Національний університет ”Львівська політехніка”
Інститут комп’ютерних наук та інформаційних технологій
Кафедра систем штучного інтелекту
Спеціальність 122 «Комп’ютерні науки»

**ПОЯСНЮВАЛЬНА ЗАПИСКА**

до магістерської кваліфікаційної роботи на тему:

**«Розпізнавання мови жестів за допомогою алгоритмів машинного навчання»**

Студента(ки) групи КНСШ-22 Гринишин А.Б.

Керівник роботи _____ (__Думин І. Б.___)
*(підпис)*

Консультанти _____ (_____)
*(підпис)*

_____ (_____)
*(підпис)*

Завідувач
кафедри СШІ _____ (__Шаховська Н. Б.___)

Львів - 2023

Національний університет “Львівська політехніка”

Інститут комп’ютерних наук та інформаційних технологій

Кафедра систем штучного інтелекту

Спеціальність 122 «Комп’ютерні науки»

**ЗАТВЕРДЖУЮ:**

Завідувач кафедри СШІ _____

“____” _____2023р.

# ЗАВДАННЯ
# НА МАГІСТЕРСЬКУ КВАЛІФІКАЦІЙНУ РОБОТУ СТУДЕНТОВІ

Гринишин Анастасії Богданівній

1. Тема роботи: «Розпізнавання мови жестів за допомогою алгоритмів машинного навчання» затверджена наказом університету № 4420-4-06 від 13.10.2023

2. Термін подання закінченої роботи: _01.12.2023 _____

3. Вихідні дані до (проекту) роботи: координати ключових точок обличчя, тіла та рук.

4. Зміст розрахунково-пояснювальної записки: (перелік питань, що належить розробити): аналітичний огляд літературних та інших джерел, методи та засоби вирішення проблеми, практична реалізація.

5. Перелік графічного матеріалу (з точним зазначенням обов’язкових креслень): графіки архітектур мереж, діаграми розподілів та графіки точностей моделей.

6. Консультанти по роботі, із зазначенням розділів проекту, що стосується їх

| Розділ | Консультант | Підпис, дата | |
|---|---|---|---|
| | | Завдання видав | Завдання прийняв |
| Консультант з іноземної мови | | | |

7. Дата видачі завдання 10.03.2023

Керівник роботи _____
*(підпис)*

Завдання прийняв до виконання _____
*(підпис)*

## КАЛЕНДАРНИЙ ПЛАН

| № п/п | Назва етапів дипломної роботи | Термін виконання етапів роботи | Примітка |
|---|---|---|---|
| 1 | Огляд літератури | 10.03.2023 – 03.04.2023 | |
| 2 | Створення архітектури | 03.04.2023 – 15.05.2023 | |
| 3 | Розробка алгоритмічного та програмного забезпечення | 15.05.2023 – 11.09.2023 | |
| 4 | Експериментальні дослідження | 11.09.2023 – 06.11.2023 | |
| 5 | Оформлення записки | 06.11.2023 – 30.11.2023 | |

Керівник роботи _____
*(підпис)*

Студент _____
*(підпис)*

# АНОТАЦІЯ

Ця дипломна робота приділяє увагу розпізнаванню американської мови жестів як актуальній та перспективній галузі, що може знайти широке застосування у спілкуванні, технологіях віртуальної реальності та підтримці людей з обмеженими можливостями. Робота описує важливість цього напрямку, аналізуючи методи та проблеми, пов'язані з розпізнаванням мови жестів.

Перший розділ присвячений критичному аналізу літератури, під час якого були ідентифіковані ключові проблеми цієї галузі та відібрано 13 релевантних джерел за допомогою методу PRISMA. Визначено потребу у подальших дослідженнях для вирішення складнощів, таких як різноманітність жестів у різних культурах та умовах.

У другому розділі роботи детально розглядаються різні методи розпізнавання мови жестів, зосереджуючись на нейронних мережах, таких як LSTM, ConvLSTM та трансформери. Аналізуються їхні переваги та використання в контексті розпізнавання жестів, включаючи важливість виявлення закономірностей між кадрами та використання функцій втрат для покращення точності системи.

У наступному розділі оцінюються результати тестування чотирьох архітектур, зокрема трансформерів, де виявлено, що модель на основі трансформера перевершує у точності попередні архітектури, хоча використання деяких функцій втрат може впливати на її ефективність. Проаналізовано потребу у подальших експериментах з регуляризації для зменшення перенавчання та покращення результатів.

Також було застосовано ансамбль моделей на основі найкращої архітектури, що показало певний ріст точності при збільшенні кількості моделей. Узагальнюючи, робота демонструє високий потенціал трансформерної архітектури в розпізнаванні жестів. У використанні різних функцій втрат, виявлені певні перспективи та важливість подальших експериментів з регуляризації для зменшення перенавчання та покращення точності. Результати свідчать про значний прогрес у цій галузі та визначають напрямки подальших

досліджень для досягнення більш точних та надійних систем розпізнавання мови жестів.

**Ключові слова:** американська мова жестів, класифікація, трансформер, ансамбль моделей, машинне навчання.

# ABSTRACT

This thesis focuses on the recognition of American Sign Language as a relevant and promising field that can find wide application in communication, virtual reality technologies and support of people with disabilities. The work describes the importance of this direction, analyzing methods and problems related to sign language recognition.

The first chapter is devoted to a critical analysis of the literature, during which the key issues of this field were identified and 13 relevant sources were selected using the PRISMA method. The need for further research to address complexities such as the diversity of gestures across cultures and settings is identified.

The second chapter of the paper deals in detail with different methods of sign language recognition, focusing on neural networks such as LSTM, ConvLSTM and transformers. Their advantages and use in the context of gesture recognition are analyzed, including the importance of detecting patterns between frames and using loss functions to improve system accuracy.

The next section evaluates the test results of four architectures, particularly transformers, where the transformer-based model is found to outperform previous architectures in accuracy, although the use of some loss functions may affect its performance. The need for further regularization experiments to reduce overtraining and improve results is analyzed.

An ensemble of models based on the best architecture was also applied, which showed some increase in accuracy as the number of models increased. In summary, the work demonstrates the high potential of transformative architecture in gesture recognition. In the use of different loss functions, some perspectives are revealed and the importance of further regularization experiments to reduce overtraining and improve accuracy. The results indicate significant progress in this field and identify directions for further research to achieve more accurate and reliable sign language recognition systems.



**Keywords:** American Sign Language, classification, transformer, model ensemble, machine learning.

# ЗМІСТ

# ПЕРЕЛІК УМОВНИХ ПОЗНАЧЕНЬ

- ASL– американська мова жестів;
- ASLLVD – відео набір даних лексики американської мови жестів;
- BERT – представлення двонапрямлених кодерів на основі трансформерів;
- BSL – британська мова жестів;
- CNN – згорткова нейронна мережа;
- DETR – детектор та описувач з використанням трансформерів;
- DNN – повнозв’язна нейронна мережа;
- FFN – позиційно-залежні мережі прямого зв'язку;
- GPT – генеративна попередньо навчена трансформерна модель;
- GRU – вентильні рекурентні вузли;
- I3D – тривимірна інтерактивна модель;
- KNN – к-найближчі сусіди;
- LSTM – довга короткочасна пам’ять;
- MLP - мережа прямого поширення;
- PRISMA – програма систематичного огляду та метааналізу;
- ReLU – функція лінійної ректифікації;
- RNN – рекурентна нейронна мережа;
- SVM – метод опорних векторів;
- TGCN – темпоральна графова згорткова мережа;
- VGG – глибокі візуальні згорткові мережі;
- WLASL – американська мова жестів на рівні слів.

# ВСТУП

У світі, де слова часто стають ключем до налагодження взаєморозуміння між людьми, існує велика група людей, для яких це завдання стає викликом. Люди зі слуховими вадами не можуть користуватися голосом як засобом комунікації, ось чому використовують альтернативний спосіб – мову жестів. Мова жестів дозволяє людям зі слуховими вадами виразити свої думки та ідеї, користуючись жестами та мімікою. Однак, не всі люди знають цю мову, незважаючи на фізичну можливість вживати жести, вони не володіють нею і зустрічають проблеми у спілкуванні з групою людей з вадами слуху.

З цією проблемою зіткнулася значна частина суспільства. За даними Всесвітньої організації охорони здоров'я, близько 466 мільйонів людей у світі мають слухову ваду [1]. Це ставить їх у нерівні умови у спілкуванні з оточенням, особливо у тих сферах, де мова є основним засобом комунікації, таких як навчання, робота, соціальні відносини тощо.

Однак, з розвитком технологій та штучного інтелекту, почали з'являтися рішення, що допомагають робити комунікацію зручнішою для людей зі слуховими вадами. Один з таких способів - це розпізнавання мови жестів. Це технологія, яка дозволяє перетворювати мову жестів в текст або мовлення, що може розв’язати проблему, допомагаючи людям з вадами слуху і тим, хто не володіє мовою жестів, зрозуміти один одного [2].

Ось чому розробка системи розпізнавання мови жестів може стати справжнім проривом у комунікації між людьми з різними здатностями до сприйняття мови. Ця система дозволить людям зі слуховими вадами використовувати мову жестів для передачі повідомлень, не залежно від того, знає їхній співрозмовник цю мову чи ні. Крім того, це може допомогти зменшити соціальну ізольованість та забезпечити можливість повноцінного включення цих людей до суспільства.

Отже, дослідження та розробка нових технологій розпізнавання мови жестів стають все більш актуальними, оскільки дозволяють розв’язати проблеми, пов'язані з комунікацією між людьми з вадами слуху та іншою частиною

населення. Застосування цих технологій може полегшити життя багатьом людям, зокрема забезпечити їм можливість ефективно спілкуватися з оточенням, зробивши комунікацію більш зрозумілою та простою.

**Мета роботи:** полягає в дослідженні та порівнянні різних наявних методів розпізнавання мови жестів, а також їх подальше удосконалення, з метою покращення ефективності та точності розпізнавання. Для досягнення цієї мети планується провести аналіз методів, виявити їх переваги та недоліки, а потім спробувати їх оптимізувати та покращити. Остаточним результатом буде збільшення точності розпізнавання мови жестів, що дозволить полегшити комунікацію між людьми з різними здатностями та внести позитивний внесок у соціальну інклюзію.

Для досягнення мети необхідно вирішити ряд задач, а саме:

1. Дослідження наявних методів розпізнавання мови жестів та їхньої ефективності в різних умовах освітлення та шуму.
2. Дослідження та збір даних, що містить різні види жестів для тренування нейронних мереж.
3. Удосконалення досліджених методів розпізнавання мови жестів.
4. Оцінення точність та ефективність розробленої моделі порівняно з наявними методами розпізнавання мови жестів та дослідити можливі шляхи покращення результатів.
5. Розроблення системи розпізнавання мови жестів для різноманітних пристроїв та платформ (наприклад, мобільних телефонів, планшетів, роботів тощо) та дослідження її ефективності та доступності для різних категорій користувачів.

**Об'єкт дослідження є** процеси розпізнавання мови жестів.

**Предмет дослідження є** методи та технології розпізнавання мови жестів на основі комп'ютерного зору та машинного навчання.

**Методами дослідження є** використання глибоких моделей, таких як глибокі нейронні мережі(DNN), Long Short-Term Memory(LSTM) та Transformer. Вибір моделей базуються на їх можливості враховувати часові зв’язки між кадрами.

**Новизна дослідження є** розробка та оптимізація алгоритму розпізнавання мови жестів, спрямованого на точну ідентифікацію різних жестів та їх перетворення в текст.

**Практична цінність** полягає у використанні систем розпізнавання мови жестів для полегшення комунікації між людьми зі слуховими вадами та без них.

**Особистий внесок магістранта це** дослідження та розробка архітектури глибокої нейронної мережі. Ця архітектура використовує комбіновані стратегії вдалих методів глибокого навчання, що спрямовані на покращення точності та ефективності розпізнавання жестів.

**Апробація результатів роботи.** На основі проведених наукових досліджень було опубліковано наукову працю – наукову статтю по темі магістерської роботи на конференції [3].

# 1. АНАЛІТИЧНИЙ РОЗДІЛ

## 1.1. Особливості мови жестів.

Мова жестів є надзвичайно важливим способом спілкування для людей з вадами слуху, а також для всієї спільноти, що взаємодіє з ними. Вона охоплює величезний спектр компонентів, включаючи жести, міміку обличчя та рухи тіла, які використовуються для передачі значень та інформації. Ця мова має власну, чітку граматичну структуру і словниковий запас, що робить її комплексною та автономною мовною системою, здатною ефективно виражати навіть найскладніші ідеї та концепції.

Важливо відзначити, що мова жестів відіграє важливу роль у поліпшенні спілкування між особами з аудітивними вадами та людьми, які мають нормальний слух. Вона дозволяє зробити комунікацію більш доступною та ефективною, сприяючи включенню та розумінню в різних соціальних ситуаціях. Ця мова постійно розвивається, адаптуючись до потреб сучасного світу, і отримує все більше визнання та підтримки як в науковому, так і в практичному плані [4, 5].

У повсякденному спілкуванні, включаючи взаємодію з родиною та друзями, люди з вадами слуху віддають перевагу мові жестів. Ця мова стала надзвичайно важливою для їхнього спілкування та виразності, і дослідження в цій області підтвердили, що жестові мови належать до справжніх мов і мають унікальні граматичні структури, словниковий запас та мовні властивості. Цей факт став ключовим для підвищення обізнаності щодо важливості мов жестів і сприяв їхньому подальшому розповсюдженню і розвитку.

Наукові дослідження, проведені такими вченими, як Сноддон [6] і Мааледж [7], детально висвітлили лінгвістичну складність мов жестів та важливу роль, яку вони відіграють у процесі забезпечення спілкування та розвитку спільноти серед глухих та людей із вадами слуху. Ці дослідження вказують на те, що жестові мови є самостійними мовними системами зі своєю власною граматикою та синтаксисом, і вони мають здатність передавати складні концепції та ідеї. Однак, варто зауважити, що існують певні відмінності між

жестовими мовами та вербальними мовами [8].

### *1.1.1 Лінгвістика жестів.*

Мови жестів — це візуальні методи спілкування, які використовують жести рук, міміку обличчя та мову тіла для передачі сенсу. Вони володіють унікальною здатністю передавати не лише мовну інформацію, але й емоційну та психологічну специфіку, роблячи їх надзвичайно різноманітним і виразним засобом комунікації.

Один із найцікавіших аспектів мов жестів полягає у їхній культурній та регіональній специфічності. Наукові дослідження, проведені видатними лінгвістами, такими як Стоко [9] і Христопулос [10], показали, що словниковий запас та синтаксис мов жестів можуть істотно відрізнятися в різних країнах і, навіть, в різних регіонах в межах однієї країни. Це свідчить про те, як мови жестів динамічно адаптуються до культурних та соціальних особливостей різних груп людей.

Для майстерного використання мови жестів необхідно враховувати кілька ключових факторів. Один із них - розпізнавання жестів, рухів, міміки та просторових варіацій, які мають велике лінгвістичне значення. Дослідження, проведені Маледжем [7] та Емморі [11], наголошують на складності перекладу жестової мови і вимагають глибокого розуміння не лише мови жестів, але й її унікальних особливостей. Вони підкреслюють важливу роль визнання жестових мов як повноцінних людських мов, які мають свою лінгвістичну цінність та значущість у полегшенні спілкування для глухих та осіб із вадами слуху.

Загалом, мови жестів відіграють надзвичайно важливу роль у розвитку комунікації та взаєморозуміння в різних соціокультурних контекстах. Їхній вплив на спілкування між різними групами людей є надзвичайно важливим, а визнання їхньої статусної цінності сприяє подальшому розвитку та розповсюдженню мов жестів в сучасному світі.

Мова жестів, всупереч своєму візуальному вираженню, є надзвичайно складною та систематичною мовною системою. Її важко розглядати просто як пантоміму, оскільки вона складається з переважно довільних знаків, які не мають

необхідного візуального відношення до того, що вони виражають. Це подібно до того, як у розмовних мовах більшість слів не є ономатопеями, а вони мають свої унікальні лексичні значення.

Важливо підкреслити, що мова жестів не обмежується візуальним відтворенням усної мови. Вона має власну складну граматику та лінгвістичну структуру, яка дозволяє їй передавати різноманітну інформацію, незалежно від її складності. Ця мовна система може бути використана для обговорення будь-якої теми, від простих і конкретних до філософських і абстрактних концепцій. Таким чином, мова жестів виявляється потужним засобом спілкування, що дозволяє висловлювати складні думки та відчуття, а також збагачує комунікацію між особами.

За змістом та структурою, мова жестів може суттєво відрізнятися від усних мов. Наприклад, щодо синтаксису, жестова мова, така як американська жестова мова (ASL), може бути більше схожою на розмовну японську мову, ніж на англійську. Це свідчить про те, що мови жестів мають свої унікальні структури та властивості, які варто досліджувати та вивчати як окрему лінгвістичну реалізацію мовного спілкування [12].

Мови жестів, подібно до усних мов, мають свою власну структуру та організацію мовних одиниць, які відповідають фонемам в усному мовленні. Колись ці мовні одиниці називали "черемами" у випадку мов жестів, і вони відіграють ключову роль у конструюванні смислових виразів. Для розуміння структури жестової мови важливо розглянути складові елементи кожного жесту, які можна узагальнити під абревіатурою HOLME.

1. Форма руки (Hand shape): Форма долоні або позиція руки може мати суттєвий вплив на те, як сприймається жест та його значення. Наприклад, відкрита долоня може вказувати на прийняття, а закрита - на відхилення чогось.
2. Орієнтація долоні (Orientation): Орієнтація долоні грає важливу роль у розумінні жесту. Вона може вказувати напрямок, об'єкт або суб'єкт жесту.
3. Місце артикуляції (Location): Це вказує на точку, де жест виражається,

наприклад, перед грудьми, на обличчі або в просторі поруч з тілом.. Місце артикуляції може бути важливим, оскільки це кардинально може змінити сенс продемонстрованого жесту.

4. Рух (Movement): Рух жесту визначає його динаміку і специфіку. Рух може бути прямолінійним, круговим, змінювати швидкість та напрямок. Це додає додатковий сенс та контекст до жесту.
5. Вираз обличчя (Facial Expression): Вираз обличчя та немовні маркери включають емоції та інші несловесні сигнали, які можуть змінювати інтерпретацію жесту. Вони надають жестовому живопису емоційну та контекстуальну глибину.

Однак важливо відзначити, що мови жестів не є алфавітом, де кожна літера має свій фіксований звуковий еквівалент. Натомість жести скоріше представляють слова або інші семантично значущі концепти. Вони вимагають комплексного сприйняття та інтерпретації, іноді залежать від контексту та мовного оточення, і можуть бути різними в різних культурах і мовних спільнотах.

Один із цікавих аспектів мови жестів полягає в тому, що її використання та інтерпретація може суттєво відрізнятися в різних країнах і регіонах світу. Різні спільноти розробляють та використовують власні жести та знаки для вираження схожих або ідентичних понять, що створює додатковий рівень складності у вивченні та розумінні мови жестів. Ця культурна різноманітність робить мову жестів цікавим об'єктом дослідження і вивчення.

### *1.1.2 Різні види мови жестів.*

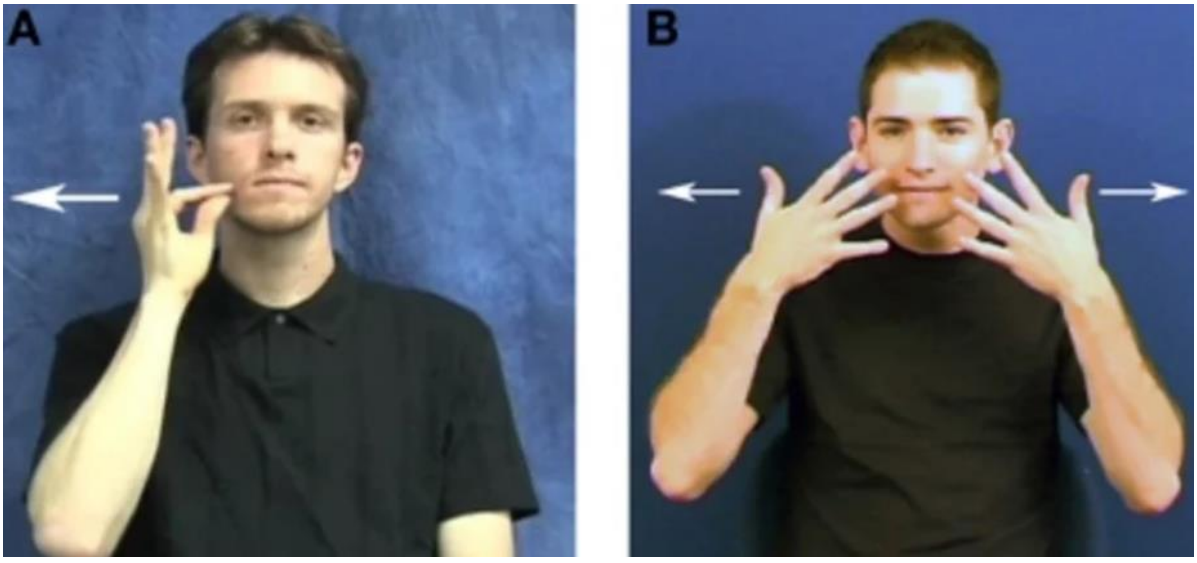
Сучасний світ вражає своєю мінливістю та різноманітністю, і ці зміни не оминають жестових мов. Зараз в усьому світі існує аж до 300 різних типів жестової мови, кожен з яких має свої власні особливості та контекст використання. Деякі з цих мов використовуються лише на локальному рівні, призначені для комунікації в маленьких спільнотах чи сільських районах, тим самим відображаючи багатоманіття мовних культур по всьому світу.

З іншого боку, існують жестові мови, якими користуються мільйони

людей. Ці мови стали інструментом спілкування для значної кількості людей і грають важливу роль в їхньому повсякденному житті. Вони дозволяють виражати думки, емоції та ідеї, а також сприяють соціальній включеності та обміну інформацією.

Однією з важливих рис більшості жестових мов є те, що вони не ставлять своєю метою безпосередньо перекладати слова із мови на знаки, які створюються руками. Кожна мова жестів - це самостійна мовна система з власним словниковим запасом, граматикою та синтаксисом. Здебільшого вони навіть не пов'язані з усними мовами, якими розмовляють у тому ж регіоні.

Наприклад, можливо багато людей мають уявлення, що оскільки британці та американці розмовляють однією мовою, мова жестів також має бути однаковою. Але це помилкова думка. Навіть між двома англомовними країнами, які здається б поділяють спільну мову, існують відмінності у жестовій мові. Наприклад, американська мова жестів (ASL) відрізняється від британської мови жестів (BSL) у використанні різних знаків, хоча деякі слова та фрази можуть бути схожими [13]. ASL використовує одну руку, тоді як BSL вимагає використання обох рук. На рис. 1.1 представлено приклад жесту «кіт» американською мовою жестів (ліворуч) та британською мовою жестів (праворуч).




***(а)*** ***(б)***

***Рис. 1.1. Приклад жесту «кіт»: (а) – американською мовою жестів; (б) – британською мовою жестів***

Також важливо відзначити, що ті самі жестові знаки можуть мати різні значення в ASL і BSL. Тому під час спілкування з людьми, які мають ваду слуху, важливо завжди запитувати, якою мовою жестів вони користуються. Це сприяє уникненню непорозумінь та плутанини в комунікації.

Відмінність більшості мов жестів обумовлене походженням цих мов. Наприклад, британська жестова мова має довгу історію, яка налічує понад три століття і відома своїм ручним алфавітом ще з 1720 року. У свою чергу американська мова жестів походить від французької мови жестів, яка була розроблена у 1700-х роках і послужила основою також багатьох європейських мов жестів. Однак деякі жестові мови, такі як китайська, японська, індо-пакистанська та левантійська арабська, мають своє власне унікальне походження, абсолютно не пов'язане з французьким чи британським корінням. Ця різноманітність свідчить про багатство та культурну різноманітність мов жестів, які сприяють підтримці та збереженню мовної різноманітності у світі [14].

### *1.1.3 Використання жестів серед людей, які чують.*

У сучасному світі мови жестів знаходяться в центрі практики спілкування осіб з вадами слуху, і вони відіграють ключову роль у забезпеченні ефективної комунікації для цієї групи людей. Проте, існують значущі складнощі, з якими стикаються глухонімі особи під час використання мови жестів у повсякденному житті.

Однією з найважливіших складнощів є недостатність розуміння мови жестів серед не глухих осіб. Багато людей, які чують, не мають відповідної підготовки чи навичок для ефективного розуміння жестової комунікації. Це може призводити до невдалих спроб спілкування, непорозумінь та відчуття відчуженості серед осіб з вадами слуху.

Додатковою проблемою є обмеженість жестових мов у вираженні абстрактних або технічних концепцій. У деяких випадках, особливо в наукових чи технічних областях, може бути важко передати складну інформацію через жести. Це може ускладнювати навчання та професійне спілкування для

глухонімих осіб.

У повсякденному житті особи з вадами слуху можуть зіткнутися зі складнощами у взаємодії з оточенням на роботі, у навчальних закладах та соціальних ситуаціях. Брак розуміння жестових мов або недостатня підготовка осіб, які чують, можуть ускладнювати процес спілкування та призводити до ізоляції глухонімих осіб від суспільства.

