УДК 520.6.05
DOI https://doi.org/10.32782/2663-5941/2022.5/03

***Лихошерстов Д.О.***
Національний технічний університет України
«Київський політехнічний інститут імені Ігоря Сікорського»

***Лебедев Д.Ю.***
Національний технічний університет України
«Київський політехнічний інститут імені Ігоря Сікорського»

# ПОРІВНЯЛЬНИЙ АНАЛІЗ ШЛЯХІВ ВИЗНАЧЕННЯ ЗОБРАЖЕННЯ НА ВІДЕОРЯДІ ЗАСОБАМИ МАШИННОГО НАВЧАННЯ

*У роботі проведено порівняльне дослідження механізмів класифікації інформації на зображені. Обґрунтовано, що жести є важливою складовою механізмів комунікації в суспільстві, керуванні сучасними технічними засобами та дозволяють покращити соціальний рівень людей із вадами слуху та мовлення за рахунок впровадження систем сурдоперекладу в побут. Зазначено, що процес обробки жестів використовує машинне навчання, котре має можливість в умовах реального часу класифікувати та виконувати ідентифікацію жестів на відеокадрах. Сформовані основні переваги та недоліки систем розпізнавання жестів та визначено задачу досліджень, котра полягає у порівнянні наявних механізмів розпізнавання та створенні власних.*

*Сформовано та обґрунтовано критерії оцінювання системи розпізнавання даних на зображені – точність ідентифікації, швидкодія, крос-платформеність, можливість апгрейду та масштабованості, легковажність системи та можливість підтримки готового рішення.*

*Створено нейромережу багатошарового персептрона із навчальної та тестової бази даних MNIST database. Проведено навчання та тестування нейромережі по визначення даних на зображені для трьох різних активаційних функцій – сигмоїдна функція, ModReLU (ReL з витоком) та тангенс. Встановлено, що ModReLU має найвищий рівень точності визначення.*

*Досліджено наявні фреймворки по розпізнаванню даних та визначено, що фреймворк MediaPipe від Google має крос-платформеність, відкритість коду та підтримку розробників, відкритість коду, а також дає високу точність визначення, особливо для класифікації жестів верхніх кінцівок в реальному часі.*

*Обґрунтовано та сформовані напрями для їх подальшого вдосконалення системи сурдоперекладу.*

***Ключові слова:*** *сурдопереклад, машине навчання, передача даних, розпізнавання жестів, нейромережа, MediaPipe, база даних MNIST database.*

**Постановка проблеми.** Передача інформації між людьми називається комунікація і може відбуватися за допомогою одного із трьох можливих способів – за допомогою тексту або мовлення або жестів. Слід зазначити, що жестова мова є повноцінним підходом для обміну інформації. Широковживані слова є можливість представити у вигляді готового набору жестів, а коли для якихось слів відсутні жести, то використовуючи дактильну абетку можна по буквах відтворити слово. Наочний приклад це американська жестова мова – American Sing Language і її база даних готових жестів – American Sing Language Online Dictionary [1].

Впровадження жестів в сучасний побут людей відбувається із розвитком технологій і на це є три об'єктивних причини. По-перше, жести виступають одним з основних механізмів передачі інформації. По-друге, розвиток доповненої (AR) [2] та віртуальної (VR) [3] реальності дозволило запровадити жестове керування девайсами. По-третє, використання жестової мови зростає при збільшенні кількості людей із вадами слуху та мовлення, а впровадження систем сурдоперекладу допомагає адаптувати таких людей до реалій суспільства.

Відстежування руху верхніх кінцівок, їхнє положення і форму – являється основним компонентом для систем детермінування жестів в реальному часі. Процес розпізнавання жестів є математичною інтерпретацією людських рухів, що обчислюються сучасними технічними засобами. Такий підхід дозволяє накопичувати, обробляти та аналізувати інформацію як і на стороні клієнтських девайсів, так і за допомогою хмарних обчислень, тобто досягати рівня крос-платформеності і це може означати, що розробка подібних процесів повинна реалізовуватися у вигляді легковажної моделі.

