DOI 10.34185/1562-9945-5-154-2024-14
УДК 004.4:004.94

В.Ю. Песчанський, Є.С. Сулема

МЕТОД СИНХРОНІЗАЦІЇ ТЕМПОРАЛЬНИХ МУЛЬТИМОДАЛЬНИХ ДАНИХ ДЛЯ СТВОРЕННЯ ЦИФРОВОГО ДВІЙНИКА ГОРТАНІ

Анотація. У статті запропоновано метод синхронізації темпоральних мультимодальних даних, який призначено для створення та оптимізацію цифрових двійників гортані людини. Метод надає можливість інтеграції діагностичних даних відеоларингостробоскопії та акустичного аналізу голосу для створення точної 3D-моделі гортані, яка відтворює фізіологічні характеристики та голосову функцію пацієнта. У статті описано метод, заснований на глибокому навчанні, який забезпечує точне вирівнювання часових маркерів фаз голосоутворення на всіх типах даних, відображаючи деталізовану біомеханіку гортані в динаміці. Результати дослідження свідчать про підвищення точності цифрового двійника. Застосування цифрового двійника гортані покликане спростити планування хірургічних втручань та інших лікувальних заходів.
Ключові слова: інженерія програмного забезпечення, програмне забезпечення, цифровий двійник, 3D-моделювання, синхронізація даних, відеоларингостробоскопія, акустичний аналіз, біомеханіка голосового апарату, персоналізована медицина.

Постановка проблеми. Створення цифрового двійника гортані є складним завданням, яке вимагає точної інтеграції різноманітних мультимодальних даних для відтворення реалістичної біомеханіки та функціональності голосового апарату. Проблема полягає у синхронізації даних, що вимірюються у різний час, надходять з різних джерел, і в різних умовах. Зокрема, у дослідженні, результати якого представлені у цій статті, вирішується задача синхронізації відеозаписів фоніації голосу та акустичних записів. Ці діагностичні дані повинні бути взаємозамінними, що дає змогу моделювати динамічні зміни у гортані під час голосоутворення. Перша складність полягає в обмеженій точності традиційних методів синхронізації, які не враховують індивідуальні варіації в анатомії та фізіології гортані. Друга – у необхідності розроблення програмних засобів, які здатні обробляти великі обсяги даних в реальному часі з високою пропускною здатністю та мінімальною помилкою. Третя – у створенні уніфікованої моделі, яка може адаптуватися до змін у фізіологічних даних пацієнта протягом часу, включаючи зміни в голосі через захворювання або вікові зміни. Ці проблеми вимагають нового підходу до обробки та аналізу даних, що передбачає використання передових технік машинного навчання для виявлення складних шаблонів в темпоральних та просторових характеристиках мультимодальних даних. Результатом дослідження має стати розроблення алгоритмічно-програмного методу, за допомогою якого можна забезпечити високу точність синхронізації темпоральних даних, які надходять з різних джерел, для створення інтегрованої і динамічної індивідуальної моделі гортані пацієнта.

Аналіз останніх досліджень і публікацій. Сфера моделювання медико-біологічних об'єктів відзначається швидким розвитком [1]. Однак, попри прогрес, задачі, пов'язані з точною синхронізацією даних різних модальностей, залишаються актуальними. Останні дослідження зі синхронізації мультимодальних даних демонструють важливість розроблення більш комплексних та інтегрованих систем, які можуть об'єднувати структурну, функціональну та біомеханічну інформацію [2]. Останні публікації з комп'ютерної томографії та магнітно-резонансної томографії зосереджені на підвищенні деталізації зображень, що дає змогу точніше визначити мікроструктури гортані [3]. У той же час, дослідження у галузі фоніатрії та ларингології акцентують важливість точного відтворення фізіологічних процесів голосоутворення. Так, інноваційні методи акустичного аналізу, такі як високошвидкісна відеоларингостробоскопія, використовуються для вивчення динаміки голосових зв'язок [4]. Синергія цих досліджень з передовими технологіями обробки зображень, такими як машинне навчання, дає змогу створювати детальні цифрові моделі гортані, що імітують реальні фізіологічні процеси. Однак, існує брак досліджень, які б зосереджувались на інтеграції даних різних модальностей у єдиний потік темпоральних даних [5], що необхідно для створення точних і динамічних цифрових двійників медико-біологічних об'єктів, зокрема гортані. Інформація про алгоритми та програмне забезпечення з вказаної галузі поки що доволі поверхнево висвітлена у наукових працях. Таким чином, рівень технологічного розвитку цієї галузі, згідно з індексом технологічної готовності TIOBE [6], перебуває на етапі розроблення прототипів. Це визначає актуальність створення нових алгоритмів та методів обробки даних, які брали б до уваги специфічні темпоральні характеристики та динаміку руху голосових зв'язок. Потреба у такому підході є особливо нагальною, оскільки вона може відкрити нові можливості для діагностики, хірургічного планування та реабілітації після операцій на гортані. В огляді літератури [7] зазначається, що незважаючи на значний прогрес у сегментації медичних зображень і аналізі голосу, залишається прогалина у дослідженнях, присвячених синхронізації цих мультимодальних даних. Наявні методи синхронізації часто вимагають ручного втручання або не враховують всіх аспектів біомеханіки голосоутворення, що обмежує їхнє застосування у клінічній практиці. З огляду на ці проблеми, необхідність розроблення вдосконаленого методу синхронізації, який би використовував переваги машинного навчання, є важливою для підвищення точності та ефективності цифрових двійників гортані. Такий метод має допомогти не тільки вдосконалити створення моделей, але й прискорити процес впровадження цих технологій у клінічну практику, сприяючи розвитку персоналізованої медицини і значно покращуючи результати лікування пацієнтів з порушеннями голосу.

