Отримані значення досить високі і з такою моделлю вже можна працювати у подальшому.

Також варто відмітити, що з аугментованими даними, розбіг для емоції «happiness» у метриці f1-score для обох датасетів значно менше. Звісно, у CREMA-D значення f1-score для усіх емоцій дуже схоже, в той час, коли для IEMOCAP

трохи нижче. Однак, значення f1-score для «happiness» у IEMOCAP теперь нижче на кілька відсотків, в той час, коли без аугментацій він був нижчий більше за 100%. Більш наочно це можна проглянути у додатку И. Там на рисунках И.1 – И.4 зображено матриці помилок для описаних моделей. Там видно, на скільки сильно впливає нерівномірність класів на якість моделі

3.4.2 GRU

Графіки навчання для наборів CREMA-D (див. рис. 3.22) та IEMOCAP (див. рис. 3.23) ідентичні до графіків навчання для BiLSTM.

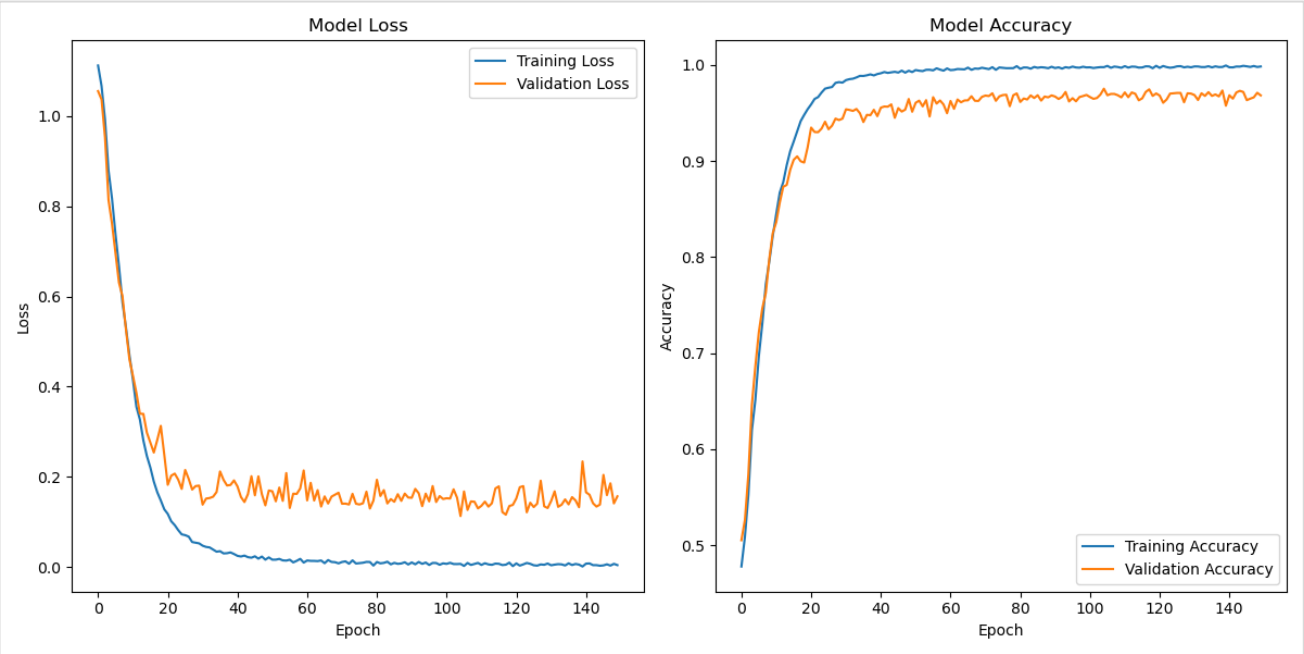

Рисунок 3.22 – Графік навчання моделі GRU на наборі CREMA-D з аугментаціями (створено самостійно)

Модель демонструє досить стабільні результати без значних стрибків у функції втрат і високі значення точності на обох наборах даних. Однак, помітна тенденція до перенавчання: хоча функція втрат не зростає, валідаційна точність залишається стабільною після 20 епохи, що може свідчити про те, що модель вчиться запам’ятовувати тренувальні дані замість того, щоб узагальнювати їх.

Тож модель на базі GRU також потребує додаткових методів для запобігання перенавчанню, таких як регуляція або налаштування архітектури самої моделі.