Сучасні підходи дозволяють визначати жести в реальному часі.Тобто в потоці вхідних даних визначати момент, коли демонструється певний жест, а далі, із залученням механізмів класифікації визначати сам жест.

**Аналіз останніх досліджень і публікацій.** В статі "Аналіз сучасних систем розпізнавання дактильно-жестової мови для системи сурдоперекладу" Лихошерстов Д.О та Лебедев Д.Ю [4] були розглянуті основні технології розпізнавання жестів і було встановлено, що метод на основі отримання зображення дозволяє ґрунтовно підходити до створення уніфікованих рішень систем сурдоперекладу.

Основна невизначеність у наявних механізмах по захоплені у відеоряді жестів і їх перекладу є алгоритми. Перевага алгоритму розпізнаванню жестів є його, доволі висока стабільність роботи і простота реалізація. Водночас недоліками таких алгоритмів в першу чергу є відсоток коректного перекладу, низька ефективність та чутливість до дії зовнішніх факторів, таких як освітленість, технічні характеристики камери, задній фон, а також віддаленість камери від об'єкта.

За останнє десятиріччя, алгоритми розпізнавання жестів в яких використовуються методи машинного навчання (Machine Learning – ML) дозволили зневолювати більшістю недоліків шляхом підвищення складності обчислень. Такий прогрес породив задачу в знаходженні балансу між точністю розпізнавання та складністю системи розпізнавання, де під складністю розуміється побудова та громісткість її архітектури, швидкодія роботи, розміри бази даних для навчання.

**Постановка технічного завдання.** У даній статті пропонується, за допомогою дослідження, провести порівняльний аналіз рішень на основі Machine Learning для визначення образу жестів і відповідно процесу їх класифікації. Вибраний спосіб буде використаний для системи сурдоперекладу, детальний опис котрої наводиться в статті [4]. Пропонується розглянути два підходи:

1. Створення власного фреймворку, котрий буде містити елементи машинного навчання, тобто самостійно буде ідентифікувати образ символу у вхідному наборі даних.

2. Використання готового рішення, що містять готову структуру нейромережі та наявну базу даних вхідних значень.

Для оцінювання кожного із підходів пропонується ввести ряд критеріїв:

1. Точність ідентифікації – найбільш пріоритетний критерій оцінки є точність визначення жестів.

2. Швидкодія – для здійснення сурдоперекладу в реальному часі.

3. Крос-платформеність – один із критеріїв універсальності сурдоперекладу.

4. Можливості апгрейду та масштабованості – для універсальної системи сурдоперекладу важливо додавати нові жестові мови і вдосконалювати переклад поточних.

5. Легковажність системи- реалізація системи, що буде підходити під велику кількість сучасних девайсів.

6. Система підтримки готових рішень – можливість фіксити баги та додавати нові фічі.

**Виклад основного матеріалу дослідження.** З врахуванням вищевизначених критеріїв пропонується спростити задачу визначення жесту у відеоряді, до класифікації введеного користувачем зображення на основу бази даних по якій навчалася нейромережа. Пропонується реалізувати нейромережу, котра буде вгадувати цифри від 0 до 9. Користувач повинен самостійно ввести символ, а система самостійно визначить цифру в діапазоні від 0 до 9.

Слід зазначити, що подальші дослідження в даній праці не ставлять на меті аналіз методів розпізнавання жестів та класифікацію жестів, алише визначення напрямку наукових досліджень – використовувати готові, більш універсальні, бібліотеки (фреймворки) чи створювати нові, власні, що будуть заточені під конкретні задачі.

Для реалізації власного фреймворку пропонується створити нейромережу моделі на основі багатошарового персептрону із вибірковою базою навчання. Наша нейромережа буде мати вхідний шар, прихований шар та шар вихідних нейронів (рис. 1).

При навчанні використаємо три активаційні функції почергово і перевіримо таким чином чи впливає функція активації на якість роботи нейромережі. Використаємо три основні функції активації – сигмоїдна функція, ModReLU(ReL з витоком) та тангенс (рис. 2). Слід звернути увагу, що ModReLU не має експонентів, а це дозволяє проводити розрахунки набагато швидше.

