

DOI: <https://doi.org/10.15276/ict.03.2026.61>

УДК 004.89:378.14

Життєвий цикл мультимодальних даних в інтелектуальній системі синхронного конспектування та семантичного аналізу матеріалів лекційних занять

Бабілунга Оксана Юріївна¹⁾

ORCID: <https://orcid.org/0000-0001-6431-3557>; babilunga@op.edu.ua. Scopus Author ID: 35316966900

Дробишевський Святослав Володимирович¹⁾

ORCID: <https://orcid.org/0009-0008-1888-4584>; clavik55582@stud.op.edu.ua

¹⁾ Національний університет «Одеська політехніка», пр. Шевченка, 1. Одеса, 65044, Україна

АНОТАЦІЯ

У статті теоретично обґрунтовано, формалізовано та розроблено наскрізний життєвий цикл мультимодальних даних в інтелектуальній системі підтримки навчального процесу. Визначено, що традиційне ручне конспектування та відокремлений аналіз записів створюють критичний когнітивний бар'єр через розрив між викладом матеріалу та його закріпленням, а також через відсутність надійного фактологічного зв'язку між відповідями штучного інтелекту та первинними матеріалами лекції. Запропоновано багатостадійну модель життєвого циклу мультимодальних даних: низькорівневе захоплення відеопотоку вікна презентації та звукового ряду мовлення викладача; адаптивну фільтрацію кадрів на основі попиксельного порівняння зображень кадру у відтинках сірого; сегментацію макета та оптичне розпізнавання тексту за гібридною нейромережевою архітектурою; просторово-часову синхронізацію аудіовізуальних потоків із похибкою менше двох секунд; координатно-часову паспортизацію знань структурованими дескрипторами; паралельне подвійне індексування у векторному та символічному просторах; синхронне потокове конспектування та організацію оперативного інтерактивного опитування під час заняття на базі реципрокного злиття рангів. На відміну від існуючих підходів, розроблена методика орієнтована на забезпечення можливості студента ставити запитання в реальному часі щодо викладеного фрагмента лекції безпосередньо в ході заняття та отримувати верифіковану відповідь із миттєвим переходом до точного слайда, візуального прямокутника та звукового інтервалу пояснення. Лабораторна стендова верифікація та розрахункові оцінки підтверджують ефективність запропонованих рішень: наскрізна затримка конвеєра обробки кадру становить триста сорок дві мілісекунди, середня затримка первинної відповіді складає чотириста двадцять мілісекунд, а архітектура забезпечує цільову повноту вилучення контексту вісімдесят дев'ять відсотків та знижує ризик формування тверджень, не підтверджених матеріалами поточного заняття.

Ключові слова: життєвий цикл даних; мультимодальні дані; синхронне конспектування; захоплення лекцій у реальному часі; оптичне розпізнавання тексту; згортково-рекурентні нейронні мережі; реципрокне злиття рангів; координатна паспортизація джерел; інтелектуальні навчальні системи

Для цитування: Бабілунга О. Ю., Дробишевський С. В. «Життєвий цикл мультимодальних даних в інтелектуальній системі синхронного конспектування та семантичного аналізу матеріалів лекційних занять». *Інформатика. Культура. Техніка*. 2026; Том 3 № 2 (4): 791–801. DOI: <https://doi.org/10.15276/ict.03.2026.61>

Актуальність. У сучасній вищій освіті спостерігається стрімке зростання інтенсивності використання мультимедійних та цифрових засобів навчання. Лекційні заняття дедалі частіше передбачають одночасне використання динамічних електронних презентацій, демонстрацій програмного коду в інтегрованих середовищах розробки, побудови архітектурних та структурних діаграм, математичних формул, графіків, таблиць і схем, а також синхронного усного коментування викладача. У результаті навчальний матеріал формується як сукупність взаємопов'язаних, але мультимодальних інформаційних потоків – аудіального, візуального, текстового, графічного та програмного.

За положеннями теорії когнітивного навантаження (англ. Cognitive Load Theory) Дж. Свеллера та його колег [1], обмеженість ресурсів робочої пам'яті слухача суттєво впливає на ефективність засвоєння нового навчального матеріалу. Водночас відповідно до теорії подвійного кодування А. Пайвіо [2], використання взаємопов'язаних вербальних і візуальних форм представлення інформації може сприяти формуванню більш стійких когнітивних структур. Проте за умов високої швидкості подання інформації одночасна обробка декількох інформаційних потоків може призводити до перевантаження когнітивних ресурсів здобувача освіти. Як зазначають Р. Майєр та Р. Морено [3], необхідність одночасно сприймати візуальний ряд, слухати пояснення викладача та вести детальний письмовий конспект може

зумовлювати виникнення ефекту розщеплення уваги (англ. Split-Attention Effect). Це, своєю чергою, підвищує ймовірність пропуску окремих пояснень, логічних переходів і причинно-наслідкових зв'язків між елементами навчального матеріалу.

Таким чином, виникає проблема не лише автоматичного створення конспекту, а й інтеграції різнорідних даних лекційного заняття в єдиний просторово-часовий та семантичний інформаційний контекст. Особливої актуальності проблема набуває в умовах необхідності оперативного in-situ аналізу навчальної інформації безпосередньо під час лекції. Якщо, наприклад, на певному етапі заняття студент не зрозумів математичний вираз, визначення або крок алгоритму, пояснення якого було наведено раніше, він повинен мати можливість швидко звернутися до попереднього контексту без переривання сприйняття поточного матеріалу. За відсутності семантичної навігації та контекстного пошуку така ситуація може призвести до накопичення нерозуміння та подальшого зниження ефективності засвоєння матеріалу.

Сучасний розвиток технологій автоматичного розпізнавання мовлення, оптичного розпізнавання символів, комп'ютерного зору, обробки природної мови, мультимодальних моделей та генеративного штучного інтелекту створює технологічні передумови для розв'язання цієї проблеми.

Отже, актуальність дослідження визначається необхідністю переходу від традиційного зберігання лекційних матеріалів у вигляді окремих аудіо-, відео- та текстових ресурсів до єдиного мультимодального інформаційного середовища, у якому різнорідні дані синхронізуються, семантично структуруються та перетворюються на доступний для інтелектуального аналізу контекст знань. Ключовим завданням є організація життєвого циклу мультимодальних даних – від їх первинного захоплення та попередньої обробки до синхронізації, семантичного структурування, формування конспекту, інтелектуального аналізу та подальшого використання здобувачем освіти.