Технологічний розвиток відіграє важливу роль у розв’язуванні проблем, пов'язаних з комунікацією осіб з вадами слуху, що використовують мову жестів. Онлайн-ресурси та додатки для мобільних пристроїв стали значущими інструментами для навчання та практики мови жестів.

Однією з важливих ініціатив є створення відеоуроків та онлайн-платформ для навчання мови жестів. Ці ресурси дозволяють людям набувати навичок жестової комунікації, вивчати мовний словник та покращувати свої здібності в будь-якому місці та в будь-який час.

Додатки для мобільних пристроїв також відіграють важливу роль у підтримці осіб з вадами слуху у їхньому повсякденному житті. Наприклад, існують додатки, які допомагають перекладати усну мову на мову жестів або навпаки. Це дозволяє глухонімим особам спілкуватися з неглухими в більш ефективний спосіб та знижує бар'єри в спілкуванні [15].

Помітною є інтеграція мови жестів у віртуальні та розширену реальність. Деякі додатки та пристрої можуть аналізувати жести та перетворювати їх на текст або мовлення, що робить комунікацію більш динамічною та доступною.

Суспільство також стає більш освіченим щодо проблем осіб з вадами слуху і важливості використання мови жестів. Інформаційні кампанії та освітні заходи сприяють зростанню розуміння та підтримки в цій сфері, що допомагає зменшити бар'єри у спілкуванні та інтеграції глухонімих осіб у суспільство.

Таким чином, технологічний розвиток вносить значний внесок у подолання складнощів у комунікації осіб з вадами слуху, сприяючи доступності та ефективності мови жестів в сучасному світі.

### 1.2. Сучасний стан досліджень предметної області.

Розпізнавання мови жестів набуває все більшої популярності з кожним роком. Ця тенденція обумовлена висхідною кількістю публікацій та досліджень в цій області, і цей ріст свідчить про великий інтерес науковців та інженерів до розробки та вдосконалення систем розпізнавання мови жестів.

Аналіз графіка публікацій за останні 10 років, представленого на рисунку 1.2, вказує на стабільний ріст кількості опублікованих статей в області розпізнавання мови жестів. Однак, варто відзначити, що в 2020 році спостерігається незначний спад кількості статей. Цей спад може бути пояснений різкими змінами в робочих умовах та призупиненням багатьох досліджень у зв'язку з початком пандемії COVID-19. Таким чином, карантинні обмеження та зміни у робочому режимі можуть вплинути на кількість публікацій.

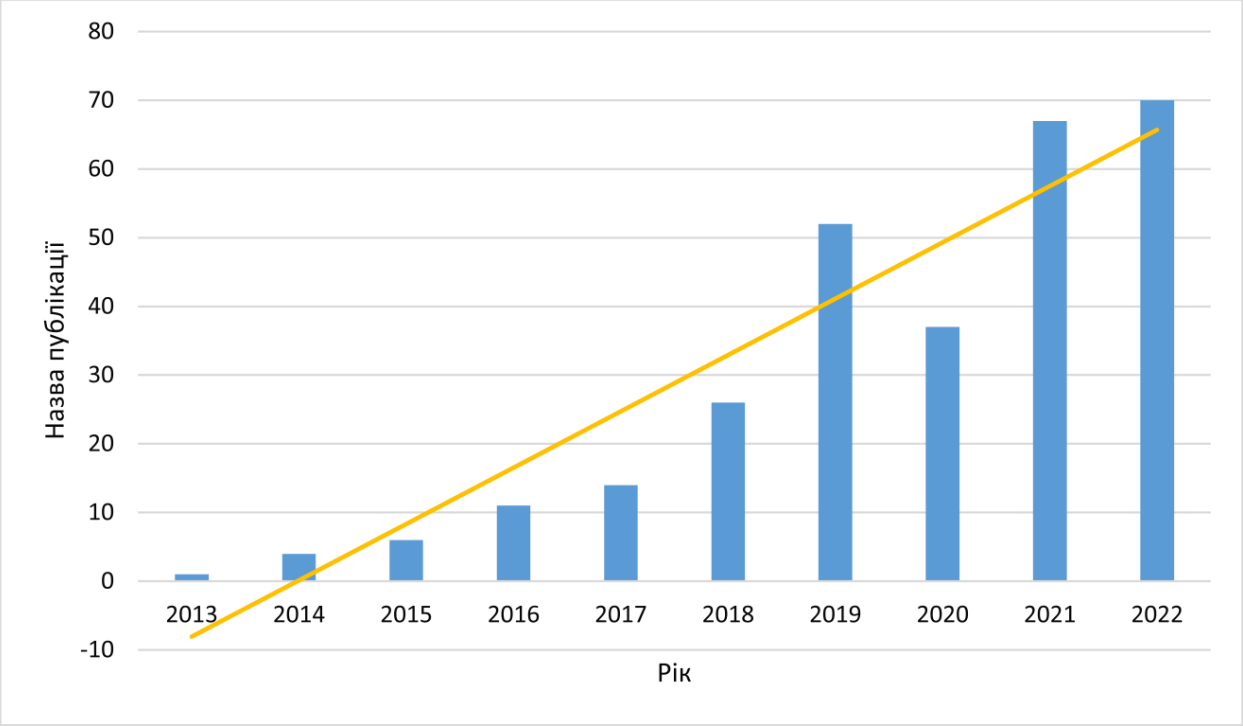



***Рис. 1.2. Діаграма розподілення досліджень за роками видання***

Однак навіть з цим незначним спадом у 2020 році, загальна тенденція залишається зростаючою. Це свідчить про стійкий інтерес до розпізнавання мови жестів як актуальної теми досліджень.

Причини цієї тенденції можуть бути різними. Однією з них є розвиток нових технологій, які дозволяють точніше аналізувати та розпізнавати мову

жестів. Крім того, зростає інтерес до робототехніки та інтерфейсів, що використовують жести для взаємодії з машинами, що стимулює дослідження в цій області. Наприклад, такі додатки, як віртуальна реальність та розширена реальність, потребують ефективних методів розпізнавання жестів для зручної інтеракції користувачів з комп'ютерами та іншими пристроями.

### *1.2.1. Підходи до систем перекладу мови жестів.*

Сучасний стан класифікації мови жестів відзначається значними досягненнями, завдяки поєднанню різних технологій і методів. Однак, після огляду публікацій за останні 5 років, основна увага приділяється двом основним методам класифікації: за допомогою датчиків і візуальним (комп'ютерний зір).

Почнемо з методу класифікації з використанням датчиків. Методи розпізнавання на основі сенсорів, також відомі як методи використання рукавичок з датчиками, збирають важливу інформацію, таку як ступінь згинання пальців, положення зап’ястя та рух руки, за допомогою вбудованих датчиків. Потім ця інформація передається на мобільний пристрій або додаток для обробки. Однак у цих систем є обмеження, такі як труднощі із записом рухів рук і пальців і нездатність зафіксувати вирази обличчя, рухи губ і рухи очей [16].

Розпізнавання мови жестів за допомогою підходу даних рукавичок передбачає використання електронних рукавичок, оснащених датчиками, які збирають і передають дані. Цю рукавичку носить підписувач, забезпечуючи канал для введення даних для передачі даних на мобільний пристрій або програму. Цей підхід популярний серед систем перекладу мовою жестів через легкість збору інформації, налаштування та перекладу даних, що потребує меншої обчислювальної потужності. Проте рукавички для передачі даних можуть бути дорогими, вартість деяких моделей перевищує 9000 доларів США. Хоча існують менш дорогі альтернативи, вони часто мають менше датчиків і більш схильні до перешкод, що призводить до втрати важливої інформації та зниження точності перекладу. Крім того, оскільки руки бувають різних розмірів, рукавичка для даних може не підходити для всіх користувачів [17].

У свою чергу, візуальний підхід до розпізнавання мови жестів ґрунтується

на використанні камер для захоплення зображень і відеоматеріалів, які послідовно обробляються для подальшого перекладу. Однією з ключових переваг цього підходу є його універсальність, оскільки він здатен аналізувати вираз обличчя, рухи голови, а також читати рухи губ, доповнюючи аналіз рухів рук.

Проте, не дивлячись на численні переваги візуального підходу до розпізнавання мови жестів, варто відзначити деякі його недоліки. Один з основних недоліків полягає у великій кількості даних, необхідних для навчання нейронних мереж та моделей комп'ютерного зору. Збір та анотування достатньої кількості відеоматеріалів може бути витратним та часоємним процесом. Окрім того, обробка великих обсягів відеоданих вимагає значних обчислювальних ресурсів та великого обсягу зберігання, що може бути фінансово накладним завданням.

Несприятливі умови освітлення або якість камери також можуть вплинути на точність системи розпізнавання мови жестів, зокрема в ситуаціях з обмеженим доступом до світла чи при низькій роздільній здатності відео. З цими викликами пов'язані також технічні труднощі у побудові надійних систем.

Хоча візуальний метод розпізнавання мови жестів може мати свої недоліки, він має потенціал для значного поліпшення за допомогою додаткового дослідження та розробки. Однією з переваг цього методу є можливість покращення часу та легкості навчання системи. Це означає, що з розвитком технологій та підвищенням точності алгоритмів розпізнавання, користувачі зможуть швидше та ефективніше взаємодіяти з системами, що використовують візуальний метод.

Своєю чергою, метод на основі датчиків вимагає додаткового обладнання, що ускладнює його доступність та інтеграцію в реальному середовищі. Це може бути фактором, який обмежує його прийняття серед користувачів і розробників. Крім того, вартість придбання та підтримки такого обладнання може бути високою, що створює фінансові бар'єри для впровадження цього методу.

З огляду на те, що тематика розпізнавання мови жестів набуває популярності та актуальності з кожним роком, це стимулює подальший розвиток

та розширення баз даних для навчання систем. Цей ріст інтересу відкриває нові можливості для покращення та оптимізації візуальних підходів до розпізнавання мови жестів. Тому для подальшого дослідження і розгляду будуть враховані саме варіанти, які базуються на візуальному методі.

### *1.2.2. Огляд літературних джерел.*

Для відбору та подальшого аналізу наукових публікацій в цьому дослідженні був використаний метод PRISMA, який є міжнародним стандартом для проведення систематичних оглядів та метааналізів [18]. Цей метод дозволяє систематично оцінювати та відбирати наукові джерела для включення в аналіз з врахуванням жорстких критеріїв.

Для проведення пошуку відповідної літератури були використані дві визначені онлайн-бази даних: Scopus [19] та Google Scholar [20]. Scopus - це один з найбільших та найвпливовіших мультисервісних баз даних наукових публікацій та цитувань, що належить компанії Elsevier. Він охоплює широкий спектр дисциплін, включаючи науки про природу, техніку, медицину, соціальні науки та гуманітарні науки. Google Scholar - це безкоштовний онлайн-пошуковий сервіс, створений Google для пошуку наукових статей, тез, книг, рецензій, дисертацій та іншої наукової літератури. Він індексує матеріали різних наукових областей та джерел, включаючи журнали, університетські репозиторії, академічні сайти та інші джерела наукової інформації. Обидва ці інструменти дозволяють вченим, дослідникам та студентам знаходити та отримувати доступ до наукової інформації для своїх досліджень та навчання. Однак, вони відрізняються за підходом до індексації, доступною інформацією та функціоналом, який вони пропонують.

Блок-схема, яка представлена на рисунку 1.3, ілюструє повний потік інформації через всі етапи систематичного огляду. Вона відображає кількість виявлених наукових записів, процес включення та виключення цих записів на основі певних критеріїв, а також раціональні причини для виключень. Ця блок-схема PRISMA є важливим інструментом для забезпечення об'єктивності та точності вибору наукових джерел у дослідженні.

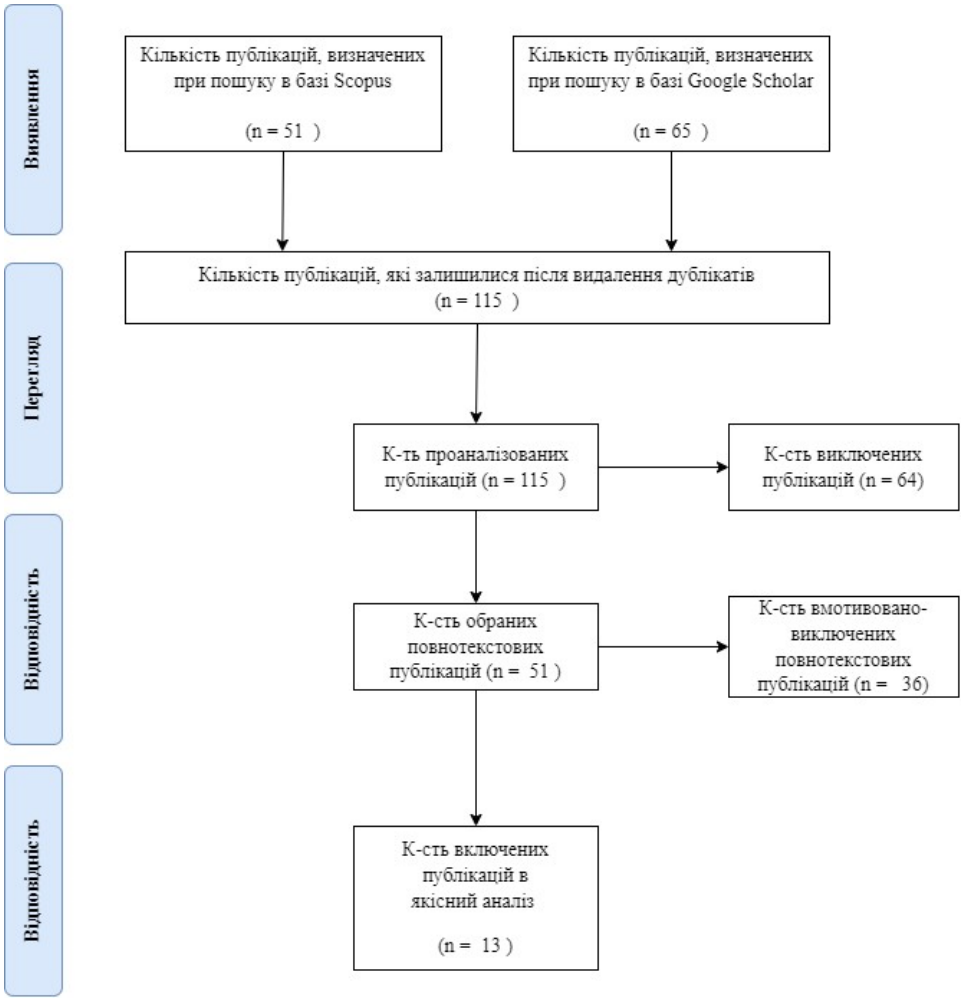



***Рис. 1.3. Схема PRISMA***

Після завершеного огляду та ретельного відбору літературних джерел було відібрано 13 статей, кожна з яких відповідає заданим критеріям і вважається релевантною для дипломної роботи, яка присвячена темі "Розпізнавання мови жестів за допомогою алгоритмів машинного навчання". Цей відбір статей був проведений з урахуванням їхнього внеску у розуміння та дослідження даної теми, а також враховуючи їхню актуальність та наукову цінність.

У таблиці 1.1 один представлена основна інформація статей, які були обрані.

**Таблиця 1.1.** Загальний опис статей

| Стаття | Класи | Дані | Метод |
|---|---|---|---|
| [21] | 2000 класів | 21,083 відео | 2DCNN + GRU, 3DCNN |

| | | | |
|---|---|---|---|
| [22] | 26 класів | 7304 відео | LSTM + Сіамська мережа |
| [23] | 28 класів | 61614 зображень | CNN |
| [24] | 100 класів | ~4200 відео | 3DCNN |
| [25] | 24 класів | 34627 зображень | CNN |
| [26] | 24 класів | 7500 зображень | CNN, SVN |
| [27] | 29 класів | 87000 зображень | CNN |
| [28] | 3300 класів | 98000 відео | Deep Forest, SVM |
| [29] | 26 класів | – | BiGRU |
| [30] | 29 класів | 87000 зображень | Random Forest, SVM, Logistic Regression, MLP, KNN, Naive Bayes |
| [31] | 9 класів | 8600 фреймів | DETR |
| [32] | 26 класів | 46 800 зображень | CNN, Hidden Markow classification, SVM |
| [33] | 25 класів | 34627 зображень | CNN |

Перейдемо до критичного опису статей. У першій роботі [21] розглядається проблема розпізнавання мови жестів за допомогою різних моделей, таких як VGG-GRU, Pose-GRU, Pose-TGCN і I3D. Стаття відзначає переваги деяких моделей, наприклад, ефективність моделі Pose-TGCN в порівнянні з Pose-GRU, оскільки вона краще захоплює просторові та часові залежності ключових точок тіла. Додатково, через більший обсяг мережі та попередню підготовку на ImageNet і Kinetics, модель I3D показує кращі результати порівняно з VGG-GRU для розпізнавання зображень. Автори також розглядають вплив розміру словника на ефективність моделей. Вони вказують,

що навчання на менших даних дозволяє досягнути кращих результатів, але у реальних умовах з великим словником, як у випадку WLASL2000, завдання стає значно складнішим. Однак, висновок зроблений авторами показує, що розпізнавання жестів на великому словнику є надзвичайно складною задачею, а для розв'язання цього завдання потрібні більш просунуті алгоритми, такі як навчання з дуже невеликої кількості прикладів.

У [22] пропонується новий підхід до розпізнавання американської мови жестів, використовуючи глибокі нейронні мережі. Вона використовує послідовності відеокадрів з тренувального набору даних для навчання нейронної мережі зі слабкокерованим навчанням, щоб автоматично генерувати мітки кадрів. Після цього використовуються сіамська мережа та ResNet-50 для навчання ефективних репрезентацій ознак жестів. Ітеративний процес тренування включає постійне оновлення набору зображень, поки не буде досягнуто максимально можливої продуктивності розпізнавання. Результати експериментів показують переваги цього методу в порівнянні з попередніми роботами у розпізнаванні символів американської мови жестів. Однак обмежена кількість символів та малий обсяг датасету можуть вплинути на точність та репрезентативність моделі у реальних умовах застосування. Також точність становить близько 50%, що є доволі низьким значенням для використання мережі на практиці.

У статті [23] представлено ефективний метод розпізнавання мови жестів, базуючись на застосуванні нейронних мереж. Система розпізнавання жестів працює в реальному часі, що робить її придатною для повсякденного використання. У цьому дослідженні досягнута висока точність розпізнавання жестів, яка складає 98.84% на перевірочних даних. Використана архітектура мережі, VGG_Net, є основою цього методу. Однак цей метод обмежується лише 28 класами, які включають 26 літер англійського алфавіту та 2 спеціальні символи. Ці дані представлені у вигляді зображень, що означає, що модель здатна розпізнавати статичні символи, а не враховує динаміку жестів. Варто відзначити, що обмежена кількість волонтерів, які створювали датасет для цього дослідження, складає всього 6 осіб. Це є досить малим числом і може вплинути

на загальну репрезентативність бази даних, можливо викликаючи негативний ефект на розпізнавання жестів, здійснених іншими особами.

Стаття [24] впроваджує новий підхід до розпізнавання американської мови жестів (ASL) за допомогою глибоких нейромереж, використовуючи спеціалізований датасет ASL-100-RGBD. Зазначена база даних містить 100 класів жестів ASL, отриманих від 15 учасників, і містить інформацію з різних каналів, таких як RGB, Depth, Optical Flow тощо. Методика розпізнавання ґрунтується на обробці відеоматеріалів з використанням тривимірних згорткових нейромереж (3DCNN) та архітектури 3D-ResNet.Результати дослідження вказують на цікавий прогрес у розпізнаванні широкого спектра класів, що відображається в роботі. Проте, важливо зауважити, що точність розпізнавання для 100 класів становить 75.9%, що може свідчити про можливість подальшого покращення в майбутніх експериментах. Ця робота дає можливість для подальшого аналізу та дослідження у цьому напрямку, пропонуючи базу для подальших вдосконалень і досліджень з метою підвищення точності розпізнавання мови жестів ASL з використанням нейромереж.

Стаття [25] описує запропоновану архітектуру мережі глибокого навчання для розпізнавання американської жестової мови. Мережа містить конволюційні та капсульні шари, які розпізнають ознаки зображень та вивчають співвідношення між ними. Хоча стаття описує детально архітектуру запропонованої моделі та її експериментальні результати, вона має деякі недоліки. По-перше, стаття не надає детальної інформації про використані гіперпараметри, що можуть бути важливими для реплікації дослідження та порівняння з іншими моделями. По-друге, автори не порівняли свою модель з іншими сучасними методами розпізнавання ASL, що знижує значущість їхнього висновку про ефективність запропонованої моделі. Отже, хоча ця стаття має свої переваги, вона потребує додаткової роботи для підтвердження ефективності методу порівняно з іншими моделями та для надання детальнішої інформації про використані гіперпараметри.

Дослідження [26] пропонує застосування найпоширенішої архітектури глибокого навчання для розпізнавання мови жестів. Запропонована архітектура

є стандартною для задач розпізнавання образів і використовується в багатьох подібних роботах. Щодо експериментів, то автори дослідження використали CUDA із бібліотекою cuDNN для навчання моделі на GPU, щоб прискорити обчислювальні завдання. Це дійсно дозволяє прискорити навчання моделі та зменшити час, необхідний для навчання. Також автори показали графік часу навчання в залежності від епох, що демонструє ефективність використання GPU. Загалом, стаття містить достатньо докладний опис архітектури, однак, у статті не пропонується нових підходів для розпізнавання жестів, а лише використання стандартної архітектури. Ця робота є корисною для початкових експериментів.

Стаття [27] присвячена розробці глибокої згорткової нейронної мережі під назвою модель E для розпізнавання мови жестів. У статті порівнюється продуктивність моделі E з моделлю AlexNet, яка часто використовується для розпізнавання жестів. Автори прийшли до висновку, що модель E ефективніша за модель AlexNet, та оптимізували модель E, збільшивши кількість епох, що привело до точності 96,83%. Однак є певні мінуси у цій роботі. Кількість класів цієї мережі є доволі малим і обмежується 26 літерами та 3 спецсимволами. Крім того, оскільки деякі жести потребують послідовності кадрів для повного представлення, використання лише окремих зображень у базі даних може ускладнити розпізнавання складних жестів та зменшити ефективність системи.

У статті [28] використовується набір даних ASLLVD та власний тестовий набір для навчання та оцінки класифікаторів мови жестів. Дослідження продемонструвало високі показники точності та повноти для класифікаторів при обробці невеликих кількостей жестів, але зі збільшенням словника жестів виявлено падіння ефективності стандартних класифікаторів. Глибокий ліс продемонстрував стабільну ефективність, що свідчить про його перевагу в порівнянні з іншими класифікаторами. Однак на практиці складніше речення чи фрази не завжди визначалися ефективно, особливо через використання ізольованих жестів для навчання. Робота вказує на потребу у подальших дослідженнях, зокрема застосуванні методів з пам'яттю для покращення роботи системи на складних послідовностях жестів та дослідженні інших баз даних для розширення набору даних.

Метою роботи [29] було розроблення техніки BODL-HGRSLC для розпізнавання мови жестів для сприяння комунікації людей з обмеженими можливостями. Робота використовує концепції комп'ютерного зору (CV) та глибокого навчання (DL), використовуючи модель ResNet для виділення ознак, оптимізацію методом Байєсівської оптимізації для гіперпараметрів і модель BiGRU для процедури розпізнавання жестів. Результати експериментів показали покращення моделі BODL-HGRSLC порівняно з іншими DL-моделями, досягаючи максимальної точності 99,75%. Однак, не зважаючи на ці високі показники точності, робота має кілька обмежень. Наприклад, не зазначено використану базу даних, що може вплинути на загальну репрезентативність результатів. Крім того, детальні характеристики моделей та їх порівняння з іншими методами DL обмежено, що ускладнює повний аналіз їхньої ефективності. Отже, хоча результати, що описуються у статті, є високими, вони повинні бути розглянуті з обережністю.

Дослідження [30] пропонує використання методу ORB для розпізнавання жестів мови. Порівняно з іншими методами, ORB є ефективним та інформативним у представленні візуальних ознак жестів. Комбінація ORB з Canny Edge Detection та Bag of Words забезпечує кращі результати та економію пам'яті порівняно з іншими методами. У порівняльному аналізі з іншими підходами, метод з ORB показує високу точність у розпізнаванні жестів. Однак ця робота обмежена тестуванням лише на статичних зображеннях жестів, не враховуючи динаміку жестів у відео. Є можливість подальшого розвитку, зокрема використання глибокого навчання та еволюційних алгоритмів для вдосконалення витягування ознак.

У статті [31] автори пропонують метод для розпізнавання мови жестів на відео з використанням комбінації Feature Pyramid Network (FPN) та Detection Transformer (DETR). Один з головних внесків цієї статті полягає у використанні DETR для визначення положення рук на відео та подальшої обробки цих даних з використанням FPN. Цей підхід дозволяє досягнути високої точності розпізнавання мови жестів на відео. Проте, стаття також має свої обмеження. Наприклад, незважаючи на те, що результати авторів показують високу точність

розпізнавання, їхня база даних містить лише 9 жестів, що може бути недостатньо для загального застосування методу на практиці.