Навчати нашу нейромережу будемо за допомогою MNIST database [5] – ресурс із готовими базами даних для навчання нейромереж. В цій базі даних знаходиться 60 000 тисяч даних для навчання та 10 000 тисяч для тестування. База даних має власну файлову систему, де є файл "train-labels-idx1-ubyte" із цифрами від "0" до "9" та файл "train-images-idx3-ubyte" із пікселями для них. Всі навчальні дані в нас будуть представлені у бінарному вигляді, де білому кольору відповідає логічне значення 1, а чорному кольору відповідає логічне значення 0. Саме зображення в базі даних має розміри 28 на 28 пікселів – отримуємо 784 вхідних нейрони.

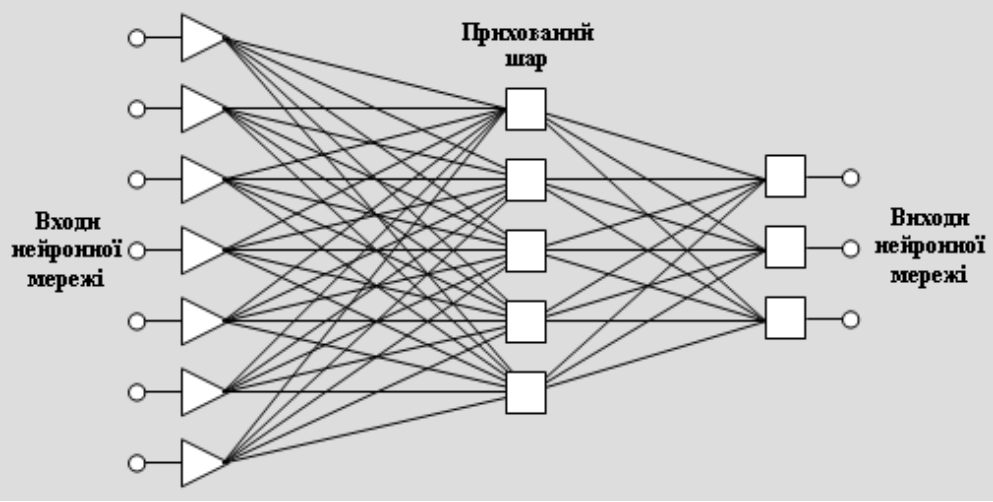


**Рис. 1. Загальна структура багатошарового персептрону**

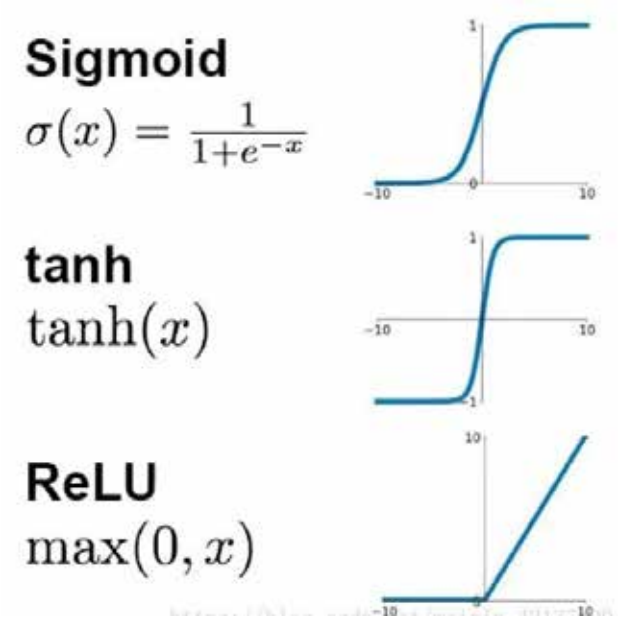


**Рис. 2. Графік та формула функцій активації**

Так як відбувся процес нормування (кольоровий діапазон представлений в діапазоні від 0 до 1), то нам необхідно внести обмеження для активаційних функцій. Для сигмоїдної функції змін немає. Для ModReLU та th(x) змінюється діапазон.