Мета досліджень. Головною метою дослідження, результати якого представлені у цій статті, є розроблення та експериментальна перевірка комплексного методу синхронізації темпоральних мультимодальних даних, які використовуються для створення і точної візуалізації цифрового двійника гортані. Основні завдання дослідження включають:

- аналіз відомих методів синхронізації даних і виявлення їхніх обмежень у контексті моделювання гортані;
- розроблення методу інтегрування динамічних зображень з відеоларингостробоскопії та акустичні дані голосу в єдиний потік темпоральних даних;
- використання методів машинного навчання для автоматизації процесу синхронізації і зменшення помилок, пов'язаних із людським фактором;

ISSN 1562-9945 (Print)
ISSN 2707-7977 (Online)

– адаптація моделі до індивідуальних особливостей пацієнтів для підвищення точності діагностики та ефективності лікувальних втручань.

Практичним результатом досягнення цієї мети має стати підвищення ефективності діагностичних та терапевтичних методів у ларингології та фоніатрії, сприяння розвитку персоналізованої медицини та покращення якості життя пацієнтів із порушеннями голосу.

Викладення основного матеріалу досліджень. У контексті створення цифрового двійника гортані аналіз мультимодальних даних відіграє ключову роль. Джерела даних включають комп'ютерну томографію, магнітно-резонансну томографію, відеоларингостробоскопію та акустичний аналіз голосу [8]. Кожне з цих джерел надає унікальний вимір структури та функціонування гортані. Найбільш поширеними способами діагностики гортані є відеоларингостробоскопія та акустичний аналіз голосу. Дослідження, результати якого наведені у цій статті, ґрунтуються саме на цих діагностичних способах, які полягають у наступному.

Відеоларингостробоскопія використовується для візуального аналізу та оцінки динамічних характеристик голосових зв'язок під час фонації, забезпечуючи візуальне уявлення про механіку голосоутворення.

Акустичний аналіз дозволяє аналізувати характеристики голосу, як-от частота, амплітуда, та тембр, що необхідно для відтворення реалістичної моделі голосу.

Для вирішення проблем, пов'язаних із синхронізацією відеоданих з відеоларингостробоскопії та аудіоданих з акустичного аналізу, пропонується метод, який ґрунтується на комбінованому використанні крос-кореляції та машинного навчання. Цей підхід дає змогу автоматично виявляти взаємозв'язок між аудіо- та відеосигналами, враховуючи їхню часову затримку та варіабельність у динаміці.

Пропонується **нормалізувати** відеопотік перед застосуванням функції *correlate*, що відіграє ключову роль у підвищенні точності та ефективності аналізу кореляції між відео- та аудіосигналами. Це допоможе усунути вплив різниці у масштабах та освітленості, а також спрощує ідентифікацію взаємозв'язків між сигналами.

Пропоновані кроки алгоритму нормалізації відеопотоку є наступними.

Перетворення у сірий колір. Цей крок зменшує обчислювальну складність розроблюваного методу, оскільки кожен кадр міститиме тільки один канал інтенсивності замість трьох колірних каналів.