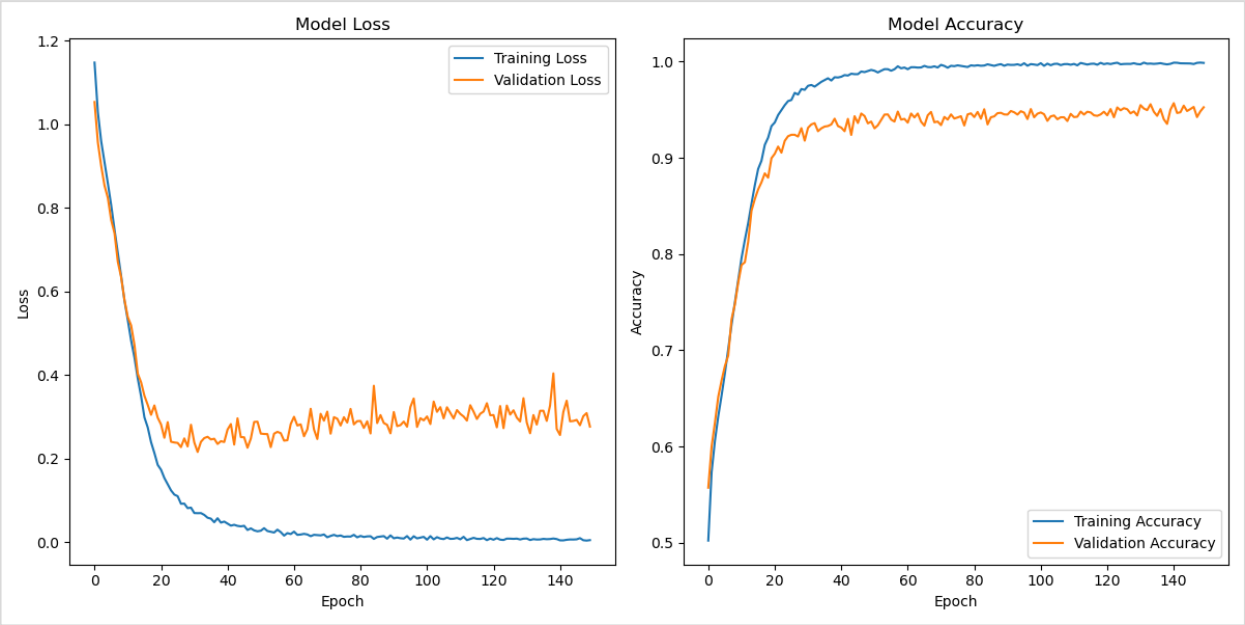

Рисунок 3.23 – Графік навчання моделі GRU на наборі IEMOCAP з аугментаціями (створено самостійно)

Тож одразу перейдемо до метрик CREMA-D (див. табл. 3.11) та IEMOCAP (див. табл. 3.12).

Таблиця 3.11 – Результати навчання моделі GRU на наборі CREMA-D з аугментаціями (створено самостійно)

	precision	**recall**	**f1-score**	**support**
anger	0,982	0,977	0,980	1 525
happiness	0,971	0,964	0,967	1 525
neutral	0,937	0,957	0,947	1 304
sadness	0,963	0,957	0,960	1 525
accuracy			0,964	
macro avg	0,963	0,964	0,963	5 880
weighted avg	0,964	0,964	0,964	5 880

Аналогічно до BiLSTM – досить високі показники точності.

Видно, що рівномірність класів доволі чітко впливає на показники метрик, зокрема f1-score.

Через трохи меншу кількість семплів класу «neutral», точність моделі для цього класу трохи нижче, однак значення метрик інших класів вище 0,96.

Таблиця 3.12 – Результати навчання моделі GRU на наборі IEMOCAP з аугментаціями (створено самостійно)

	precision	**recall**	**f1-score**	**support**
anger	0,979	0,970	0,974	1 324
happiness	0,939	0,904	0,921	714
neutral	0,947	0,959	0,953	2 050
sadness	0,946	0,954	0,950	1 301
accuracy			0,953	
macro avg	0,953	0,947	0,950	5 388
weighted avg	0,954	0,953	0,953	5 388

Одразу видно, що модель з GRU демонструє схожий із BiLSTM результат.

Однак відрив між наборами зменшився. Якщо для BiLSTM на наборі IEMOCAP точність складала ~94%, то з GRU вона складає вже ~95%

3.4.3 CNN

Графіки навчання для однієї з кращих моделей представлені на рисунках нижче для CREMA-D (див. рис. 3.24) та IEMOCAP (див. рис. 3.25).