Враховуючи всі особливості було створено власний фреймворк, де користувач курсором миші вводить символ, а нейромережа видає відповідь. Проект був створений на C++ в середовищі програмування Visual Studio 2022 із використанням бібліотеки Qt для створення крос-платформового застосунку (рис. 3). За рахунок бібліотеки Qt та інтегрованого середовища розробки Qt Creator є можливість з легкістю адаптувати створений застосунок до інших операційних систем. Інтерфейс застосунку зображений на рис. 4 та лістинг програми був завантажено на GitHub [6]. Відео роботи застосунку можна знайти на платформі Google [7].

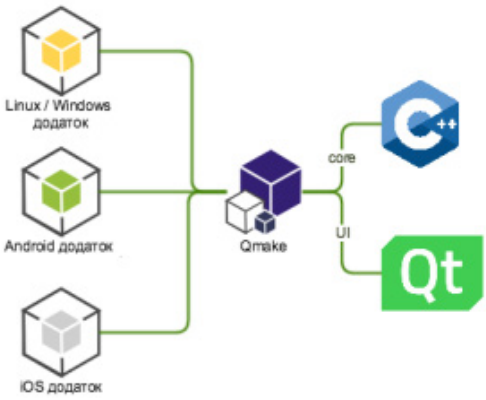


**Рис. 3. Функціонально-структурна схема додатку**

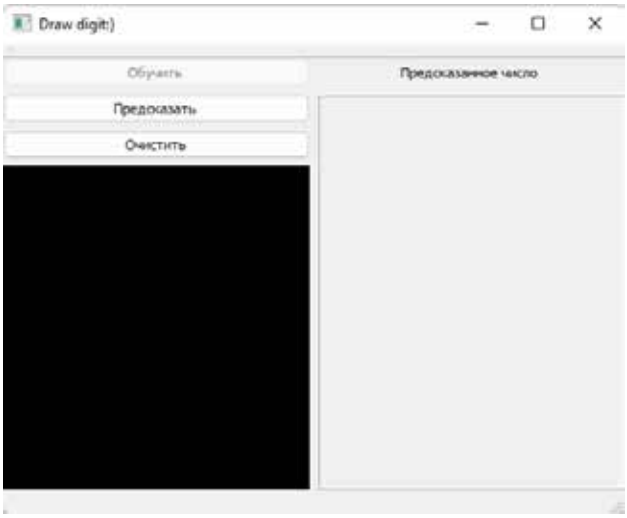


**Рис. 4. Загальний вигляд користувацького інтерфейсу додатку по визначенню цифр на вхідному зображені**

Таблиця 1

**Значення точності визначення для активаційних функцій**

| № | Активаційна функція | Ra навчання | Ra тестування |
|---|---|---|---|
| 1 | Сигмоїдна функція | Від 85.8946 до 79.181 | 68.79 |
| 2 | ModReLU | Від 85.6881 до 99.28391 | 97.26 |
| 3 | th(x) | Від 8.86124 до 9.02778 | 8.92 |