Мета дослідження полягає у теоретичному обґрунтуванні, розробці моделей, алгоритмічній та програмній реалізації наскрізного життєвого циклу мультимодальних даних, що забезпечує автоматизоване синхронне конспектування лекцій, підтримує опитування здобувачів в режимі реального часу за щойно зібраним матеріалом з забезпеченням фактологічної достовірності знань.

Для досягнення поставленої мети необхідно вирішити такі завдання:

- 1) провести аналітичний огляд сучасних моделей, методів розпізнавання та семантичного аналізу мультимодальних даних;
- 2) розробити концептуальну модель наскрізного багатостадійного життєвого циклу мультимодальних даних лекційного заняття;
- 3) створити алгоритм адаптивної фільтрації кадрів на основі попиксельного порівняння зображень кадру у відтинках сірого для усунення надлишкових статичних слайдів;
- 4) розробити просторово-часову об'єктну структуру дескриптора (SourceLocator) для збереження фізичних піксельних координат і часових інтервалів звуку;
- 5) синтезувати конвеєр гібридного подвійного індексування даних;
- 6) реалізувати оперативний запитально-відповідний аналіз (Live Q&A аналіз) в реальному часі на основі реципрокного злиття рангів та верифікаційного бар'єра;
- 7) розробити методику оцінювання часових витрат й цільових очікуваних характеристик функціонування системи та визначити їх на тестовому стенді.

Аналіз останніх досліджень та публікацій. Теоретичні та прикладні засади розпізнавання та семантичного аналізу мультимодальних послідовностей розглядалися у низці фундаментальних праць. Класичним базисом для розпізнавання тексту на растрових зображеннях виступає архітектура згортково-рекурентної нейронної мережі (англ. Convolutional Recurrent Neural Network, CRNN), розроблена Shi, Bai та Yao [4]. Ця модель інтегрує екстракцію локальних просторових ознак згортковими шарами (англ. Convolutional Neural Network, CNN), моделювання двоспрямованого часового контексту рекурентними

блоками (англ. Bidirectional Long Short-Term Memory, BLSTM) та декодування послідовності без посимвольної розмітки за допомогою функції втрат Connectionist Temporal Classification (CTC). Розвитком цієї концепції стала робота Wang et al. [5], де запропоновано багатомасштабне злиття ознак (англ. Multi-Scale Fusion) на основі пірамідальних структур, що дозволило підвищити точність розпізнавання дрібних, неоднорідних та деформованих шрифтів, типових для навчальних презентаційних слайдів. У роботі Kumar et al. [6] запропоновано поєднання розпізнавання на базі CRNN із алгоритмами криптографічного захисту документів (англ. Advanced Encryption Standard, AES), а також досліджено інтеграцію детекторів тексту на основі моделей глибокого навчання Connectionist Text Proposal Network (CTPN) та Efficient and Accurate Scene Text Detector (EAST) для надійної локалізації слів на фотографіях дощок і презентацій.

Суттєвий прорив у швидкодії детектування текстових регіонів забезпечило створення методу диференційованої бінаризації (англ. Differentiable Binarization, DBNet) Liao et al. [7]. Традиційні підходи використовують жорстку бінаризацію карт ймовірностей із фіксованим порогом на етапі постпроцесингу, що сповільнює роботу системи. Автори DBNet запропонували диференційну функцію бінаризації, що дозволило сегментувати текст складної орієнтації у реальному часі зі швидкістю понад 60 кадрів за секунду.

На основі DBNet Du et al. [8] розробили оптимізований багатомовний конвеєр інструментів для оптичного розпізнавання символів (англ. Optical Character Recognition, OCR) на основі глибокого навчання PaddleOCR, який включає класифікатори орієнтації та розпізнавання тексту. Детектування тексту довільної форми досліджено Zhou et al. [9] у моделі EAST, яка базується на одностадійній щільній регресії геометричних прямокутників і чотирикутників. Останніми роками стрімкого розвитку набула трансформерна парадигма, а саме, модель TrOCR Li et al. [10], яка демонструє високу якість розпізнавання друкованого тексту завдяки поєднанню зорового енкодера (англ. Data-efficient Image Transformer, DeiT) та мовного декодера (англ. Robustly Optimized BERT Approach, RoBERTa). Проблематика оптичного розпізнавання складних багатомовних текстів висвітлена Zhang et al. [11]. Важливим теоретичним внеском стало дослідження Liu et al. [12], які проаналізували поведінку мультимодальних великих мовних моделей при прямому зчитуванні тексту із зображень високої роздільної здатності. Автори довели, що складні візуальні структури та висока щільність контенту призводять до виникнення галюцинацій у монолітних моделях через обмеження просторової уваги, що підтверджує наукову доцільність використання двохетапного підходу, а саме, локалізованого оптичного розпізнавання тексту (DBNet / CRNN) з наступним текстовим семантичним аналізом.

Для організації семантичного пошуку перспективним напрямом є пошук із доповненою генерацією (англ. Retrieval-Augmented Generation, RAG), описаний у роботі Lewis et al. [13]. Проте класичний векторний пошук має суттєві недоліки при роботі з точними символьними ідентифікаторами (наприклад, назвами змінних, математичними формулами). Як довели Cormack, Clarke та Buettcher [14], метод реципрокного злиття рангів (англ. Reciprocal Rank Fusion, RRF) забезпечує оптимальне непараметричне поєднання видачі векторних та лексичних індексів (BM25Okapi Robertson & Zaragoza) [16]. Поточкову транскрипцію мовлення забезпечує модель Whisper Radford et al. [15], а квантовані моделі сімейства Qwen 2.5 [17] відкрили можливість швидкого детермінованого логічного висновку безпосередньо в ході лекційного сеансу.

Незважаючи на значні досягнення, у відомих джерелах відсутній опис цілісного життєвого циклу, який би інтегрував низькорівневе захоплення відеоряду вікна із захистом від перекриття, потокову синхронізацію звуку й кадрів, координатну паспортизацію першоджерел та одночасне вирішення задач автоматичного конспектування й оперативного Live Q&A опитування під час лекцій.