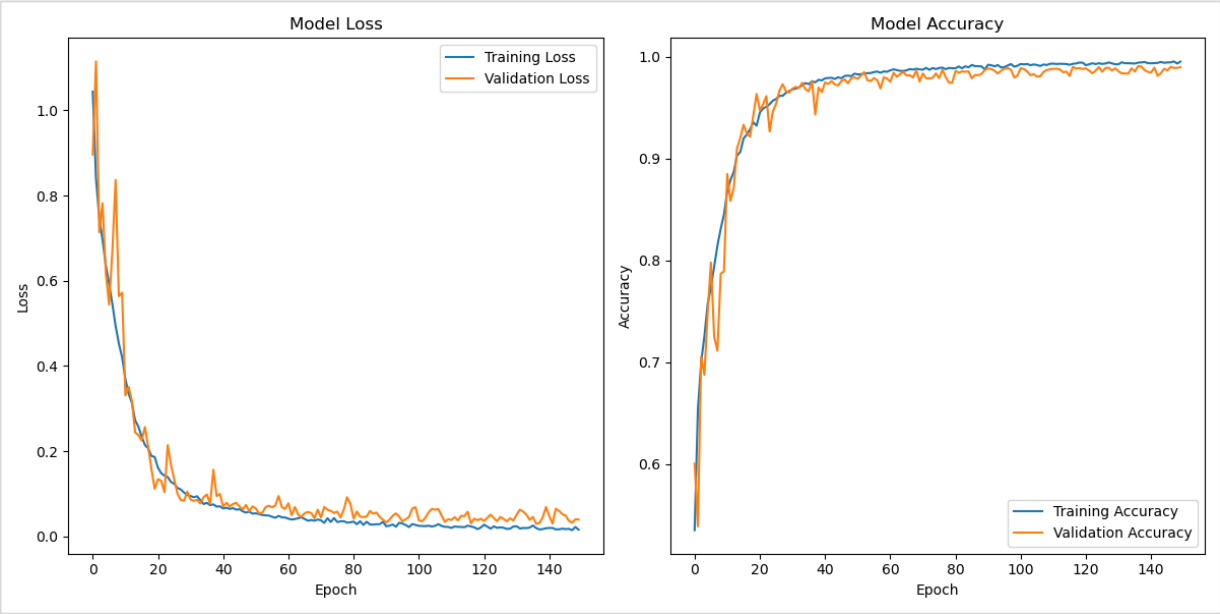

Рисунок 3.24 – Графік навчання моделі CNN на наборі CREMA-D з аугментаціями (створено самостійно)

Вони демонструють досить високу стабільність під час навчання, а також якісне розпізнавання без перенавчання, оскільки графіки навчання та валідації для обох наборів дуже близько розташовані.

Розглянемо метрики для наборів CREMA-D (див. табл. 3.13) та IEMOCAP (див. табл. 3.14).

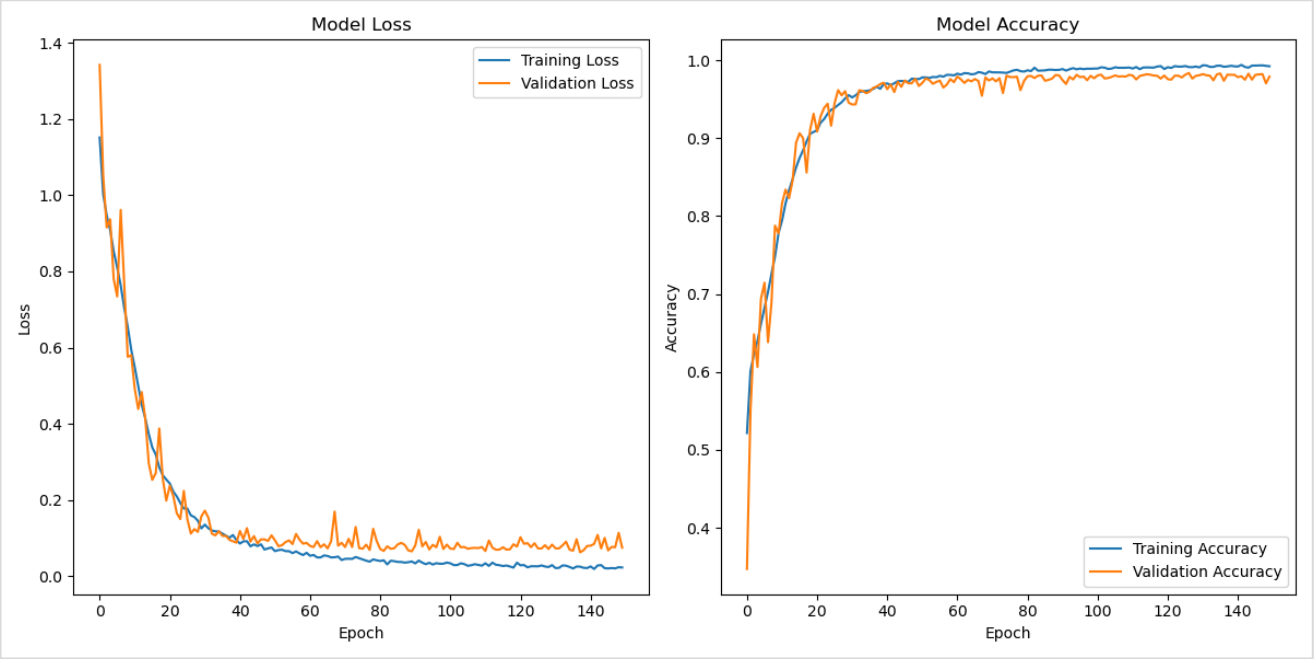

Рисунок 3.25 – Графік навчання моделі CNN на наборі IEMOCAP з аугментаціями (створено самостійно)