Лістинг 1. Приведення кадру відеосигналу до сірого кольору.

```
    Import cv2
gray_video = cv2.cvtColor(video_frame, cv2.COLOR_BGR2GRAY)
```

Масштабування інтенсивності. Нормалізація інтенсивності кольору пікселів у відеосигналу до діапазону [0, 1] або [-1, 1] має на меті уніфікацію рівня освітленості та контрастності між різними кадрами. Цього можна досягти, наприклад, з використанням бібліотеки *ffmpeg* [9].

Лістинг 2. Нормалізація відеосигналу.

```
    Import cv2
ffmpeg-normalize input.mp4 -o output.mp4 -c:a aac -b:a 192k
```

Виділення характеристичних ознак. Для подальшого виявлення кореляції відео- з аудіосигналами доцільно використовувати методи комп'ютерного зору для виділення характеристичних ознак, які відображають рух голосових зв'язок. Пропонується використовувати функцію бібліотеки cv2 [10].

Лістинг 3. Виділення характеристик для визначення кореляції.

```
    import cv2
import numpy as np
cap = cv2.VideoCapture('path/to/your/video.mp4')
while True:
    ret, frame = cap.read()
    if not ret:
        break  # Якщо кадри закінчились, завершити цикл
    # Виявлення "хороших" точок
   corners = cv2.goodFeaturesToTrack(gray_frame,maxCorners=50,
       qualityLevel=0.01, minDistance=10)
```

Крос-кореляція між аудіо- та відеосигналом даними визначає ступінь взаємозв'язку між ними при різних затримках часу. Це дозволяє визначити оптимальну затримку між аудіо- та відеосигналом, максимізуючи їхню кореляцію. У випадку моделювання гортані, аудіосигнали відповідають за звуки, що генеруються голосовими зв'язками, тоді як відеосигнали демонструють фізичний рух цих зв'язок.

Для двох сигналів *x(t)* та *y(t)* дискретна крос-кореляція [11] визначається як:

$$R(\tau) = \sum_{t=0}^{N-1} x[t] \cdot y[t + \tau],$$

де $R(\tau)$ – функція крос-кореляції при затримці;

x(t) – аудіо сигнал;

y(t+ τ*)* – відеосигнал зі зміщенням на τ;

N – довжина сигналу;

t – часовий індекс.

Досить часто тривалість аудіо- та відеосигналу не збігається. Це вимагає спеціального підходу для їх синхронізації. Для аналізу кореляції між сигналами різної довжини можна використати зсув (lag) аудіо- відносно відеосигналу. Це дозволяє "просканувати" взаємозв'язок між сигналами на різних інтервалах часу з метою виявлення найбільшої кореляції. З практичної точки зору доцільніше зсувати коротший сигнал, обчислюючи **кореляцію** для кожного зсуву.

ISSN 1562-9945 (Print)
ISSN 2707-7977 (Online)

Лістинг 4. Виявлення кореляції.

```
    import numpy as np
from scipy.signal import correlate
import matplotlib.pyplot as plt
    # Обчислюємо крос-кореляцію
correlation = correlate(video, audio, mode='full')
    lags = np.arange(-len(x) + 1, len(x))
plt.figure(figsize=(10, 5))
plt.plot(lags, correlation)
```

Після виявлення кореляції між відео- та аудіосигналами, наступним кроком є збереження темпоральних даних, отриманих у процесі кореляційного аналізу.

Доцільно зберігати дати та часові мітки, значення затримок, індекси кадрів відеосигналу, де виявлено максимальну кореляцію, та відповідні аудіофрагменти.

Лістинг 5. Приклад пропонованої структури даних.

```
    import json

    data = {
        'max_correlation_values': [0.95, 0.89, ...],
        'video_frame_indices': [150, 300, ...],
        'audio_timestamps': ['00:01:05', '00:02:10', ...]
    }
```

Це дає змогу використовувати ці дані для подальшого аналізу, відстеження, відновлення чи інших завдань. Наприклад, крім побудови цифрового двійника, можна використовувати машинне навчання для детального аналізу взаємозв’язку між рухом голосових зв’язок та характеристиками голосу. Також можна застосовувати елементи штучного інтелекту для виявлення специфічних ознак в аудіо- та відеоданих, які корелюють з певними фізіологічними станами.

Розглянемо результати експериментальної частини дослідження.

На рис. 1 показано знайдені точки для трекінгу. Зміщення таких точок відносно першого кадру можна використовувати як нормалізоване представлення відеопотоку. За нормалізоване значення аудіосигналу візьмемо його частоту. В експерименті використовувалися фрагменти довжиною в 1.2 та 1 секунду.