У [32] використовується два типи вхідних даних – відфільтровані зображення та зображення з використанням сегментації. Застосовується CNN для розпізнавання жестів з можливістю інтерпретації речення на їх основі. Тестування проводилися на даних, які охоплюють відео жестів, що виконані однією людиною. За результатами експериментів, середня точність розпізнавання американської жестової мови складає 96.59%. Виявлено деякі помилки через схожість форм жестів. Найвища точність спостерігається для символу 'Y', а найнижча для 'M'. Порівняно з іншими методами, запропонована система виявляється точнішою. Водночас існують недоліки, такі як обмежена кількість класів та відсутня можливості розпізнавання складніших жестів. У подальшій роботі планується збір більшого обсягу даних та покращення системи для розпізнавання динамічних жестів.

Автори наступної статті [33] розробили моделі для розпізнавання мови жестів з використанням ASL алфавіту за допомогою глибоких згорткових нейронних мереж. Використовувався набір даних MNIST, що містить 24 класи (виключаючи J і Z) та використовувались метрики оцінки такі як точність та F1-оцінка для аналізу продуктивності. Модель навчалась на 34 627 зображеннях і показала високу точність на тестових даних. Після ітераційного навчання та підбору параметрів досягнута найвища оцінка на валідації 99,96% на 16 епохах, і 99,97% на тренувальних даних на 20-й епосі. На тестових даних було досягнуто точності близько 99%. Порівняно з іншими підходами до розпізнавання мови жестів на ASL-датасеті, запропонована модель показала найкращі результати, перевершуючи їхні точності. Однак, деякі обмеження цієї роботи включають обмеження кількості класів та використання статичних зображень, що може ускладнити розпізнавання складних жестів. У майбутньому дослідники планують розширити роботу на розпізнавання реального часу за допомогою контролера Leap Motion та розпізнавання жестів через відеокадри, що є складним завданням.

### 1.3. Постановка проблеми та її обґрунтування.

На основі проведеного критичного аналізу статей, можна виділити наступний ряд проблем:

1. Обмежена кількість даних: У більшості статей зазначено, що для тренування моделей використовують обмежену кількість даних, що може негативно вплинути на точність розпізнавання.
2. Різні види мови жестів: Існує велика різноманітність жестів у різних мовах, що ускладнює розпізнавання.
3. Різноманітність людських жестів: Навіть у межах однієї мови жестів, люди можуть виконувати жести по-різному, що також може ускладнити розпізнавання.
4. Обмеження апаратного забезпечення: Деякі алгоритми розпізнавання можуть вимагати великої кількості обчислювальних ресурсів, що може бути обмежено для певних пристроїв.
5. Недостатня точність: Навіть з використанням глибоких нейронних мереж, точність розпізнавання може бути недостатньою для деяких випадків використання.
6. Стійкість до шуму та перешкод: Розпізнавання може стикатися з проблемами, коли жести затемнені, перекриті або частково приховані.
7. Швидкість роботи: Для деяких застосувань необхідна швидкість розпізнавання жестів у режимі реального часу.
8. Розпізнавання жестів в контексті речення: У деяких випадках жести можуть мати різні значення в залежності від контексту речення, що може бути важким для розпізнавання.

Ці проблеми вимагають подальших досліджень, щоб забезпечити точніше та ефективніше розпізнавання мови жестів. Наприклад, є необхідність у вдосконаленні алгоритмів обробки зображень, що дозволяють виявляти руки та жести на різному тлі та в різних умовах освітлення.

Також, є необхідність у вдосконаленні методів навчання глибоких нейронних мереж для розпізнавання мови жестів, щоб забезпечити більш точні

та швидкі результати. Наприклад, можна розглядати можливості використання механізмів уваги та рекурентних нейронних мереж для забезпечення кращої роботи з послідовністю жестів. Також, проблеми з розпізнаванням жестів, що мають схожі форми та рухи, можуть вимагати використання додаткових датчиків, таких як датчики жестів та позиції рук, або використання складніших алгоритмів для розпізнавання динаміки жестів.

Нарешті, є необхідність у вдосконаленні процесу збору та анотування даних, що використовуються для навчання моделей розпізнавання мови жестів. Це може включати створення нових датасетів з більш різноманітними жестами та умовами використання, а також розробку більш ефективних методів анотування даних.

### 1.4. Висновки

Розпізнавання мови жестів - це тема, яка вже давно привертає увагу науковців та інженерів, оскільки вона має великий потенціал у різних сферах життя. Вона може стати допоміжним інструментом для людей з обмеженими можливостями, які не можуть спілкуватися за допомогою мови, а також використовуватися у віртуальній та доповненій реальності, інтерактивних пристроях та інших технологіях.

З огляду на проведений критичний аналіз, можна зробити висновок про те, що розпізнавання мови жестів є актуальною та перспективною науковою галуззю, яка має значний потенціал у різних сферах життя, від медицини до розваг та комунікації. Однак, в цій сфері ще є досить багато проблем, які потребують вирішення, такі як складність розпізнавання різних мов жестів, відмінності в жестах між різними культурами та індивідами, а також змінність жестів в залежності від умов освітлення та інших факторів.

У цьому розділі було проведено критичний аналіз літератури та методів, що використовуються в розпізнаванні мови жестів, а також були визначені головні проблеми, з якими стикається ця наукова галузь. Відбір статей здійснювався за допомогою схеми PRISMA, де за допомогою критерій відбору, було відсіяно не релевантні роботи та список публікацій скоротився із 115 до 13.

Цей спосіб допоміг обрати кількість робіт, які якнайкраще підходять для аналізу сучасного стану цієї предметної області.

Загалом, проведений аналіз літературних джерел дозволив зробити висновок про важливість досліджень у сфері розпізнавання мови жестів, які можуть виявитися корисними для людей з обмеженими можливостями та для покращення комунікації в цілому. Також, проведений критичний аналіз наголошує на тому, що ця сфера потребує додаткових досліджень та вдосконалень у майбутньому.

# 2. ДОСЛІДНИЦЬКИЙ РОЗДІЛ

У світі сучасних технологій, розпізнавання мови жестів стає важливою складовою для покращення спілкування та взаєморозуміння між людьми, особливо в ситуаціях, де вимовлення слів є викликом або неможливим. Це завдання зазвичай охоплює аналіз послідовностей рухів та поз, які створюють мову жестів, і вимагає від системи здатності розрізняти різні жести та асоціювати їх із відповідними значеннями або командами.

Розробка ефективної системи розпізнавання мови жестів вимагає використання передових методів та технологій у галузі обробки зображень та природної мови. Одними з сучасних архітектур, які проявили себе дуже ефективно у різних завданнях обробки послідовностей, є LSTM (Long Short-Term Memory), ConvLSTM (Convolutional LSTM) та трансформери. LSTM відомі своєю здатністю моделювати довгострокові залежності в послідовностях, ConvLSTM комбінує конволюційні шари з механізмом LSTM для обробки зображень та послідовностей, тоді як трансформери використовують механізм самоуваги для ефективної обробки послідовностей довільної довжини. У порівнянні з трансформерами, LSTM та ConvLSTM мають свої унікальні переваги та можливості в контексті розпізнавання мови жестів, що робить їх потенційно цікавими для використання в даному дослідженні.

## 2.1. Довга короткочасна пам’ять.

Звичайні рекурентні нейронні мережі (RNN) мають обмеження в здатності запам'ятовувати та працювати з довгостроковими залежностями у послідовностях даних. LSTM (Long Short-Term Memory) виникла як відповідь на ці обмеження. Ця архітектура мережі є однією з ключових технологій в глибокому навчанні, спроєктована для роботи з послідовностями даних, зокрема в мові, часових рядів та інших областях.

Основна ідея LSTM полягає в управлінні потоком інформації через спеціальні вентилі, які регулюють потік даних усередині комірки пам'яті. Вона складається з комірок, які замість одного шару мають чотири основних компоненти: ворота забуття (forget gate), ворота входу (input gate), новий вміст

(cell state) і ворота виходу (output gate). На рис. 2.1 представлена загальна архітектура.

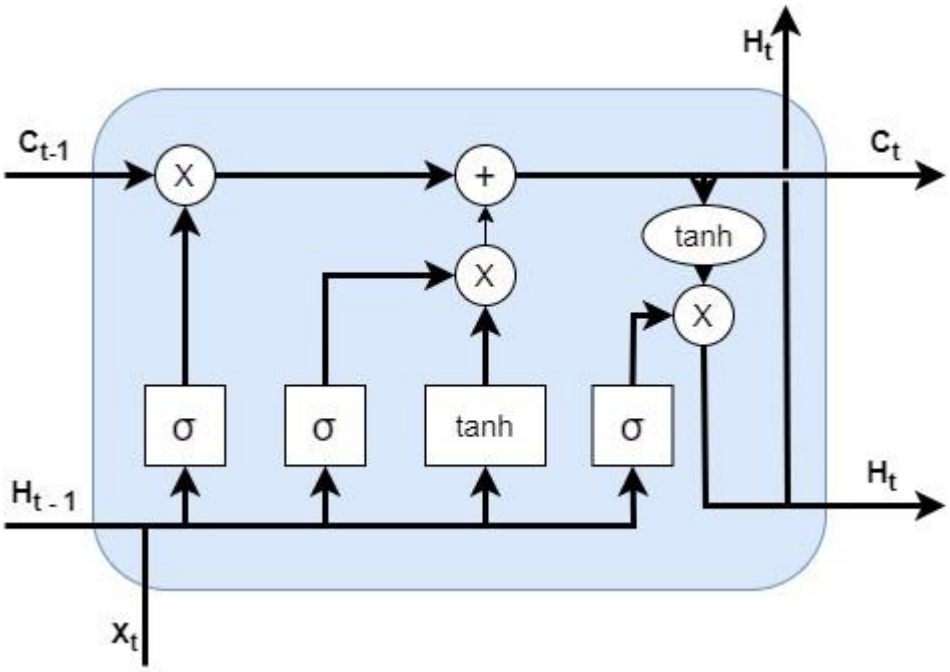



***Рис. 2.1 Архітектура комірки довго короткочасної пам'яті***

Блок LSTM приймає три вектори у вигляді вхідних даних: стан комірки (C) і прихований стан (H), які надходять з попереднього моменту часу (t - 1), а також вхідний вектор (X), що надходить ззовні у момент часу t. Ці вектори дозволяють LSTM регулювати потік інформації та оновлювати значення стану комірки та прихованого стану, які стануть вхідними даними для моменту t + 1.

Контроль потоку інформації здійснюється шляхом використання воріт. Стан комірки виступає як довгострокова пам'ять, що дозволяє зберігати важливі дані на тривалий час, тоді як прихований стан слугує короткочасною пам'яттю, яка визначає вихід мережі на кожному кроці. Це дозволяє LSTM ефективно керувати потоком інформації, зберігаючи ключові залежності та уникати проблем, таких як зникання чи вибування градієнта, які часто виникають у звичайних RNN [34]. Перейдемо до опису кроків роботи LSTM.

Першим кроком є ворота забуття. Цей шар визначає, яка інформація має бути відкинута з попереднього стану. Вона вирішує, яка частина пам'яті повинна бути забута, оскільки не вся інформація однаково важлива. Це досягається за допомогою сигмоїдального шару, який вирішує, що тримати або забути.

Формула 2.1 представляє роботу цього шару.

$$f_t = \sigma\big(W_f \cdot [H_{t-1}, x] + b_f\big) \tag{2.1}$$

Наступним кроком є ворота входу. Цей шар розглядає нову інформацію, яку потрібно зберегти. Він складається з двох частин: сигмоїдального шару, який вирішує, які значення оновлювати(формула 2.2), та tanh-шару, який створює новий вектор кандидатів для оновлення стану(формула 2.3).

$$i_t = \sigma(W_i \cdot [H_{t-1}, x_t] + b_i) \tag{2.2}$$

$$\tilde{C_t} = \tanh(W_C \cdot [H_{t-1}, x_t] + b_C) \tag{2.3}$$

Далі відбувається оновлення стану. Це об'єднання інформації з воріт входу та воріт забуття для оновлення попереднього стану комірки. Це дає можливість враховувати нові важливі дані та забувати неважливі. Формула 2.4 демонструє цей крок.

$$C_t = f_t * C_{t-1} + i_t * \tilde{C_t} \tag{2.4}$$

І останнім кроком є вивід інформації через вихідні ворота. Цей шар вирішує, яку частину стану ми виведемо. Спочатку дані проходять через сигмоїдальний шар, визначаючи, які частини стану будуть виведені(формула 2.5), а потім ці значення проходять через tanh-шар(формула 2.6) [35].

$$o_t = \sigma(W_0[H_{t-1}, x_t] + b_0) \tag{2.5}$$

$$H_t = o_t * \tanh(C_t) \tag{2.6}$$

LSTM (Long Short-Term Memory) в контексті розпізнавання мови жестів має значний потенціал завдяки своїй здатності ефективно моделювати та управляти складними послідовностями даних. Цей метод може виявити особливу корисність в аналізі жестів, оскільки він дозволяє враховувати довгострокові залежності та важливі контекстуальні відношення між різними рухами. Наприклад, в контексті розпізнавання мови жестів, LSTM може відтворювати складні послідовності рухів, дозволяючи системі враховувати важливі етапи виконання жесту, послідовно зв'язуючи їх для зрозуміння в цілому. Це може бути корисним для розробки систем, що вимагають докладного визначення певних жестів або їх послідовностей.

### 2.2. Конволюційний LSTM.

Крім LSTM, одним з методів, який також заслуговує уваги у цьому контексті, є ConvLSTM (Convolutional LSTM). ConvLSTM поєднує в собі переваги згорткових нейронних мереж (CNN) та LSTM, дозволяючи аналізувати зображення або послідовності даних у вигляді часових шарів. Цей підхід може бути корисним у випадках, коли важлива просторова інформація на зображеннях жестів, оскільки ConvLSTM може ефективно працювати зі сполученням просторової та часової інформації, що може бути корисним для розпізнавання послідовних рухів та їх взаємодії.

Одна з основних відмінностей ConvLSTM від звичайної LSTM полягає в тому, що замість того, щоб оперувати лише з одновимірними часовими послідовностями даних, ConvLSTM працює з двовимірними просторовими послідовностями, які можуть бути зображеннями або відео [36]. Принцип роботи ConvLSTM представлений на рис. 2.2.

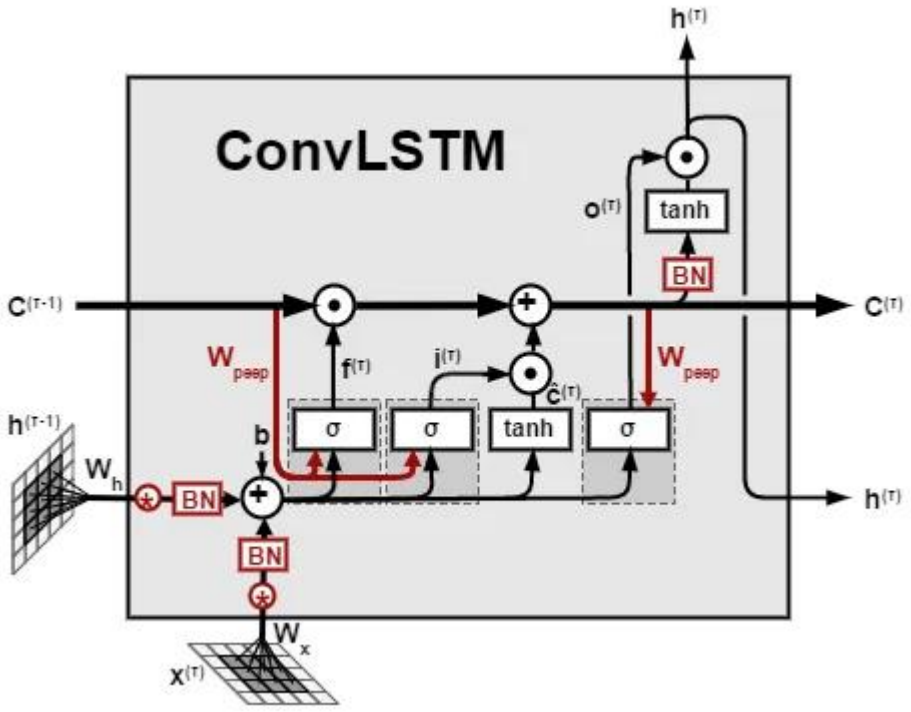



***Рис. 2.2. Загальний принцип роботи конволюційної LSTM***

Згорткова довготривала короткочасна пам’ять (ConvLSTM) є унікальним типом нейронної мережі, яка злагоджує в собі потужність моделей згорткової нейронної мережі (CNN), відомих своєю ефективністю у розпізнаванні та

витяганні ознак із зображень, та довготривалої короткочасної пам’яті (LSTM), що використовується для роботи з послідовностями та зберігання інформації в часі. ConvLSTM було розроблено з метою об'єднати ці дві моделі, створивши потужний механізм обробки даних, який враховує як просторові, так і часові залежності.

Архітектура ConvLSTM працює, зберігаючи просторову інформацію свого входу, а також вивчаючи тимчасові залежності. Зазвичай, згорткові нейронні мережі втрачають просторову інформацію під час обробки вхідних зображень через проходження їх через згорткові та пулінгові шари. Моделі LSTM, з іншого боку, потребують збереження тимчасової інформації, тобто вони працюють з впорядкованими послідовностями даних. Модель ConvLSTM використовує комбінацію згорткових і LSTM шарів для збереження як просторової, так і тимчасової інформації.

Архітектура моделі ConvLSTM складається з чотирьох основних компонент:

1. Згорткові шари (Convolutional Layers).
2. Ворото забуття (Forget Gate).
3. Ворото входу (Input Gate).
4. Ворото виходу (Output Gate) [37].

Згорткові шари у моделі виконують ту саму роль, що й у стандартній згортковій нейронній мережі, витягаючи відповідні особливості з вхідних даних. Вони зберігають просторову інформацію вхідних даних протягом роботи моделі, на відміну від стандартної LSTM моделі. Згорткові шари, що використовуються у моделі ConvLSTM, зазвичай аналогічні тим, які використовуються у стандартній згортковій нейронній мережі.

Формула для згорткового шару може бути виражена так:

$$Conv(x) = \sigma(W * X + b) \tag{2.7}$$

де:

- X – вхідні дані;
- W – ваги згорткового ядра;

- b – зміщення;
- $\sigma$ – функція акттивації, наприклад ReLU.

Згортковий шар пропускає вхідне зображення через ядро згортки, додає зміщення та застосовує функцію активації для отримання остаточного результату. У подальшому виконанні процедури ConvLSTM, відбувається передача даних у ворота забуття, ворота входу та ворота виходу. Принцип функціонування цих елементів у ConvLSTM відображає загальні принципи роботи звичайної LSTM, які вже були розглянуті раніше.

### 2.3. Опис алгоритму роботи трансформера.

Крім зазначених раніше методів, варто звернути увагу на ще одну сучасну архітектуру, яка проявила високу ефективність у різних завданнях обробки послідовностей — трансформер. Використання трансформерів для класифікації мови жестів обґрунтовується кількома ключовими перевагами. По-перше, трансформери здатні враховувати довгострокову залежність між різними елементами послідовності, що особливо важливо в контексті розпізнавання мови жестів. Жести часто мають складні залежності та вимагають аналізу контексту, наприклад, попередніх рухів, щоб правильно інтерпретувати їх значення.

По-друге, трансформери дозволяють враховувати просторову і тимчасову інформацію одночасно, завдяки використанню механізму уваги. Це дозволяє моделі аналізувати, як об'єкти рухаються в просторі та часі, що є ключовим для класифікації жестів.

Крім того, трансформери можуть бути навчені на великих обсягах даних, що робить їх ідеальними для завдань розпізнавання мови жестів, оскільки для навчання ефективних моделей потрібна велика кількість даних, що включають різні жести та їх варіації.

Використання трансформерів для класифікації мови жестів відкриває нові можливості для покращення точності та ефективності системи розпізнавання мови жестів, роблячи її більш адаптивною та потужною у виявленні та інтерпретації жестів у різних ситуаціях та контекстах. Перейдемо до розгляду алгоритму трансформера, що дасть можливість зрозуміти, як саме ця передова

архітектура працює та як вона може бути використана для ефективного розпізнавання мови жестів.

Трансформери є важливою архітектурою нейронних мереж, і вони відрізняються від інших традиційних архітектур, таких як рекурентні нейронні мережі (RNN) та згорткові нейронні мережі (CNN). Їхній винятковий успіх в багатьох завданнях глибокого навчання пояснюється наявністю трьох ключових елементів:

- Самоувага;
- Позиційне вбудовування;
- Багатоголова увага.

Ці компоненти були вперше представлені у 2017 році у статті «Attention Is All You Need» [38] Васвані та іншими.

Самоувага є фундаментальним механізмом трансформерів. Він дозволяє моделі виявляти зв’язки між різними елементами в послідовності, навіть якщо ці елементи знаходяться на великій відстані один від одного. Самоувага оцінює важливість цих зв’язків і дозволяє моделі краще розуміти контекст вхідних даних. Як сказав Васвані, "Значення є результатом взаємозв’язків між речами, а самоувага є загальним способом вивчення цих взаємозв’язків."

Позиційне вбудовування впроваджують інформацію про позиції слів у послідовності. Це важливо, оскільки трансформери не мають вбудованого розуміння порядку слів у тексті, і позиційне вбудовування допомагає моделі усвідомити цей аспект послідовності. Вони додаються до ембедінгів слів, щоб вказати їхні позиції в послідовності.

Багатоголова увага дозволяє трансформерам працювати з різними аспектами взаємозв’язків у даному вхідному сигналі. Він дозволяє моделі фокусуватися на різних частинах вхідної послідовності та взаємодіяти з ними паралельно. Багатоголова увага допомагає розширити потужність трансформерів та забезпечує їхню здатність адаптуватися до різних завдань глибокого навчання.

Ці ключові компоненти роблять трансформерів дуже потужними і дозволяють їм досягати вражаючих результатів у багатьох завданнях машинного навчання, зокрема в обробці природної мови. Моделі, які базуються на

архітектурі трансформерів, такі як BERT і GPT, встановили нові стандарти продуктивності в цьому напрямку та допомогли трансформерів стати основною архітектурою для роботи з текстом і мовою. З їхнім розвитком ми можемо очікувати більше захоплюючих досягнень у майбутньому[39].

### *2.3.1. Архітектура трансформера.*

Зважаючи на важливість трансформерів у сучасному глибокому навчанні та їхню величезну роль у досягненнях в області обробки природної мови, розглянемо детальніше архітектурну основу цих потужних моделей. У цьому розділі розглянемо трансформер як нейронну мережу та розглянемо її ключові компоненти, які роблять її особливою та дозволяють їй вирішувати складні завдання з обробки послідовностей. На рис. 2.3 представлена структура архітектури трансформера.

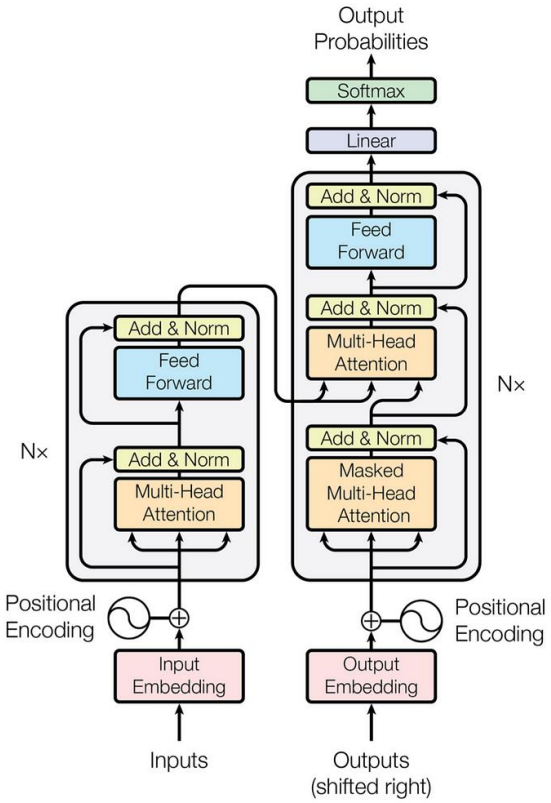



***Рис. 2.3. Архітектура трансформера***

Енкодер розташований ліворуч, а декодер — праворуч. Як енкодер, так і декодер складаються з модулів, які можна стекати один на одного кілька разів, що позначається як Nx на малюнку. Як видно, модулі в основному складаються

з багатоголової увага та шарів прямого поширення. Вхідні та вихідні дані (цільові речення) спершу вбудовуються в n-вимірний простір, оскільки ми не можемо використовувати дані без попередньої обробки.

Одна з важливих деталей цієї архітектури – позиційне кодування слів. Оскільки не використовуються рекурентні мережі, які могли б запам’ятовувати послідовність вводу в модель, потрібно певним чином надати кожному елементу у послідовності відносну позицію. Ці позиції додаються до вбудованого представлення (n-вимірний вектор) кожного слова.

### *2.3.2. Механізм уваги.*

Для більш детального огляду, розглянемо рис. 2.4, на якому зображено алгоритм роботи уваги.