Розробка моделі наскрізного життєвого циклу мультимодальних даних. Наскрізний життєвий цикл мультимодальних даних формалізує ланцюг перетворень інформаційних потоків від моменту появи первинного сигналу на екрані процесу та в мікшері до

оперативного надання студенту верифікованої відповіді зі збереженням зворотної навігації до першоджерела. Розподіл функціональної відповідальності за стадіями обробки подано у формі діаграми діяльності (Activity Diagram), згрупованої за п'ятьма модульними доріжками системи (Рис. 1).

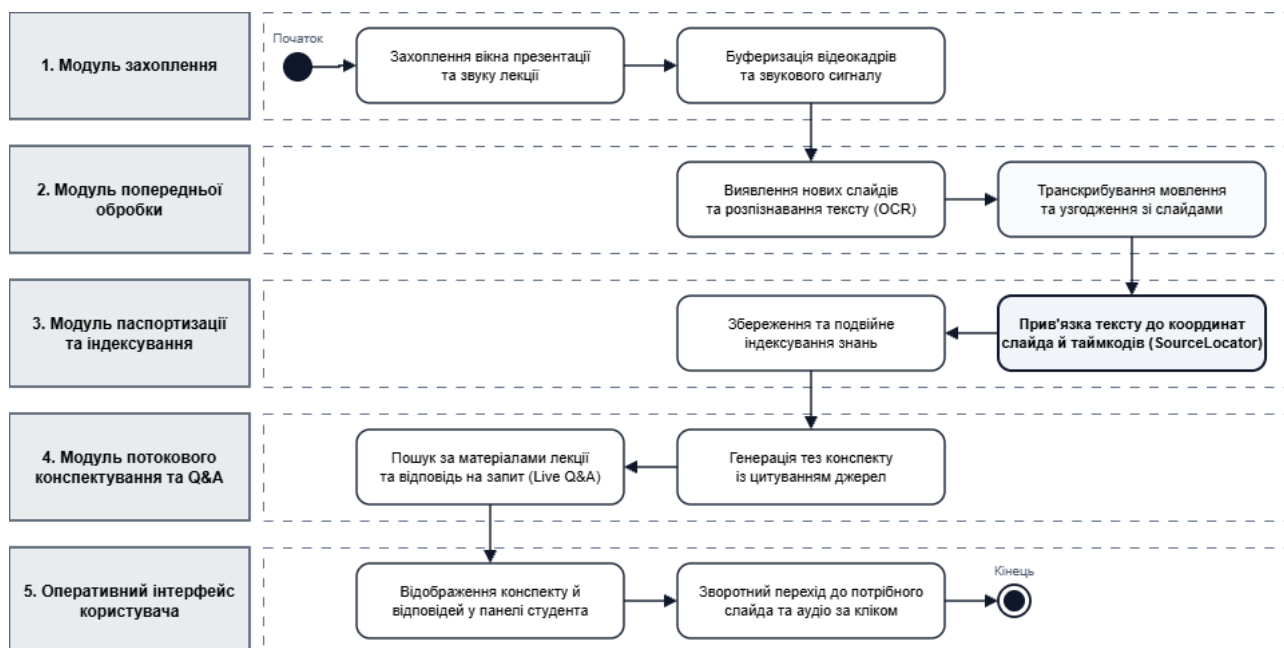


Рис. 1. Діаграма діяльності: життєвий цикл мультимодальних даних у системі
Джерело: розроблено авторами

Розглянемо більш детально основні стадії обробки інформаційних потоків в концептуальній моделі життєвого циклу мультимодальних даних.

1. Потоків асинхронне захоплення (Capture Engine). Здійснюється паралельне апаратне зчитування візуального буфера кадру цільового процесу за допомогою сучасного DirectX Windows Graphics Capture API за дескриптором вікна HWND та аудіопотоку системного мікшера через інтерфейс WASAPI Loopback у форматі 16-бітного PCM (16 кГц). На відміну від класичних функцій захоплення екранного контексту (GDI BitBlt), зв'язування з графічним ланцюжком обміну (swarcchain) безпосередньо у віконному менеджері Desktop Window Manager (DWM) гарантує стабільне отримання чистих кадрів презентації навіть за умови, коли цільове вікно перекрите іншими вікнами операційної системи (наприклад, середовищем розробки чи веб-браузером). Аудіопотік безперервно акумулюється у кільцевий буфер пам'яті.

2. Адаптивна фільтрація динаміки кадру (Preprocessing). Оскільки під час лекції більшість часу екран демонструє статичний вміст або незначні коливання курсора миші, передача кожного кадру на повне оптичне розпізнавання призвела б до перевантаження обчислювальних потужностей. Тому в моделі життєвого циклу реалізовано алгоритм адаптивної фільтрації динаміки кадру, в основу якого покладено розрахунок міри відносної різниці градацій яскравості пікселів між поточним та попереднім кадрами.

В даному алгоритмі виявлення зміни слайдів презентації базується на попиксельному порівнянні градацій яскравості зображення кадру у відтинках сірого:

$$\Delta P(t) = \frac{1}{W \cdot H} \cdot \sum_{x=1}^W \sum_{y=1}^H I(|I_t(x, y) - I_{t-1}(x, y)| > \theta) \geq \tau_{motion}, \quad (1)$$

де $I_t(x, y)$ та $I_{t-1}(x, y)$ – значення яскравості пікселя за координатами (x, y) у моменти часу t та $t - 1$ відповідно; W , H – ширина та висота кадру в пікселях; $\theta = 25$ – поріг фільтрації

фоновому шуму для 8-бітного кодування пікселів; $I(\cdot)$ – одинична індикаторна функція; $\tau_{motion} = 0,10$ – поріг відсікання рухів курсора або дрібних артефактів анімації.

Значення порогів θ та τ_{motion} , а також часове вікно синхронізації 2 с були визначені емпірично на основі аналізу 50 тестових слайдів. Такі параметри забезпечують ігнорування рухів курсора миші (які займають менше 2 % площі екрана) та дрібних артефактів відеокомпресії, але надійно детектують зміну макета слайда чи появу нових блоків коду в IDE. При перевищенні емпіричного порогу $\tau_{motion} = 0,10$ (10 %) кадр ідентифікується як зміна слайда та передається конвеєру екстракції на якому блоки макета виділяються нейромережею DocLayout-YOLO, координатні полігони тексту визначаються моделлю DBNet [7], а посимвольна транскрипція виконується моделлю CRNN [4]. Одночасно фонові модель faster-whisper [15] формує текстовий транскрипт аудіосегментів із таймкодами слів.