Таблиця 3.13 – Результати навчання моделі CNN на наборі CREMA-D з аугментаціями (створено самостійно)

	precision	**recall**	**f1-score**	**support**
anger	0,998	0,992	0,995	1 525
happiness	0,992	0,987	0,989	1 525
neutral	0,977	0,986	0,981	1 304
sadness	0,985	0,987	0,986	1 525
accuracy			0,988	
macro avg	0,988	0,988	0,988	5 880
weighted avg	0,988	0,988	0,988	5 880

Унікально високі результати серед усіх розглянутих моделей у моделі із використанням згорткових шарів

Таблиця 3.14 – Результати навчання моделі CNN на наборі IEMOCAP з аугментаціями (створено самостійно)

	precision	**recall**	**f1-score**	**support**
anger	0,993	0,992	0,993	1 324
happiness	0,993	0,953	0,973	714
neutral	0,981	0,983	0,982	2 050
sadness	0,967	0,987	0,977	1 301
accuracy			0,982	
macro avg	0,984	0,979	0,981	5 388
weighted avg	0,982	0,982	0,982	5 388

Лідери, виявлені під час тестування без аугментацій, збереглися.

Точність на обох наборах досягає вище 98%. До того ж розбіжність між моделями на CREMA-D та IEMOCAP тепер складає в межах одного відсотка.

3.4.4 CRNN

Остання модель для тестування на наборах із аугментаціями – CRNN. Графіки навчання ідентичні до графіків навчання моделі CNN, тож одразу розглянемо метрики CREMA-D (див. табл. 3.15) та IEMOCAP (див. табл. 3.16).

Таблиця 3.15 – Результати навчання моделі CRNN на наборі CREMA-D з аугментаціями (створено самостійно)

	precision	**recall**	**f1-score**	**support**
anger	0,996	0,994	0,995	1 525
happiness	0,993	0,981	0,987	1 525
neutral	0,981	0,984	0,983	1 304
sadness	0,982	0,992	0,987	1 525
accuracy			0,988	
macro avg	0,988	0,988	0,988	5 880
weighted avg	0,988	0,988	0,988	5 880

Також видно дуже високий результат відповідно до відображених в таблицях метри.

Таблиця 3.16 – Результати навчання моделі CRNN на наборі IEMOCAP з аугментаціями (створено самостійно)

	precision	**recall**	**f1-score**	**support**
anger	0,997	0,989	0,993	1 324
happiness	0,990	0,962	0,975	714
neutral	0,984	0,980	0,982	2 050
sadness	0,960	0,989	0,974	1 301
accuracy			0,982	
macro avg	0,983	0,980	0,981	5 388
weighted avg	0,982	0,982	0,982	5 388

Як можна бачити з метрик, результати для CRNN ідентичні до результатів CNN. Отже обидві моделі показують однаково якісний результат більше 98% точності.

3.5 Аналіз результатів експериментальних досліджень

У додатку К можна побачити зведені таблиці експериментів для кожного набору даних з аугментаціями та без.

Звернемося до узагальнених результатів дослідження (див. рис. 3.26).

