DOI 10.31891/2307-5732-2024-339-4-29
УДК 004.8:025.171:681.3:930.25

ЛИПАК ГАЛИНА
Тернопільський національний технічний університет імені Івана Пулюя
https://orcid.org/0000-0001-9187-5758
e-mail: halyna.lypak@gmail.com
ЛИПАК ТАРАС
Тернопільський національний технічний університет імені Івана Пулюя
e-mail: taraslypak.work@gmail.com
КУНАНЕЦЬ НАТАЛІЯ
Національний університет «Львівська політехніка»
https://orcid.org/0000-0003-3007-2462
e-mail: neklviv@gmail.com

ПРОЄКТУВАННЯ ІНФОРМАЦІЙНОЇ СИСТЕМИ НА ОСНОВІ МАШИННОГО НАВЧАННЯ ДЛЯ ЗБЕРЕЖЕННЯ ТА КЛАСИФІКАЦІЇ АРТЕФАКТІВ ДОКУМЕНТАЛЬНОЇ СПАДЩИНИ

Численні виклики, з якими традиційно стикається сфера охорони культурної спадщини, вимагає сучасних підходів до їх подолання. Найбільшої актуальності в останні роки набувають підходи, що базуються на штучному інтелекті (ШІ), зокрема методи машинного навчання.

Дослідження в області машинного навчання та штучного інтелекту знаходяться на передовій сучасної науки. Застосування цих методів до класифікації документів культурної спадщини є актуальним напрямком, що поєднує передові технології з необхідністю збереження і вивчення історичної та культурної інформації та дозволяє використовувати найновіші наукові досягнення для вирішення практичних задач.

В статті проаналізовано ряд найактуальніших методів штучного інтелекту, що застосовуються сьогодні для збереження культурної спадщини, узагальнено їх характеристики та сфери прикладного застосування, описано переваги застосування машинного навчання в розпізнаванні документів культурної спадщини, розкрито механізм класифікації документів з використанням алгоритму випадкового лісу (Random Forest) як одного з найдієвіших інструментів навчання моделі; подано концептуальну модель інформаційної системи машинного навчання для класифікації документів.

Ключові слова: інформаційна система; штучний інтелект; машинне навчання; збереження документальної спадщини.

LYPAK HALYNA, LYPAK TARAS
Ternopil I. Pulyuy National Technical University
KUNANETS NATALIA
Lviv National Technical University

DESIGNING A MACHINE LEARNING-BASED INFORMATION SYSTEM FOR PRESERVING AND CLASSIFYING DOCUMENTARY HERITAGE ARTIFACTS

The modern world faces numerous challenges in the field of cultural heritage protection, such as natural disasters, climate change, wars, neglect, and limited resources. These factors can lead to irreversible losses of cultural sites and monuments.

Against the backdrop of these challenges, the use of artificial intelligence (AI) is becoming an increasingly relevant and promising area in the field of cultural heritage preservation. AI provides unique opportunities to effectively address a number of pressing issues that will help ensure the preservation and transmission of cultural heritage to future generations, improving the efficiency and quality of protection measures.

Research in machine learning and artificial intelligence is at the forefront of modern science. The application of these methods to the classification of cultural heritage documents is a relevant area that combines advanced technologies with the need to preserve and study historical and cultural information and allows the use of the latest scientific achievements to solve practical problems.

The article analyzes a number of the most relevant artificial intelligence methods used today for the preservation of cultural heritage, summarizes their characteristics and areas of application, in particular, such methods as machine and deep learning, computer vision, natural language processing, robotics, big data processing, semantic networks, and 3D modeling. The advantages of using machine learning in the recognition of cultural heritage documents and examples of their successful application are described. The mechanism of document classification using the Random Forest algorithm as one of the most effective model training tools is described in detail, and a detailed sequence of processes for applying the Random Forest algorithm for document classification is presented. A conceptual model of a machine learning information system for document classification is developed and presented, in particular, an activity diagram for such an information system is developed and further directions of research in this area are outlined.

Keywords: information system; artificial intelligence; machine learning; preservation of documentary heritage.