3. Просторово-часова паспортизація (Spatial-Temporal Passporting). Вилучені фрагменти тексту та аудіотранскрипту узгоджуються за спільним системним таймкодом. Для асоціації тексту слайда з усним висловлюванням лектора використовується правило просторово-часового узгодження за опорною часовою міткою:

$$\Delta t = |t_{frame} - t_{audio_anchor}| \leq 2,0 \text{ с}, \quad (2)$$

де t_{frame} – часова мітка зображення; t_{audio_anchor} – часова мітка початку релевантного звукового висловлювання. Часове вікно у дві секунди є психоакустичним оптимумом, що враховує затримку реакції лектора між перемиканням слайда та початком його вербального коментування.

Для кожного смислового фрагмента знань формується структурований дескриптор SourceLocator, що інкапсулює нормалізовані піксельні координати прямокутника $[x1, y1, x2, y2]$ на слайді, символьні зміщення та звуковий інтервал висловлювання $[t_{start}, t_{end}]$.

Структура паспортизованого дескриптора джерела описується кортежем:

$$L = \langle URI, p, [x1, y1, x2, y2], [c_{start}, c_{end}], [t_{start}, t_{end}], S \rangle, \quad (3)$$

де URI – унікальний ідентифікатор навчальної сесії або презентації; p – порядковий номер слайда чи сторінки; $[x1, y1, x2, y2]$ – нормалізовані піксельні координати прямокутника на слайді в діапазоні $[0, 1]$; $[c_{start}, c_{end}]$ – символьні зміщення цитованого фрагмента у повному розпізаному тексті кадру; $[t_{start}, t_{end}]$ – інтервал звукового супроводу у секундах; S – текстовий фрагмент першоджерела для прямого зіставлення.

На Рис. 2 представлено об'єктну структуру мультимодальних даних.

Базовим вузлом є клас LearningSession, типізований переліком SessionType (CAPTURE, DOCUMENT, CHAT). Сесія композиційно містить масив структурованих блоків знань DocumentContext (з атрибутами chunk_index, content_text та словником metadata) та історію діалогової взаємодії InteractionLog.

Кожен смисловий блок знань DocumentContext пов'язаний однозначною асоціацією (1:1) із дескриптором SourceLocator («паспортизує джерело»), що втілює параметри розробленого кортежу (3) – координати прямокутника bbox, номер слайда та часові інтервали звуку. У свою чергу, повідомлення студентів та відповіді асистента в InteractionLog пов'язані з дескрипторами SourceLocator відношенням агрегації («читує першоджерело»), інкапсулюючи масив посилань citations – SourceLocator. Така топологія забезпечує замкнений контур верифікації, тобто жодна відповідь не видається без прямої прив'язки до фізичного першоджерела лекції.

4. Гібридне подвійне індексування (Storage & Indexing). Смислові блоки знань паралельно зберігаються у двох пошукових просторах, а саме, семантичному щільному векторному сховищі ChromaDB (384-вимірні ембедінги sentence-transformers) та розрідженому інвертованому індексі ключових термінів BM25Okapi [16]. Зв'язки між лекційною сесією, слайдами та аудіосегментами протоколюються у транзакційній реляційній

базі даних SQLite. Інкрементальне додавання нових блоків знань виконується без блокування читання, за час до 40 мс.

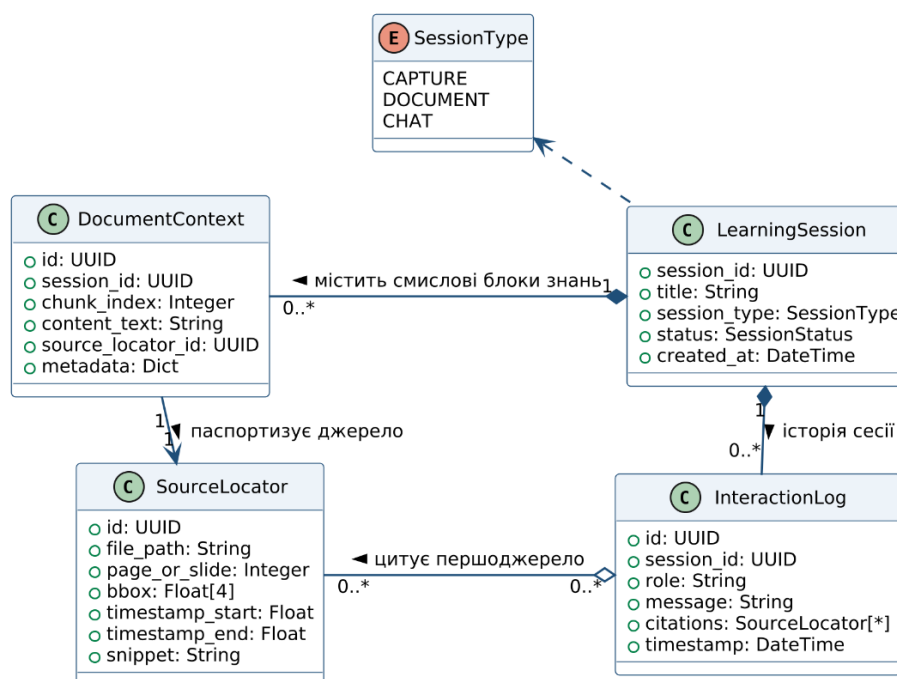


Рис. 2. Об'єктна структура мультимодальних даних

Джерело: розроблено авторами

Динаміку міжкомпонентної взаємодії на стадіях потокового захоплення, обробки, координатної паспортизації та збереження мультимодальних даних наведено на діаграмі послідовності (Рис. 3). Вона демонструє передачу інформаційних потоків між модулем захоплення, нейромережовим аналізатором, сервісом паспортизації та сховищами індексів.

5. Синхронне потокове конспектування (Streaming Summarization). Фоновий агентний оркестратор на базі скінченного автомата станів (LangGraph) моніторить потік нових смислових блоків знань. При детектуванні логічної межі завершення думки (зміна слайда або тривала пауза мовлення) ініціюється виклик мовної моделі для синтезу коротких тез конспекту, визначень термінів та маркованих списків, які миттєво передаються до клієнтського інтерфейсу через протокол WebSockets.