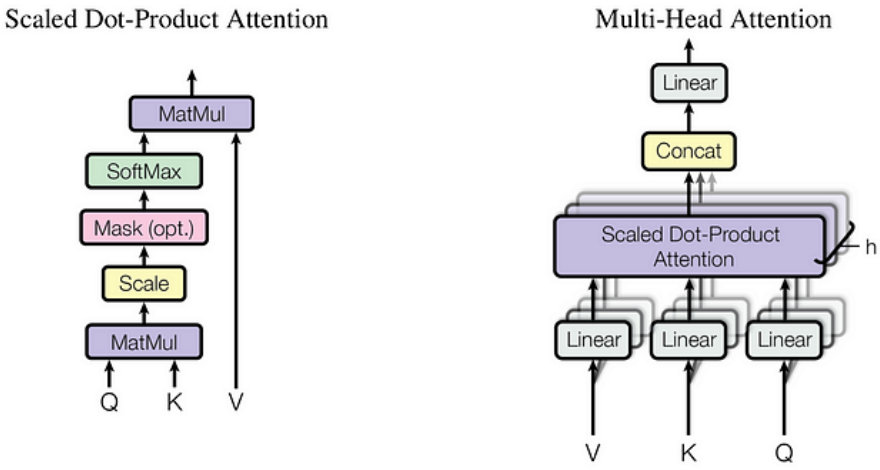



***Рис. 2.4. Scaled Dot-Product Attention (ліворуч) та Multi-Head Attention (праворуч)***

Scaled Dot-Product Attention - це ключовий компонент у трансформерах і багатьох інших моделях обробки послідовностей. Цей механізм дозволяє моделі фокусуватися на різних частинах вхідних послідовностей та визначати важливість їхніх взаємозв'язків.

Основна ідея полягає в обчисленні "уваги" для кожної пари векторів у

вхідній послідовності. Ця увага визначає, наскільки важливою є кожна інша частина послідовності для даного елемента. Щоб вирахувати увагу, використовується добуток між матрицями запиту (query) та ключа (key) для кожної пари елементів. Результат добутку піддається функції softmax для отримання нормалізованого розподілу уваги.

Алгоритм роботи Scaled Dot-Product Attention можна описати наступною формою:

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (2.8)$$

- Q (Query) - матриця запитів (query), де кожен рядок відповідає запиту для кожного елемента в послідовності;
- K (Key) - матриця ключів (keys), де кожен рядок відповідає ключу для кожного елемента в послідовності;
- V (Value) - матриця значень (values), де кожен рядок відповідає значенню для кожного елемента в послідовності;
- $d_k$ - розмірність (кількість функціональних вимірів) для запитів та ключів.

Увага обчислюється для кожного елемента окремо. Параметр $\sqrt{d_k}$ важливий для стабільності та контролю над градієнтами під час навчання.

Отриманий нормалізований розподіл уваги використовується для зважування значень (V) кожного елемента або позиції. Результат обчислення - це вагова сума значень, що відповідають важливим частинам вхідної послідовності для поточного елемента або позиції. Ця операція допомагає моделі концентруватися на важливих аспектах вхідної інформації під час обробки тексту або послідовностей.

Права частина малюнку описує використання багатоголового механізму уваги. Багатоголова увага - це розширений механізм уваги, який дозволяє трансформерам фокусуватися на різних частинах вхідної послідовності одночасно та незалежно. Він створює кілька "голів" уваги, кожна з яких навчається фокусуватися на різних аспектах даних і визначає важливість різних

частин послідовності. Формула, яка описує цей алгоритм представлена нижче:

$$MultiHead(Q,K,V) = Concat(head_1, \dots, head_h)W^0$$
$$where\ head_i = Attention\left(QW_i^Q, KW_i^K, VW_i^V\right) \quad (2.9)$$

- $W_i^Q, W_i^K, W_i^V$ – матриці ваг для «голів», що проєкціонують вхідні послідовності в просторі запитів, ключів та значень відповідно;
- h – кількість «голів»;
- $W^0$ - матриця ваг для проєкційного шару.

Якщо розібрати покроково, то спочатку для кожної «голови» готуються окремі вектори запиту, ключа та значень. Далі для кожної з «голови» обчислюється увага (Attention) за допомогою векторів запиту, ключа та значення, що були підготовлені на попередньому кроці. Після обчислення уваги для кожної голови, їх результати конкатенуються в одну послідовність. Завершальним етапом є проходження через лінійний проєкційний шар $W^0$.

Отже, багатоголова увага використовує кілька "голів" уваги, кожна з яких фокусується на різних аспектах даних, і результати цих голів об'єднуються та проходять через проєкційний шар, щоб отримати остаточний вихід моделі. Ця операція дозволяє трансформерам моделювати важливі взаємозв'язки між різними частинами послідовностей і підвищує їхню потужність у завданнях обробки природної мови та інших послідовностей.

### *2.3.3. Позиційне кодування.*

Оскільки трансформер не має рекурентних шарів або явної внутрішньої пам'яті для визначення порядку, потрібно додатково ввести інформацію про позицію кожного елемента у послідовності. Це робиться за допомогою позиційного кодування.

Основна ідея полягає в тому, щоб кожному елементу в послідовності додати вектор, який кодує його позицію у послідовності. В цьому контексті використовується функція синуса та косинуса з різними частотами для створення цих позиційних векторів:

$$PE(pos, 2\text{i}) = \sin\left(pos \,/\, 10000^{2\text{i}/d_{model}}\right) \quad (2.10)$$

$$PE(pos, 2\text{i}+1) = \cos\left(pos \,/\, 10000^{2\text{i}/d_{model}}\right) \quad (2.11)$$

- $PE(pos, 2\text{i})$ - позиційне кодування для парних позицій;
- $PE(pos, 2\text{i}+1)$ - позиційне кодування для непарних позицій;
- $pos$ - позиція елемента в послідовності;
- i - індекс в рядку позиційних векторів;
- $d_{model}$- розмірність позиційних векторів.

Отже, кожна позиція $pos$ у послідовності отримує свій унікальний позиційний вектор, який обчислюється за допомогою функцій синуса та косинуса з різними частотами. Ці вектори додаються до векторів входу перед подачею на вхід трансформера. Це дозволяє моделі розрізняти різні позиції у послідовності та враховувати їх в обробці даних без необхідності рекурентних шарів.

### *2.3.4. Позиційно-залежні мережі прямого зв'язку.*

На додаток до рівня уваги, кожен з шарів енкодера та декодера містить повнозв'язну мережу прямого зв'язку, яка застосовується для кожного елемента окрему. Позиційно-залежні мережі прямого зв'язку (FFN) - це компоненти трансформерів, які додають нелінійність і обробляють кожен елемент вхідної послідовності окремо. Мережа складається з двох лінійних перетворень та з активацією ReLU між ними. Формула FFN виглядає наступним чином:

$$FFN(x) = \max(0, xW_1 + b_1)W_2 + b_2 \quad (2.12)$$

- x- вхідний вектор;
- $W_1$ і $b_1$ – ваги та зсув для першого лінійного шару;
- $W_2$ і $b_2$ - ваги та зсув для другого лінійного шару;
- $\max(0, \dots)$ – операція активації ReLU, яка відсікає негативні значення і залишає позитивні.

Цей алгоритм операцій відбувається окремо для кожного елемента вхідної послідовності. Це означає, що кожен елемент вхідної послідовності обробляється

і перетворюється незалежно від інших елементів. Position-wise Feed-Forward Network допомагає моделі виявляти та враховувати позиційно-залежні ознаки у вхідних даних та забезпечує необхідну нелінійність для моделювання складних взаємозв'язків між елементами послідовності.

### 2.4. Функції втрат.

Функції втрат є важливою складовою процесу навчання глибоких нейронних моделей. Вони визначають різницю між передбачуваними та фактичними значеннями, що дозволяє моделі коригувати свої параметри для досягнення кращої точності. У контексті розпізнавання мови жестів, де точність та ефективність грають ключову роль у роботі моделі, розгляд функцій втрат є важливим кроком для досягнення кращих результатів. У цьому розділі розглянемо принципи роботи перехресної ентропії зі згладжуванням міток та ArcFace.

#### *2.4.1. Перехресна ентропія зі згладжуванням міток.*

Функція втрат зі згладжуванням міток є методом регуляризації, який застосовують у задачах навчання моделей. Використовується ця техніка для боротьби із зайвою самовпевненістю та покращення узагальнення в глибоких нейронних мережах. Ідея згладжування міток полягає в тому, щоб внести невелику кількість невизначеності в навчальні мітки замість використання традиційного одноразового кодування. Замість призначення єдиної 1 справжній мітці класу та 0 усім іншим класам, згладжування міток замінює 1 на значення трохи менше 1 і рівномірно розподіляє залишкову масу ймовірності між іншими класами [40].

При класифікації модель намагається передбачити правильний клас для кожного вхідного зразка. Зазвичай це відбувається шляхом використання функції втрат, наприклад, категоріальної крос-ентропії, для порівняння передбаченого розподілу з реальним.

Класична крос-ентропія має такий вигляд:

$$L(p,q) = -\sum_{i} p(i)\log\big(q(i)\big) \tag{2.13}$$

де:

- $p$ – реальний розподіл(вихідні мітки);
- $q$ – передбачений розподіл.

Під час тренування нейронної мережі, модель намагається максимально точно визначити класи об'єктів у вхідних даних. Проте це може призвести до перенавчання - коли модель стає дуже специфічною для навчального набору даних, і втрачає здатність узагальнювати на нові, реальні дані. Функція втрат із згладжуванням міток спрямована на зменшення цього перенавчання. Вона використовує параметр згладжування (α), який дозволяє моделі не впевнюватися абсолютно в реальних мітках.

Основна ідея полягає у створенні нового розподілу міток, який поєднує реальні мітки з рівномірним розподілом. Новий розподіл $\tilde{p}$ для кожного класу визначається так:

$$\tilde{p}(i) = (1-\alpha)\cdot p(i) + \alpha \cdot u(i) \tag{2.14}$$

де:

- α – це параметр згладжування, який визначає відносний внесок оригінального та рівномірного розподілів;
- $u(i)$ – це ймовірність для кожного класу, яка розподілена рівномірно.

Після цього зміненого розподілу, формула крос-ентропії виглядає наступним чином:

$$L(\tilde{p},q) = -\sum_{i} \tilde{p}(i)\log\big(q(i)\big) \tag{2.15}$$

Це дозволяє моделі враховувати як реальні мітки, так і деякий рівень неузгодженості, що допомагає уникнути перенавчання на конкретних мітках та зробити модель більш робастною до шумів або неточностей у даних [41].

### *2.4.2. Функція втрат ArcFace.*

Функція втрат ArcFace використовується в задачах класифікації для

підвищення точності і універсальності моделі. Вона базується на ідеї внутрішньокласової відстані, що максимізує кут між векторами репрезентації класів, тим самим забезпечуючи кращу роздільну здатність між класами.

Основна мета ArcFace полягає в тому, щоб змусити модель генерувати репрезентації об'єктів у просторі таким чином, щоб вектори, що відповідають різним класам, були далеко один від одного, а вектори в межах одного класу були якомога ближче до центру класу.

Функція втрат ArcFace визначається наступним чином:

$$L_{ArcFace} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{e^{s*\cos\left(\theta_{y_i}+m\right)}}{e^{s*\cos\left(\theta_{y_i}+m\right)} + \Sigma_{j=1,i\neq y_i}^{C} e^{s^*\cos(\theta_j)}} \quad (2.16)$$

де:

- N – кількість прикладів у наборі даних;
- C – кількість класів;
- $\theta_j$ - кут між вектором ознак $x_i$та вагами класу j (embeddings);
- $\theta_{y_i}$ - кут між вектором ознак $x_i$ та вагами для правильного класу;
- s - параметр шкали, який називають softmax temperature;
- m - параметр маржі (margin), який дозволяє збільшити відстань між векторами різних класів, що покращує роздільну здатність класів.

Ця функція втрат допомагає моделі змінювати свої параметри таким чином, щоб максимізувати косинусні значення між векторами, що відповідають правильним класам, і при цьому знижувати косинусні значення між векторами, що відповідають неправильним класам [42].

Це сприяє збільшенню внутрішньокласових подібностей та зменшенню міжкласових подібностей у просторі репрезентацій, що веде до покращення універсальності та роздільної здатності моделі у задачах класифікації.

### 2.5. Структура та організація даних.

Розпізнавання мови жестів стає все більш популярним напрямком досліджень. Кількість статей, присвячених цій темі, постійно зростає, і це

створює потребу у якісних базах даних для подальших досліджень. Однак доступність і збереження великих баз даних, які складаються з відеоматеріалів, може бути проблемою. У даному дослідженні використовується база даних “Isolated Sign Language Recognition” [43] дані, які були перетворенні на табличні, шляхом перетворення відеоданих за допомогою бібліотеки Mediapipe. Ця база даних представляє собою колекцію орієнтирів рук і обличчя, які були згенеровані за допомогою бібліотеки Mediapipe версії 0.9.0.1. Вона створена на основі приблизно 100 000 відеозаписів жестів, виконаних 21-ма добровольцями, які спілкуються американською мовою жестів. У цій базі даних враховано 250 різних жестів.

Застосування Mediapipe дозволило ефективно обробляти відеодані та конвертувати їх у табличний формат, що спростило подальший аналіз та роботу з даними для завдання розпізнавання мови жестів. Такий підхід став важливою складовою цього дослідження та дозволив створити базу даних, яку значно легше зберігати через зменшення розміру даних.

### *2.5.1. Застосування mediapipe для оптимізації зберігання та покращення розпізнавання мови жестів.*

Mediapipe [44] є відкритою бібліотекою, розробленою Google, яка надає зручний і потужний інструментарій для розпізнавання та аналізу різних аспектів людського обличчя та рук у відеопотоці. Ця бібліотека використовує глибоке навчання для виділення ключових точок та розпізнавання певних жестів, що може бути корисним для аналізу поведінки та комунікації.

Для визначення пози обличчя, рук та їхніх рухів, Mediapipe використовує власний набір алгоритмів, що дозволяють точно визначити ключові точки, такі як положення очей, носа, рота, а також кінцівок рук. За допомогою цих ключових точок можна відстежувати зміни в позі, руках та жестах людини.

У цьому дослідженні була використана бібліотека Mediapipe для отримання точок, які характеризують особливості пози, яку представляє користувач. Як сказано вище, точки виділяються на обличчі, руках та тулубі, які повною мірою демонструють основні відмінності у позах і дають точно

розпізнавати рухи. Таким чином, бібліотека Mediapipe стала ключовим інструментом для реалізації системи розпізнавання мови жестів, що дозволило в реальному часі визначати та аналізувати рухи рук та обличчя людини з метою виявлення мовних жестів.

Однією з ключових особливостей бібліотеки Mediapipe є її можливість використовувати не одну, а кілька нейронних мереж для визначення ключових точок на різних областях людського тіла. Кожна з цих нейронних мереж спеціалізується на визначенні точок в конкретній області, що дозволяє детально аналізувати рухи та пози.

Зокрема, для визначення ключових точок на обличчі, Mediapipe використовує окремий алгоритм, який здатний точно визначити положення очей, носа, рота та інших важливих ділянок обличчя. Це дозволяє виявляти вирази обличчя та зміни в позі голови. Приклад відображення точок на лиці людини представлений на рис. 2.5.

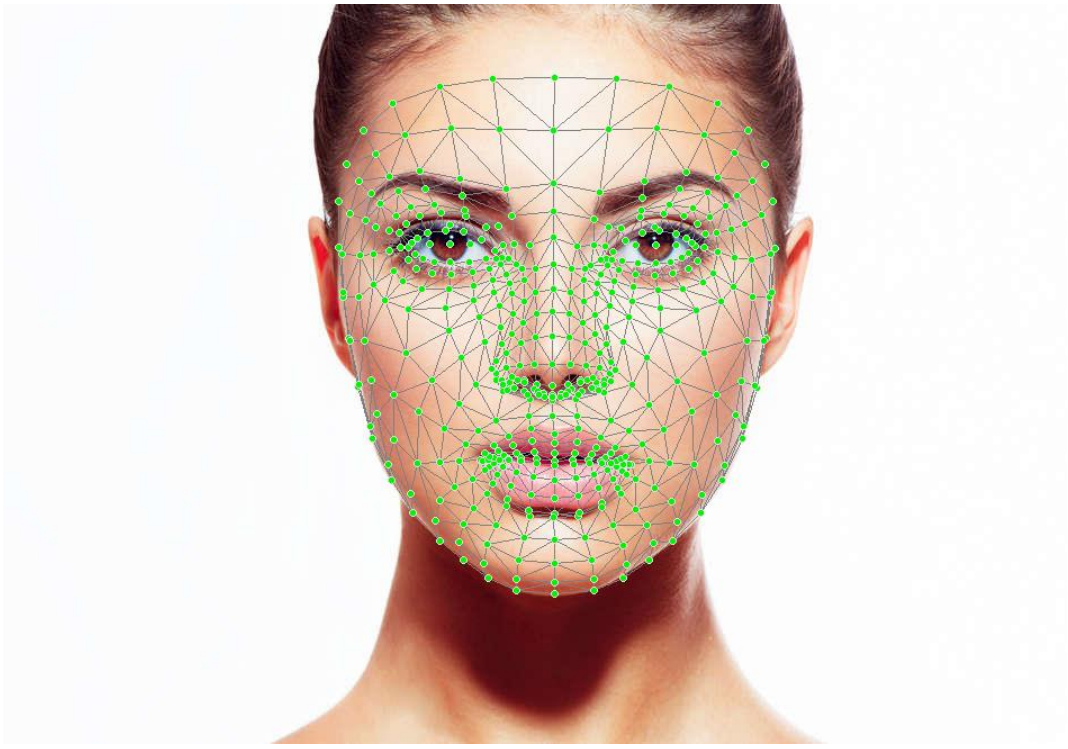


***Рис. 2.5. Ключові точки, які виділяються за допомогою MediaPipe Face Mesh***

Face Mesh [45] використовує конвеєр двох нейронних мереж для визначення 3D-координат 468 орієнтирів обличчя. Перша мережа (BlazeFace 1 ) обчислює розташування обличчя на основі повного зображення. Потім друга

мережа працює лише з обрізаною ділянкою, щоб визначити орієнтири. Весь конвеєр побудовано для швидкості та вимагає менше 10 мс на кращих смартфонах.

Для визначення точок на руках бібліотека використовує окрему мережу. Вона забезпечує високу точність при виявленні положень пальців, кінцівок рук та їх рухів. Це дозволяє відстежувати жести рук, такі як рухи пальцями, виконання певних рухів та інші комунікативні жести. Накладання точок на руку відображено на рис. 2.6.

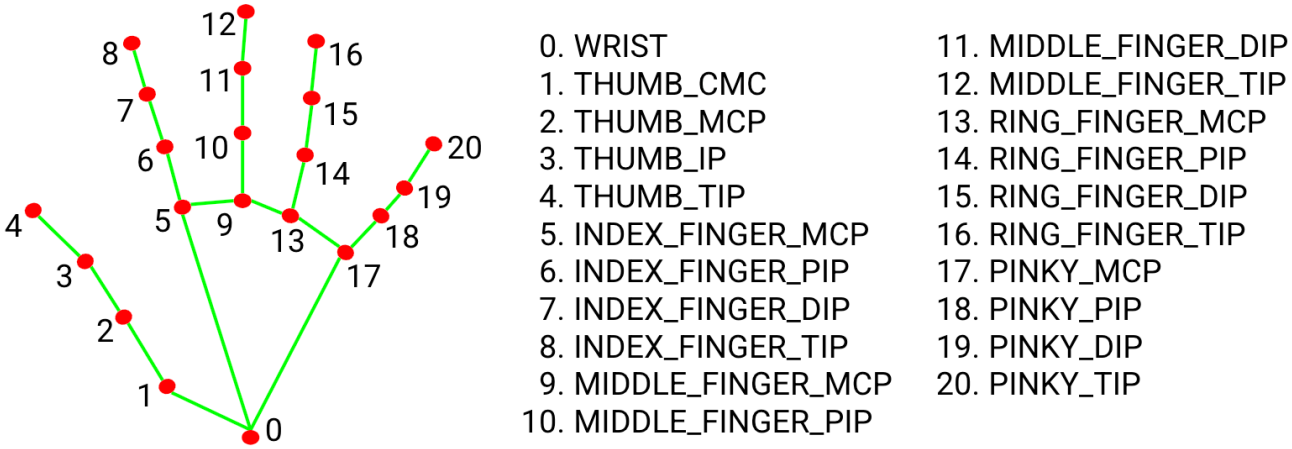



***Рис. 2.6. Ключові точки, які виділяються за допомогою MediaPipe Hands Landmarker***

MediaPipe Hands [46] використовує структуру глибинного навчання, що складається з кількох моделей, які співпрацюють між собою: модель виявлення долоні, яка працює з повним зображенням і повертає орієнтовану рамку долоні та модель локації ключових точок долоні, яка працює на області обрізаного зображення, визначеної детектором долоні, і повертає 21 точку руки у 3D просторі. Ця стратегія схожа на ту, яку використовується в MediaPipe Face Mesh, де використовується детектор обличчя разом з моделлю локації ключових точок обличчя.

Щодо детекції поз, Mediapipe використовує ще одну нейронну мережу, яка дозволяє визначати ключові точки на тулубі. Це може бути корисно для аналізу змін у положенні тіла, орієнтації та загальної пози людини. Представлення ключових точок зображена на рис. 2.7.

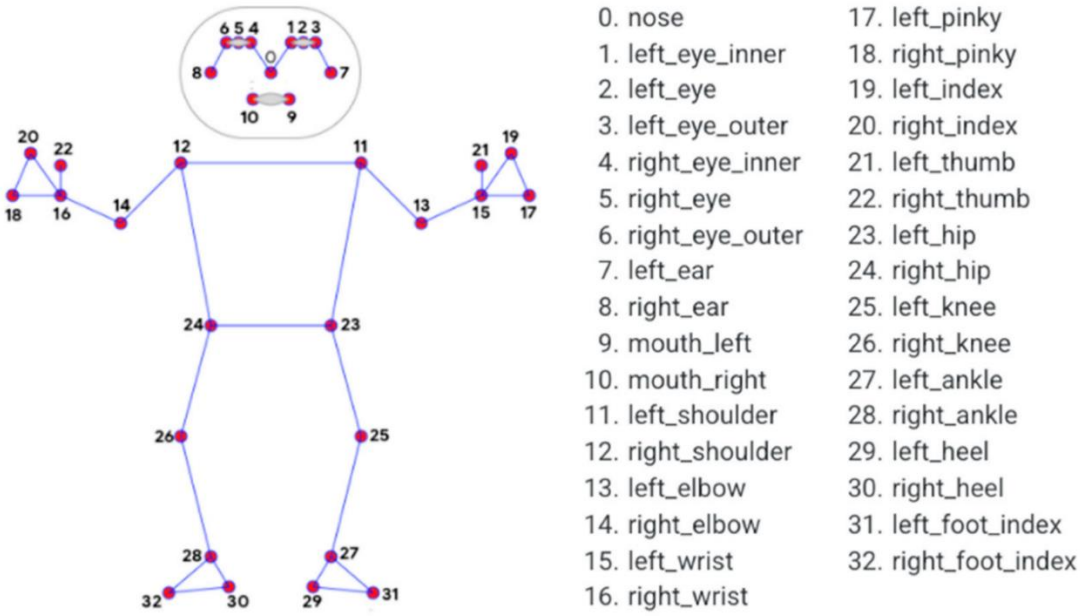



***Рис. 2.7. Ключові точки, які виділяються за допомогою MediaPipe Pose Landmarker***

Як і попередні моделі, для визначення ключових точок використовується дві моделі, одна з них визначає присутність тіла у кадрі, а друга модель визначає 33 ключові точки. Цей комплект використовує згорткову нейронну мережу, подібну до MobileNetV2, і оптимізований для використання на пристроях у реальному часі. Цей варіант моделі BlazePose використовує GHUM, технологічний процес тривимірного моделювання людської форми, для оцінки повної тривимірної пози тіла особи на зображеннях або відео [47].

Отже, за допомогою свого потужного інструментарію, Mediapipe надає можливість виявляти та аналізувати різні аспекти обличчя, тулуба та рук у відеопотоці. Використовуючи глибоке навчання, бібліотека здатна визначити ключові точки та розпізнавати певні жести, що відкриває широкі можливості для аналізу поведінки та комунікації. Це стало ключовим фактором для реалізації системи розпізнавання мови жестів і зберіганням великого обсягу відеоданих.

Таким чином, бібліотека Mediapipe стала фундаментальним інструментом для успішної реалізації системи розпізнавання мови жестів у цьому дослідженні. Її здатність визначати ключові точки та аналізувати рухи рук, позиції тулуба та обличчя з високою точністю і в реальному часі робить її незамінним ресурсом для вивчення мови жестів та її застосувань.

### *2.5.2. Аналіз даних та їх розподіл.*

Дані включають номер кадру (frame), ідентифікатор рядка (row_id), тип (type), індекс точки (landmark_index), координати x, y та z точки на обличчі. Кожен елемент складається з кількох кадрів, оскільки для демонстрації більшості жестів, потрібно ряд кадрів. Середня тривалість демонстрації жестів рівна 37 кадрів. На рис. 2.8 зображено десять класів у яких найбільша тривалість жесту.