Постановка проблеми

Документи культурної спадщини мають велике історичне, наукове та культурне значення, а отже питання їх оцифрування та довгострокового збереження не втрачатиме актуальності. Систематична класифікація таких документів сприяє їх збереженню та полегшує доступ до них для дослідників, істориків та широкої громадськості. З розвитком цифрових технологій оцифрування документів набуває все ширших масштабів, що створює великі масиви даних. Традиційні методи класифікації стають неефективними при роботі з такими обсягами інформації, а отже доводиться звертатися до новітніх методів обробки інформації, базованих на штучному інтелекті. Ускладнює проблеми класифікації і той факт, що документи культурної

спадщини можуть бути представлені у різних форматах (тексти, зображення, аудіо, відео) та різними мовами, а точність класифікації може впливати на дослідницькі висновки та збереження історичної інформації.

Більшість з цих викликів здатне подолати машинне навчання. Воно дозволяє автоматизувати процес класифікації, значно зменшуючи час і ресурси, необхідні для обробки даних. Такі методи машинного навчання, як глибоке навчання, можуть ефективно працювати з різноманітними типами даних та мовними ресурсами. Сучасні методи машинного навчання, такі як нейронні мережі та алгоритми глибокого навчання, демонструють високу точність у задачах класифікації, а також вони можуть бути інтегровані з іншими сучасними технологіями, такими як обробка природної мови (NLP), розпізнавання зображень та автоматичний переклад. Це створює нові можливості для аналізу та інтерпретації документів культурної спадщини.

Аналіз досліджень та публікацій

У сфері охорони культурної спадщини дослідження останніх десятиліть спрямовувалися на створення цифрових ресурсів для керування колекціями історичної документальної спадщини, що взяло початок з масового оцифрування, а згодом переросло у розроблення складних систем, здатних швидко реагувати на різний контент та навіть вилучати його, наприклад, із соцмереж [1]. З огляду на важливість класифікації оцифрованих документів та їх анотування, для вирішення цих завдань почали широко застосовуватися методи машинного навчання, а особливо глибинне навчання та згорткові нейронні мережі (CNN) [2, 3]. Окремі дослідники [4] в межах експериментів інтегрують методи глибинного навчання для розпізнавання пам'яток за допомогою мобільних застосунків. Висока точність класифікації зображень архітектурної спадщини була досягнута фахівцями [3], які тестували згорткові нейронні мережі, задавши десять категорій архітектурних елементів. Інше дослідження [5], хоч і продемонструвало невисоку точність результатів прогностичних моделей, проте значно зменшило потребу використання людського ресурсу при ідентифікації публікацій у соціальних мережах авторів з постраждалих від стихійних лих місцевостей.

У 2016 р. група дослідників [6] успішно застосувала глибинне навчання для пошуку зображень в класичних латинських і грецьких написах, що забезпечило правильну ідентифікацію у понад 90% випадків пошуку в цифрових архівах. Найбільша кількість досліджень з використанням глибинних нейронних мереж стосується сфери образотворчого живопису – як для класифікації візуальних елементів, так і для визначення стилю живопису [7–9].

Не стоять осторонь використання методів ШІ і інституційні установи, серед яких варто відзначити Метрополітен-музей та його співпрацю з Microsoft і Массачусетським технологічним інститутом у сфері розроблення нових високотехнологічних продуктів, які дозволять виявляти приховані візерунки в творами мистецтва та автоматизувати процес тегування, що значно покращить для користувачів роботу з колекціями [10]. Разом з тим, для досягнення вищої якості опрацювання контенту, інформаційні системи на основі штучного інтелекту поступово впроваджуються і в бібліотеках [11].

Слід зазначити, що, на відміну від медичної галузі, в якій методи ШІ використовуються для діагностики захворювань, використання глибинного навчання при опрацюванні цінних зображень культурної спадщини ще не знайшло широкого використання.

Формулювання цілей статті

Метою роботи є розроблення концептуальної моделі інформаційної системи, що використовує методи машинного навчання для автоматизації процесів збереження, аналізу та класифікації артефактів документальної спадщини. Стаття також спрямована на дослідження ефективності використання сучасних алгоритмів машинного навчання в архівній справі та збереженні культурних цінностей, а також на оцінку можливостей їх інтеграції в існуючі архівні та музейні системи.