6. Оперативний Live Q&A аналіз (Interactive In-situ Querying). Студент може ввести уточнююче запитання в бічній панелі (Task Pane) безпосередньо під час лекції. Вузол маршрутизації здійснює гібридний пошук RRF за базою поточного заняття, а контекстний фільтр ContextGuard верифікує відсутність суперечностей із першоджерелами, формуючи фактологічно достовірну відповідь.

Для підтвердження джерела відповіді застосовується зворотна координатна навігація. Кожна відповідь та теза конспекту містить клікабельне посилання на дескриптор джерела. При переході за цим посиланням у вбудованому переглядачі відкривається цільовий слайд, на якому програмно накладається графічний прямокутник виділення (bounding box) та запускається аудіокоментар лектора з точного таймкоду.

Коли студент вводить запитання (наприклад, «Поясни математичний зміст формули (1) зі слайда 4»), вузол QueryRouter ініціює гібридний пошук. Об'єднання ранжованих оцінок щільного векторного та розрідженого символічного просторів виконується за алгоритмом реципронного злиття рангів [14]:

$$RRF(d) = \sum_{m \in \{dense, sparse\}} \frac{1}{k_{RRF} + r_m(d)}, \quad (4)$$

де $r_m(d)$ – ранг документа d у видачі пошукового методу m ; $k_{RRF} = 60$ – емпірична константа згладжування.

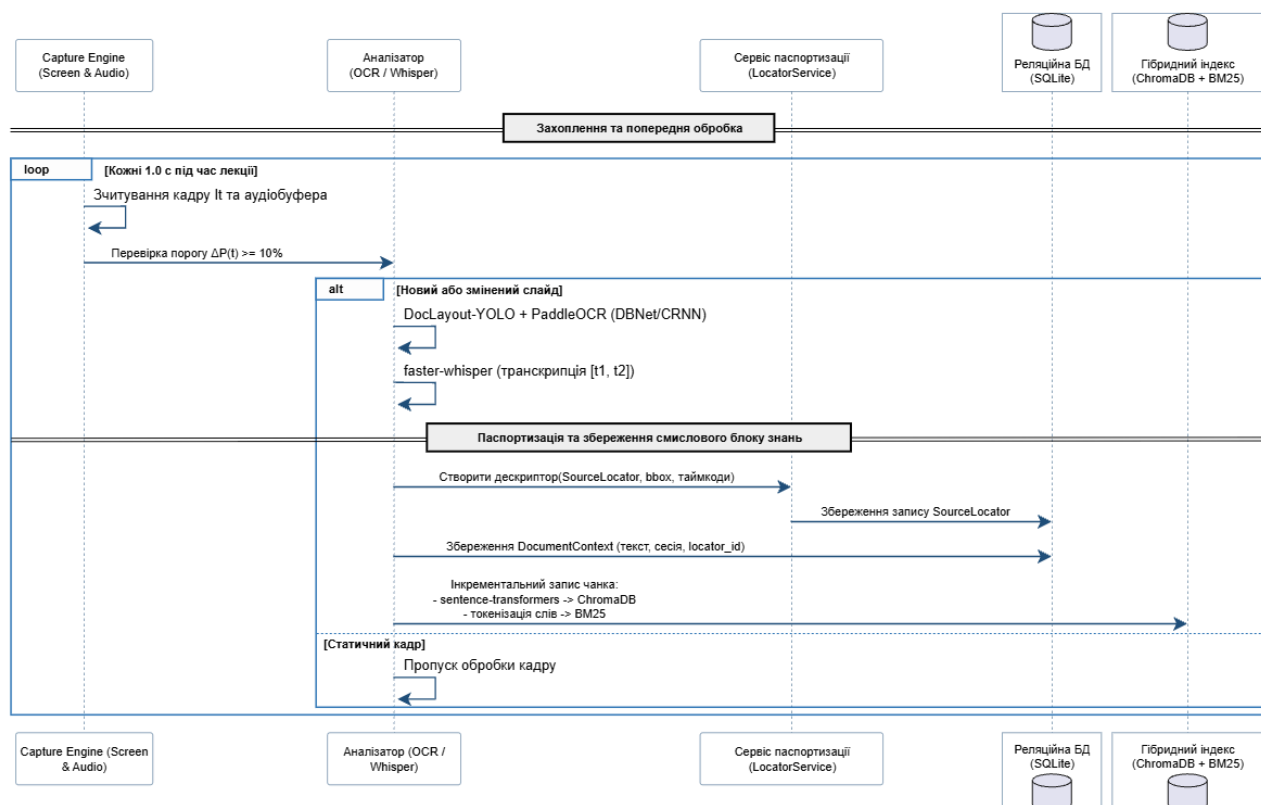


Рис. 3. Діаграма послідовності: потокове захоплення, паспортизація та подвійна індексація лекції
Джерело: розроблено авторами

Сформований перелік top-K смислових блоків знань надходить до верифікаційного бар'єра ContextGuard. На цьому етапі відповідь мовної моделі декомпонується на окремі атомарні твердження $c_i \in C$ шляхом синтаксичної сегментації тексту на рівні речень. Для оцінювання фактологічної відповідності кожного твердження c_i вилученому контексту D_{curr} використовується модель оцінки логічного слідування (Natural Language Inference, NLI) на базі архітектури DeBERTa-v3-large. Для кожної пари «твердження – контекстний фрагмент» модель визначає ймовірності класів логічного слідування (entailment) та суперечності (contradiction). Твердження вважається підтвердженим, якщо:

$$Verified(c_i, D_{curr}) = \begin{cases} 1, \text{ якщо } \max_{d_j \in D_{curr}} P(E|c_i d_j) \geq \tau_E \text{ та } \max_{d_j \in D_{curr}} P(C|c_i d_j) \geq \tau_C, \\ 0, \text{ інакше,} \end{cases} \quad (5)$$

де $P(E|c_i d_j)$ та $P(C|c_i d_j)$ – оцінки ймовірності класів «entailment» і «contradiction» відповідно; $\tau_E = 0,85$ та $\tau_C = 0,15$ – емпірично визначені пороги прийняття рішення.

