УДК 004.85 https://doi.org/10.15407/pp2020.04.081

*А.П. Жиркова, О.П. Ігнатенко*

# АНАЛІЗ МЕТОДІВ МАШИННОГО НАВЧАННЯ В ЗАДАЧІ КЛАСИФІКАЦІЇ ДОКУМЕНТІВ

Публікація досліджує методи класифікації документів за наявністю печатки. Для цього, по-перше, проаналізовано вже існуючі методи вирішення поставленої проблеми; по-друге, запропоновано модель згорткової нейронної мережі для класифікації документів; по-третє, відображено залежність коректності роботи нейронної мережі від кількості вхідних даних, на яких навчається модель. В результаті отримано нейронну мережу, що класифікує документи за наявністю печатки з точністю трохи більше ніж 88 %.

Ключові слова: машинне навчання, класифікація, згорткові нейронні мережі, послідовна модель.

## Вступ

Дана робота досліджує актуальну проблему класифікації відсканованих документів за наявністю печатки. Ця проблема виникає у багатьох областях діяльності, пов’язаних з документообігом, оскільки поточна нормативна база України орієнтується, в основному, на паперові документи. Особливо важливим напрямком застосування є класифікація документів у системі публічних закупівель “Прозорро”, в якій у 2019 р. успішно відбулось 1,238 млн. публічних закупівель з очікуваною вартістю 581,3 млрд грн. Кількість активних організаторів закупівель склала 28 850, кількість активних учасників – 159 980 (дані за 2019 рік). З огляду на успішність даного проекту планується подальше розширення сфери застосування системи Прозорро на інші напрямки.

Розпізнавання паттернів на відсканованих документах, зокрема визначення наявності печатки, підпису, реквізитів та інших шаблонізованих частин є критично важливим для автоматичної перевірки коректності завантажених документів. Кількість щоденних тендерів змушує шукати алгоритмічні шляхи розв’язання проблеми, і тут природнім напрямком пошуку є методи машинного навчання. Машинне навчання – вже не новий, але дуже популярний напрям для досліджень та розробки, орієнтований на роботу з різними видами даних, розуміння їх структури та взаємозв’язків.

Обсяг даних, представлений у вигляді зображень, значно виріс з розвитком технологій та популяризацією фотографій як способу поділитися певною інформацією. Присутність камери на всіх мобільних пристроях, що випускаються останнім часом, та поліпшення якості фотографій, є значним рушієм для поширення використання даних такого виду. А отже, з’являються і методи обробки подібної інформації, де центральним поняттям є термін “computer vision”, який на українську перекладається як “комп’ютерний зір” та означає процес обробки графічних даних, який має на меті, наприклад, розпізнавання об’єктів певного класу.

Можливість сканувати або фотографувати документи дозволяє зберігати їх у вигляді зображень, а отже, і застосовувати до них відповідні методи обробки. Наявність або відсутність печаток на зображенні є типовим представником задачі класифікації. При складанні документів виникають помилки різного типу, які важко відстежити та навіть при дуже уважному перегляді можна пропустити. Тоді виникає потреба обробляти документи в автоматизованому режимі, щоб відслідковувати помилки та мінімізувати їхню присутність у документах. Використання машинного навчання для розв'язання задач цього класу є актуальним способом вирішення подібних проблем. В даній роботі

пропонується метод вирішення описаної задачі за допомогою навченої нейронної мережі. Розробка моделі згорткової нейронної мережі для коректної класифікації документів за наявністю або відсутністю печаток на ньому потребує вирішення наступних завдань, а саме: проведення аналізу існуючих методів вирішення поставленої проблеми, проведення збору та обробки даних, побудови нейронної мережі для класифікації документів, дослідження залежності точності моделі від кількості даних для її навчання.