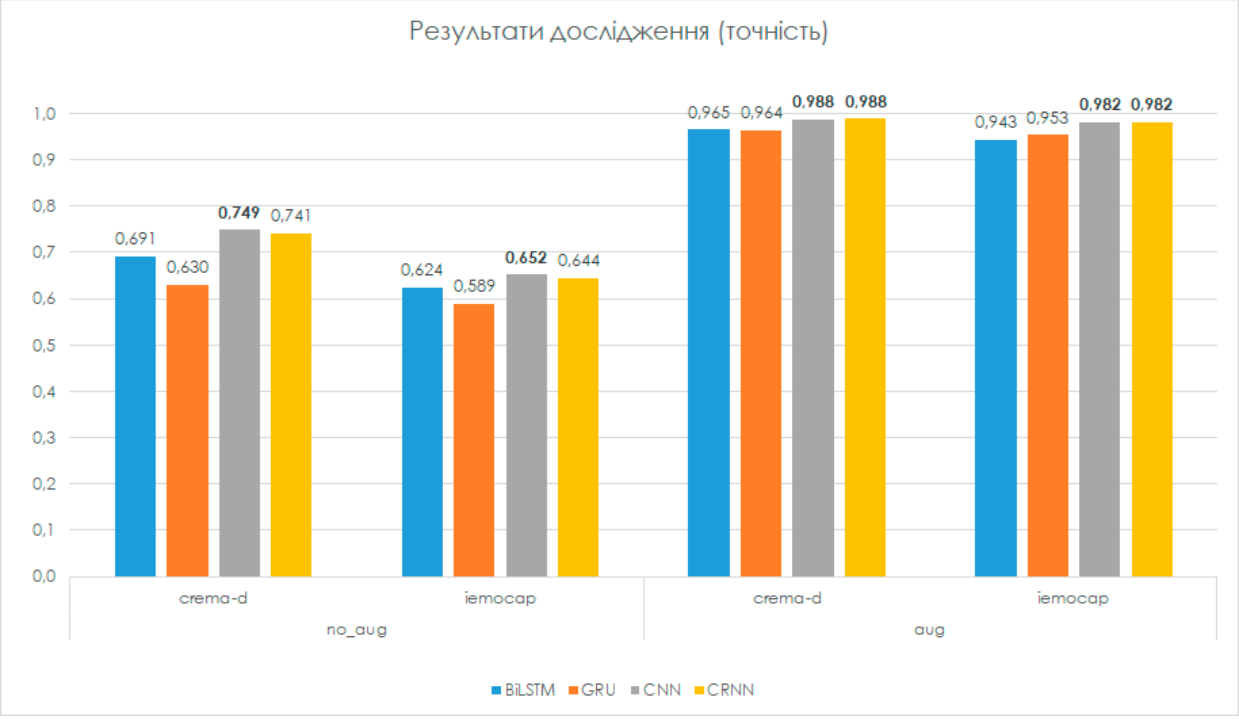

Рисунок 3.26 – Графік результатів дослідження (точність моделей) (створено самостійно)

Видно, що неоднорідність класів у наборі даних IEMOCAP дещо негативно впливає на якість моделей, особливо коли розмір такого набору даних відносно невеликий. Для всіх моделей на наборі даних IEMOCAP точність в середньому на 7% нижча.

Дивлячись на графіки навчання, можна також зробити висновок, що рекурентні мережі навчаються на таких даних набагато швидше, ніж згорткові, але графік функції втрат починає зростати після певної точки, що свідчить про перенавчання. Згорткові мережі також демонструють нестабільний тренд функції втрат, що свідчить про менший обсяг даних для вивчення закономірностей.

Тим не менш, згорткова модель показала кращі результати на обох наборах даних, і ця тенденція зберіглася в експериментах з розширеними даними.

На розширених наборах даних усі моделі показали приблизно однакову точність. Однак, як BiLSTM, так і GRU демонструють гірші результати порівняно з моделями згортки шарів. Крім того, ймовірний вплив неоднорідності класів нівелюється, і GRU показує навіть вищу точність, ніж BiLSTM (у випадку набору даних IEMOCAP).

Також модель CNN виявилася найбільш якісною в описаному підході та видає найвищу точність розпізнавання при усіх умовах експериментів. Хоча усі моделі, що були досліджені, мають високу точність і можуть бути використані у подальшій роботі.

Графік демонструє приріст точності чотирьох різних моделей (BiLSTM, GRU, CNN, CRNN) при додаванні аугментації даних у задачах розпізнавання емоцій на наборі даних CREMA-D (див. рис. 3.27). Приріст точності склав 23-31%. Найбільший приріст отримала модель із використанням GRU, де точність збільшилась на 31.1%. Це свідчить про те, що GRU модель особливо добре реагує на аугментацію даних, що може бути пов'язано з її здатністю ефективно захоплювати тимчасові залежності у аудіо сигналах.

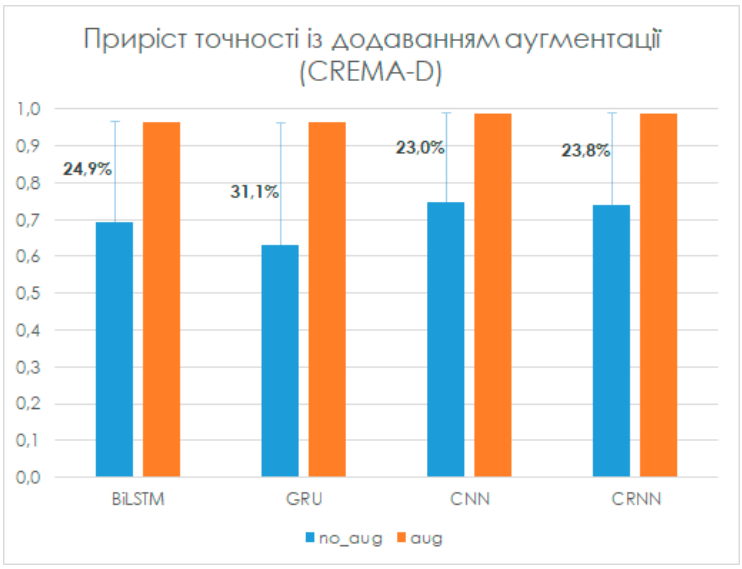

Рисунок 3.27 – Приріст точності від аугментації на наборі CREMA-D (створено самостійно)

Найбільш цікавим є приріст точності на наборі IEMOCAP (див. рис. 3.28). Нагадаю, що у цьому наборі досить нерівномірний розподіл класів серед семплів.