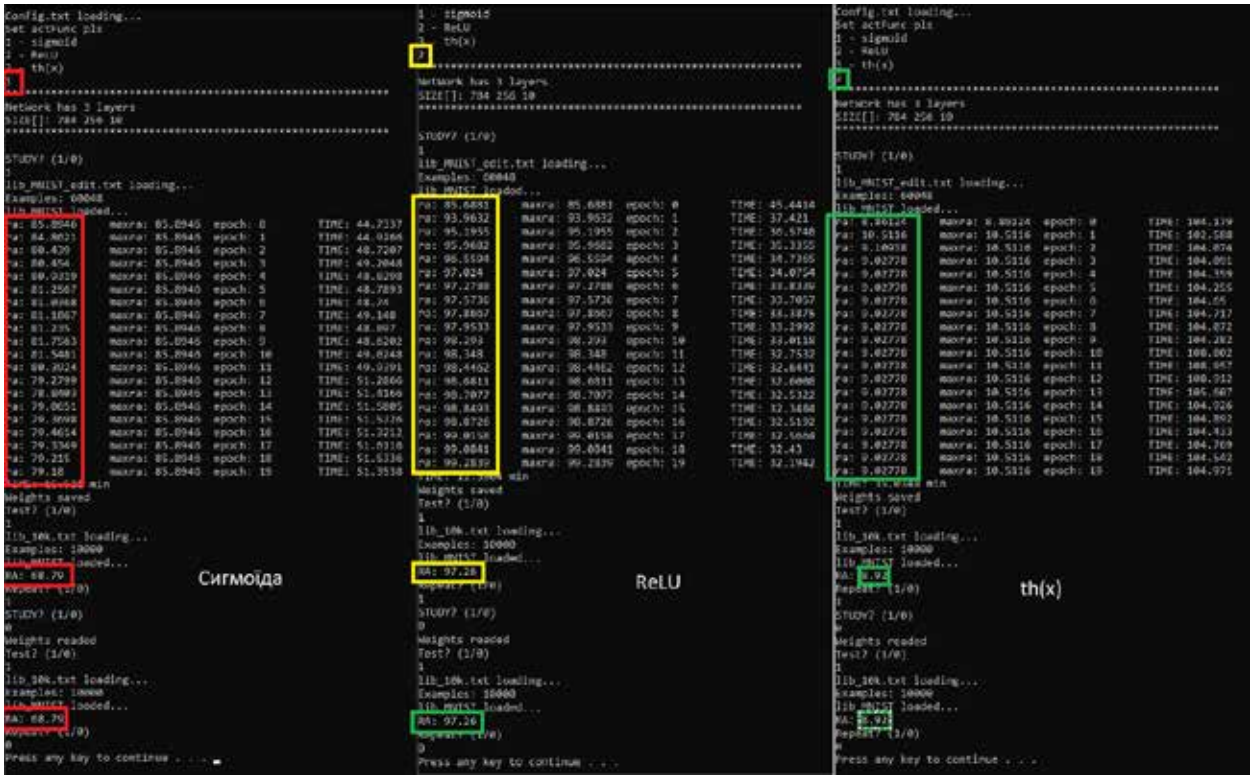


**Рис. 5. Результати роботи нейромережі для різних активаційних функцій**

В ході тестування фреймворку було встановлено, що:

1. Точність система із функцією активації ModReLU має найбільшу точність (рис. 5) в межах 85% – 96% в залежності від навчального датасету. Детальніше на табл. 1 та рис. 5.

2. Система не має жодних сповільнень в роботі на системі із процесором AMD Ryzen 3900X та 32GB пам'яті DDRR4.

3. За рахунок використання бібліотеки Qt наявна система крос-платформеності.

4. Систему можна вдосконалити за рахунок використання інших функцій активації. В даному досліджені ModReLU продемонструвала найвищу точність та найбільшу швидкодію, а отже можна спробувати знайти й більш швидку функцію, котра не містить експоненту.

5. Реалізована система сумарно займає 1.59GB, але самий білд файл 194КБ.

6. На поточний момент в даному рішені не представлено системи підтримки.

Як ми можемо бачити, дане рішення має доволі високе значення похибки ( > 10% ) та досить чутливе до якості вхідних даних. А якщо врахувати, що рука людини має до тридцяти геометричних особливостей (рис. 6) визначених на основі будови пальців та долоні, то можна стверджувати, що збільшення складності вхідних даних не приведе до збільшення точності розрахунків та в цілому не дозволить коректно визначати жести в реальному часі[8]. Отже, розробка власного фреймворку не має перспектив та потрібно розглянути наявні варіанти.

Існує багато фреймворків машинного навчання і це пояснюється трендами на Machine Learning (скорочено ML – машине навчання) в поточний момент (рис. 7).