Для досягнення поставленої мети, необхідно вирішити наступні завдання:

- Провести аналіз існуючих методів і технологій збереження та класифікації артефактів документальної спадщини
- Обрати оптимальний метод навчання моделі та розроблення алгоритму його застосування для класифікації артефактів;
- Розроблення концептуальної моделі інформаційної системи для класифікації артефактів, що базується на обраному методі.

Виклад основного матеріалу

Аналіз публікацій теоретично-наукового та прикладного спрямування дозволив виокремити основні напрями сучасних досліджень, спрямованих створення методів штучного інтелекту для розпізнавання, класифікування артефактів культурного надбання. Для зручності їх подано в таблиці 1.

Розглянемо детальніше, у який спосіб можна застосовувати методи машинного навчання для класифікації документів, в тому числі історичних і тих, що належать до культурної спадщини. Зазначимо, що класифікація документів — це процес організації та групування документів на основі певних характеристик або критеріїв, таких як зміст, формат, тип, або призначення. Цей процес допомагає в упорядкуванні документів для їх легшого пошуку, зберігання та управління. Класифікація може бути ручною або автоматизованою, з використанням методів машинного навчання для аналізу та сортування документів за певними категоріями або темами. Машинне навчання для класифікації документів

використовує певні алгоритми, щоб автоматично визначати категорії або теми документів на основі їх вмісту. Опишемо, як це працює.

Таблиця 1

Методи ШІ, що використовуються в галузі культури [12]

Метод ШІ	Характеристики	Приклади застосування
Машинне навчання (ML)	Автоматичне виявлення шаблонів у даних; аналіз на основі отриманих даних	Класифікація культурних об'єктів, створення анотацій документів
Глибинне навчання (DL)	Використання методів навчання нейронних мереж при аналізі даних	Розпізнавання зображень, відновлення втрачених фрагментів
Комп'ютерний зір (CV)	Розпізнавання та опрацювання візуальної інформації	Віртуальна реконструкція; автоматична анотація зображень
Обробка природної мови (NLP)	Аналіз та генерація текстової інформації	Семантична анотація історичних документів; переклад текстів
Робототехніка	Використання роботів для фізичної взаємодії з культурними об'єктами	Автоматичне сканування та збереження артефактів
Аналіз великих даних (Big Data Analytics)	Обробка та аналіз великих обсягів даних для виявлення закономірностей	Моніторинг стану культурних об'єктів; прогнозування ризиків
Семантичні мережі	Представлення знань у вигляді графів для полегшення пошуку та аналізу	Організація та пошук інформації про культурні об'єкти
3D-моделювання	Створення тривимірних моделей культурних артефактів	Віртуальна реконструкція архітектурних пам'яток

Спочатку відбувається збір даних та їх поділ на категорії. Далі здійснюється попередня обробка даних, що включає очистку тексту від зайвих символів та знаків, розбиття тексту на окремі слова або фрази (т. зв. токенізація), а також зведення слів до їх базових форм та видалення стоп-слів, які не несуть значущого змісту. Наступний етап – перетворення тексту в числові дані за допомогою таких підходів: модель мішка слів – текст перетворюється в вектор, який представляє частоту кожного слова в документі; частота терміну зворотно-зважена до частоти документів – призначений для оцінки важливості терміну в межах документа та в контексті всього корпусу документів; алгоритми, такі як Word2Vec або GloVe, перетворюють слова в багатовимірні вектори, які зберігають семантичні відношення між словами. Далі відбувається розподіл на навчальний набір і тестовий набір. Перший використовується для навчання моделі. В цьому наборі містяться вхідні дані та відповідні їм цільові значення. Другий призначений для оцінки продуктивності моделі. Для вибору та навчання моделі застосовують один з алгоритмів (як-от Naive Bayes, Logistic Regression, Decision Tree, Random Forest тощо) і далі модель навчається на навчальному наборі, намагаючись мінімізувати помилки та підвищити точність класифікації. Після навчання модель перевіряється на тестовому наборі, тобто відбувається її оцінка за такими метриками як точність, влучність, повнота та F-міра. Після оцінки модель може використовуватися для класифікації нових, невідомих документів. Новий документ проходить через той самий процес попередньої обробки та перетворення тексту в числові дані, а потім модель визначає його категорію.