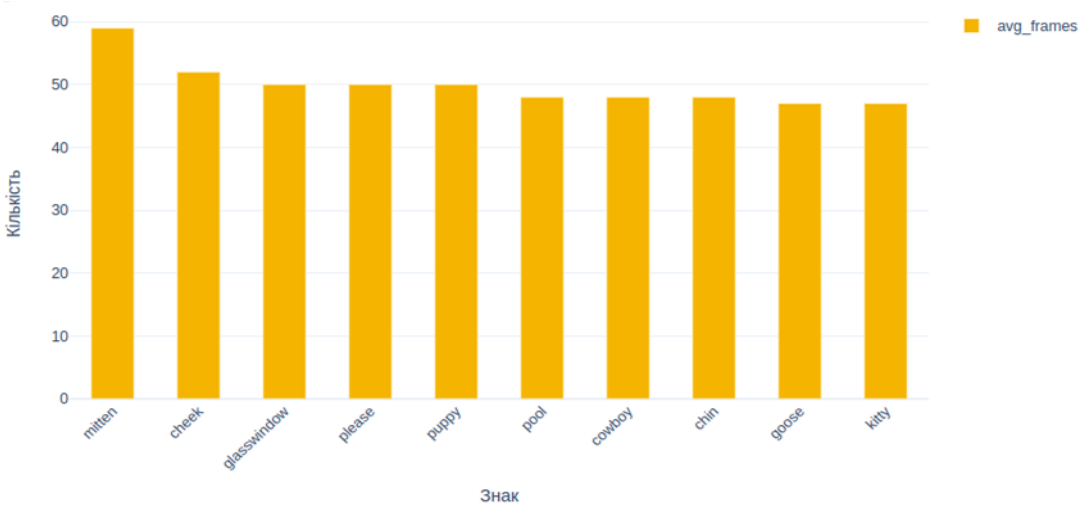



***Рис. 2.8. 10 класів з найбільшою кількістю фреймів***

Як видно, є класи, в яких середня кількість кадрів більше ніж 40, для кращої демонстрації розподілу, представлений коробчастий графік на рис. 2.9.

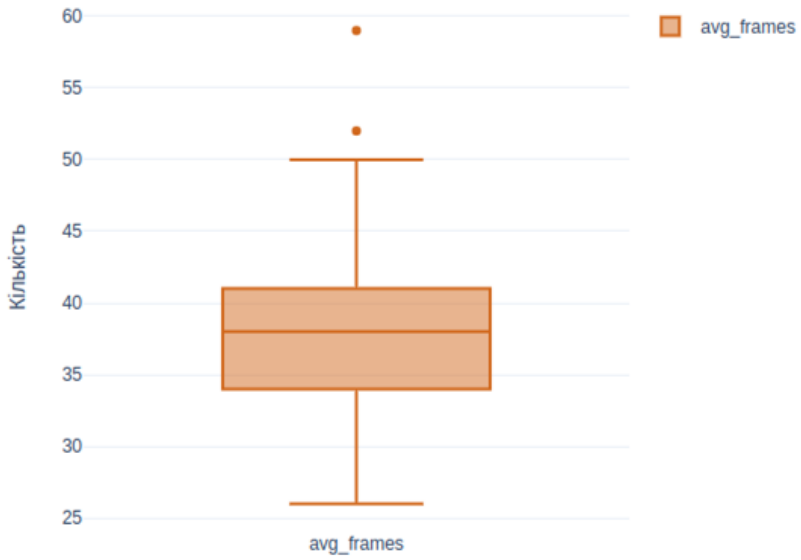


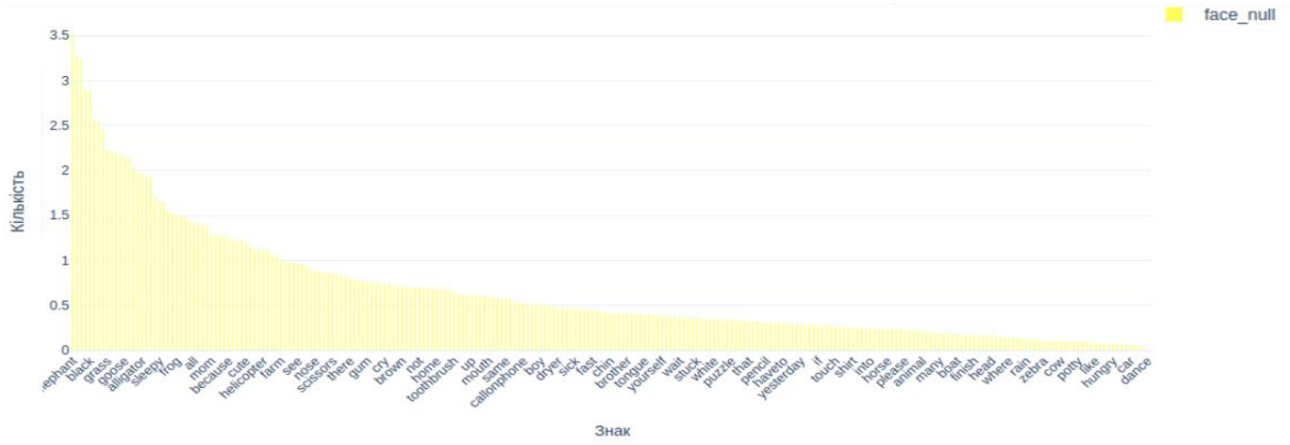

***Рис. 2.9. Розподіл середньої кількості фреймів для всіх класів***

З цього графіка видно, що в основному тривалість демонстрації жесту є від 34 до 41 кадру, також наявні викиди, а саме класи «mitten» та «cheek». Окрім даної характеристики, для навчання є важливим кількість екземплярів кожного класу. Середня кількість становить 377, для візуалізації на рис. 2.10 представлений коробчастий графік.

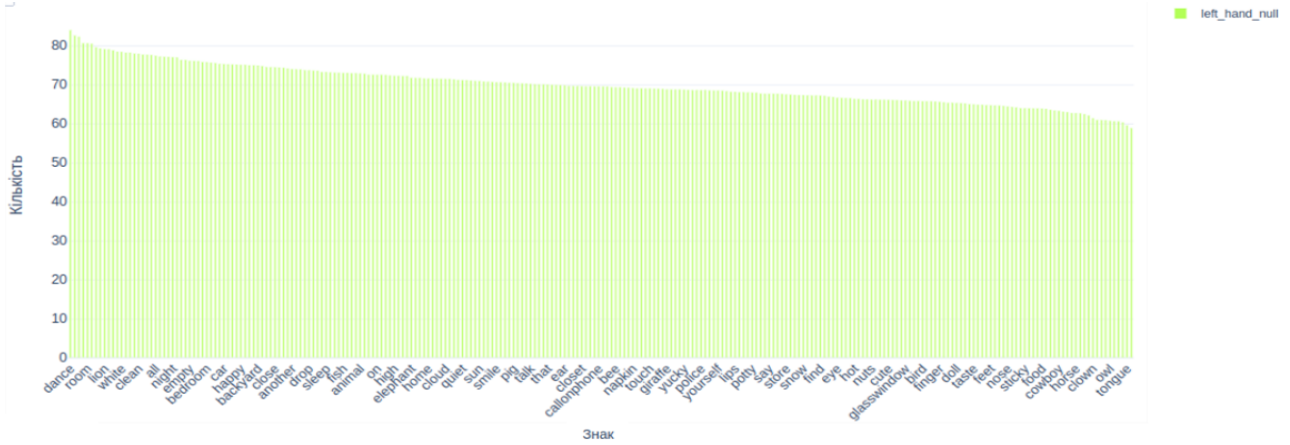

***Рис. 2.10. Розподіл кількості екземплярів для всіх класів***

Як видно з рис. 2.10, кількість екземплярів в основному є в межах від 369 до 391 одиниць. Але також є викиди, які показують що в деяких класах є менша кількість відео. Найменша кількість екземплярів, а саме 299, присутня для класу «zipper». Однак, слід зауважити що більшість класів мають однакову кількість екземплярів, що свідчить про рівномірно розподілену базу даних.

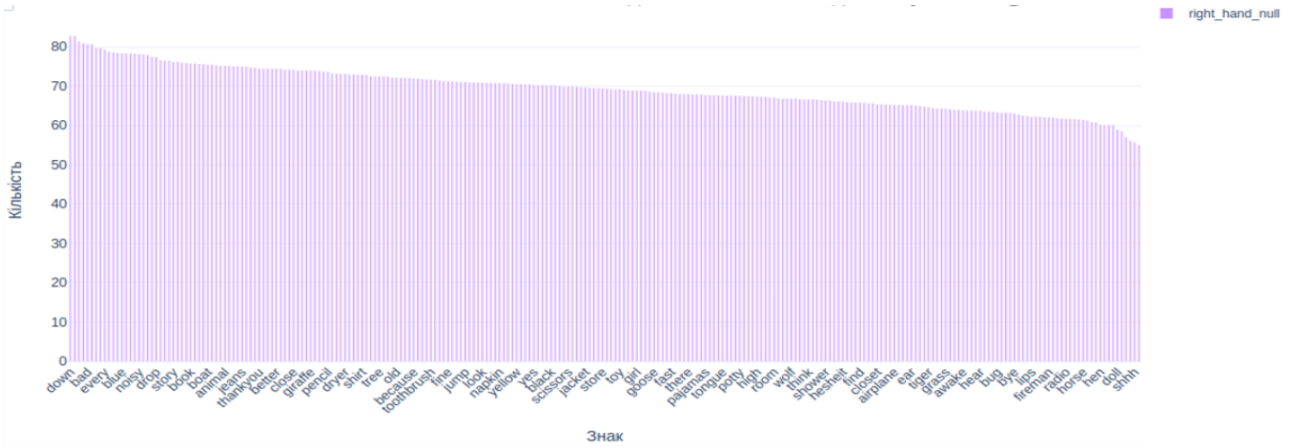
Також було досліджено відсоткова кількість «None» значень для кожного типу точок. У цій базі даних є чотири типу точок, а саме: лице, поза, права та ліва руки. На рис. 2.11 представлено результат аналізу.




*(а)*




*(б)*




*(в)*

***Рис. 2.11. Відсоткова кількість відсутніх значень для типу точок: (а) – лиця; (б) – лівої руки; (в) – правої руки***

Хоч вище було зазначено про чотири типи точок, одна подано три графіки. Графік для типу точок pose є відсутнім, оскільки цей тип точок ніколи не має «None» значень. Face має не велику кількість відсутніх значень, це пов’язано з перекриванням швидше за все рукою лице і зменшенням кількості розпізнаваних

точок. Як видно з поданих графіків вище, в основному часто відсутні точки на правій та лівій руках.  Точки на лівій руці частіше відсутні ніж на правій. Також клас у якому мало присутніх точок певної руки, не збігаються з іншою. Отже, для багатьох символів потрібно використовувати лише одну руку, що й утворює велику кількість «None» для іншої.

Для кращого розуміння як виглядає певний жест, на рис 2.12 представлено приклад демонстрації класу face.

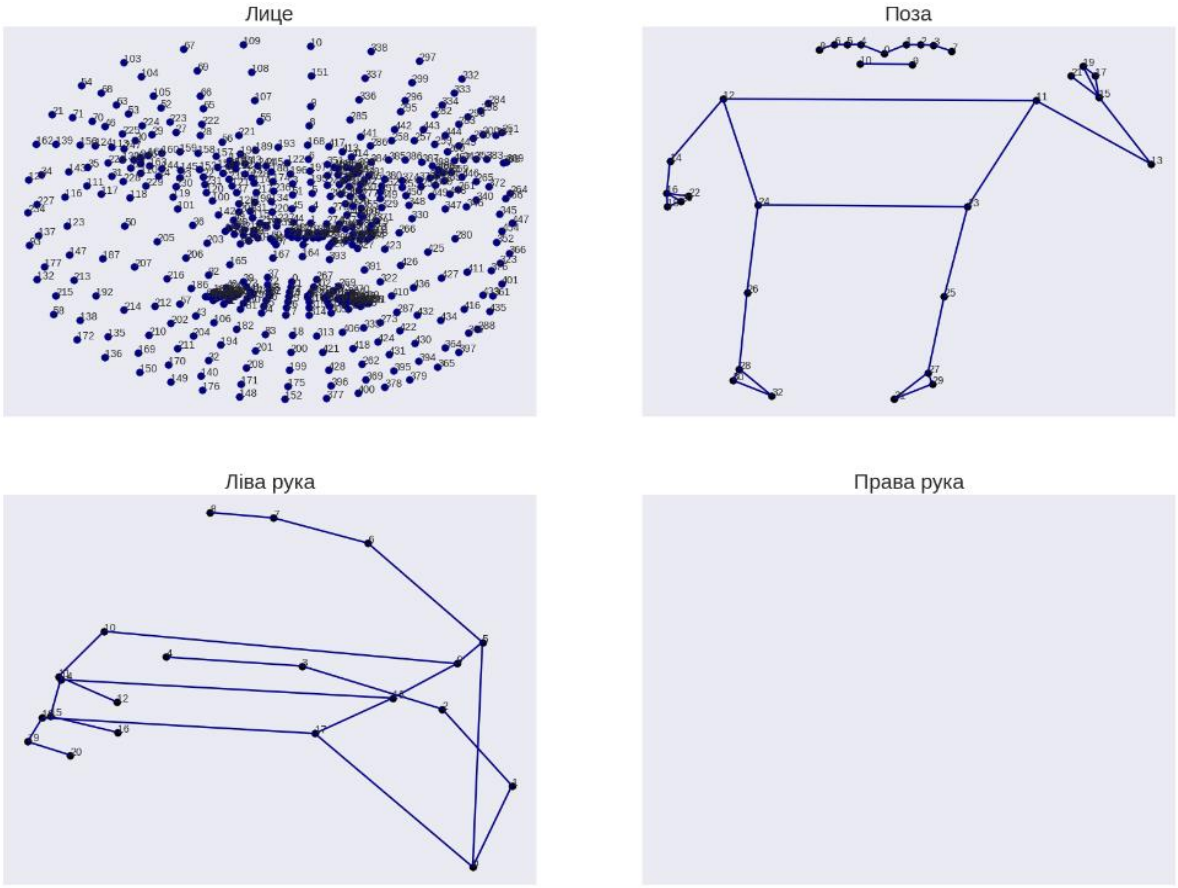



***Рис. 2.12. Графічне відображення ключових точок жесту «лице» на різних частинах тіла***

З зображення 2.11 видно, що велика кількість точок використовується для опису обличчя - 468, для пози — 31, для рук — 21. Представлення правої руки відсутнє, оскільки для демонстрації цього жесту, вистачає лише однієї руки. Загальне представлення цього кадру, продемонстроване на рис. 2.13.

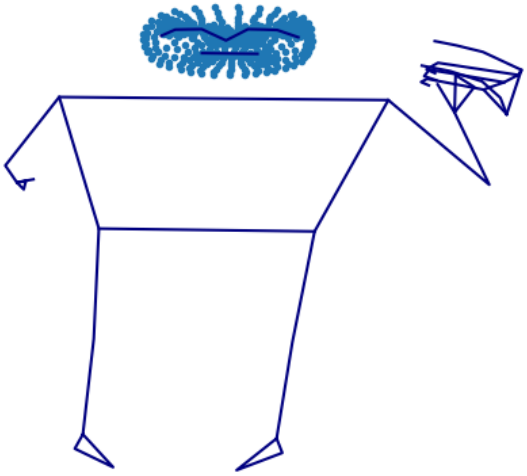


***Рис. 2.13. Сумісне представлення жесту «лице»***

Цей датасет відіграє важливу роль у вивченні та розробці систем розпізнавання ASL, дозволяючи дослідникам навчати та валідувати свої моделі на великій кількості реальних жестів. Це сприяє подальшому розвитку технологій розпізнавання жестів та покращенню комунікації для осіб, які використовують ASL як мову спілкування.

## 2.6. Висновки

Зважаючи на постійний розвиток технологій у сфері машинного навчання та штучного інтелекту, розпізнавання мови жестів стає ключовим напрямком досліджень. У цьому розділі дипломної роботи вивчалися та проаналізовані різні методи, що використовуються для цієї мети. Особливу увагу приділено нейронним мережам, таким як LSTM, ConvLSTM та трансформери, що демонструють значні переваги у впровадженні систем розпізнавання жестів.

Значний акцент був зроблений на вивченні основних принципів роботи цих методів та їх практичного застосування в контексті розпізнавання мови жестів. Досліджено важливість виявлення закономірностей між різними кадрами, оскільки динамічні жести часто потребують зв'язку між послідовними кадрами для точного розпізнавання.

Крім того, особлива увага приділена вивченню функцій втрат, таких як перехресна ентропія зі згладжуванням міток та ArcFace. Ці функції виявилися ключовими у покращенні точності системи розпізнавання, особливо у випадках, коли кількість класів є значною, і перенавчання стає проблемою.

Окрім цього, проведено детальний аналіз структури даних та використання бібліотеки Mediapipe для представлення відеоданих у вигляді ключових точок. Це дозволило зберігати та обробляти лише необхідну інформацію, що є важливим у вдосконаленні роботи моделей розпізнавання жестів.

Аналіз розподілу даних підтвердив рівномірний характер бази даних, однак підкреслив важливість попередньої обробки даних, особливо у зв'язку з варіативною кількістю фреймів для різних класів.

Цей розділ підкреслює важливість обрання ефективних методів та стратегій для створення систем розпізнавання мови жестів. Використання потужних алгоритмів нейронних мереж та якісне представлення даних забезпечують високу точність та надійність у функціонуванні цих систем.

# 3. РОЗДІЛ АПРОБАЦІЙ ТА РЕЗУЛЬТАТІВ

## 3.1. Попередня обробка даних.

Під час попередньої обробки даних для дипломної роботи важливо враховувати нюанси датасету. Цей етап є фундаментальним, оскільки від нього залежить точність та ефективність подальшого аналізу та класифікації. Нижче представлено докладний опис методів та процедур, що використовуються на цьому етапі.

Починаючи з основи нашої дослідницької роботи, необхідно звернутися до структури бази даних. База складається з великої кількості файлів, які містять інформацію про ключові точки, які характеризують різні аспекти об'єктів на кожному кадрі. При цьому ця інформація додатково супроводжується номером фрейму, оскільки дані представляють послідовність кадрів. Зображення структури бази даних подано на рис 3.1.

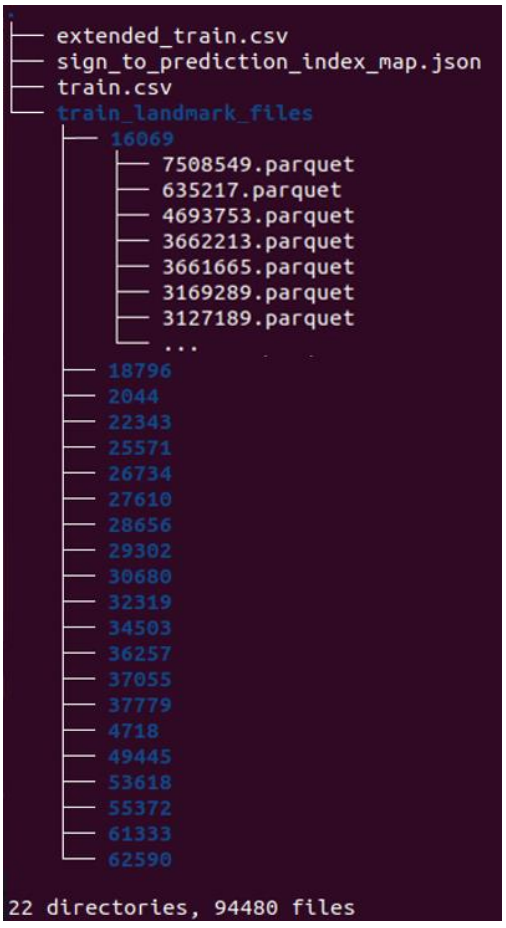



***Рис. 3.1. Структура бази даних***

Структура датасету передбачає розділення даних на 22 окремі папки,

кожна з яких відповідає конкретному учаснику. Кожна папка містить набір екземплярів жестів, які були записані певним спостерігачем. Важливо відзначити, що ця структура датасету є інформаційною та організаційною складовою, яка дозволяє зберегти ідентичність та особливості кожного учасника окремо. Кожен з них має свій унікальний "ID," який визначає папку, де зберігаються їх жести. Окрім цього, є файли у корневій папці, які представляють інформацію про структуру та класи датасету.

Для подальшого аналізу та обробки даних важливо спростити їхню структуру та зменшити розмірність. На Рис. 3.2 наведено кроки попередньої обробки даних.

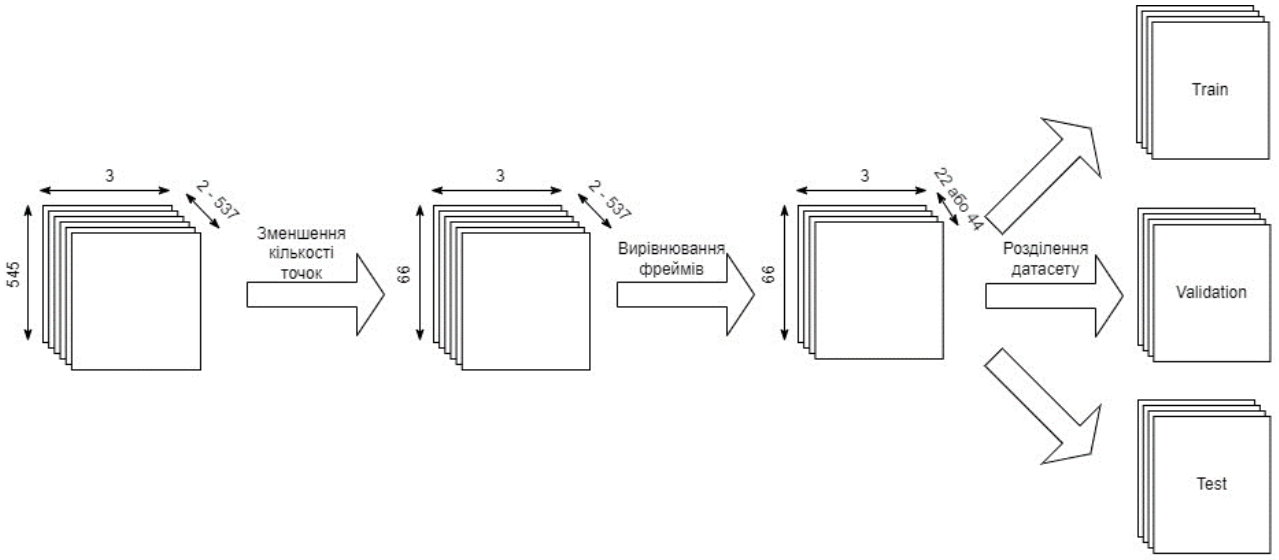



***Рис. 3.2. Представлення покрокової обробки даних***

На першому етапі обробки даних важливо ретельно розглянути кількість точок, які відображаються в датасеті, і здійснити необхідну редукцію. У цьому випадку, кількість точок зменшується із початкових 545 до 66, що є кроком в сторону оптимізації обробки та аналізу даних. Ця редукція охоплює етап видалення зайвих точок, які не несуть суттєвої інформації для класифікації жестів. Насамперед зберігаються лише ті точки, які надають інформацію про форму та положення губ, які є фундаментальними для розпізнавання жестів. Також вибрана необхідна кількість точок для опису руки. На Рис. 3.3 зображено точки, які були збережені.

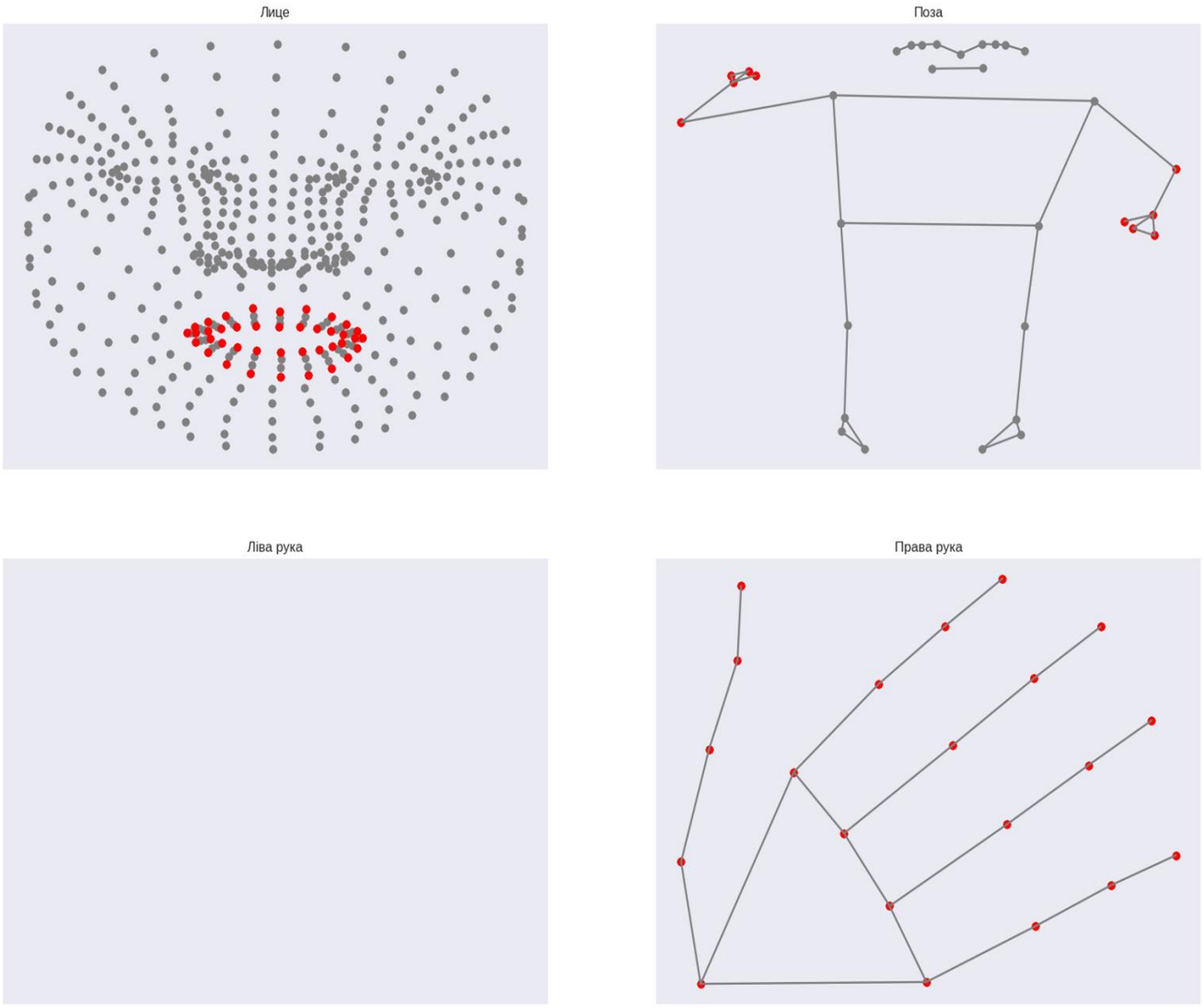



***Рис. 3.3. Вибрані ключові точки для опису кожної з частин тіла***

Червоним кольором позначені точки, які зберігаються для подальшого аналізу. Однак навіть після цієї операції кількість точок залишається значною. Для подальшого скорочення кількості ознак вводиться етап розділення на праву та ліву сторони. Кожен символ у наборі даних репрезентується лише однією з рук, залежно від того, чи є учасник правшою чи лівшою. Ця інформація визначається відносно кількості нульових значень (None) і вибирається лише одна з сторін для подальшого аналізу. Проте для того, щоб забезпечити стабільність та об'єктивність результатів аналізу, проводиться додаткова обробка. На рис 3.4 представлений коробчастий графік, який ілюструє розподіл точок для правої сторони (від точки 21 до 40) та лівої сторони (від точки 0 до 20).