Розглянемо сучасних гігантів ринку машинного навчання і почнемо із гігантів хмарних обчислень – рішення, де розрахунки відбуваються не на стороні користувача. Рішення від IBM, Google, Microsoft, Amazon мають ML сервіси, що

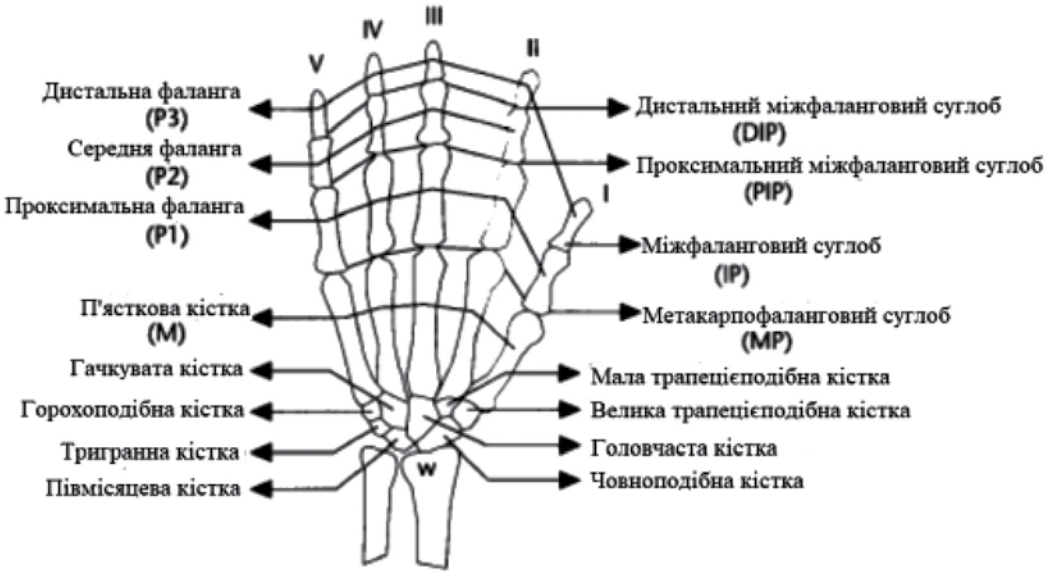


**Рис. 6. Скелет руки**

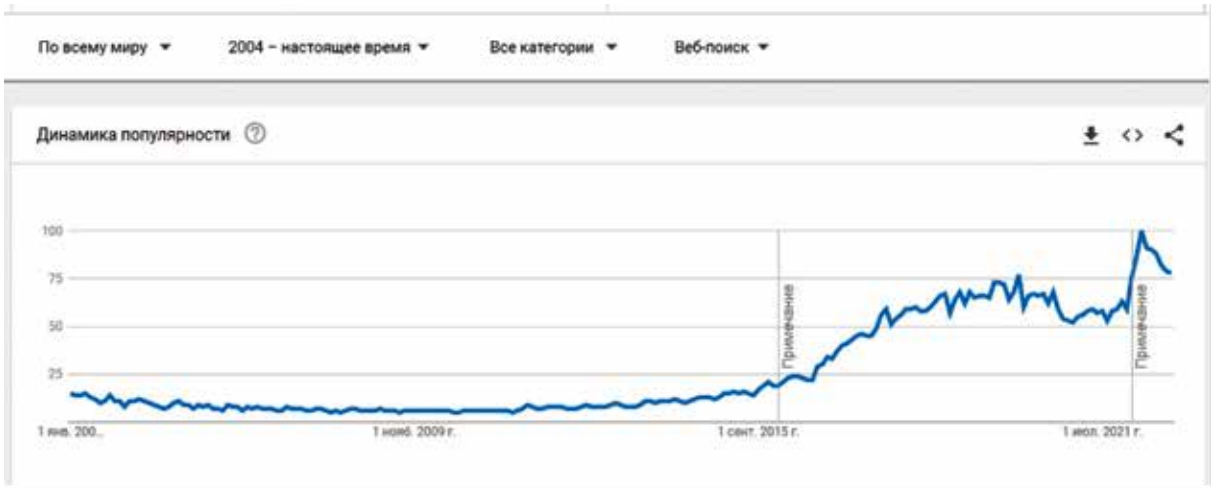


**Рис. 7. Залежність запистів Machine Learning в часі**

дозволяють розпізнавати зображення або запускати різноманітні нейромережі для визначення інших даних. Всі розрахунки відбуваються в хмарних системах і майже не можливо зробити реал тайм систему – тому вони нам не відповідають нашим критеріям.