Щодо вибору алгоритму для навчання моделі, вважаємо за доцільне використовувати алгоритм випадкового лісу — це ансамблевий метод машинного навчання, для задач класифікації та регресії. Він складається з множини дерев рішень, кожне з яких генерується на випадково обраному підмножині даних та ознак.

При побудові моделі випадкового лісу виконуються наступні кроки:

- Вибір випадкової підвибірки: З навчального набору даних вибирається випадкова підвибірка з поверненням. Це означає, що деякі зразки можуть бути вибрані більше одного разу, а деякі ні.

- Вибір випадкових функцій: Також здійснюється випадковий вибір функцій (або ознак) з загального набору ознак. Це допомагає забезпечити різноманітність у деревах рішень, що створюються, і уникнути перенавчання.

- Побудова дерева рішень: Для кожної випадкової підвибірки та випадкового підмножини функцій будується дерево рішень. Дерева рішень побудовані таким чином, щоб максимізувати інформативність при розділенні даних.

- Комбінування результатів: Коли всі дерева рішень побудовані, їх прогнози комбінуються для отримання кінцевого результату. У випадку класифікації це може бути рішення більшості, а для регресії - середнє значення прогнозів.

Ця процедура дозволяє створити модель, яка зазвичай має кращу узагальнюючу здатність порівняно з одним деревом рішень, оскільки вона враховує різноманітність даних та функцій. Крім того, вона відносно стійка до перенавчання і може працювати добре з різними типами даних. Як приклад такого підходу, наведемо досвід авторів [13], які ставили за мету відділити римований текст від неримованого в межах проекту «Smart Document Analysis Using Machine Learning».

Застосування алгоритму випадкового лісу для класифікації документів відбувається в наступній послідовності:

1. Підготовка даних. На цьому етапі збирають текстові документи, які потрібно класифікувати, а далі перетворюють текст документів на числові вектори за допомогою методів векторизації, таких як TF-IDF або Bag of Words.
2. Створення випадкового лісу. Спочатку обирають кількість дерев у лісі і гіперпараметрів моделі (наприклад, глибина дерев, критерій розбиття), потім будують кожне дерево рішень на випадковій підвибірці даних та випадковій підмножині ознак.
3. Навчання моделі: кожне дерево навчається на своїй випадковій підвибірці даних і випадковій підмножині ознак.
4. Прогнозування класу. Коли модель навчена, можна застосувати її для класифікації нових документів. Кожне дерево дає свій власний прогноз для класу кожного документа. Після цього використовується голосування більшості або інші методи комбінування результатів для визначення кінцевого класу кожного документа.
5. Оцінка точності моделі. Проводиться на тестовому наборі даних, які не використовувалися під час навчання., а далі визначають метрики (такі як точність, відгук, специфічність та інші) для оцінки якості моделі.

На підставі описаних вище кроків можна зобразити схему застосування алгоритму випадкового лісу для класифікації документів (рис. 1):

Рис. 1. Алгоритм Random Forest для класифікації документів

Інформаційна система для класифікації документів на основі машинного навчання може складатися з декількох ключових компонентів, які працюють разом для автоматизації процесу класифікації. Опишемо основні компоненти та характеристики такої системи:

1. Збір та управління даними. Цю дію виконує користувач, використовуючи в якості джерел даних електронну пошту, новинні статті, наукові публікації, звіти, соціальні медіа тощо. Зібрані документи, а також їх метадані (наприклад, дату створення, автора, тип документа) зберігаються у базі даних.
2. Попереднє опрацювання даних здійснюється інформаційною системою та передбачає очищення тексту (видаляються зайві символи, пунктуація, HTML-теги, тощо), токенізацію (розбивання тексту на окремі слова або фрази), стемінг та лематизацію (зведення слів до їх базових або кореневих форм), видалення стоп-слів (які не несуть значущого змісту).
3. Перетворення тексту в числові дані здійснюється інформаційною системою за допомогою технологій BoW – Bag of Words (представлення тексту як вектору частот слів), TF-IDF – Term Frequency-Inverse Document Frequency) (вимірювання важливості слів у документі відносно всього корпусу документів), Word Embeddings (використання методів, таких як Word2Vec або GloVe, для перетворення слів у багатовимірні вектори).
4. Розподіл даних на навчальний набір і тестовий набір, які використовуються в процесі машинного навчання для побудови та оцінки моделей.
5. Навчання моделі. Для цього можуть застосовуватися різні алгоритми машинного навчання (наївний Байєсів класифікатор, логістична регресія, метод опорних векторів (SVM), глибокі нейронні мережі тощо). У нашому випадку обрано і вище обґрунтовано алгоритм випадкового лісу. Модель навчається на навчальному наборі, налаштовуючи свої параметри для мінімізації помилок класифікації.

6. Оцінка моделі проводиться на тестовому наборі даних з використанням метрик оцінки точності (accuracy), відповідності (precision), повноти (recall) та F-міри (F1-score).

7. Класифікація нових документів. Нові документи проходять той самий процес попереднього опрацювання та перетворення в числові вектори. Модель визначає категорію для кожного нового документа на основі навченої моделі.

8. Моніторинг та оновлення здійснюється адміністратором і передбачають регулярний контроль точності моделі на нових даних та періодичне перенавчання моделі на нових даних для підтримки актуальності та точності.

9. Інтерфейс користувача. Для завантаження нових документів, перегляду результатів класифікації та керування інформаційною системою використовуватиметься відповідний веб-інтерфейс, а для інтеграції з іншими системами або автоматизації процесу класифікації – інтерфейс програмування (API).

Виходячи з описаного, розроблено діаграму діяльності інформаційної системи класифікації документів на основі машинного навчання, яка відображає послідовність кроків, необхідних для виконання завдання класифікації документів (рис. 2).

Ця діаграма діяльності дає загальне уявлення про кроки, необхідні для створення та використання системи класифікації документів на основі машинного навчання, а сама інформаційна система для класифікації документів на основі машинного навчання є потужним інструментом для автоматизації та оптимізації процесу управління документами. Вона дозволяє ефективно організовувати великі обсяги текстової інформації, забезпечуючи швидкий доступ до потрібних даних та підвищуючи продуктивність роботи.