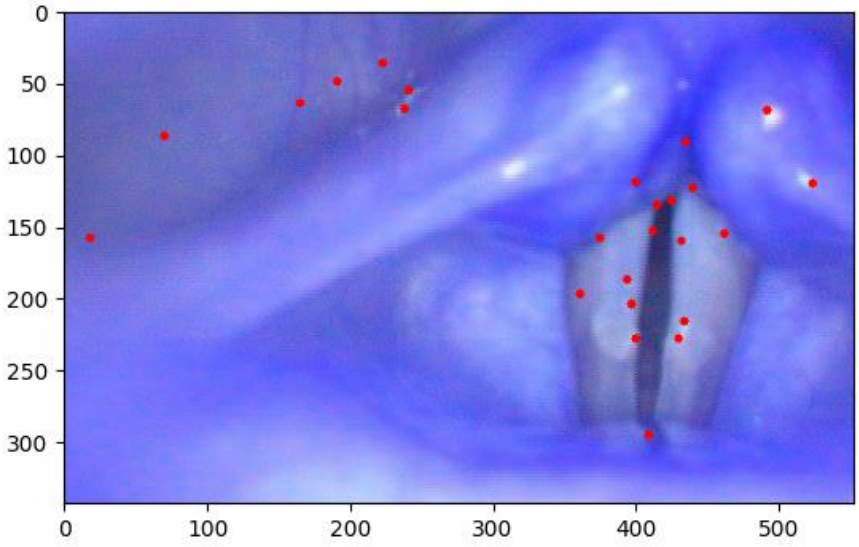

Рисунок 1 - Візуалізація знайдених точок для трекінгу

Показаний на рис. 2 графік демонструє візуалізацію крос-кореляції між нормованими відео- та аудіосигналами. Точка на графіку, яка позначена червоним кольором, є маркером зміщення, при якому кореляція між сигналами досягає свого максимуму. Це зміщення і є шуканим оптимальним значенням синхронізації між аудіо- та відеосигналом.

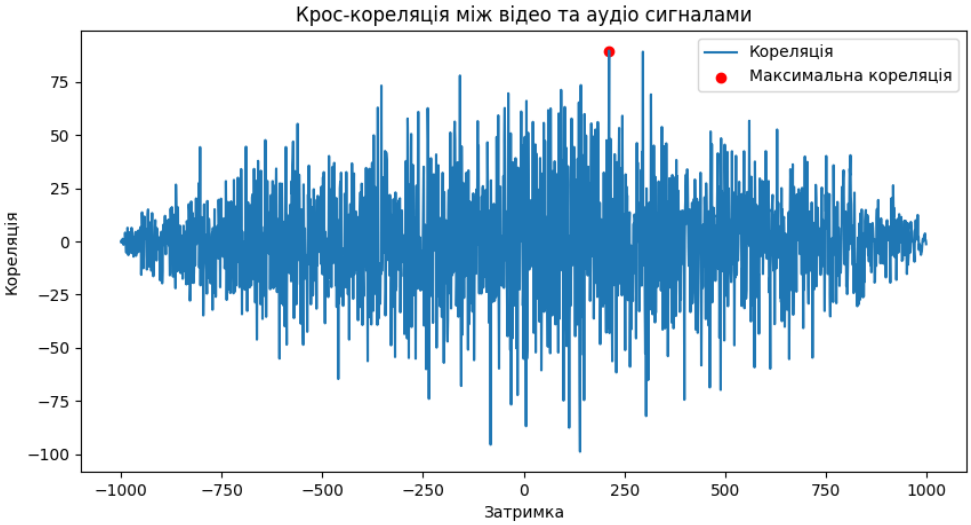

Рисунок 2 - Візуалізація обчисленої максимальної кореляції

Побудуємо графіки обох сигналів і накладемо їх один на одного з урахуванням зміщення для кращої візуалізації. Отриманий результат наведено на рис. 3.