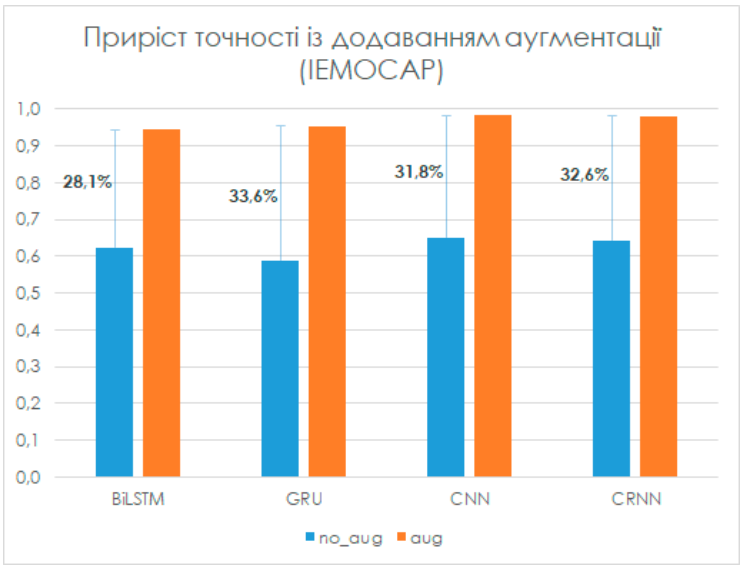

Рисунок 3.28 – Приріст точності від аугментації на наборі IEMOCAP (створено самостійно)

Отже, з рисунку 3.28 видно, що приріст вже склав 28-33%. Тобто можна зробити висновок, що так, звісно, нерівномірність класів не йде на користь якості моделі, однак аугментація дозволяє нівелювати цей негативний ефект.

Тож, незважаючи на відносно нижчі показники точності у моделі на нерівномірному наборі без аугментації, ця ж модель змогла вийти на рівень більш репрезентативного набору завдяки додаванню шуму, розтягуванні у часі та зміні висоти тону семплів.

Ці результати можуть бути використані в якості основи для подальших досліджень (описані у наступному підрозділі), а також для розробки програмних систем, для яких необхідно визначати емоційний стан людини за мовленням.

3.6 Рекомендації щодо подальших досліджень

В якості подальших досліджень можна запропонувати кілька напрямків, які можуть значно покращити результати розпізнавання емоцій за мовленням.

Інтеграція візуальних даних, таких як вирази обличчя, жести і мова тіла, може покращити точність розпізнавання емоцій. Поєднання аудіо- та відеоданих у мультимодальній моделі може забезпечити більш комплексне уявлення про емоційний стан людини. Це дозволить моделі вивчати кореляції між мовою, інтонацією та виразами обличчя, що зробить систему більш стійкою до шумів та інших перешкод.

Варто дослідити ефективність інших типів моделей глибокого навчання, таких як трансформери. Ці моделі вже показали високу ефективність в інших задачах обробки природної мови та можуть бути корисними для задач SER.

Додавання більшої кількості емоцій може зробити систему більш детальною і корисною для практичних застосувань. Дослідження додаткових аудіо характеристик може допомогти виявити нові індикатори емоційного стану.

Також подальші дослідження можуть включати інтеграцію із XAI (eXplainable AI) для інтерпретації результатів.

На основі отриманих моделей варто розробити повноцінну програмну систему, яка буде здатна працювати в реальному часі. Це може включати інтеграцію з додатками для відеоконференцій, системами дистанційного навчання, медичними діагностичними інструментами та іншими областями, де розпізнавання емоцій може бути корисним.

ВИСНОВКИ

В ході дослідження була проаналізована предметна галузь дослідження методів обробки аудіо запису за допомогою штучного інтелекту для виявлення емоційного стану людини. Були проаналізовані наукові роботи попередників, особливості їх досліджень за вказаною темою. В ході аналізу, були визначені найбільш популярні та ефективні методи для цієї задачі. Ці методи були ретельно розглянуті з описом їх особливостей для поточного дослідження.