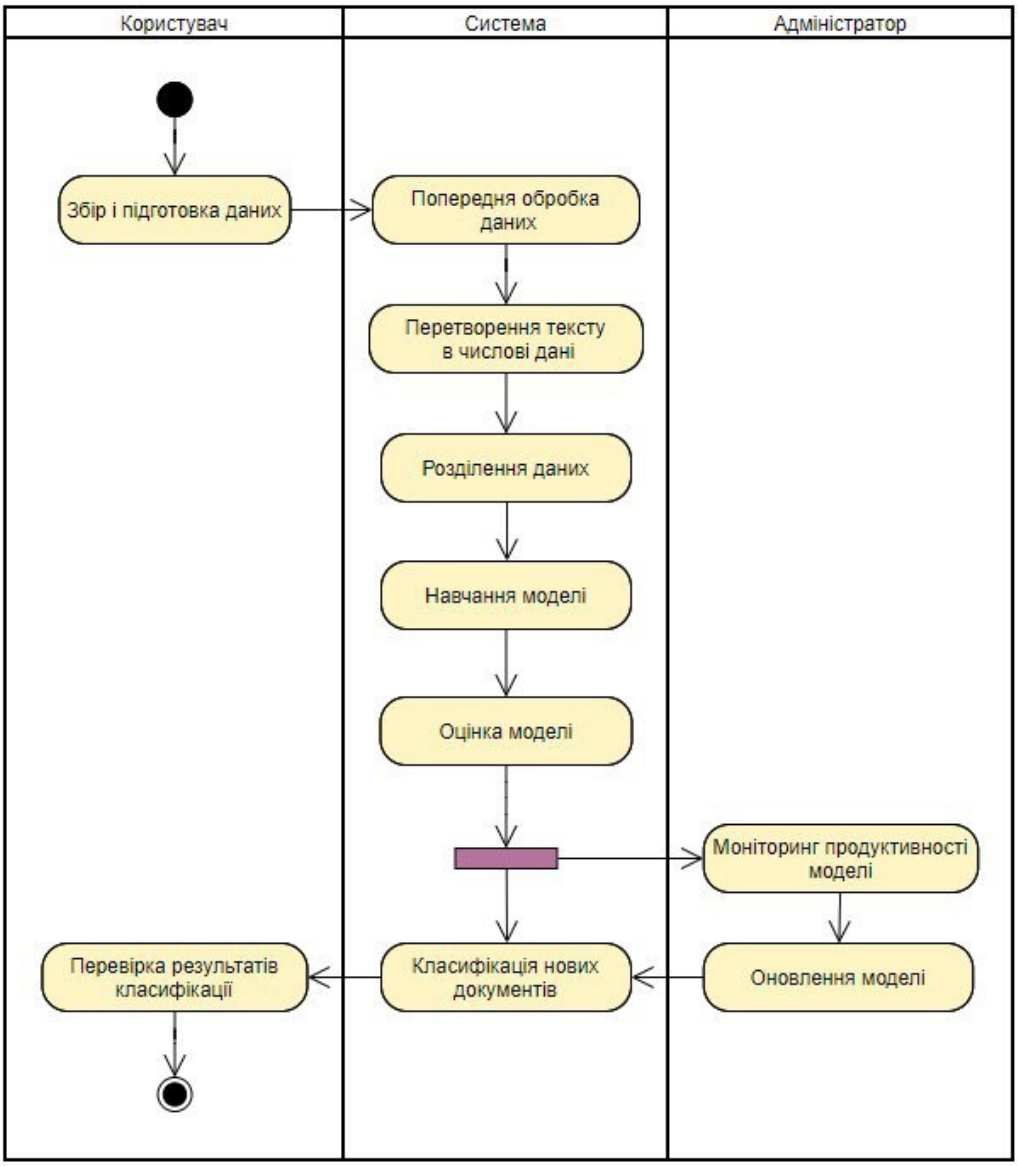

Рис. 2. Діаграма діяльності ІС класифікації документів на основі машинного навчання

Незважаючи на перелічені вище переваги та доцільність, застосування методів машинного навчання для класифікації документів, однак, стикається з рядом проблем. Окреслимо найвагоміші з них:

- Для ефективного навчання моделей машинного навчання необхідні великі обсяги даних, а документи культурної спадщини часто унікальні, і доступні обсяги оцифрованих документів можуть бути обмеженими. Крім того, документи можуть бути пошкоджені або містити помилки, що ускладнює їх обробку.

- Документи культурної спадщини можуть бути представлені в різних форматах, таких як рукописи, друковані тексти, зображення, аудіозаписи та відеоматеріали. Кожен формат вимагає спеціальних підходів до обробки і класифікації, що ускладнює створення універсальних моделей.

- Документи можуть бути написані різними мовами, включаючи стародавні та мертві мови. Багато сучасних моделей машинного навчання добре працюють лише з поширеними мовами, і їх потрібно адаптувати для роботи з менш поширеними або спеціалізованими мовами.

- Документи культурної спадщини часто містять складний контекст і термінологію, що може бути важко зрозуміти і правильно інтерпретувати машиною. Моделі машинного навчання можуть мати труднощі з розумінням історичного, культурного чи соціального контексту документів.

- Використання машинного навчання для обробки документів культурної спадщини може піднімати етичні питання, пов'язані з правами на інтелектуальну власність, приватність та культурну чутливість. Необхідно забезпечити, щоб моделі використовувалися з дотриманням відповідних етичних стандартів і правових норм.

- Сучасні методи машинного навчання, зокрема глибокі нейронні мережі, часто є "чорними скринями" і важко інтерпретувати результати їхньої роботи. Це може бути проблематично в академічних та культурних контекстах, де важливо розуміти, як і чому було прийнято певне рішення.

- Навчання та застосування моделей машинного навчання, особливо великих і складних моделей, вимагає значних обчислювальних ресурсів. Це може бути викликом для організацій з обмеженим бюджетом, включаючи музеї, архіви та бібліотеки.