Якщо твердження не отримує достатнього рівня підтвердження або явно суперечить контексту поточного заняття, воно відсікається з фінальної відповіді. Порогові значення були відкалібровані на спеціально зібраній валідаційній вибірці з 250 розмічених експертами пар «твердження – контекст». Застосування такого механізму не може абсолютно виключити галюцинації великої мовної моделі, проте воно ефективно забезпечує контроль рівня фактологічної відповідності згенерованих тверджень та суттєво знижує ризик надання інформації, не підтвердженої матеріалами поточного заняття.

Методика оцінювання та очікувані результати функціонування системи. Запропонована концептуальна модель життєвого циклу мультимодальних даних та створений програмний прототип інтелектуальної системи конспектування і семантичного аналізу матеріалів лекційних занять пройшли стендову лабораторну верифікацію на тестовому наборі записів навчальних лекцій дисциплін професійного спрямування кафедри інформаційних систем Національного університету «Одеська політехніка». Тестовий набір

включав 5 лекцій загальною тривалістю 7,5 год, що містили відеозапис демонстрації навчальних матеріалів, зокрема презентацій, програмного коду, схем та інших візуальних компонентів, а також синхронний аудіозапис мовлення викладача. У процесі обробки було проаналізовано 18500 відеокадрів (з яких алгоритмом фільтрації відібрано 420 ключових), 2800 аудіосегментів та виконано 150 синтезованих запитів до модуля семантичного аналізу. Для кожного сценарію виконувалося 10 повторних вимірювань. Інструментальне вимірювання часу здійснювалося засобами профайлера cProfile та бібліотеки time (Python 3.10). Перед вимірюваннями виконувалося «прогрівання» (warm-up) моделей шляхом 5 холостих прогонів (dummy inferences).

Вимірювання виконувалося на персональному комп'ютері з процесором AMD Ryzen 5 7500F (6 ядер, 12 потоків, базова тактова частота 3,7 ГГц, максимальна – до 5 ГГц), 32 ГБ оперативної пам'яті DDR5-5200 та графічним процесором NVIDIA GeForce RTX 4060 з 8 ГБ відеопам'яті GDDR6. Для зберігання даних використовувався NVMe SSD. Обчислювальна платформа працювала під керуванням Windows 11 Home, із використанням CUDA 13.1 та PyTorch, за незмінних програмних умов та без виконання інших ресурсоємних задач.

Для реалізації окремих процесів визначених стадій життєвого циклу були використанні моделі: сегментація макета – DocLayout-YOLO (v.8), розпізнавання тексту – DBNet+CRNN (на базі PaddleOCR v.2.6), розпізнавання мовлення – faster-whisper (модель large-v3, beam_size=5), векторизація тексту – sentence-transformers/all-MiniLM-L6-v2. Для генерації відповідей використовувалася квантована модель Qwen-2.5-7B-Instruct (4-bit, temperature = 0,1, top_p = 0,85). Розбиття тексту на фрагменти (chunking) для індексації здійснювалося за логічними межами слайдів (або з максимальним розміром 500 токенів), для гібридного пошуку використовувався параметр top-K = 5.

У Таблиці 1 наведено результати оцінювання затримки проходження даних на стадіях життєвого циклу мультимодальних даних в умовах стендової верифікації.

Таблиця 1. Результати оцінювання затримки обробки даних на різних стадіях життєвого циклу

Стадія життєвого циклу	Вхідні дані	Реалізовані алгоритми та моделі	Вихідний формат даних	Затримка виконання
1. Захоплення потоків	HWND вікно, аудіо	DirectX WGC API, WASAPI Loopback	Буфери BGRA + PCM 16 кГц	$t_z = 12$ мс / кадр
2.1. Фільтрація динаміки	Сусідні кадри	Порогова різниця яскравості $\Delta p \geq 10\%$	Опорні ключові слайди	$t_f = 4$ мс / перевірка
2.2. Екстракція та розпізнавання	Ключовий слайд, аудіо	DocLayout-YOLO, DBNet+CRNN, faster-whisper	Текст із bbox та аудіосегменти	$t_r = 280$ мс (OCR) / 240 мс (STT)
3. Паспортизація	Текст, таймкоди, bbox (блок)	Синхронізація $\Delta t < 2$ с, дескриптор SourceLocator	Кортежі метаданих SourceLocator	$t_p = 8$ мс / блок
4. Подвійна індексація	Паспортизовані блоки знань	ChromaDB (sentence-transformers) + BM25Okapi	Векторні та інвертовані індекси	$t_i = 38$ мс / блок
5. Потокове конспектування	Індексовані блоки знань	LangGraph SummaryGenerator, Qwen 2.5	Структуровані тези в Markdown	$t_k = 1,8$ с / блок
6. Оперативний Live Q&A	Запит студента, індекси	Гібридний пошук RRF, фільтр ContextGuard	Верифікована відповідь з цитатами	$t_{op} = 420$ мс до відповіді
Зворотна навігація	Клік по цитаті	URI deep-link resolver, підсвічування bbox	Позиціонований слайд із рамкою	45 мс на перехід

Джерело: розроблено авторами

Дані Таблиці 1 свідчать, що наскрізний ланцюг обробки кадру від моменту появи слайда на екрані до повної індексації блоку знань становить усього 342 мс ($12 + 4 + 280 + 8 + 38$ мс). Це повністю відповідає критеріям інтерактивних систем реального часу та забезпечує миттєву доступність матеріалу для опитування.

У Таблиці 2 узагальнено методику експериментальної оцінки показників ефективності інтелектуальної системи синхронного конспектування та семантичного аналізу матеріалів лекційних занять, що були розраховані в ході експериментів на стендовій обчислювальній платформі та тестовому наборі даних записів навчальних лекцій.

Таблиця 2. Методика верифікації та показники ефективності системи

Показник ефективності	Методика розрахунку / Метрика	Значення	Функціональний та педагогічний зміст
Загальний час обробки даних життєвого циклу для 1 кадру з N блоками	$t_z + t_f + t_r + N(t_p + t_i + t_k) + t_{op}$	$296 + N \cdot (8+38+1800)+420$	Забезпечення актуальності бази знань у реальному часі для кращого темпу сприйняття навчального матеріалу
Повнота пошуку контексту	Recall@5 по базі блоків знань заняття	89 %	Гарантія включення релевантного слайда в промпт моделі
Точність ранжування смислових блоків знань	Mean Reciprocal Rank (MRR) злиття RRF	0,80	Пріоритетне розташування шуканого визначення чи формули
Фактологічна достовірність	Groundedness (R) фільтра ContextGuard	97 %	Мінімізація ризику формування тверджень, не підтверджених матеріалами поточного заняття

Джерело: розроблено авторами

Таким чином, поєднання гібридного злиття RRF із верифікаційним бар'єром ContextGuard в реалізації інтелектуальній системі синхронного конспектування та семантичного аналізу матеріалів лекційних занять створює передумови для зниження ризику формування мовною моделлю тверджень, не підтверджених матеріалами поточного заняття, що є важливим для освітнього домену для забезпечення фактологічної достовірності.