## Огляд літератури

Опишемо існуючі методи, які використовуються для розпізнавання печаток на документах. Двоетапний підхід до вилучення візуальних об'єктів з паперових документів, про який розповідається у [1], серед інших задач вирішує і поставлену у даній роботі. Двоетапний підхід працює наступним чином: спочатку застосовується певний алгоритм для розпізнавання об'єкту на вхідному зображенні, після чого застосовується інший метод, заснований на добуванні з зображення низько-рівневих характеристик (з англ. "features") – таким чином перевіряється правильність роботи попереднього етапу. Отже, перший етап – каскадне навчання та розпізнавання на основі класифікатора AdaBoost. На другому етапі проводиться оцінка низько-рівневих характеристик зображення за допомогою різних алгоритмів машинного навчання. Зображення представляється даними, отриманими з нього за допомогою різних функцій, однією з яких є статистика першого порядку, що означає представлення зображення у вигляді таких характеристик як середня інтенсивність пікселів, дисперсія, асиметрія, центральний момент та ентропія. Наступним методом представлення зображення є його опис за допомогою статистики довжини сірого, дані про яку надають інформацію щодо текстури зображення. Також серед таких методів є гістограма напрямлених градієнтів, яка допомагає розрізняти об'єкти різних типів, та локальні бінарні патерни, які є універсальними дескрипторами текстури. На другому етапі до усіх цих представлень застосовуються алгоритми машинного навчання, такі як метод k-найближчих сусідів (k=1), наївний Байєс, метод опорних векторів, бінарне дерево рішень та інші. Використання двоетапного підходу до вилучення візуальних об'єктів з паперових документів у випадку розпізнавання печаток дало середню точність 53.3 %.

Наступний підхід до виявлення печаток у документах, представлений у [2], використовує поєднання деяких простих характеристик зображення. Алгоритми машинного навчання (такі як метод k-найближчих сусідів, метод опорних векторів, випадкові ліси), що використовуються для виявлення печаток, обробляють інформацію про зображення, в якому закодували початкову модель RGB у модель, що представляє зображення як поєднання Y, Cb, Cr, де кожна з компонент є сумою значень RGB, перемножених на сталі коефіцієнти, після чого зображення бінаризується. В результаті навчання та валідації зображень, виявилося, що середня точність передбачень становить близько 70 %.

Також існує підхід до розпізнавання печаток, який в цілому фокусується на розпізнаванні геометричних форм, притаманних їм. Для цього використовується перетворення Хафа, оскільки його метою є виявлення кругів та квадратів. Також тут застосовується алгоритм згладжування, який прибирає шуми. Після усіх перетворень метод опорних векторів класифікує документи за наявністю/відсутністю печаток. За словами авторів, вони досягли 92 % точності роботи алгоритму [3]. Але тут варто зауважити, що даний підхід фокусується на розпізнаванні печаток, представлених у формі кругу чи квадрату, що є лише підмножиною усіх можливих

форм печаток. А тому результати дослідження є ідеалізованими і не відповідають реальним практичним задачам.

Наступним варіантом вирішення задачі розпізнавання печаток у документах може бути підхід, викладений у роботі [4]. Він полягає у наступному: відскановані зображення розбиваються за кольорами, які в свою чергу розділяються на кандидатів на печатку, після чого з зображення виділяються деякі його характеристики, які допомагають визначитися, чи об'єкт є печаткою, чи ні. Даний метод добре працює, коли документ є кольоровим і колір тексту відрізняється від кольору печатки, в інакшому випадку він не дає задовільних результатів.

Одним з методів, запропонованих останнім часом, є використання FCN (Fully Convolutional Neural Network). Автори представляють підхід, який розпізнає печатки на картинках документів, під назвою D-StaR [5]. Для кращої роботи нейронної мережі вони використовують переднавчену модель VGG-Net, вихідний результат роботи якої є вхідними даними для FCN. Такий підхід використовується, оскільки при його реалізації автори зіштовхнулися з нестачею даних для навчання і валідації (в доступі було 400 картинок), до того ж поділ на навчальну та тренувальну вибірки зроблений як 90 % і 10 % від усього обсягу даних. Тобто, валідація роботи методу проходила на зовсім малій вибірці, що варте зауваження. В результаті роботи D-StaR точність становить 87 %, але при розпізнаванні печаток, які накладаються на текст, вона падає до 74 %. Також можна відзначити роботу [6], де методи навчання застосовуються до розпізнавання традиційних монгольських печаток. Ключова ідея полягає у поєднанні аналізу головних компонент (PCA) та рекурентних нейронних мереж. Автори декларують високу точність, хоча можливо це пояснюється невеликими розмірами датасету.

Цікавим напрямком розвитку є логічне продовження ідеї перевірки документів, яке полягає у розпізнаванні підпису. Цій нетривіальній задачі присвячена робота [7], де пропонується спочатку "очищувати" сліди печатки (як правило підпис і печатка мають перекриватись, забезпечуючи додатковий захист документу) за допомогою генеративних нейромереж (GAN), а потім розпізнавання підпису виконується з використанням тих же згорткових мереж.