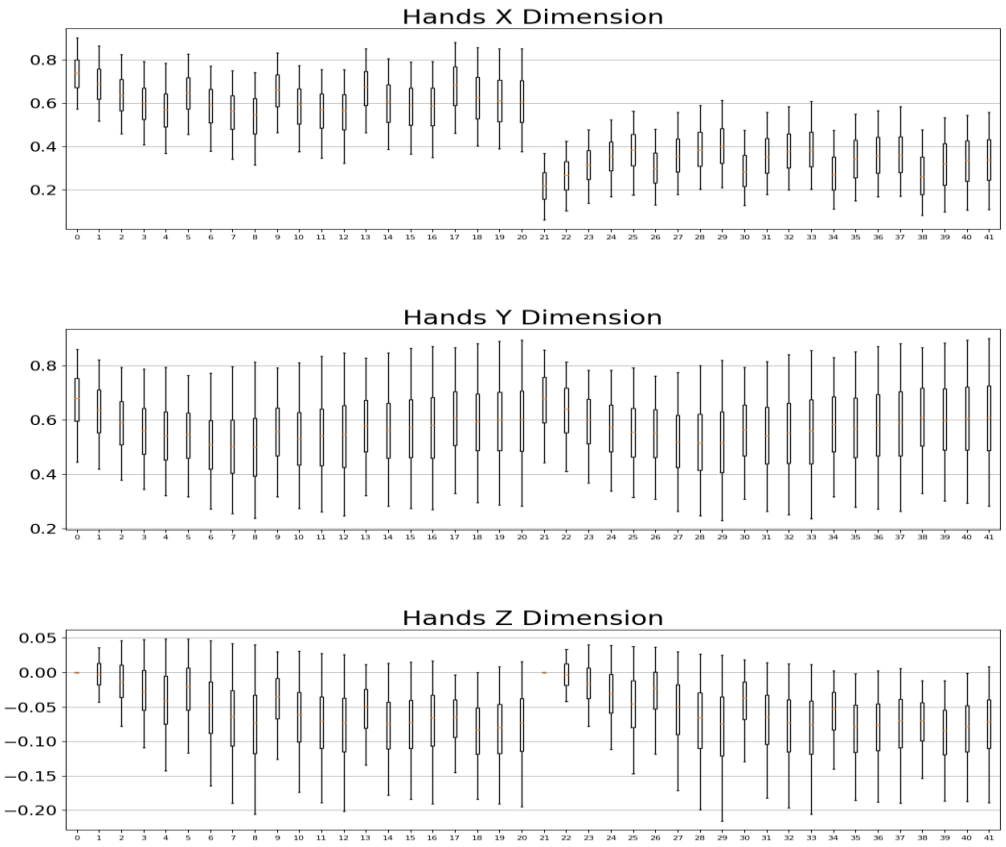



***Рис. 3.4. Розподіл ключових точок лівої та правої руки для кожної з осі***

Щоб нормалізувати зразки потрібно віддзеркалити праву сторону відносно осі x. Важливо відзначити, що ця операція віддзеркалення застосовується лише до координат вздовж осі X, оскільки розподіл на рис. 3.4 вказує на те, що координати вздовж осей Y та Z залишаються незмінними для обох лівої та правої руки. Центр координат у просторі даних розташований у точці 0.5. Для відзеркалення правої сторони додаткові 0.5 додаються до значень координат усіх 21 ключової точки, що відповідають правій руці. Важливо зазначити, що ця операція віддзеркалення застосовується лише до семплів, в яких права рука домінує, тобто до семплів, де більшість рухів виконується правою рукою.

Після віддзеркалення обидві руки стають подібними, і нейронна мережа може більш ефективно вивчати патерни жестів, оскільки вона не залежить від того, чи є рука правою, чи лівою. Це допомагає підвищити робастність та універсальність моделі, забезпечуючи більш точні результати класифікації незалежно від сторони руки, з якої виконується жест.

У підсумку, результуючий набір даних містить 40 точок для опису обличчя (зокрема, губ), 21 точку для опису однієї з рук, та 5 точок для опису правої чи

лівої сторони тулуба, що складає загалом 66 точок.

Наступним важливим кроком у процесі обробки даних є вирівнювання кількості фреймів. Кількість фреймів у наборі даних суттєво змінюється в залежності від конкретного символа та учасника, який демонструє його. Ця різниця в кількості фреймів може впливати на подальший аналіз та обробку даних, тому важливо провести вирівнювання. На рис. 3.5 можна побачити графік, який ілюструє розподіл кількості фреймів у цьому датасеті. Цей графік показує, що кількість фреймів коливається в діапазоні від 2 до 537.

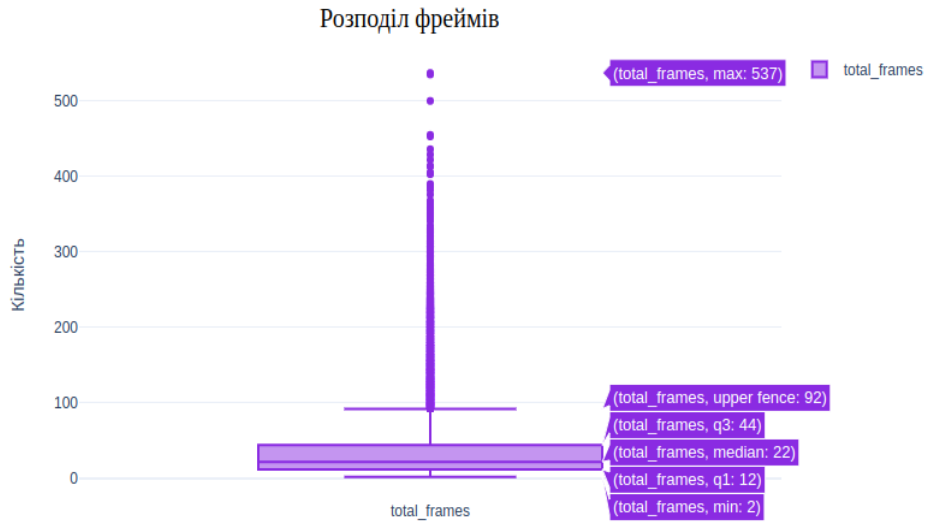



***Рис. 3.5. Коробчастий графік розподілу кількості фреймів***

Для забезпечення стабільності та зручності подальшої обробки, приведено кількість фреймів до медіани – 22 фрейми та третього квартиля - 44 фрейми. Цей вибір обумовлений розглянутою статистикою, де виявлено, що 22 фрейми охоплює 50% значень, а 44 фрейми охоплюють 75% всього об'єму даних в датасеті. Обрані значення для експериментів не є надмірно великим, що дозволяє уникнути зайвої кількості нульових значень, які можуть виникнути при заповненні бракуючих фреймів. З іншого боку, обрані значення не є занадто малим, щоб не обрізати важливу інформацію для символів. Враховуючи природу жестів та необхідність належного опису символів були обрані саме ці кількості фреймів, які повинні дозволити зберегти змістовну інформацію і забезпечити належну структуру для подальшого аналізу та класифікації. У наступному

розділі цієї робити, будуть представлені результати порівняння для цих кількостей фреймів.

Останнім кроком є розподіл даних на навчальну (80227) та валідаційну (9592) та тестову(4656) вибірки. Це розділення виконується з урахуванням ідентифікаційних номерів учасників з метою уникнути перетину даних між цими двома вибірками. Цей підхід гарантує об'єктивну оцінку точності моделі та запобігає змішуванню впливу особистих характеристик учасників на результати класифікації символів.

Усі ці операції разом створюють надійний та готовий до подальшого аналізу набір даних, який допоможе досягти об'єктивних та точних результатів у класифікації символів.

### 3.2. Представлення архітектури мереж.

Перейдемо до розгляду архітектури нейронних мереж для розв’язання цієї задачі. Завдяки аналізу літературних джерел було вибрано підходи, які продемонстрували високі результати. Серед них - повнозв’язні шари, LSTM, ConvLSTM та трансформери.

#### *3.2.1. Архітектура вбудування.*

Однак, перед застосуванням цих підходів, хорошою практикою є використання вбудування. Основна перевага використання Embedding полягає в тому, що воно здатне враховувати контекст та взаємозв'язки між різними елементами даних, підвищуючи точність та швидкість обробки інформації нейронною мережею. На рис. 3.6 представлено архітектуру вбудування.

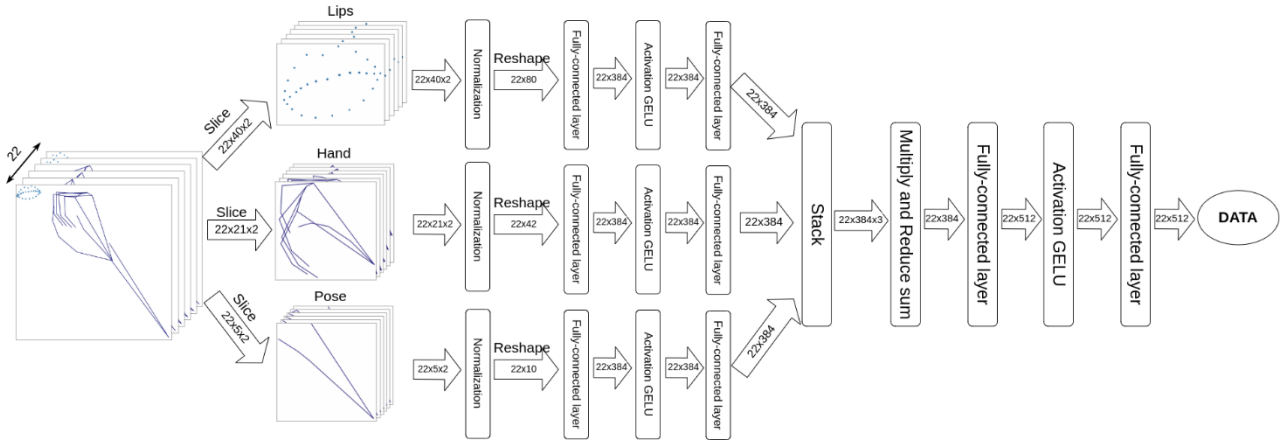



***Рис. 3.6 Архітектура вбудування для розпізнавання мови жестів***

У контексті обробки даних мови жестів спочатку проводиться сегментація даних залежно від типів точок, таких як губи, руки та поза. Наступним кроком є нормалізація даних для кожної групи точок. Цей процес полягає в перетворенні значень ключових точок лиця, тіла або рук у відповідність до середнього значення та середнього відхилення всіх точок цієї групи. Нормалізація здійснюється за наступною формулою:

$$z = \frac{x - \mu}{\sigma} \tag{3.1}$$

- $x$ – це матриці значень ключових точок лиця, тіла або рук.
- $\mu$ – середнє значення всіх точок лиця, тіла або рук
- $\sigma$ – середнє відхилення всіх точок лиця, тіла або рук

Як видно, процес нормалізації застосовується не універсально до всіх точок, а відносно кожної групи точок, визначаючи, до якої частини тіла належать ці точки. Це дозволяє зберегти правильний розподіл даних і запобігає втраті інформації.

Після нормалізація, дані передаються у вбудування. Спочатку кожна точка, яка має початкові розміри, еквівалентні кількості кадрів помноженій на кількість точок конкретного типу на кількість вимірів, проходить через повнозв'язний шар. Потім застосовується активація GELU, а результат знову проходить повнозв'язний шар. Після цих операцій всі три типи даних мають однакові розміри, а саме кількість кадрів на 384.

Після об'єднання трьох типів даних у матрицю розмірністю кількість фреймів на 384 на 3, наступним етапом є зменшення розмірності цієї матриці. Це досягається через використання операцій, таких як reduce sum, які спрямовані на зменшення обсягу даних шляхом обчислення суми значень по певних вимірах матриці. Після цього використовуються повнозв'язні шари, які проводять операції над даними для подальшого скорочення розмірності.

В результаті цих операцій розмірність даних зменшується до кількості фреймів на 512. Ця розмірність підготовлює дані для передачі до нейронних мереж у вигляді, який краще підходить для подальшого аналізу та обробки в мережах глибокого навчання. Цей процес змінення розмірності даних є

ключовим у підготовці вхідних даних для нейронних мереж, допомагаючи забезпечити оптимальну продуктивність та ефективність в процесі розпізнавання мови жестів.

### *3.2.2. Мережа на основі повнозв'язних шарів.*

Одним з перших типів мереж у цій роботі є повнозв'язна мережа. Зазначена нейронна мережа використовувалася для створення базового рівня або базової моделі, що слугує стартовим пунктом в дослідженні. Вона є вихідною точкою для порівняння з іншими складнішими чи покращеними архітектурами. Це дозволяє оцінити ефективність нових методів чи модифікацій, порівнявши їх з цією базовою моделлю. Архітектура використаної мережі представлена на рис. 3.7.

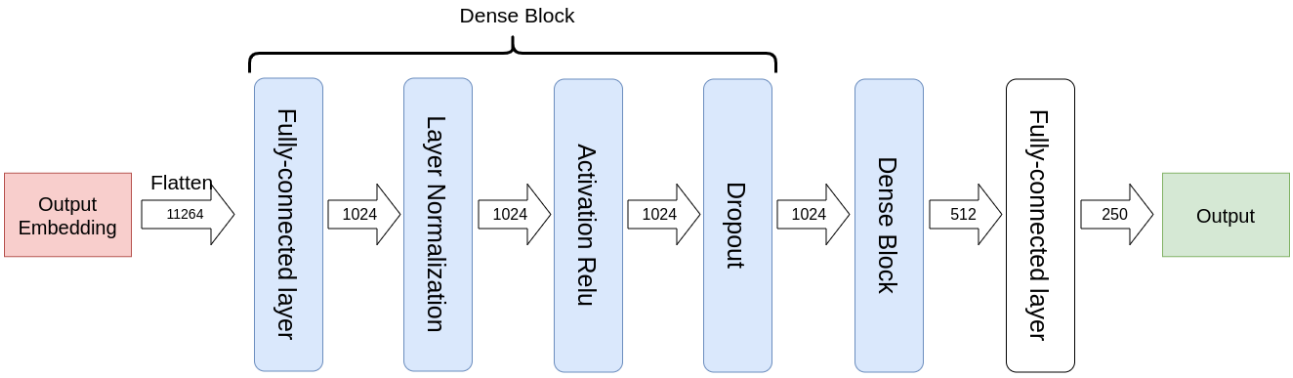



***Рис. 3.7 Архітектура повнозв'язної нейронної мережі***

Ця нейронна мережа приймає на вхід вже оброблені дані через вбудування, яке перетворило початкові дані жестів у вектори чисел фіксованої довжини. Після вбудування дані проходять через декілька шарів для подальшої обробки та аналізу.

Перший шар цієї мережі - Dense (повнозв'язний) - має 1024 нейрони. Після цього шару використовується шар нормалізації, що допомагає у стандартизації вихідних значень та регулюванні їх діапазону. Після шару нормалізації застосовується активаційна функція, яка активує нейрони в шарі, і Dropout, який випадково вимикає окремі нейрони для запобігання перенавчанню. Потім іде наступний повнозв'язний шар з 512 нейронами, після якого знову використовується шар нормалізації, активаційна функція та Dropout.

Останній шар у цій моделі - це Dense шар з 250 нейронами, що відповідає кількості класів жестів у базі даних. Цей шар генерує остаточний вихід, який представляє ймовірності належності вхідних даних до різних класів жестів. Кожен нейрон у цьому останньому Dense шарі відповідає конкретному класу жесту, а його активаційна функція надає інформацію про ймовірність того, що вхідні дані належать до цього класу.

У цілому, ця архітектура мережі складається з послідовного розміщення Dense шарів, шарів нормалізації, активаційних функцій та dropout шарів, які поетапно обробляють вхідні дані для отримання остаточного результату - передбачення класів жестів.

### *3.2.3. Мережа на основі LSTM.*

Наступна архітектура мережі побудована з використанням LSTM, через її можливість пошуку залежностей між фреймів, яке є важливим у передбачуванні жестів, де жест являють собою набір кадрів. На рис. 3.8 зображена архітектура цієї мережі.

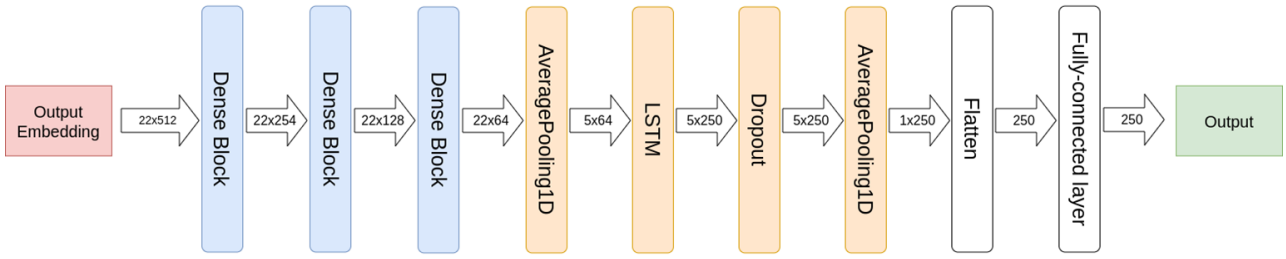



***Рис. 3.8 Архітектура мережі на основі LSTM***

Наведена нейронна мережа складається з кількох шарів, що виконують операції обробки даних для класифікації американської мови жестів після вбудування. Починаючи з Dense шару з 254 нейронами та розмірністю (Batch size, 22, 254), який слідує за вбудованим шаром. Потім використовується шар нормалізації, цей шар застосовується для балансування значення нейронів, далі застосовується активаційна функція, що надає нелінійність даним. Після активації в мережі використовуються dropout шари, які "вимикають" певні нейрони для уникнення перенавчання. Далі використовується ще два Dense шари зі 128 та 64 нейронами відповідно, зі своїми шарами нормалізації та dropout.

Після останнього dropout шару, дані проходять через шари з пониженням розмірності, такі як average_pooling1d, які виконують операцію середнього згортання для зменшення розмірності даних. Потім використовується LSTM шар, що є рекурентним шаром для роботи з послідовностями даних. Цей шар призначений для зберігання інформації про послідовність жестів, відображаючи їх залежності в часі. Він допомагає у врахуванні динаміки жестів, оскільки дані можуть представляти послідовність рухів.

Останніми шарами в мережі є dropout, який зменшує перенавчання для LSTM шару, і шар flatten, що перетворює тривимірний вихід з LSTM у вектор фіксованої довжини. Нарешті, останній повнозв'язний шар має 250 нейронів, що відповідають кількості класів жестів, і генерує остаточний результат передбачення класів жестів.

Ця архітектура має декілька переваг. Використання LSTM дає можливість моделі працювати з послідовностями даних, що дозволяє зберігати динаміку жестів у часі. Рекурентний підхід також поліпшує розуміння контексту жестів та їх зв'язку в часі. Проте, є деякі недоліки в цій архітектурі. Обчислювальна складність з використанням LSTM та додаткових шарів зниження розмірності може збільшити обчислювальний обсяг мережі, що потребує більше обчислювальних ресурсів та часу для тренування. Крім того, цей підхід може призвести до перенавчання моделі на великій кількості даних, що вимагає застосування додаткових стратегій регуляризації для уникнення цього ефекту.

### *3.2.4. Мережа на основі ConvLSTM.*

Наступна мережа є подібна до попередньої, однак має свої відмінності. Архітектура мережі представлена на рис. 3.9.

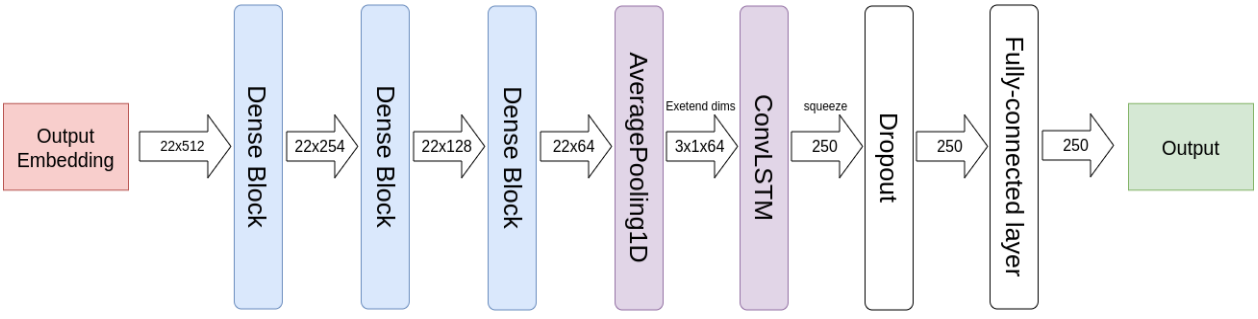



***Рис. 3.9 Архітектура мережі на основі ConvLSTM***

Наведена нейронна мережа також призначена для класифікації американської мови жестів після вбудування даних. Ця модель має послідовну структуру, яка містить кілька Dense шарів, шари нормалізації, функції активації та dropout шари.

Починаючи з Dense шару з 254 нейронами та розмірністю (Batch size, 22, 254), вхідні дані проходять через нормалізаційний шар, який вирівнює значення нейронів, та застосовується активаційна функція, що надає нелінійність даним. Після цього використовуються dropout шари для уникнення перенавчання. Потім в мережі використовуються ще два Dense шари зі 128 та 64 нейронами відповідно, зі своїми шарами нормалізації та dropout.

Після останнього dropout шару, дані проходять через average_pooling1d, який виконує операцію середнього згортання для зменшення розмірності даних. Потім застосовується tf.expand_dims, що розширює розмірність даних для використання у ConvLSTM1D.

Останнім шаром є ConvLSTM1D, який являє собою комбінацію згорткового шару і шару LSTM. Цей шар здатний враховувати просторові та часові взаємозв'язки в даних. Після цього використовується tf.compat.v1.squeeze для стискання розмірності даних. Останній повнозв’язний шар має 250 нейронів, що відповідають кількості класів жестів у базі даних, і генерує остаточний результат передбачення класів жестів.

Ця архітектура нейронної мережі, що використовує шари ConvLSTM1D після вбудування, має деякі специфічні позитивні та негативні аспекти, аналогічні до LSTM-моделі. Щодо переваг, варто відзначити, що використання ConvLSTM1D шарів дозволяє моделі враховувати як просторові, так і часові залежності в даних. Це може бути корисним для розпізнавання жестів, які мають важливу послідовність або залежать від динаміки в часі. Крім того, комбінація різних типів шарів (Dense, normalization, dropout) дозволяє моделі вивчати складніші залежності у вхідних даних.

Однак, така архітектура може мати певні недоліки. Обчислювальна складність зростає через використання ConvLSTM1D та додаткових шарів зменшення розмірності, що може призвести до збільшення вимог до

обчислювальних ресурсів під час тренування та використання моделі. Крім того, підвищений обсяг даних, а також недостатня кількість даних для тренування, можуть спричинити перенавчання моделі, особливо у випадку складних архітектур з великою кількістю параметрів.

### *3.2.5. Мережа на основі архітектури трансформер.*

Остання архітектура, яка розглядається для цієї робота це трансформер, на основі лише енкодера. Архітектура цієї мережі представлена на рис. 3.10.