Виходячи із результатів аналізу, була поставлена задача на кваліфікаційну роботу дослідити нейронні мережі, які демонструють кращі результати серед інших методів штучного інтелекту для задачі розпізнавання емоційного стану та виявити найбільш ефективну архітектуру мережі серед:

- BiLSTM;
- GRU;
- CNN;
- CRNN (CNN + LSTM).

Також було проведено аналіз наявних наборів даних у відкритих джерелах, що підходять для даного дослідження. Після порівняння низки датасетів, було визначено, що CREMA-D та IEMOCAP є одними з кращих наборів даних для задачі SER.Саме вони і були обрані для проведення експериментів дослідження.

Також були проаналізовані найпоширеніші характеристики аудіо, їх опис та вплив на сприйняття людиною. Було розглянуто методи для обробки аудіо записів, такі як аугментація, розбиття на фрейми, використанні динамічного розміру фрейму, використання матричного те векторного підходу обробку характеристик, що дозволяють отримати звукові характеристики, які далі можна використовувати як вхідні дані до нейронної мережі. Було виявлено, що найбільш ефективними характеристиками є MFCCs, які представляють часово-частотний аспект аудіо сигналу. Саме ці характеристики були використані, та продемонстрували дуже високі результати. Також були описані метрики для оцінки ефективності моделей, на основі яких було створено моделі нейронних мереж.

Після повного аналізу ту теоретичної підготовки до проведення експериментів, було розроблено програмні компоненти, необхідні для створення та навчання моделей, зняття метрик та інших показників при навчанні. Після цього було проведено низку експериментальних досліджень на двох наборах даних (CREMA-D та IEMOCAP) із аугментацією та без на чотирьох моделях нейромереж.

Результати експериментів показали наступне:

- серед досліджуваних моделей, найкращу точність демонструє модель із використанням згорткових шарів (CNN);
- набір даних CREMA-D показує в середньому дещо вищу точність у всіх проведених експериментах;
- аугментація даних відіграла значну роль, покращивши точність всіх моделей в середньому на 25-30%.

Результати цього дослідження показують перевагу описаного підходу над проаналізованими роботами попередників, а саме точність та f1-score дорівнюють 0,988. Ключові ролі у досягненні таких результатів також зіграло використання динамічного розміру фрейму при вилученні характеристик та матричного підходу до обробки характеристик нейронними мережами.

Чим більше даних і рівномірніше розподілені класи в наборі даних, тим ефективнішою може бути розроблена модель. Незважаючи на збільшення використання ресурсів, велика кількість вибірок є перевагою. Хоча аугментація покращує якість моделі, її ефективність на реальних даних може бути лише приблизною.

Описані методи можна вдосконалити і дослідити в інших умовах, розширивши дослідження, включивши візуальну інформацію в SER, провести експерименти з моделями з більш ніж чотирма класами у наборі даних, іншими характеристиками аудіо. Також подальші дослідження можуть включати інтеграцію із XAI (explainable AI) для інтерпретації результатів. Вже зараз отримані результати можуть бути використані при розробці програмного забезпечення у різних напрямках, зокрема для служб екстреної допомоги, лікарень та розумних будинків.

ПЕРЕЛІК ДЖЕРЕЛ ПОСИЛАННЯ

1. What is speech emotion recognition? – klu. Design, Deploy, and Optimize LLM Apps with Klu – Klu.ai. URL: https://klu.ai/glossary/speech-emotion-recognition (дата звернення: 28.05.2024).

2. Nazarenko D. S., Afanasieva I. V., Golian N. V. Neural network approach for emotional recognition in text. Bionics of intelligence. 2019. Т. 1, № 92. С. 9–13. URL: https://doi.org/10.30837/bi.2019.1(92).02 (дата звернення: 29.05.2024).

3. Investigation of the deep learning approaches to classify emotions in texts / D. Nazarenko та ін. CEUR workshop proceedings. 2021. Т. 2870. С. 206–224. (дата звернення: 17.03.2024).

4. D S., I A., K O. Research of audio recording processing methods using ai to detect emotional state. Zenodo. URL: https://zenodo.org/records/11049575 (дата звернення: 28.05.2024).

5. Emotional speech recognition using deep neural networks. PubMed Central (PMC). URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8877219/ (дата звернення: 30.12.2023).

6. Speech emotion recognition using deep learning. Blog - dataiku. URL: https://blog.dataiku.com/speech-emotion-recognition-deep-learning (дата звернення: 30.12.2023).

7. Speech emotion recognition with deep learning. ScienceDirect. URL: https://www.sciencedirect.com/science/article/pii/S1877050920318512 (дата звернення: 30.12.2023).