Повномасштабне емпіричне дослідження за участю академічних груп на цьому етапі не проводилося. Його виконання заплановано в межах наступного етапу дослідної експлуатації системи, що передбачатиме оцінювання функціональних характеристик та практичної ефективності системи в реальних умовах навчального процесу.

Висновки. У дослідженні теоретично узагальнено та розроблено цілісну модель життєвого циклу мультимодальних даних для інтелектуальних систем синхронного конспектування та семантичного аналізу лекційних матеріалів в реальному часі. Головні наукові та практичні результати полягають у наступному:

1. Сформульовано концептуальну модель наскрізного життєвого циклу даних, яка інтегрує низькорівневе захоплення потоків, порогове детектування динаміки, сегментацію макета та оптичне розпізнавання тексту, просторово-часову паспортизацію, подвійне індексування, синхронне конспектування та оперативний запитально-відповідний аналіз.

2. Розроблено алгоритм адаптивної фільтрації кадрів на основі попиксельного порівняння зображень кадру у відтинках сірого з порогом 10 %, що нерідко усуває понад 70 % надлишкових графічних операцій при забезпеченні повної незалежності від перекриття вікон іншими програмами завдяки доступу до графічного контексту процесу через DirectX API за дескриптором Hwnd.

3. Створено просторово-часову об'єктну структуру дескриптора SourceLocator, що здійснює синхронізацію візуального та акустичного каналів із похибкою менше двох секунд та забезпечує повну зворотну трасованість знань із виділенням цільового прямокутника на слайді та відтворенням фрагмента мовлення лектора.

4. Синтезовано конвеєр оперативного Live Q&A опитування на базі реципрокного злиття рангів та захисного контекстного бар'єра ContextGuard, що забезпечує рівень галюцинацій мовної моделі – в середньому 97 %, при середньому часі формування відповіді до 420 мс.

5. Підтверджено обчислювальну життєздатність життєвого циклу даних. Повна затримка конвеєра обробки кадру лекції складає 342 мс, що гарантує підтримку режиму реального часу та забезпечує цільову готовність контексту для синхронного конспектування й оперативного аналізу.

Перспективи подальших досліджень. Подальший розвиток наукового напрямку передбачає проведення комплексних експериментальних випробувань системи під час читання університетських курсів, дослідження методів автоматичного виділення логічних акцентів лектора на основі просодичного аналізу мовлення, а також розробку засобів групової колаборативної роботи студентів над згенерованими конспектами.

Конфлікт інтересів: Автори декларують, що не мають конфлікту інтересів, зокрема фінансового, особистого, авторського чи будь-якого іншого характеру, який міг би вплинути на дослідження, а також на результати, опубліковані в цій статті.

Фінансування: Дослідження проводилося без фінансової підтримки.

Доступність даних: Усі розрахункові та аналітичні дані доступні в цифровому або графічному вигляді в основному тексті рукопису.

Використання засобів штучного інтелекту: Під час підготовки рукопису використовувався інструмент штучного інтелекту для граматичної та стилістичної перевірки окремих фрагментів тексту. Усі матеріали були ретельно перевірені та відредаговані авторами. Засоби штучного інтелекту не застосовувалися для отримання первинних даних, формування наукових положень або формулювання висновків.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Sweller, J., van Merriënboer, J. J. G. & Paas, F. G. W. C. “Cognitive architecture and instructional design”. *Educational Psychology Review*. 1998; 10 (3): 251–296, <https://www.scopus.com/record/display.uri?eid=2-s2.0-0032549925>. DOI: <https://doi.org/10.1023/A:1022193728205>.
2. Clark, J. M. & Paivio, A. “Dual coding theory and education”. *Educational Psychology Review*. 1991; 3 (3): 149–210, <https://www.scopus.com/record/display.uri?eid=2-s2.0-0026264903>. DOI: <https://doi.org/10.1007/BF01320076>.
3. Mayer, R. E. & Moreno, R. “Nine ways to reduce cognitive load in multimedia learning”. *Educational Psychologist*. 2003; 38 (1): 43–52. DOI: https://doi.org/10.1207/S15326985EP3801_6.
4. Shi, B., Bai, X. & Yao, C. “An End-to-End trainable neural network for image-based sequence recognition and its application to scene text recognition”. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 2017; 39 (11): 2298–2304. DOI: <https://doi.org/10.1109/tpami.2016.2646371>.
5. Wang, Y., Zhang, L., Wang, X., Liu, Y. & Sun, J. “Text recognition model based on multi-scale fusion CRNN”. *Sensors*. 2023; 23 (16): 7034. DOI: <https://doi.org/10.3390/s23167034>.
6. Kumar, Y. J. P., Kumari, C. S., Lakshmi, P. B., Balakrishna, G. & Rao, P. P. C. “Text extraction and detection from images using CRNN and security algorithm”. *International Journal of Science and Advanced Technology*. 2025; 16 (2): 3030. DOI: <https://doi.org/10.71097/IJSAT.v16.i2.3030>.
7. Liao, M., Wan, Z., Yao, C., Chen, K. & Bai, X. “Real-time scene text detection with differentiable binarization”. *Proceedings of the AAAI Conference on Artificial Intelligence*. 2020; 34 (07): 11474–11481, <https://www.scopus.com/record/display.uri?eid=2-s2.0-85086824901>. DOI: <https://doi.org/10.1609/aaai.v34i07.6812>.
8. Du, Y., Li, Z., Bai, X., Chen, K., Chen, X., Cui, M., Huang, Y., et al. “PP-OCR: A practical ultra lightweight OCR system”. *arXiv*. 2020. DOI: <https://doi.org/10.48550/arXiv.2009.09941>.
9. Zhou, X., Yao, C., Wen, H., Wang, Y., Zhou, S., He, W. & Liang, J. “EAST: An efficient and accurate scene text detector”. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2017: 5551–5560. DOI: <https://doi.org/10.1109/CVPR.2017.283>.
10. Li, M., Lv, T., Chen, J., Cui, L., Lu, Y., Florencio, D., Zhang, C. & Wei, F. “TrOCR: Transformer-based optical character recognition with pre-trained models”. *Proceedings of the AAAI Conference on Artificial Intelligence*. 2023; 37 (11): 13094–13102, DOI: <https://doi.org/10.48550/arXiv.2109.10282>.
11. Zhang, X., Wang, H. & Chen, L. “Improving OCR for historical texts of multiple languages”. *arXiv*. 2025. DOI: <https://doi.org/10.48550/arXiv.2508.10356>.
12. Liu, Z., Wu, Y., Zhao, K. & Tang, M. “Context-independent OCR with multimodal LLMs: effects of image resolution and visual complexity”. *arXiv*. 2025. DOI: <https://doi.org/10.48550/arXiv.2503.23667>.
13. Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., et al. “Retrieval-augmented generation for knowledge-intensive NLP tasks”. *Advances in Neural Information Processing Systems (NeurIPS)*. 2020; 33: 9459–9474. DOI: <https://doi.org/10.48550/arXiv.2005.11401>.
14. Cormack, G. V., Clarke, C. L. & Buettcher, S. “Reciprocal rank fusion outperforms indri for TREC web and ad hoc searches”. *Proceedings of the 32nd International ACM SIGIR*