Також можна розглянути різноманітні готові бібліотеки. Вони дозволяють робити все те саме, що і клауд сервіси, але локально. Більшість із них написані на Python і створені у вигляді готового шаблону на основі котрого є можливість розробити саме власну власний алгоритм. Такий підхід дозволяє готові рішення адаптувати під власні задачі, але недоліком цих бібліотек є часткова закритість вихідного коду та заборона на використання їх в комерційних проектах. Кожні дії потрібно погоджувати із розробниками бібліотеки.

Серед різноманітних фреймворків існує рішення, котре в середині 2019 року компанія Google представила на конференції з комп'ютерного зору та розпізнавання образів – MediaPipe. Крос-платформний фреймворк для машинного навчання з моделями для розпізнавання обличчя, рук, волосся та різних об'єктів навколишнього світу. При цьому, вихідний код знаходиться у відкритому доступі і кожен розробник має можливість інтегрувати MediaPipe у власні рішення. На рис. 8 представлено класифікацію жестів руки за допомогою MediaPipe.

Фреймворк MediaPipe містить три моделі штучного інтелекту, що працюють у взаємозв'язку.

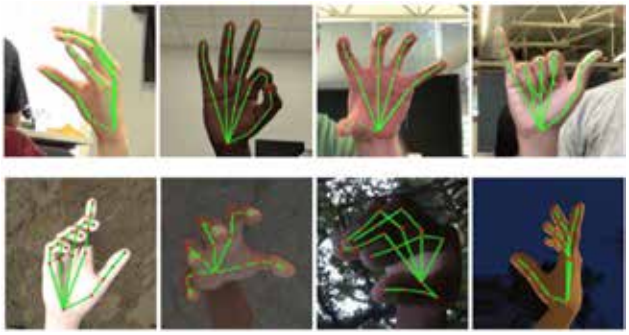

**Рис. 8. Робота крос-платформеного фреймворку MediaPipe**

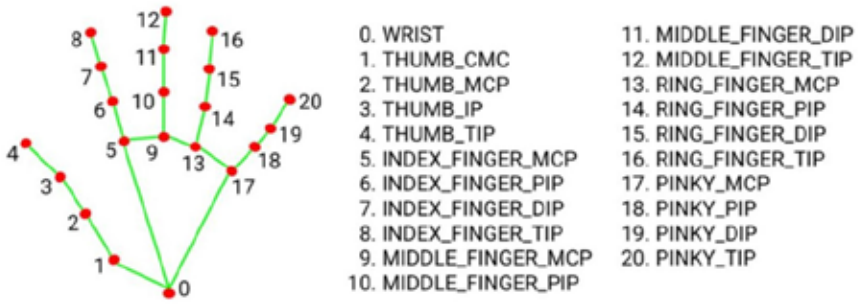


**Рис. 9. Модель для розмітки долоні**

Детектор долоні BlazePalm – аналізує кадри відеоряду і повертає прямокутні ділянки, в яких знаходяться долоні. Модель для розмітки долоні – аналізує прямокутну область зображення від BlazePalm та повертає 21 координату, що відповідає розташуванню суглобів та пальців і таким чином формують скелет руки (рис. 9). Модель, що класифікує отриману конфігурацію точок та ідентифікує їх із вхідним набором жестів;

Даний фреймворк дозволяє відрізняти закриті та відкриті положення та володіючи інформацією про точне розташування маркерних точок, можна ефективно будувати різноматні моделі. Процес детектування верхніх кінцівок можна налаштовувати в широкому діапазоні (від максимальної точності до визначення кількості рук на кадрі).

В результаті детального вивчення характеристик на офіційному сайті [9] та тестування MediaPipe [10] було встановлено, що даний


**Рис. 10. Структурно-функціональна організація системи розпізнавання дактильно-жестової мови**

фреймворк відповідає всім сформованим вище критеріям та дозволяє гнучко видозмінювати систему сурдоперекладу.