Висновки з даного дослідження
і перспективи подальших розвідок у даному напрямі

В статті здійснено дослідження сучасних методів машинного навчання, що застосовуються в секторі культурної спадщини для збереження та класифікації різних документів, узагальнено їх характеристики та сфери найефективнішого застосування; розроблено та представлено схему застосування машинного навчання для класифікації документів культурної спадщини та доведено ефективність застосування алгоритму випадкового лісу для навчання моделей розпізнавання та класифікації документів; подано концептуальну модель інформаційної системи класифікації документів на основі машинного навчання. Подальші дослідження будуть спрямовані на розробку веб-інтерфейсу та програмного модуля інформаційної системи, базованої на досягненнях штучного інтелекту, що здійснює відбір, категоризацію та класифікацію історичних документів.

References

1. Karterouli K., Batsaki Y. AI and Cultural Heritage Image Collections: Opportunities and challenges, Proceedings of EVA, 2021, doi: 10.14236/ewic/EVA2021.33.

2. LeCun Y., Bengio Y., Hinton G. Deep learning. Nature, 2015, 521, P. 436–444. https://doi.org/10.1038/nature14539.

3. Castillo Lamas Alberto, Tabi, Siham, Cruz Policarpo, Montes Rosana, Martinez-Sevilla Alvaro, Cruz Sánchez Teresa, Herrera Francisco. MonuMAI: Dataset, deep learning pipeline and citizen science based app for monumental heritage taxonomy and classification. Neurocomputing, 2021, № 420, P. 266-280. https://doi.org/10.1016/j.neucom.2020.09.041.

4. Palma Valerio. Towards deep learning for architecture: a monument recognition mobile app. International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, 2019, Bergamo, Italy, 6-8 February 2019, P. 551-556. https://doi.org/10.5194/isprs-archives-XLII-2-W9-551-2019.

5. Kumar P., Ofli F., Imran M., Castillo C. Detection of Disaster-Affected Cultural Heritage Sites from Social Media Images Using Deep Learning Techniques. Journal on Computing and Cultural Heritage, 2020, № 13, P. 23, 10.1145/3383314.

6. Amato Giuseppe, Falchi Fabrizio, Vadicamo Lucia. Visual Recognition of Ancient Inscriptions Using Convolutional Neural Network and Fisher Vector. Journal on Computing and Cultural Heritage, 2016, № 9, P. 1-24. 10.1145/2964911.

7. Belhi A., Bouras A., Foufou S. Leveraging Known Data for Missing Label Prediction in Cultural Heritage Context. Applied Sciences. 2018; 8(10):1768. https://doi.org/10.3390/app8101768.

8. Elgammal Ahmed, Mazzone Marian, Liu Bingchen, Kim Diana, Elhoseiny Mohamed. The Shape of Art History in the Eyes of the Machine. Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, New Orleans, Louisiana, USA, February 2-7, 2018, New Orleans: AAAI Press, 2018

9. Sandoval Catherine, Pirogova Elena, Lech Margaret. Two-Stage Deep Learning Approach to the Classification of Fine-Art Paintings. IEEE Access, 2019. P. 41770-41781, https://doi.org/7. 10.1109/ACCESS.2019.2907986.

10. Choi J. Exploring Art with Open Access and AI: What's Next? 2019. URL: https://www.metmuseum.org/blogs/now-at-the-met/2019/met-microsoft-mit-exploring-art-open-access-ai-whats-next.

11. Cordell R. C. Machine Learning + Libraries. A Report on the State of the Field. LC Labs, Library of Congress. https://labs.loc.gov/static/labs/work/reports/Cordell-LOC-ML-report.pdf.

12. Lypak T. A. Investigation of prospects and opportunities to use artificial intelligence to preserve cultural heritage: qualification work for the master's degree in specialty "122 - Computer Science". Ternopil: TNTU, 2024. 88 p.

13. Kumar P., Ramakanth, Smart Document Classification Using AI-ML (2019). International Journal of Innovative Research in Computer Science & Technology (IJIRCST), Volume-7, Issue-3, May-2019, SSRN: https://ssrn.com/abstract=3527533.