ISSN 1562-9945 (Print)
ISSN 2707-7977 (Online)

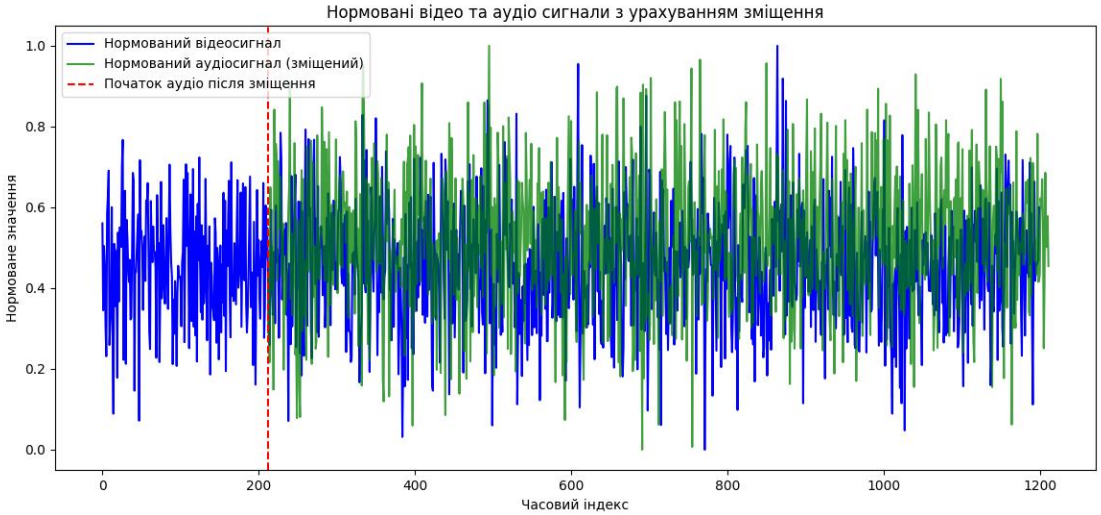

Рисунок 3 - Накладені нормалізовані значення

Отримана візуалізація ілюструє, як два сигнали відповідають один одному у часі, враховуючи визначене оптимальне зміщення.

Хоча розроблений метод синхронізації темпоральних мультимодальних даних демонструє високу ефективність, він не позбавлений недоліків. Один з основних недоліків полягає у потребі у доволі значних обчислювальних ресурсів, особливо при обробці великих даних. Крім того, алгоритм реалізації розробленого методу може бути чутливим до шуму в даних, що може впливати на точність синхронізації. Виходячи з вищезазначеного, шляхи подальшого вдосконалення можуть включати розроблення більш ефективних алгоритмів обробки, які оптимізують використання обчислювальних ресурсів. Також корисним буде впровадження додаткових методів фільтрації шуму та покращення чутливості до помилок у даних. Використання технік машинного навчання для автоматичного виправлення помилок та ідентифікації оптимальних параметрів синхронізації також може значно підвищити точність і надійність методу.

Висновки. У статті розглянуто передові підходи та методики для синхронізації темпоральних мультимодальних даних, зокрема в контексті моделювання гортані. Встановлення точної кореляції між відео- та аудіосигналами є вирішальним для точного аналізу динаміки голосоутворення, що може сприяти покращенню діагностики та лікування розладів голосу.

Запропоновано удосконалений метод синхронізації темпоральних мультимодальних даних, який інтегрує алгоритми крос-кореляції з техніками машинного навчання для точного виявлення зв’язку між відео- та аудіосигналами. Цей підхід дає змогу не тільки виявляти затримки між сигналами з високою точністю, але й ефективно обробляти мультимодальні дані з різноманітними характеристиками. Основною особливістю запропонованого методу є використання комп’ютерного зору для автоматичного та коректного виявлення ключових точок на зображенні та встановлення їхньої кореляції з аудіопараметрами, що відкриває нові можливості для аналізу динаміки голосоутворення та дослідження інших медико-біологічних об’єктів. Пропонований метод забезпечує високу стійкість до зовнішніх перешкод та шумів, а також надає можливість адаптації до довільної роздільності вхідного відеопотоку. Це є

важливим для точної обробки темпоральних даних, необхідних для подальшого моделювання цифрового двійника або інших маніпуляцій.

Автори висловлюють подяку Шидловській Т.А., доктору медичних наук, професору, завідувачці лабораторії голосу і слуху Державної установи «Інститут отоларингології ім. проф. О.С. Коломійченка АМН України», за надання медичних зображень та консультаційну допомогу. Медичні зображення надані з дотриманням вимог медичної етики і конфіденційності інформації.