Основним подальшими дослідженнями є модифікація структурно-функціональної організації системи [4] розпізнавання дактильно-жестової мови (рис. 10) і реалізація прикладного результату на основі фреймворку MediaPipe.

**Висновки.** Проведено порівняльний аналіз фреймворків для визначення даних на зображені та встановлено, що фреймворк MediaPipe від компанії Google дозволяє реалізувати систему сурдоперекладу на основі структурно-функціональної схеми, представленої в праці "Аналіз сучасних систем розпізнавання дактильно-жестової мови для системи сурдоперекладу" Лихошерстов Д.О та Лебедев Д.Ю.

Створено та проведено дослідження над власним фреймворком. Встановлено, що поточний рівень реалізації багатошарового персептрона не дозволяє якісно визначати певні дані на зображені. Експериментальним шляхом встановлено, що не експоненціальні функції активації мають більшу швидкодію та точність визначення даних на зображені.

**Список літератури:**

1. American Sign Language Dictionary – https://www.signasl.org/.
2. AR окуляри від Facebook – https://tech.liga.net/gadgets/novosti/reagiruet-na-jesty-i-impulsy-facebook-razrabatyvaet-gadjet-dlya-upravleniya-ochkami-ar.
3. Керування Sony PlayStation VR за допомогою жестів – https://www.playstation.com/uk-ua/ps-vr/.
4. Стаття Лихошерстов Та Лебедев – http://www.tech.vernadskyjournals.in.ua/32-71-6.
5. База даних MNIST database – http://yann.lecun.com/exdb/mnist/.
6. Лістинг фреймворку – https://github.com/date7887/Multilayer-perceptron.
7. Відео роботи застосунка – https://drive.google.com/drive/folders/19-RL--k3L91iW2-1JPjoTAEIaxpxeA8y?usp=sharing.
8. Y. Bulatov, "Hand recognition using geometric classifiers," / S. Jambawalikar, P. Kumar, and S. Sethia // in Biometric Authentication. Springer, –2004, – pp. 753 759.
9. Офіційний сайт MediaPipe та приклади роботи – https://google.github.io/mediapipe/.
10. Процес налагодження і тестування фреймворка MediaPipe – https://www.youtube.com/watch?v=pG4sUNDOZFg&t=757s/.

**Lykhosherstov D.O., Lebedev D.Yu. COMPARATIVE ANALYSIS OF THE WAYS OF IMAGE DETERMINATION ON A RANGE OF VIDEOS USING MACHINE LEARNING TOOLS**

*In work, a comparative study of the mechanisms of classification of information on the image was carried out. It is substantiated that gestures are an important component of communication mechanisms in society, management of modern technical means and allow for the improvement of the social level of people with hearing and speech impairments due to the introduction of sign language translation systems into everyday life. It is noted that the gesture processing process uses machine learning, which has the ability to classify and identify gestures on video frames in real-time. The main advantages and disadvantages of gesture recognition systems are formed, and the research task is defined, which consists of comparing existing recognition mechanisms and creating their own.*

*The criteria for evaluating the image data recognition system were formed and substantiated – identification accuracy, speed, cross-platform compatibility, the possibility of upgrading and scalability, the lightness of the system, and the possibility of supporting a ready-made solution.*

*A multilayer perceptron neural network was created from the training and test database MNIST database. The training and testing of the neural network were carried out to determine the data on the image for three different activation functions – sigmoid function, ModReLU (ReL with leakage), and tangent. ModReLU was found to have the highest level of detection accuracy.*

*Existing data recognition frameworks were investigated and found that Google's MediaPipe framework has cross-platform, open code and developer support, available code, and high detection accuracy, especially for real-time upper limb gesture classification.*

*Directions for their further improvement of the sign language translation system have been substantiated and formed.*

***Key words:*** *sign translation, machine learning, data transfer, gesture recognition, neural network, MediaPipe, MNIST database.*