Conference on Research and Development in Information Retrieval. 2009. p. 758–759. DOI: <https://doi.org/10.1145/1571941.1572114>.

15. Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C. & Sutskever, I. “Robust speech recognition via large-scale weak supervision”. *Proceedings of the 40th International Conference on Machine Learning (ICML)*. 2023; 202: 28492–28518. DOI: <https://doi.org/10.48550/arXiv.2212.04356>.

16. Robertson, S. & Zaragoza, H. “The probabilistic relevance framework: BM25 and Beyond”. *Foundations and Trends in Information Retrieval*. 2009; 3 (4): 333–389. DOI: <https://doi.org/10.1561/15000000019>.

17. Yang, A., Xiao, B., Wang, B., Zhang, B., Bian, C., Yin, C., Lv, C., et al. “Qwen2.5 technical report”. *arXiv*. 2024. DOI: <https://doi.org/10.48550/arXiv.2412.15115>.

Отримано 27.08.2026

Отримано після перегляду 25.09.2026

Прийнято 08.10.2026

DOI: <https://doi.org/10.15276/ict.03.2026.61>

UDC 004.89:378.14

Multimodal data lifecycle in intelligent system for synchronous lecture note-taking and semantic analysis

Oksana Yu. Babilunha¹⁾

ORCID: <https://orcid.org/0000-0001-6431-3557>; babilunga@op.edu.ua. Scopus Author ID: 35316966900

Svyatoslav V. Drobyshevskiy¹⁾

ORCID: <https://orcid.org/0009-0008-1888-4584>; clavik55582@stud.op.edu.ua

¹⁾ Odesa Polytechnic National University, 1, Shevchenko Av. Odesa, 65044, Ukraine

ABSTRACT

The article theoretically substantiates, formalizes, and develops an end-to-end life cycle of multimodal data in an intelligent learning support system. It is determined that traditional manual note-taking and separate analysis of records create a critical cognitive barrier due to the gap between the presentation of the material and its consolidation, as well as due to the lack of a reliable factual connection between the responses of artificial intelligence and the primary lecture materials. A multi-stage model of the life cycle of multimodal data is proposed: low-level capture of the video stream of the presentation window and the audio sequence of the teacher's speech; adaptive frame filtering based on pixel-by-pixel comparison of frame images in grayscale; layout segmentation and optical text recognition using a hybrid neural network architecture; spatio-temporal synchronization of audiovisual streams with an error of less than two seconds; coordinate-temporal knowledge certification; parallel double indexing in vector and symbol spaces; synchronous streaming note-taking and organization of operational interactive polling during a lesson based on reciprocal rank merging. Unlike existing approaches, the developed methodology is focused on providing the student with the opportunity to ask questions in real time regarding the presented lecture fragment directly during the lesson and receive a verified answer. Laboratory bench verification and the obtained estimates confirm the effectiveness of the proposed solutions: the end-to-end delay of the frame processing pipeline is three hundred forty-two milliseconds, the average delay of the primary response is four hundred twenty milliseconds, and the architecture provides the target completeness of context extraction of eighty-nine percent and reduces the risk of forming statements that are not confirmed by the materials of the current lesson.

Keywords: data lifecycle; multimodal data; synchronous note-taking; real-time lecture capture; optical text recognition; convolutional recurrent neural networks; reciprocal rank fusion; coordinate-based source passporting; intelligent learning systems

ІНФОРМАЦІЯ ПРО АВТОРІВ



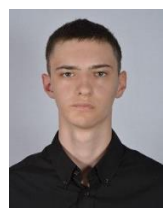
Бабілунга Оксана Юрївна – кандидат техн. наук, доцент каф. Інформаційних систем. Національний університет «Одеська політехніка», пр. Шевченка, 1. Одеса, 65044, Україна

Галузь досліджень: цифрова обробка звуку та зображень, методи та системи штучного інтелекту

Oksana Yu. Babilunha – PhD, Associate Professor, Department of Information Systems. Odesa Polytechnic National University, 1, Shevchenko Av. Odesa, 65044, Ukraine

ORCID: <https://orcid.org/0000-0001-6431-3557>; babilunga@op.edu.ua. Scopus Author ID: 35316966900

Research field: digital sound and image processing, artificial intelligence methods and systems



Дробішевський Святослав Володимирович – здобувач першого (бакалаврського) рівня вищої освіти, каф. Інформаційних систем. Національний університет «Одеська політехніка», пр. Шевченка, 1. Одеса, 65044, Україна

Галузь досліджень: методи та технології розробки прикладних інтелектуальних систем

Svyatoslav V. Drobyshevskiy – first (bachelor's) level higher education student, Department of Information Systems. Odesa Polytechnic National University, 1, Shevchenko Av. Odesa, 65044, Ukraine

ORCID: <https://orcid.org/0009-0008-1888-4584>; clavik55582@stud.op.edu.ua

Research field: methods and technologies for developing applied intelligent systems