## Постановка задачі

При роботі з документами деякі компанії стикаються з проблемами, які необхідно автоматизувати, оскільки їх вирішення співробітниками займає багато часу та не звільняє від помилок, які важко контролювати. Прикладами таких задач є розпізнавання печаток (чи їхню відсутність) у документах, розпізнавання підписів, виявлення слів, написаних однією мовою, у документі, написаному іншою.

Всі ці задачі можливо вирішити засобами машинного навчання.

Наразі проведено роботу з розпізнавання печаток у документах. Для цього оброблено документи, вивантаженні з сайту Prozorro [8], які є дозволеними для їх подальшого використання та розповсюдження. Тут варто враховувати, що документи представлені не тільки в форматі зображень (з розширенням ".png", ".jpg" або ".jpeg"), а також у вигляді документів PDF. Тому при формуванні навчальної вибірки (а також при подальшій класифікації документів, що мають печатку, або не мають її взагалі) слід перетворити усі документи, представленні у форматі PDF, на зображення.

Для навчання обраної моделі машинного навчання, на вхід подаються зображення документів, які вона класифікує як 0, якщо на зображенні немає печатки, та як 1, якщо має хоча б одну (рис. 1).

12.2. ПІДРЯДНИК несе персональну відповідальність за наявність ліцензій та інших дозвільних документів, необхідних для виконання робіт, визначених нормами чинного законодавства.
12.3. ПІДРЯДНИК є платником податку на загальних підставах.

13. ЮРИДИЧНІ АДРЕСИ СТОРІН.

**ЗАМОВНИК:**
**Маріупольська загальноосвітня санаторна школа-інтернат І-ІІ ступенів №11 Донецької обласної ради**
Юридична адреса: 87518, Донецька область, м. Маріуполь, вул. Бахчиванджи, 29
р/р
Держказначейська служба України м.Київ
ЄДРПОУ

Директор

М.П.

**ПІДРЯДНИК:**
**ТОВ «АУДИТ-ПРОМ»**
87400, Донецька область, Мангушський район смт. Мангуш, вул. Котанова, б.100
р/р
в МФ ПАТ «Приватбанк»
ЄДРПОУ
Витяг з реєстра платників ПДВ
№ від
Тел. +380.

Директор ТОВ «АУДИТ - ПРОМ»

М.П.

**4. Доплати і нарахування**

1. Компенсація витрат на відрядження експертів на засідання ДКЗ (без ПДВ):
2. Разом витрат:
3. ПДВ (20%):
4. ВСЬОГО:
5. Надбавка за виконання робіт, не передбачених планом (%):
6. Всього по договору:

грн. кон. (. грн. коп.), у т.ч. ПДВ – грн. коп.

Начальник управління
Начальник відділу
Головний економіст

*Примітка:** Коефіцієнт повних затрат враховує витрати, пов'язані з організаційно-господарською діяльністю ДКЗ. Величина коефіцієнту повних витрат визначається як відношення загальних витрат ДКЗ до вартості експертизи. (Порядок визначення вартості робіт по проведенню державної експертизи та оцінки запасів корисних копалин, затверджений наказом ДКЗ України від 02.08.1996р. №29 та зареєстрований в Міністерстві юстиції України 05.11.1996 за № 652/1677).
Загальні витрати складаються з фонду оплати праці працівників ДКЗ, податків і обов'язкових платежів, витрат на комунальні послуги, зв'язок та інші послуги, використання та обслуговування техніки, організаційне та інформаційне забезпечення експертних робіт і визначаються як правило один раз на рік за фактичними економічними даними минулого року і тими, що плануються у поточному році. Згідно з виконаними підрахунками коефіцієнт повних витрат на 2019 р. дорівнює 7,7 (додаток до наказу ДКЗ України від 28.02.2019 р. № 108).

Рис. 1. Приклад документів з печатками (зліва) та без печаток (справа)

## Опис методу вирішення задачі

Розв'язувати задачу розпізнавання печаток у документах було вирішено за допомогою нейронних мереж. В ході виконання роботи було застосовано декілька різних моделей, в результаті чого підтвердилося припущення, що згорткові нейронні мережі є найбільш ефективним методом роботи з зображеннями (звичайні послідовні нейронні мережі дали 59.9 % точності, а рекурентні – 59.8 %, водночас як згорткові нейронні мережі на тих самих даних мають 88.03 % точності). Тому далі більш детально розглянуто саме згорткові нейронні мережі.