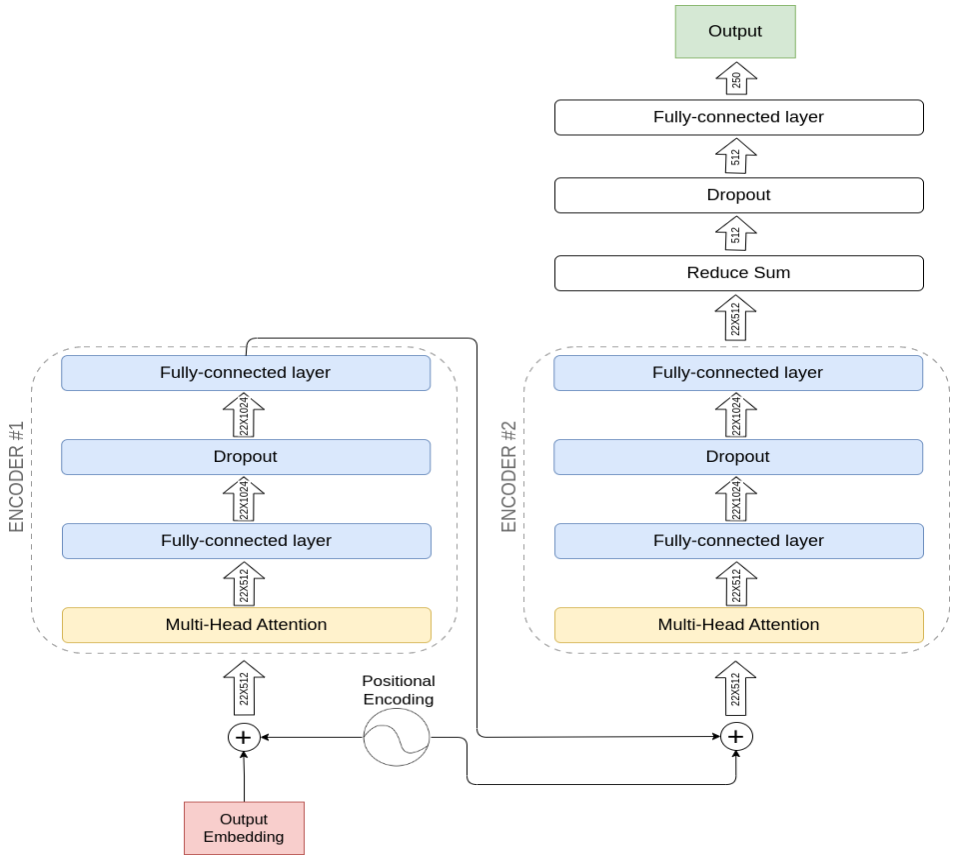



***Рис. 3.10 Архітектура трансформ***

Пройдемо до детального огляду архітектури нейронної мережі. Відзначимо наявність двох типів вхідних даних, які спершу піддаються спеціальній обробці, а саме — positional encoding та output embedding. Перший тип даних служить для передачі інформації про наявність даних. Якщо значення цього параметра рівне -1, то цей кадр складається виключно з нулів, що свідчить про відсутність даних для даного фрейму. Така ситуація виникає внаслідок штучного заповнення нулями, яке використовується для вирівнювання розмірностей даних. Вхідні дані, що містять інформацію про наявність або

відсутність даних, необхідні для подальшої маскування. Маскування є методом, за допомогою якого шари послідовної обробки повідомляються про відсутність даних в певних часових кроках вхідних даних, і тому ці кроки слід пропустити під час подальшої обробки. На рис. 3.11 зображена попередня обробка позиційного кодування.

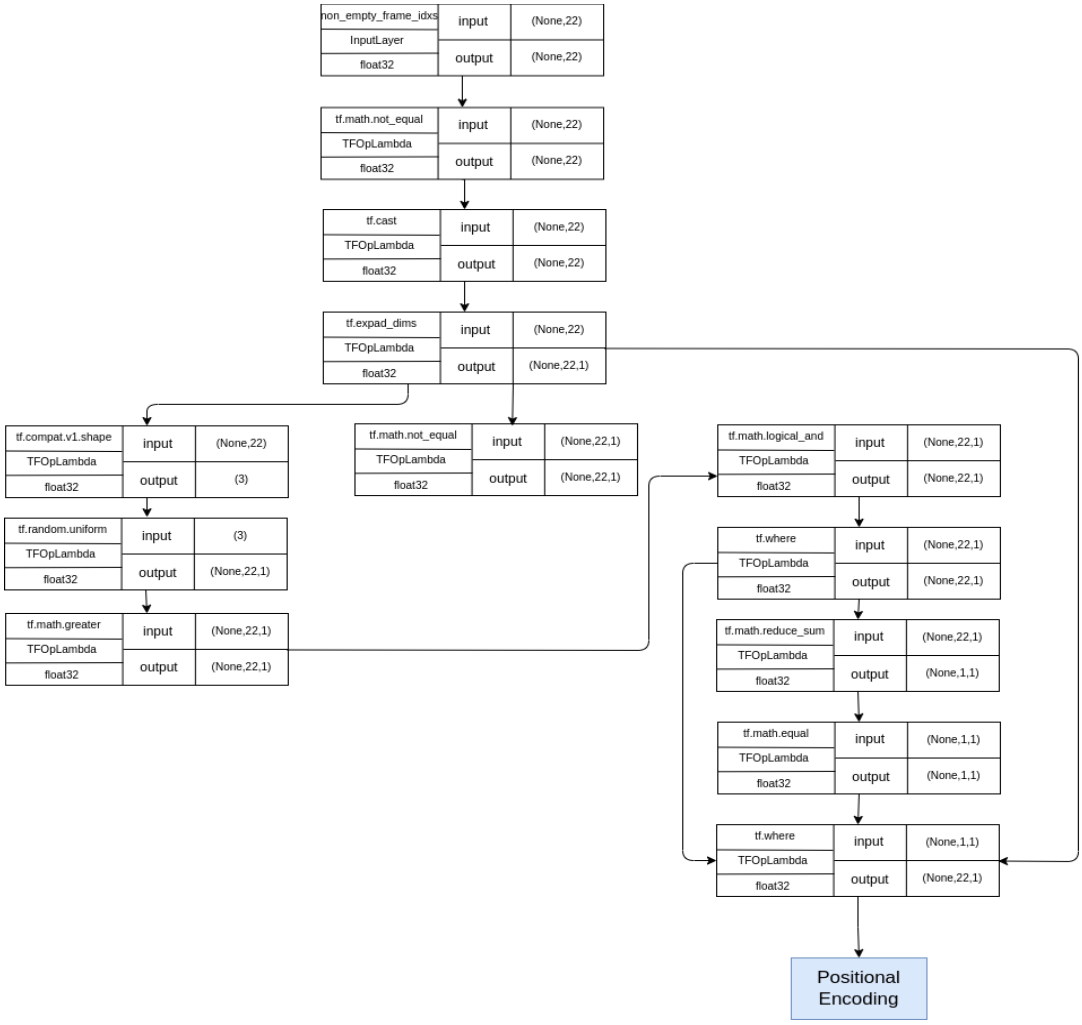



***Рис. 3.11 Попередня обробка позиційного кодування***

Першим етапом є процес рандомізація фреймів для маскування. Це потрібно для відключення деяких кадрів у процесі навчання моделі. Це може бути корисним для регуляризації моделі та запобігання перенавчання. Основна мета полягає в тому, щоб навчити модель працювати з частково відсутніми або випадково відключеними кадрами, що може покращити її загальну здатність до узагальнення на нових даних.

Останнім кроком перед використанням трансформера для цих даних є перевірка, чи всі кадри вибірки були випадково замасковані через певну

ймовірність. Це може відбутися, оскільки один з кроків вище, є рандомізований вибір кадрів, які не мають відсутніх значень. Через це, може статися випадок, коли усі кадри стали зашумленні, тобто рівні 0. Це є проблематичним, оскільки таке маскування може призвести до втрати важливої інформації в зразку. Тому проводиться додаткова перевірка за допомогою tf.math.reduce_sum, яка обчислює суму значень в масці для кожного зразка. Якщо сума для якого-небудь зразка дорівнює 0 (тобто всі кадри для цього зразка зашумлені), то такий зразок буде відновлено до своїх оригінальних значень.

Після ретельного аналізу позиційного кодування, яке насамперед спрямована на процес створення маски, варто відзначити, що вбудування для трансформера також має враховувати позиційне кодування.

Перейдемо до аналізу самої архітектури трансформера. Трансформер є критичною компонентою моделі, відповідальною за обробку та взаємодію з вхідними векторами. Він складається з таких основних компонентів:

- Multi-Head Attention (Багатоголовий механізм уваги): Ця складова є однією з ключових частин трансформера та використовується для моделювання взаємодії між різними частинами вхідних даних. У структурі трансформера існують два таких блоки, кожен з яких відповідає за пошук закономірностей між кадрами;
- Sequential (Послідовний блок): Ця частина трансформера відповідає за послідовну обробку даних після блоків механізму уваги. Вона включає кілька шарів для обробки та структуризації інформації. Кожен послідовний блок містить два Dense шари та шар Dropout з метою регуляризації та запобігання перенавчанню. Також, аналогічно до попереднього компонента, трансформер включає два блоки цієї частини, кожен з яких відповідає за обробку вихідної інформації відповідно до першої або другої "голови" уваги.

Загальною метою цього трансформера є виконання операцій оптимізації та взаємодії між векторами, що подаються на вхід, з метою підготовки їх до фінального класифікатора та передачі результатів для обчислення остаточних прогнозів моделі.

Останні кроки в моделі відповідають за обробку результатів трансформера, їх нормалізацію, регуляризацію, та генерацію остаточних класифікаційних прогнозів для вхідних даних.

### 3.3. Експерименти.

На підставі розглянутих архітектур проведено серію експериментів, в яких варіювалася кількість фреймів – 22 та 44 – а також застосування глибини (врахування координати z) та відсутність цієї характеристики. Окрім варіації даних, були випробувані різноманітні функції втрат, які були детально описані у розділі 2. У таблиці 3.1 представлені результати, отримані за допомогою мережі, заснованої на повнозв'язних шарах. Такий підхід дозволив детально проаналізувати вплив різних параметрів на результативність моделі та зробити висновки щодо її ефективності у різних умовах.

**Таблиця 3.1.** Результати експериментів з повнозв'язною моделю

| | 22 Фрейми без z координатою | 22 Фрейми з z координатою | 44 Фрейми без z координатою | 44 Фрейми з z координатою |
|---|---|---|---|---|
| Повнозв'язна нейронна мережа | 59.47% | 60.18% | 56.83% | 57.07% |
| Повнозв'язна нейронна мережа + ArcFace функція втрат | **61.64%** | 61.00% | 57.73% | 57.45% |
| Повнозв'язна нейронна мережа + функція втрат зі згладжуванням міток | 59.89% | 60.87% | 58.93% | 59.97% |
| Повнозв'язна нейронна мережа + функція втрат зі згладжування міток + ArcFace функція втрат | 58.63% | 60.78% | 57.97% | 55.36% |

Загалом, спостерігається, що використання більшої кількості фреймів (44 замість 22) призводить до певного зниження ефективності моделі навіть у випадках з врахуванням та без урахування координати z. Однак при використанні деяких функцій втрат, збільшення кількості фреймів може

призвести до покращення результатів.

У порівнянні зі звичайною повнозв'язною нейронною мережею, додавання ArcFace функції втрат зазвичай призводить до підвищення точності класифікації в усіх режимах. Функція втрат зі згладжуванням лейблів також демонструє певне покращення результатів в порівнянні з базовою моделлю, особливо у випадку більшої кількості фреймів. Крім того, комбінація цієї функції втрат з ArcFace може зменшити ефект перенавчання, сприяючи збалансованійшій результативності моделі при використанні обох функцій втрат одночасно. Графіки точності при навчанні для різних моделей для 22 фреймів без координатою z представлені на рис. 3.12.

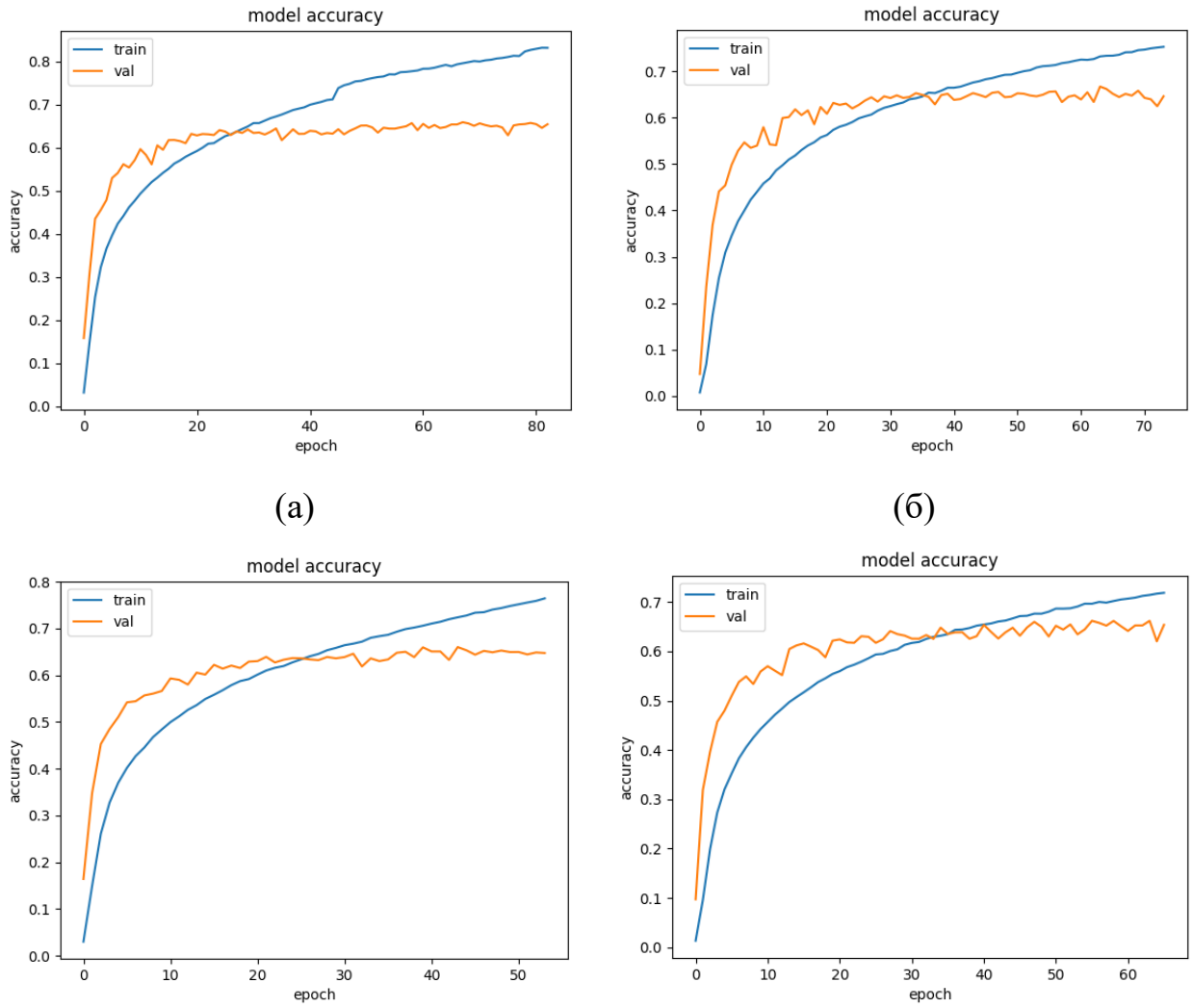



***Рис. 3.12 Графік точності під час тренування для повнозв'язної нейронної мережі на даних з 22 фреймами та без координати z: (а) – перехресна***

***ентропія; (б) – ArcFace функція втрат; (в) - функція втрат зі згладжування міток; (г) – змішана функція втрат***

Отже, використання відповідних функцій втрат дозволяє покращити точність моделі, а їх комбінація сприяє зменшенню ефекту перенавчання та підвищенню її стабільності. Перейдемо до розгляду результатів наступної мережі.

**Таблиця 3.2.** Результати експериментів з LSTM моделю

| | 22 Фрейми без z координатою | 22 Фрейми з z координатою | 44 Фрейми без z координатою | 44 Фрейми з z координатою |
|---|---|---|---|---|
| LSTM | 57.2% | 56.19% | 58.23% | 57.02% |
| LSTM мережа + ArcFace функція втрат | 57.02% | 56.85% | 58.08% | 58.25% |
| LSTM мережа + функція втрат зі згладжування лейблів | 57.37% | 57.04% | **59.39%** | 56.72% |
| LSTM мережа + функція втрат зі згладжування лейблів + ArcFace функція втрат | 57% | 56.57% | 59.17 | 57.78% |

Порівнюючи результати цієї моделі з попередньою, видно, що застосування LSTM-мереж у цьому контексті показує деяку стабільність, але меншу точність порівняно з розглянутою раніше архітектурою на основі повнозв’язних шарів у всіх конфігураціях.

Додавання змінної z (глибини) при обробці фреймів показує певне погіршення результатів у всіх випадках застосування LSTM-мережі, проте цей спад не є значним. Найкращі результати для всіх функцій втрат спостерігаються при використанні 44 фреймів без координат глибини. Перейдемо до розгляду функцій втрат. На рис. 3.13 представлено графіки навчання.

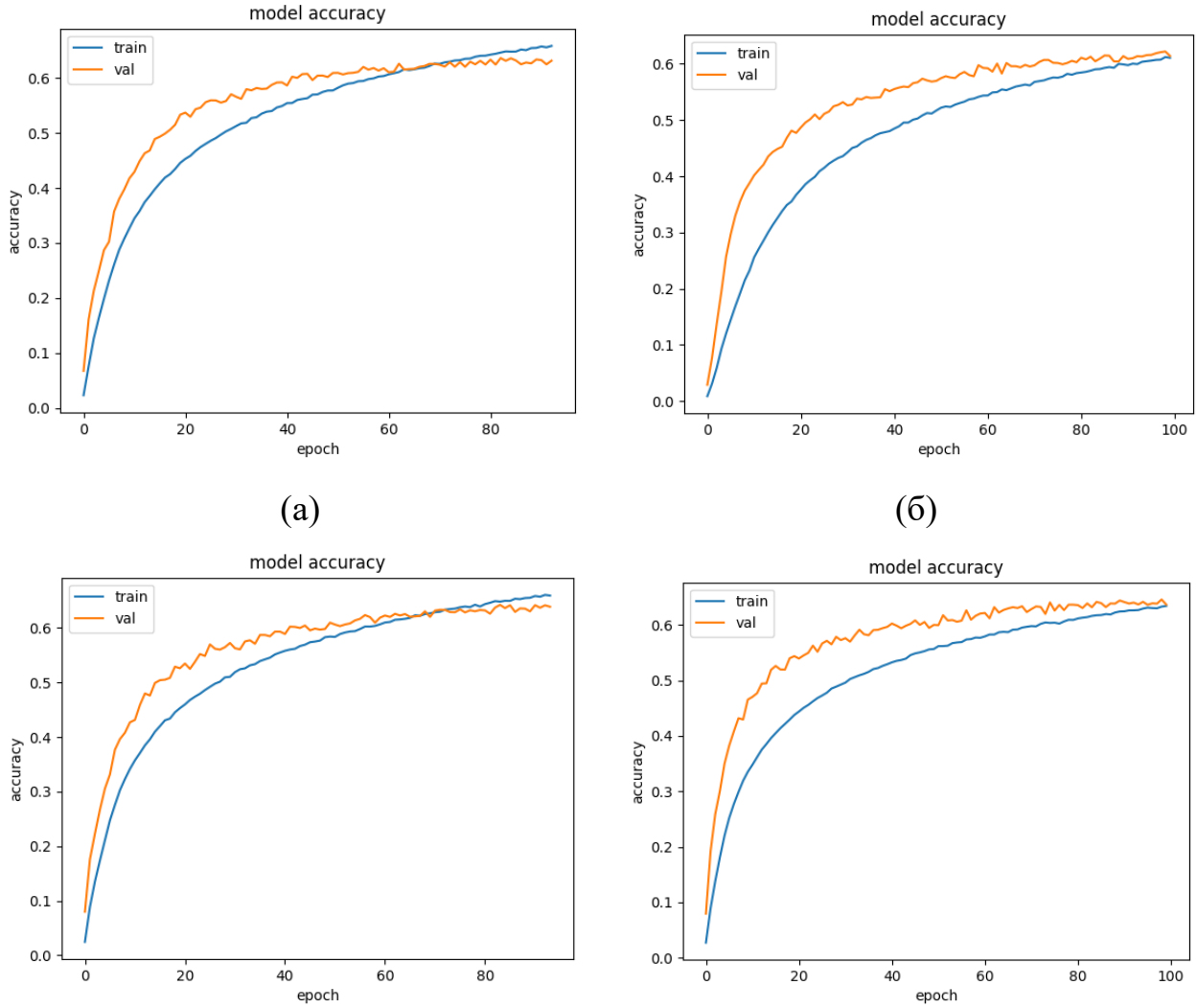



***Рис. 3.13 Графік точності під час тренування для LSTM мережі на даних з 44 фреймами та без координати z: (а) – перехресна ентропія; (б) – ArcFace функція втрат; (в) - функція втрат зі згладжування міток; (г) – змішана функція втрат***

Як видно з графіків та таблиці, використання ArcFace функції втрат у поєднанні з LSTM-мережею виявилося корисним у запобіганні перенавчанню. Ця функція допомагає зберігати стабільність моделі і запобігає надмірному підгонюванню до тренувальних даних, що може бути спостережено при використанні інших видів функцій втрат.

Функція втрат зі згладжуванням лейблів також демонструє деякий контроль над перенавчанням, зменшуючи його ефект при використанні разом з LSTM-мережею. Однак, при поєднанні з ArcFace функцією втрат спостерігається більш стабільна робота мережі, що проходить всі 100 епох навчання без

вираженого перенавчання.

Загально, LSTM-мережа спільно з функціями втрат виявляє себе більш стійкою до перенавчання при збалансованому врахуванні параметрів, таких як кількість фреймів та використання різних функцій втрат. Однак точність є дещо меншою. Причиною цього може бути гірший підбір параметрів порівняно з попередньою моделю.

У таблиці 3.3 представлені результати наступної мережі.

**Таблиця 3.3.** Результати експериментів з ConvLSTM моделю

| | 22 Фрейми без z координатою | 22 Фрейми з z координатою | 44 Фрейми без z координатою | 44 Фрейми з z координатою |
|---|---|---|---|---|
| ConvLSTM | 56.21% | 54.02% | 57.95% | 53.72% |
| ConvLSTM мережа + ArcFace функція втрат | 57.75% | 54.55% | **58.98%** | 56.53% |
| ConvLSTM мережа + функція втрат зі згладжування лейблів | 55.86% | 55.73% | 57.03% | 54.12% |
| ConvLSTM мережа + функція втрат зі згладжування лейблів + ArcFace функція втрат | 55.22% | 54.89% | 56.65% | 54.94% |

З аналізу представлених результатів можна зробити декілька спостережень у порівнянні з попередніми моделями LSTM та простою мережею на основі повнозв'язних шарів.

ConvLSTM-мережі виявилися менш ефективними у випадку класифікації американської мови жестів порівняно з попередніми моделями. Їхні показники точності залишаються на нижчому рівні у всіх варіаціях даних, незалежно від кількості фреймів чи включення z-координати.

Комбінація ConvLSTM з ArcFace функцією втрат показала деяке покращення у результативності, особливо при 22 фреймах без z-координати. Проте, це покращення все ще не дозволило досягти рівня точності, який має проста повнозв'язна модель або LSTM-мережа.

Використання функції втрат зі згладжуванням лейблів, окремо чи разом з ArcFace, не виявилося вирішальним для покращення результатів ConvLSTM-моделі. Моделі залишилися менш ефективними в порівнянні з аналогічними підходами до LSTM та простих повнозв'язних мереж.

Такі результати вказують на те, що в даній задачі архітектура з використанням ConvLSTM має певні обмеження у здатності до вивчення відображення між жестами та класифікацією в порівнянні з більш традиційними підходами, такими як LSTM чи прості повнозв'язні моделі. Перейдемо до результатів наступної моделі.

**Таблиця 3.4.** Результати експериментів з трансформером

| | 22 Фрейми без z координатою | 22 Фрейми з z координатою | 44 Фрейми без z координатою | 44 Фрейми з z координатою |
|---|---|---|---|---|
| Трансформер | 69.63% | 69.16% | 68.08% | 67.31% |
| Трансформер + ArcFace функція втрат | 67.20% | 67.37% | 66.65% | 52.58% |
| Трансформер + функція втрат зі згладжування лейблів | **71.92%** | 71.65% | 70.36% | 70.19% |
| Трансформер + функція втрат зі згладжування лейблів + ArcFace функція втрат | 67.80% | 68.36% | 66.32% | 66.17% |

Ці нові дані демонструють, що архітектура трансформера значно перевершує у точності всі попередні моделі, такі як ConvLSTM, LSTM і прості повнозв'язні мережі, навіть при різних умовах (кількість фреймів та наявність координати z). Трансформер показав стабільну та високу точність, що може бути пов'язане з його здатністю ефективно моделювати взаємодії між даними в часових послідовностях.

Застосування ArcFace функції втрат до трансформера показує деяке зниження точності в порівнянні з базовою моделлю без цієї функції. Це може вказувати на те, що в деяких випадках використання ArcFace може вводити

більше шуму або перевантаження моделі.