8. Speech emotion recognition using deep learning techniques: a review. Researchgate. URL: https://www.researchgate.net/publication/335360469_Speech_Emotion_Recognition_Using_Deep_Learning_Techniques_A_Review (дата звернення: 30.12.2023).

9. A deep learning approach for speech emotion recognition optimization using meta-learning. MDPI. URL: https://www.mdpi.com/2079-9292/12/23/4859 (дата звернення: 23.02.2024).

10. Srinidhi P. Speech based emotion recognition. International journal for research in applied science and engineering technology. 2022. Т. 10, № 6. С. 3160–3165. URL: https://doi.org/10.22214/ijraset.2022.44583 (дата звернення: 12.05.2024).

11. Introduction to convolutional neural networks (CNN). Analytics Vidhya. URL: https://www.analyticsvidhya.com/blog/2021/05/convolutional-neural-networks-cnn/ (дата звернення: 30.12.2023).

12. Basic CNN architecture: explaining 5 layers of convolutional neural network | upgrad blog. upGrad blog. URL: https://www.upgrad.com/blog/basic-cnn-architecture/ (дата звернення: 22.03.2024).

13. What are recurrent neural networks? | IBM. IBM in Deutschland, Österreich und der Schweiz | IBM. URL: https://www.ibm.com/topics/recurrent-neural-networks (дата звернення: 30.12.2023).

14. Dancker J. A brief introduction to recurrent neural networks. Towards Data Science. URL: https://towardsdatascience.com/a-brief-introduction-to-recurrent-neural-networks-638f64a61ff4 (дата звернення: 01.03.2024).

15. StatQuest with Josh Starmer. Long Short-Term Memory (LSTM), Clearly Explained, 2022. YouTube. URL: https://www.youtube.com/watch?v=YCzL96nL7j0 (дата звернення: 30.12.2023).

16. Shopynskyi M., Golian N., Afanasieva I. Long short-term memory model appliance for generating music compositions. 2020 IEEE international conference on problems of infocommunications. science and technology (PIC S&T), м. Kharkiv, Ukraine, 6–9 жовт. 2020 р. 2020. URL: https://doi.org/10.1109/picst51311.2020.9468088 (дата звернення: 29.05.2024).

17. Kostadinov S. Understanding GRU networks. Medium. URL: https://towardsdatascience.com/understanding-gru-networks-2ef37df6c9be (дата звернення: 30.12.2023).

18. Krish Naik. Bidirectional RNN indepth intuition - deep learning tutorial, 2020. YouTube. URL: https://www.youtube.com/watch?v=D-a6dwXzJ6s (дата звернення: 30.12.2023).

19. Ceshine L. Understanding Bidirectional RNN in PyTorch. Towards Data Science. URL: https://towardsdatascience.com/understanding-bidirectional-rnn-in-pytorch-5bd25a5dd66 (дата звернення: 14.03.2024).

20. Surrey Audio-Visual Expressed Emotion (SAVEE) Database. University of Surrey. URL: http://kahlan.eps.surrey.ac.uk/savee/ (дата звернення: 30.12.2023).

21. R L. S., A R. F. The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS). Zenodo. URL: https://zenodo.org/records/1188976 (дата звернення: 30.12.2023).

22. Toronto Emotional Speech Set (TESS) | TSpace Repository. URL: https://tspace.library.utoronto.ca/handle/1807/24487 (дата звернення: 30.12.2023).

23. GitHub - CheyneyComputerScience/CREMA-D: Crowd Sourced Emotional Multimodal Actors Dataset (CREMA-D). GitHub. URL: https://github.com/CheyneyComputerScience/CREMA-D (дата звернення:30.12.2023).

24. IEMOCAP: Interactive Emotional Dyadic Motion Capture Database / C. Busso та ін. Language resources and evaluation. 2008. Т. 42, № 4. С. 335–359. URL: https://doi.org/10.1007/s10579-008-9076-6 (дата звернення: 28.05.2024).

25. Bevor Sie zu YouTube weitergehen. URL: https://www.youtube.com/@ValerioVelardoTheSoundofAI (дата звернення: 30.12.2023).

26. Valerio Velardo - The Sound of AI. How to extract audio features, 2020. YouTube. URL: https://www.youtube.com/watch?v=8A-W1xk7qs8 (дата звернення: 13.03.2024).

27. Windowing signals – telecommunication engineering. Telecommunication Engineering – My WordPress Blog. URL: https://telecommunicationengineering.softecks.in/535/ (дата звернення: 03.04.2024).

28. How to train MFCC using machine learning algorithms. Tutorialspoint. URL: https://www.tutorialspoint.com/how-to-train-mfcc-using-machine-learning-algorithms (дата звернення: 30.12.2023).

29. librosa – librosa 0.10.1 documentation. Librosa. URL: https://librosa.org/doc/latest/index.html (дата звернення: 30.12.2023).