Перед етапом навчання моделі необхідно сформувати тренувальну та тестову вибірки, для чого: усі документи було перетворено у формат зображень, стандартизовано їх розмір та розбито на тренувальну та тестову вибірки випадковим чином так, щоб перша містила 70 % усіх зображень, а друга – відповідно, 30 %.

Для класифікації документів використовується послідовна модель, яка у своїй структурі має три згорткових шари для обробки зображень, та використовує ще три для їхньої класифікації.

Числові значення, показані на рис. 2, є розмірністю вихідного простору відповідного шару. До усіх шарів, окрім вихідного, застосовується активаційна функція "relu", останній же використовує "sigmoid". При заміні активаційної функції у вихідному шарі (зокрема, на активаційну функцію "softmax") точність роботи моделі зменшується приблизно у два рази.

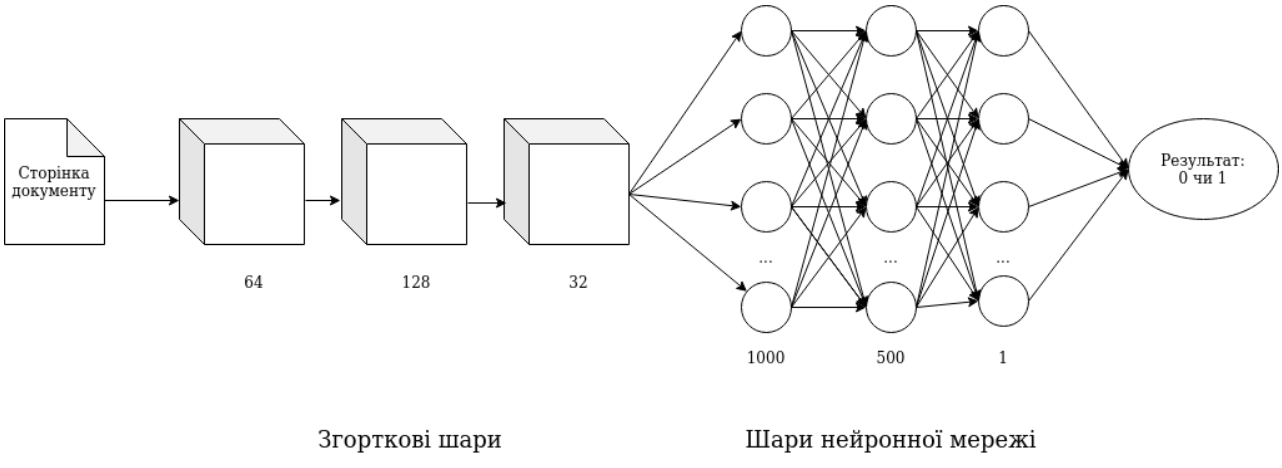


Рис. 2. Реалізація нейронної мережі для класифікації документів за наявністю або відсутністю печаток

## Результати експериментів

Для порівняння правильності роботи моделі в залежності від кількості тренувальних даних, тестування проводилося на одному й тому самому наборі зображень. Це дає можливість побачити вплив більшої кількості даних на більш правильне налаштування моделі під час навчання.

Після отримання даних для навчання, набір поділяється на тренувальну та валідаційну вибірки, які використовуються для навчання моделі. Набір даних для валідації роботи моделі складає близько 25 % від загального обсягу тренувальних даних (рис. 3, 4).

Під час тестування роботи моделі, яку було навчено на 435 зображеннях, виявилося, що правильно класифіковано приблизно 81 % тестових даних. А після навчання нейронної мережі на 3216 об'єктах, доля правильних відповідей становила трохи більше 88 % (табл. 1).

Також варто зауважити, що на другій і третій ітерації точність роботи моделі на тестових даних трохи погіршилась, після чого досить різко виросла на двох останніх. Такий ефект можна приписати якості самих даних та їхньої попередньої класифікації, оскільки вона проводилася вручну, а при цьому можлива невелика похибка. Щодо самих даних, то при класифікації було виявлено велику кількість погано відсканованих документів, на яких відображалися печатки з наступних сторінок. В результаті було вирішено класифікувати такі зображення як ті, що не містять печаток, хоча їх на документі гарно видно.

При збільшенні розміру тренувальної вибірки час на навчання моделі зростає, а час роботи на тестових даних починає істотно збільшуватися тільки на останніх ітераціях, що зображено у табл. 2.