Варто зауважити, що використання функції втрат зі згладжуванням лейблів покращує точність трансформера у всіх випадках, що свідчить про ефективність цього методу регуляризації та поліпшення узагальнюючої здатності моделі.

Найнижчі результати у випадку комбінації функцій втрат зі згладжуванням лейблів і ArcFace функції втрат. Це може вказувати на деяку несумісність або взаємодію між цими двома методами, що може погіршувати загальну ефективність моделі. Розглянемо також графіки навчання для даних, на яких модель продемонструвала найвищу точність, а саме 22 фрейми та без використання координати z.

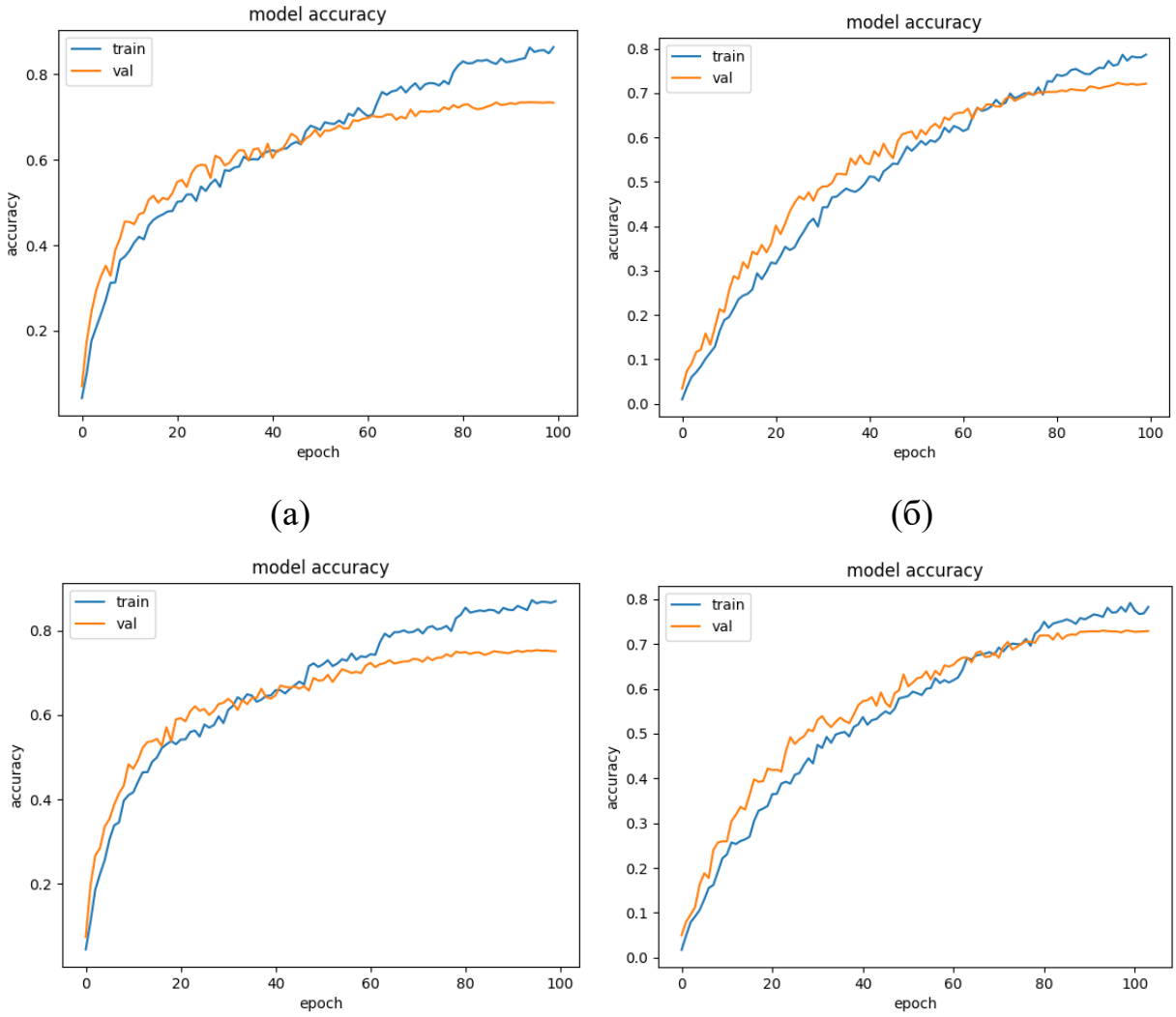



***Рис. 3.14 Графік точності під час тренування для тансформер мережі на даних з 22 фреймами та без координати z: (а) – перехресна ентропія; (б) –***

*ArcFace функція втрат; (в) - функція втрат зі згладжування міток; (г) – змішана функція втрат*

Загальна тенденція відносно функцій втрат у всіх моделях є наступною: просте використання крос ентропії та ентропії зі згладжуванням лейблів призводить до виявлення ознак перенавчання. З іншого боку, ці функції можуть забезпечити хороші результати на тестових даних. Для покращення результатів варто провести додаткові експерименти з використанням різних методів регуляризації, для зменшення перенавчання та покращення точності.

В цілому, трансформерна архітектура демонструє високий потенціал у розпізнаванні жестів порівняно з попередніми моделями. Проте, оптимізація використання функцій втрат та їх комбінацій може бути ключовою для підвищення її ефективності та уникнення перенавчання.

### 3.4. Ансамбль моделей.

Використання ансамблю моделей є важливим компонентом для покращення точності та робастності системи розпізнавання мови жестів. Ансамбль моделей є групою або комбінацією різних індивідуальних моделей машинного навчання, які працюють разом для прийняття остаточного рішення. Цей підхід дозволяє використовувати різноманітність та комбіновану експертизу декількох моделей для отримання кращого результату.

Ансамбль може використовувати як групу різних моделей з різними архітектурами та параметрами, так і кілька копій однієї архітектури моделі зі зміненими параметрами або навчанням на різних підмножинах даних. Кожна модель може мати свої сильні та слабкі сторони у вирішенні конкретних завдань, та поєднання їх усіх може призвести до покращення загального результату.

Техніка голосування або узгодження рішень дозволяє ансамблю приймати остаточне рішення на основі багатьох індивідуальних прогнозів. Це може бути голосування більшості (коли більшість моделей підтримує певний клас) або зважене голосування, де кожна модель має свою вагу в прийнятті рішення.

У таблиці 3.5 представлені результати точності натренерованих моделей на даних з 22 фреймів та без координати z. Їх конфігурації є подібними і не мають

сильних відмінностей. Архітектура цих мереж — трансформери, які продемонстрували найвищі результати.

**Таблиця 3.5.** Результати експериментів з трансформером зі зміною кількістю розміру пакета та регуляризацією

| | Accuracy | Top-5-accuracy | Top-10-accuracy |
|---|---|---|---|
| Модель 1 (Розмір пакету 64) | 71.59% | 90.7% | 93.9% |
| Модель 2 (Розмір пакету 128) | 72.47% | 90.57% | 93.47 |
| Модель 3 (Розмір пакету 128 + регуляризація L2) | 73.3% | 90.49% | 93.54% |
| Модель 4 (Розмір пакету 258) | 71.8% | 89.84% | 93.3% |
| Модель 5 (Розмір пакету 258 + регуляризація L2) | 71.43% | 90.14% | 93.34% |

Окрім просто точності, у таблиці представлено результати top-5-accuracy та top-10-accuracy. Top-5-accuracy вимірює відсоток прикладів, для яких правильна мітка (клас) знаходиться серед п'яти (top-5) найбільш ймовірних прогнозів моделі. Це означає, що модель може не вгадати точний клас, але якщо правильний клас знаходиться серед п'яти найбільш ймовірних варіантів, це все ще вважається правильним.

Top-10-accuracy аналогічно вимірює точність моделі, але враховує верхні десять (top-10) найбільш ймовірних прогнозів моделі. Оцінка top-n для кожної моделі в ансамблі підтверджує їхню високу точність у виборі правильних класів серед найбільш ймовірних варіантів. Це підкреслює потенціал ансамблю моделей для поєднання їх сильних сторін та досягнення ще кращих результатів у класифікації.

На рис 3.15 представлено графік залежності точності ансамблю моделей відносно кількості моделей у ньому.

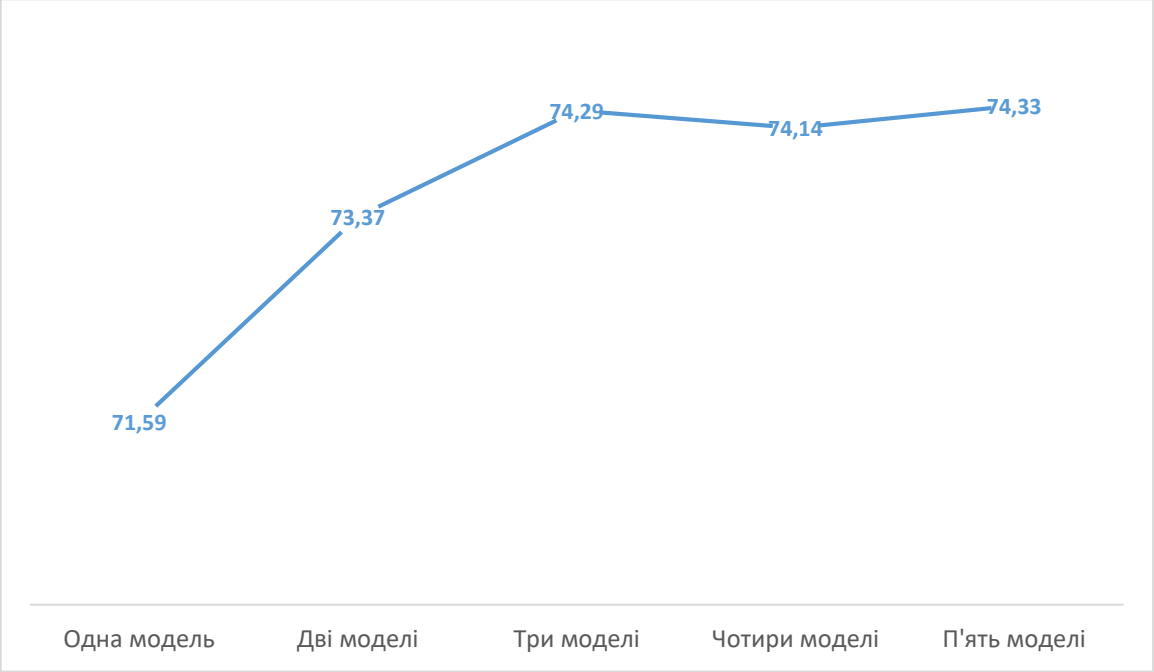



***Рис. 3.15 Графік залежності кількості об’єднань моделей до точності передбачення***

Зазначені точності представляють результати ансамблю моделей з різними кількостями входящих моделей. Ці результати показують, як точність змінюється зі збільшенням кількості об'єднаних моделей у вигляді ансамблю.

- Об’єднання двох моделей (73.37%): Точність ансамблю з двох моделей склала 73.37%, що вказує на певне підвищення точності порівняно з окремими моделями, але не настільки значуще.
- Об’єднання трьох моделей (74.29%): За додавання третьої моделі точність ансамблю зросла до 74.29%, виявивши певне поліпшення у прогнозуванні.
- Об’єднання чотирьох моделей (74.14%): При включенні ще однієї моделі точність ансамблю зменшилася до 74.14%. Це може бути пов’язано з певними особливостями та взаємодією моделей у складі ансамблю.
- Об’єднання п’яти моделей (74.33%): Додавання п’ятої моделі призвело до збільшення точності ансамблю до 74.33%, що може вказувати на певне покращення точності через додаткову варіативність або сильні сторони цієї моделі.

Невідомо, чи варто збільшувати кількість моделей у складі ансамблю далі,

оскільки зростання кількості моделей може збільшити точність, але разом з цим значно збільшить час обчислень. Тому, хоча збільшення кількості моделей може підвищити точність, важливо зберігати баланс між точністю та швидкодією моделі, враховуючи потреби конкретного застосування.

### 3.5. Висновки

Розпізнавання мови жестів є складним завданням, яке передбачає кілька критичних етапів для досягнення точності моделі. Один із важливих кроків - обробка даних, яка забезпечує відповідність та підготовку даних для подальшого використання у моделі. У цьому розділі було проведено низку операцій для адаптації даних до необхідного формату та розміру. Одним із важливих аспектів було зменшення обсягу даних, оскільки не всі точки, що описують риси обличчя та пози, є вирішальними для розпізнавання жестів.

Після обробки даних розглядалися різні архітектури моделей, такі як повнозв’язні мережі, LSTM, ConvLSTM та трансформери. Кожна з цих архітектур має свої особливості та складові, які були детально описані. Внаслідок експериментів, проведених з цими моделями, було виявлено, що трансформери проявили себе краще у розпізнаванні мови жестів у порівнянні з іншими архітектурами. Ймовірно, це пов’язано з їхньою здатністю моделювати довгострокові залежності та взаємодії між жестами завдяки використанню механізмів уваги.

Експерименти з різними функціями втрат показали, що просте використання крос-ентропії та ентропії зі згладжуванням лейблів може призвести до перенавчання моделей. Однак, ці функції можуть надати високі результати на тренувальних даних. Для подальшого покращення точності моделей рекомендується використовувати методи регуляризації з метою зменшення перенавчання.

У контексті ансамблю моделей трансформерів відзначено, що з збільшенням кількості моделей у складі ансамблю зросла точність прогнозування. Однак, варто враховувати, що збільшення кількості моделей може збільшити обчислювальні витрати. Тому, для збереження балансу між

точністю та ефективністю моделі, необхідно ураховувати специфіку конкретного застосування.

# ВИСНОВКИ

Розпізнавання мови жестів - складна і актуальна тема, яка має великий потенціал у сучасному світі. Ця область досліджень привертає увагу науковців та інженерів через свої перспективи в медицині, технологіях спілкування, віртуальній реальності та інших сферах. Незважаючи на цільові переваги, розпізнавання мови жестів стикається з численними викликами, такими як варіативність жестів у різних культурах, складність розпізнавання рухів в різних умовах освітлення та індивідуальних особливостей.

Враховуючи розмаїття викликів, що виникають у процесі розпізнавання мови жестів, важливо визначити ключові аспекти, що впливають на його результативність. У першому розділі дослідження акцент був зроблений не лише на важливості розпізнавання мови жестів для покращення комунікації людей з обмеженими можливостями, а й на вивченні актуальних проблем, що виникають у цій галузі та способах їх вирішення. Проведений критичний аналіз літератури та методів розпізнавання мови жестів дозволив виокремити декілька ключових аспектів. Однією з цих проблем є вплив зовнішніх умов, таких як освітлення та інші фактори, на точність розпізнавання жестів. Це може призвести до нестабільності у визначенні жестів та погіршення якості розпізнавання в різних умовах.

У рамках розділу було розглянуто існуючі методи розпізнавання мови жестів та їх ефективність у вирішенні зазначених проблем. Виявлено, що хоча існують певні досягнення у цій галузі, проте є потреба у подальших дослідженнях та вдосконаленнях, особливо з урахуванням вищезгаданих викликів. Таким чином, перший розділ не лише підкреслив важливість теми розпізнавання мови жестів для поліпшення комунікації, а й розкрив існуючі проблеми і шляхи їх вирішення, що свідчить про актуальність і потенціал подальших досліджень у цій області.

У другому розділі дослідження було проведено глибокий аналіз та порівняння різних методів розпізнавання жестів, зокрема тих, що ґрунтуються на нейронних мережах, таких як LSTM, ConvLSTM та трансформери. Визначено, що виявлення закономірностей між кадрами є критичним аспектом для точного

розпізнавання динамічних жестів. Використання бібліотеки Mediapipe виявилося важливим для підвищення точності системи розпізнавання мови жестів. Ця бібліотека дозволяє ефективно виявляти ключові точки у відеоданих, що представляють жести, та зменшувати об'єм даних, які використовуються для подальшого аналізу. Це стає важливим у випадках, коли обчислювальні ресурси обмежені, а також допомагає у зменшенні впливу зовнішніх умов, таких як освітлення, на процес розпізнавання жестів.

У третьому розділі дослідження було проведено обробку та адаптацію даних для моделювання, а також порівняльний аналіз різних архітектур моделей та функцій втрат. Вибір трансформерів як більш ефективної архітектури для розпізнавання жестів був обумовлений їхньою здатністю моделювати довгострокові залежності між жестами.

Під час експериментів проводилися тести з різною кількістю фреймів (22 та 44) та з використанням координати z та без неї. Виявлено, що повнозв'язна мережа та трансформер показали кращі результати при використанні 22 фреймів та без координати глибини, досягнувши точності 61.64% та 71.92% відповідно. У той же час, LSTM та ConvLSTM показали найвищі результати при використанні 44 фреймів та без координати глибини, однак продемонстрована точністю є лише 59.39% та 58.98% відповідно.

Разом з тим, варто згадати про ансамбль трансформерів, що продемонстрував точність на рівні 74.33%. Це перевищило результат від використання однієї моделі. Однак, важливо відзначити, що значна кількість моделей у складі ансамблю негативно впливає на швидкість отримання результатів. Такий підхід має свої переваги щодо точності, але потребує додаткових обчислювальних ресурсів та часу для отримання прогнозів.

У контексті подальшого вдосконалення системи розпізнавання мови жестів, можливі наступні кроки для підвищення точності та оптимізації процесу. Вдосконалення архітектури моделей через застосування регуляризації та проведення більшої кількості експериментів з метою вибору оптимальних параметрів. Додатково, розгляд можливості зміни стратегій попередньої обробки даних та тестування точності з різною кількістю ключових точок та фреймів

може сприяти у досягненні кращих результатів. Проте важливо зазначити, що обмежені обчислювальні ресурси ускладнюють проведення більшого обсягу експериментів та збільшення об'єму даних, необхідного для покращення моделей.

Вирішення проблеми розпізнавання мови жестів має велике значення. Подальший прогрес у цій галузі може принести істотні переваги для людей з обмеженими можливостями, розвиток технологій віртуальної реальності, медицинських технологій та інших сфер. Розробка ефективних систем розпізнавання мови жестів відкриває шлях до більш ефективного спілкування та інтеграції різних груп користувачів у сучасному світі.

# СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ


1. Deafness and hearing loss [Електронний ресурс] – Режим доступу https://www.who.int/news-room/fact-sheets/detail/deafness-and-hearing-loss. .
2. Papastratis I., Chatzikonstantinou C., Konstantinidis D., Dimitropoulos K., Daras P. Artificial Intelligence Technologies for Sign Language. *Sensors (Basel, Switzerland)*. 2021. Вип. 21, № 17. С. 5843.
3. Boyko N., Hrynyshyn A. Using Recurrent Neural Network to Noise Absorption from Audio Files. .
4. Sutton-Spence R., Woll B. The Linguistics of British Sign Language: An Introduction. Cambridge University Press, 1999. ISBN 9781139167048.
5. Dubey P., Shrivastav M. P. Iot based sign language conversion. *International Journal of Research in Engineering and Science (IJRES)*. 2021. Вип. 9, № 2. С. 84–89.
6. Snoddon K. Wendy Sandler & Diane Lillo-Martin, Sign language and linguistic universals. Cambridge: Cambridge University Press, 2006. Pp. xxi, 547. Pb $45.00. *Language in Society*. 2008. Вип. 37, № 4. С. 628–628.
7. LINGUIST List 13.1631: Cognitive Science: Emmorey (2002). 2002.
8. Tervoort B. T. Sign language: the study of deaf people and their language: J.G. Kyle and B. Woll, Cambridge, Cambridge University Press, 1985. ISBN 521 26075. ix+318 pp. *Lingua*. 1986. Вип. 70, № 2. С. 205–212.
9. Stokoe W. C. Sign Language Structure: An Outline of the Visual Communication Systems of the American Deaf. *Journal of Deaf Studies and Deaf Education*. 2005. Вип. 10, № 1. С. 3–37.
10. Christopoulos: Sign language - Google Академія. .
11. Emmorey K. Language, Cognition, and the Brain: Insights From Sign Language Research. Psychology Press, 2001. 402 с. ISBN 978-1-135-66481-7.
12. Stokoe W. C., Casterline D. C., Croneberg C. G. A Dictionary of American Sign Language on Linguistic Principles. Linstok Press, 1976. 396 с. ISBN 978-0-932130-01-3.
13. Why are there different types of sign language? 2023.
14. S R. Blog - Different Types of Sign Language Used Around the World. 2020.

15. Papatsimouli M., Lazaridis L., Kollias K.-F., Skordas I., Fragulis G. F. Speak with signs: Active learning platform for Greek Sign Language, English Sign Language, and their translation. arXiv, 2020.
16. Pezzuoli F., Tafaro D., Pane M., Corona D., Corradini M. L. Development of a New Sign Language Translation System for People with Autism Spectrum Disorder. *Advances in Neurodevelopmental Disorders*. 2020. Вип. 4, № 4. С. 439–446.
17. Shubankar B., Chowdhary M., Priyaadharshini M. IoT Device for Disabled People. *Procedia Computer Science*. 2019. Вип. 165. С. 189–195.
18. Ізонін І.В., Мочурад Л.І., Ткаченко Р.О., Шиманський В.М. Методичні вказівки до виконання контрольної роботи для студентів спеціальності 122 «Комп'ютерні науки» першого (бакалаврського) рівня вищої освіти. – Львів: Видавництво Національного університету «Львівська політехніка», 2023. – 15 с. 2020.
19. Scopus. 2023.
20. Google Scholar. 2023.
21. Li D., Opazo C. R., Yu X., Li H. Word-level Deep Sign Language Recognition from Video: A New Large-scale Dataset and Methods Comparison. Snowmass Village, CO, USA:IEEE, 2020. ISBN 978-1-72816-553-0.
22. Kumwilaisak W., Pannattee P., Hansakunbuntheung C., Thatphithakkul N. American Sign Language Fingerspelling Recognition in the Wild with Iterative Language Model Construction. *APSIPA Transactions on Signal and Information Processing*. 2022. Вип. 11, № 1.
23. Amer Kadhim R., Khamees M. A Real-Time American Sign Language Recognition System using Convolutional Neural Network for Real Datasets. *TEM Journal*. 2020. С. 937–943.
24. Jing L., Vahdani E., Huenerfauth M., Tian Y. Recognizing American Sign Language Manual Signs from RGB-D Videos. arXiv, 2019.
25. Jalal M. A., Chen R., Moore R. K., Mihaylova L. American Sign Language Posture Understanding with Deep Neural Networks. 2018.
26. Jaiswal P. Automatic ASL Gesture Recognition System Using Convolutional

Neural Network. *Asian Journal For Convergence In Technology (AJCT) ISSN-2350-1146*. 2018.

27. Pratama Y., Marbun E., Parapat Y., Manullang A. Deep convolutional neural network for hand sign language recognition using model E. *Bulletin of Electrical Engineering and Informatics*. 2020. Вип. 9, № 5. С. 1873–1881.
28. Xue Q., Li X., Wang D., Zhang W. Deep Forest-Based Monocular Visual Sign Language Recognition. *Applied Sciences*. 2019. Вип. 9, № 9. С. 1945.
29. Hand Gesture Recognition for Disabled People Using Bayesian Optimization with Transfer Learning.
30. Sharma A., Mittal A., Singh S., Awatramani V. Hand Gesture Recognition using Image Processing and Feature Extraction Techniques. *Procedia Computer Science*. 2020. Вип. 173. С. 181–190.
31. Liu Y., Nand P., Hossain M. A., Nguyen M., Yan W. Q. Sign language recognition from digital videos using feature pyramid network with detection transformer. *Multimedia Tools and Applications*. 2023. С. 1–13.
32. Rahim M. A., Shin J., Yun K. S. Hand Gesture-based Sign Alphabet Recognition and Sentence Interpretation using a Convolutional Neural Network. *Annals of Emerging Technologies in Computing (AETiC)*. 2020. Вип. 4, № 4. С. 20–27.
33. Mannan A., Abbasi A., Javed A. R., Ahsan A., Gadekallu T. R., Xin Q. Hypertuned Deep Convolutional Neural Network for Sign Language Recognition. *Computational Intelligence and Neuroscience*. 2022. Вип. 2022. С. e1450822.
34. Calzone O. An Intuitive Explanation of LSTM. 2022.
35. Understanding LSTM Networks -- colah’s blog.
36. Xavier A. An introduction to ConvLSTM. 2019.
37. www.aionlinecourse.com What is Convolutional Long Short-Term Memory (ConvLSTM) | Ai Basics | Ai Online Course.
38. Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A. N., Kaiser L., Polosukhin I. Attention Is All You Need. arXiv, 2023.
39. Transformers in ML: What They Are and How They Work.
40. Rahman M. Enhancing Deep Learning Performance with Label Smoothing.

2023.
41. Mao L. Label Smoothing. 2019.
42. Deng J., Guo J., Xue N., Zafeiriou S. ArcFace: Additive Angular Margin Loss for Deep Face Recognition. 2019.
43. Google - Isolated Sign Language Recognition.
44. MediaPipe.
45. Face landmark detection guide | MediaPipe | Google for Developers.
46. Hand landmarks detection guide | MediaPipe.
47. Pose landmark detection guide | MediaPipe.

# Додаток А.

https://github.com/anastasiia43/Recognition-of-Sign-Language