ЛІТЕРАТУРА / REFERENSES

1. Yanase, J. and Triantaphyllou, E., 2019. A systematic survey of computer-aided diagnosis in medicine: Past and present developments. *Expert Systems with Applications*, *138*, p.112821.
2. Lynn, L.A., 2019. Artificial intelligence systems for complex decision-making in acute care medicine: a review. *Patient safety in Surgery*, *13*(1), p.6.
3. Bailly, L., Cochereau, T., Orgéas, L., Henrich Bernardoni, N., Rolland du Roscoat, S., McLeer-Florin, A., Robert, Y., Laval, X., Laurencin, T., Chaffanjon, P. and Fayard, B., 2018. 3D multiscale imaging of human vocal folds using synchrotron X-ray microtomography in phase retrieval mode. *Scientific reports*, *8*(1), p.14003.
4. Woo, P., 2021. *Stroboscopy and high-speed imaging of the vocal function*. Plural publishing.
5. Soltanisehat, L., Alizadeh, R., Hao, H. and Choo, K.K.R., 2020. Technical, temporal, and spatial research challenges and opportunities in blockchain-based healthcare: A systematic literature review. IEEE Transactions on Engineering Management, 70(1), pp.353-368.
6. Steneker, M., 2016. Towards an empirical validation of the TIOBE Quality Indicator (Doctoral dissertation, Eindhoven University of Technology).
7. White, A.C. and Carding, P., 2022. Pre-and postoperative voice therapy for benign vocal fold lesions: factors influencing a complex intervention. Journal of Voice, 36(1), pp.59-67.
8. Rast, C., Unteregger, F., Honegger, F., Zwicky, S. and Storck, C., 2023. An Old Myth: Prediction of the Correct Singing Voice Classification. True or not?. Journal of Voice, 37(6), pp.968-e13.
9. Wu, X., Qu, P., Wang, S., Xie, L. and Dong, J., 2021. Extend the FFmpeg framework to analyze media content. arXiv preprint arXiv:2103.03539.
10. Demidenko, O.M., Aksionova, N.A., Varuyeu, A.V. and Kucharav, A.I., 2021, November. 3D-modeling of Augmented Reality objects using Shi-Tomasi corner detection algorithms. In Journal of Physics: Conference Series (Vol. 2091, No. 1, p. 012058). IOP Publishing.
11. Sun, Z., Sarma, P., Sethares, W. and Liang, Y., 2020, April. Learning relationships between text, audio, and video via deep canonical correlation for multimodal language analysis. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 34, No. 05, pp. 8992-8999).

Received 24.06.2024.
Accepted 27.06.2024.

Methods of creating digital twins of medico-biological objects using otolaryngology as an example

The paper presents an in-depth analysis on the most suitable tools and techniques for the formulation of a digital twin, specifically focusing on internal organs. The larynx, a pivotal organ in the human respiratory and vocal systems, is highlighted as the primary case study. The basis of this digital twin generation is the video stream sourced directly from a medical device's onboard camera, which provides real-time visual data of the organ. Two methodologies are put under the

ISSN 1562-9945 (Print)
ISSN 2707-7977 (Online)

microscope in this study: one that harnesses the power of neural networks and another grounded in algorithmic reconstruction using crucial points or markers within the video feed. Each method is evaluated against a set of well-defined criteria. These benchmarks include the efficiency of the tool, the accuracy of the digital representation it produces, the speed of its response, and its overall applicability within the confines of a typical medical environment.

After a rigorous comparative analysis, the research gravitates towards neural network-based approaches, spotlighting them due to several standout features. Neural networks, as elucidated in the paper, exhibit remarkable adaptability, ensuring that they can be tailored to diverse medical scenarios. Their accuracy, even when confronted with "noisy" or fragmented data, is another standout feature. This is paramount, especially in real-world scenarios where the data might not always be pristine. The ability of neural networks to sieve through such data and still produce accurate digital representations is a significant advancement in the field.

In conclusion, by affirming the superior potential of neural networks in crafting precise digital avatars of internal organs, the research not only provides a blueprint for enhanced diagnostic and therapeutic methodologies but also underscores a paradigm shift in how medical professionals can leverage technology for better patient outcomes. This synthesis of medical expertise and cutting-edge technology is poised to redefine the boundaries of medical science, heralding a new era of advanced diagnostics and treatment modalities.

Песчанський Владислав Юрійович – аспірант кафедри програмного забезпечення комп'ютерних систем Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського», Київ, vladpeschansky@gmail.com.
Сулема Євгенія Станіславівна – доктор технічних наук, завідувачка кафедри програмного забезпечення комп'ютерних систем. Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського», Київ.

Vladyslav Peschanskii – Post-Graduate Student of Computer Systems Software Department National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute", Kyiv, vladpeschansky@gmail.com.
Yevgeniya Sulema – DSc, Head of Computer Systems Software Department. National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute", Kyiv.