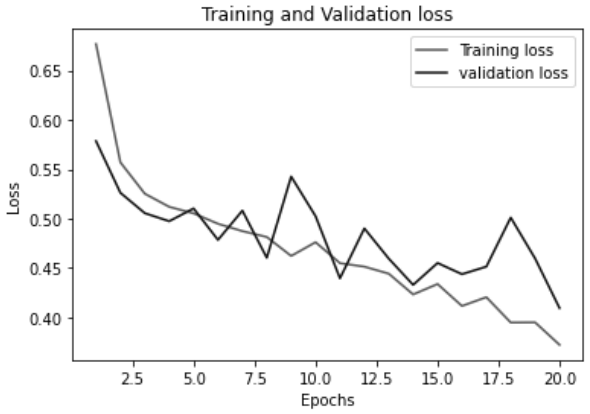


Рис. 3. Втрати в процесі тренування та валідації моделі

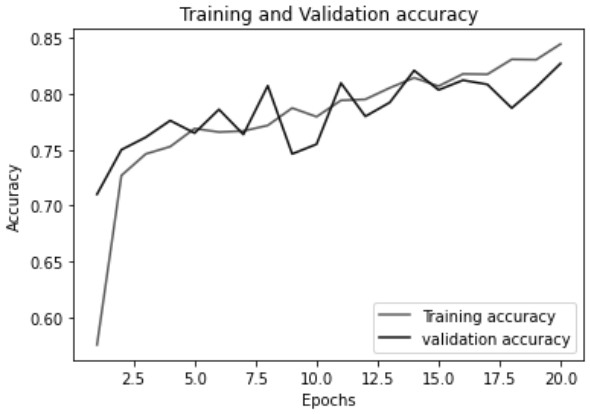


Рис. 4. Точність в процесі тренування та валідації моделі

Таблиця 1. Вплив розміру тренувального набору даних на точність роботи нейронної мережі

| Розмір тренувальної вибірки | Розмір валідаційної вибірки | Розмір тестової вибірки | Точність роботи моделі |
|---|---|---|---|
| 435 | 109 | 234 | 81.19 % |
| 1164 | 291 | | 80.76 % |
| 1896 | 474 | | 80.76 % |
| 2596 | 650 | | 86.32 % |
| 3216 | 804 | | 88.03 % |

Таблиця 2. Вплив розміру тренувального набору даних на час роботи нейронної мережі

| **Розмір тренувальної вибірки** | **Розмір валідаційної вибірки** | **Розмір тестової вибірки** | **Час, необхідний на навчання моделі (у хв)** | **Час роботи мережі на тестових даних (у сек)** |
|---|---|---|---|---|
| **435** | 109 | 234 | 0,66 | 0,23 |
| **1164** | 291 | | 1,78 | 0,22 |
| **1896** | 474 | | 2,86 | 0,21 |
| **2596** | 650 | | 3,86 | 0,27 |
| **3216** | 804 | | 4,81 | 0,37 |

## Висновки

Машинне навчання широко використовується для роботи з зображеннями, особливо ефективним підходом до їхньої обробки є використання згорткових нейронних мереж. Через структуру архітектури мереж даного типу вони гарно працюють з зображеннями, тому широко використовуються для роботи з цим типом даних.

Автоматизоване розпізнавання різних типів помилок у документах є актуальною темою, що потребує вирішення.

Прикладом таких помилок є відсутність печаток на документах, для розпізнавання чого в рамках даної курсової роботи було побудовано нейронну мережу, здатну класифікувати дані, що подаються на вхід у вигляді зображень, за наявністю або відсутністю печаток.

Початкова точність роботи даної моделі складала 81.19 %, при цьому для її навчання використовувалося 435 зображень (для тренування мережі) і 109 (для валідації її роботи). Наступні дві спроби навчити модель на більшій кількості даних, ніж при попередній спробі, дали трохи гірші результати, а саме точність роботи моделі на тестовому наборі даних складала 80.76 % в обох випадках. Спочатку для навчання мережі використовувалося 1164 зображення (і 291 для валідації), після чого на вхід моделі було подано 1896 зображень для навчання (і 474 для валідації). Останні два процеси навчання на більшій кількості даних (2596 для навчання і 650 для валідації в першому випадку, 3216 для навчання і 804 для валідації – в другому), показали покращення при класифікації тестового набору, що дає 86.32 % і 88.03 % правильних відповідей.

## Література


1. Forczmanski P., Markiewicz A. Two-stage approach to extracting visual objects from paperdocuments. *Machine Vision and Applications*. 2016. N 27. P. 1243–1257.
2. Forczmanski P., Markiewicz A. Stamps Detection and Classification Using Simple Features Ensemble. *Mathematical Problems in Engineering*. 2015.
3. Roy P., Pal U., Lladós J. Seal Detection and Recognition: An Approach for Document Indexing [Електронний ресурс]. 10th International Conference on Document Analysis and Recognition. 2009. Режим доступу до ресурсу: https://www.researchgate.-net/publication/220861099_Seal_Detection_and_Recognition_An_Approach_for_Document_Indexing.
4. Micenkova B., van Beusekom J., Shafait F. Stamp Verification for Automated Document Authentication [Електронний ресурс]. Режим доступу до ресурсу: http://pure.au.dk/-portal/files/51730044/Barbora_Stamp_Verification_IWCF12.pdf.
5. D-StaR: A, Younas M., Afzal M., Malik та ін. Generic Method for Stamp Segmentation from Document Images [Електронний ресурс]. 2017. Режим доступу до ресурсу: https://tukl.seecs.nust.edu.pk/members/project

s/conference/D-StaR-A-Generic-Method-for-Stamp-Segmentation-from-Document-Images.pdf.
6. Gantuya P., Mungunshagai B., Suvdaa B. "Mongolian Traditional Stamp Recognition using Scalable kNN." International journal of advanced smart convergence 4.2 (2015): 170–176.
7. Engin Deniz, et al. "Offline Signature Verification on Real-World Documents." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. 2020.
8. Офіційний портал оприлюднення інформації про публічні закупівлі України [Електронний ресурс]. Режим доступу до ресурсу: https://prozorro.gov.ua.

## References

1. Forczmanski P., Markiewicz A. Two-stage approach to extracting visual objects from paperdocuments. *Machine Vision and Applications*. 2016. N 27. P. 1243–1257.
2. Forczmanski P., Markiewicz A. Stamps Detection and Classification Using Simple Features Ensemble. *Mathematical Problems in Engineering*. 2015.
3. Roy P., Pal U., Lladós J. Seal Detection and Recognition: An Approach for Document Indexing [Електронний ресурс]. 10th International Conference on Document Analysis and Recognition. 2009. Режим доступу до ресурсу: https://www.researchgate.-net/publication/220861099_Seal_Detection_and_Recognition_An_Approach_for_Document_Indexing.
4. Micenkova B., van Beusekom J., Shafait F. Stamp Verification for Automated Document Authentication [Електронний ресурс]. Режим доступу до ресурсу: http://pure.au.dk/-portal/files/51730044/Barbora_Stamp_Verification_IWCF12.pdf.
5. D-StaR: A, Younas M., Afzal M., Malik та ін. Generic Method for Stamp Segmentation from Document Images [Електронний ресурс]. 2017. Режим доступу до ресурсу: https://tukl.seecs.nust.edu.pk/members/projects/conference/D-StaR-A-Generic-Method-for-Stamp-Segmentation-from-Document-Images.pdf.
6. Gantuya P., Mungunshagai B., Suvdaa B. "Mongolian Traditional Stamp Recognition using Scalable kNN." International journal of advanced smart convergence 4.2 (2015): 170–176.
7. Engin Deniz, et al. "Offline Signature Verification on Real-World Documents." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. 2020.
8. Official portal for publishing information on public procurement in Ukraine [Electronic resource]. Access to the resource: https://prozorro.gov.ua.

Одержано 04.11.2020

***Про авторів:***

*Жиркова Анастасія Павлівна*,
студентка Національного університету "Києво-Могилянська Академія".
Кількість наукових публікацій в українських виданнях – 1.
https://orcid.org/0000-0002-4604-1137,

*Ігнатенко Олексій Петрович*,
доктор фізико-математичних наук,
провідний науковий співробітник.
Кількість наукових публікацій в українських виданнях – 27.
Кількість наукових публікацій в зарубіжних виданнях – 7.
http://orcid.org/0000-0001-8692-2062.

***Місце роботи авторів:***

Національний університет
"Києво-Могилянська Академія",
вулиця Григорія Сковороди, 2, Київ.

Інститут програмних систем
НАН України, 03187, Київ-187,
проспект Академіка Глушкова, 40.

E-mail: nastia.nastia.zh@gmail.com,
o.ignatenko@gmail.com