Державна служба України з надзвичайних ситуацій
Львівський державний університет безпеки життєдіяльності
Навчально-науковий інститут цивільного захисту
Кафедра інформаційних технологій та систем електронних комунікацій

«Допущено до захисту»
Начальник кафедри ІТтаСЕК
підполковник служби цивільного захисту
_____ Олександр ПРИДАТКО
“____” _____2024 року

КВАЛІФІКАЦІЙНА РОБОТА МАГІСТРА

на тему «**Прогнозування, виявлення та обробка аномалій даних журналу серверів методами машинного навчання**»

Виконав:
здобувач VI курсу, групи КН-61мз
спеціальності (освітньої програми)
122 «Комп’ютерні науки» (Комп’ютерні науки)
(шифр і назва спеціальності (освітньої програми)
Андрій ПЕТРИКОВСЬКИЙ
(ім’я та прізвище)
Керівник Олександр ХЛЕВНОЙ
(ім’я та прізвище)
Рецензент _____
(ім’я та прізвище)

Львів – 2024 року

Державна служба України з надзвичайних ситуацій
Львівський державний університет безпеки життєдіяльності
Навчально-науковий інститут цивільного захисту

Кафедра інформаційних технологій та систем електронних комунікацій
Освітній ступінь магістр
Спеціальність 122 “Комп’ютерні науки”
Освітня програма Комп’ютерні науки

ЗАТВЕРДЖУЮ
Начальник кафедри ІТтаСЕК
підполковник служби цивільного захисту
_____ Олександр ПРИДАТКО
“_____”_____2024 року

ЗАВДАННЯ

на кваліфікаційну роботу

Здобувач Андрій ПЕТРИКОВСЬКИЙ
(ім’я, прізвище)

1.Тема Прогнозування, виявлення та обробка аномалій даних журналу серверів методами машинного навчання
керівник роботи Олександр Хлевной, к.т.н.
(ім’я, прізвище, науковий ступінь, вчене звання)

затверджені наказом ЛДУ БЖД від “____”_____20___року №_____

2. Термін подання здобувачем роботи _____

3. Початкові дані до роботи:

1. Відомості про архітектуру корпоративної мережі ДСНС України, лог-файли роботи серверного обладнання
2. M. Mohri, A. Rostamizadeh, A. Talwalkar. (2018). “Foundations of Machine Learning, second edition”, The MIT Press, pp. 1-5.
3. I. Goodfellow, Y. Bengio, A. Courville. (2016). “Deep Learning”, The MIT Press, pp. 152-160, 2016.
4. Методичні вказівки до виконання кваліфікаційної роботи магістра для здобувачів другого рівня вищої освіти спеціальності 122 «Комп’ютерні науки». Укл. Ольга Смотр, Олександр Придатко, Назарій Бурак. Львів: Вид-во ЛДУ БЖД, 2023. – 36 с.

4. Зміст кваліфікаційної роботи/проекту (перелік питань, які потрібно розробити)

Вступ

Розділ 1. Виявлення аномалій методами машинного навчання

Розділ 2. Моделі машинного навчання

Розділ 3. Прогнозування та аналіз виявлених аномалій з використанням моделей машинного навчання

Висновки

Список використаних джерел

Додатки

5. Консультанти розділів роботи

Розділ	Ім'я, Прізвище та посада консультанта	Підпис, дата	
		завдання видав	завдання прийняв

7. Дата видачі завдання _____

КАЛЕНДАРНИЙ ПЛАН

№ з/п	Назва етапів виконання кваліфікаційної роботи	Термін виконання етапів роботи	Примітка
1	Огляд літературних джерел машинного навчання.	02.11.2023 - 15.11.2023	
2	Огляд та класифікація завдань машинного навчання	15.11.2023 - 25.11.2023	
3	Збір необхідного набору даних	25.11.2023 - 3.12.2023	
4	Аналіз та обробка даних	3.12.2023 - 15.01.2024	
5	Виявлення аномальних записів у журналу роботи серверу	15.01.2024 - 24.01.2024	
6	Застосування виявлених аномалій методами кібербезпеки	24.01.2024 - 31.01.2024	

Здобувач _____ Андрій ПЕТРИКОВСЬКИЙ
(підпис) (ім'я та прізвище)

Керівник роботи _____ Олександр ХЛЕВНОЙ
(підпис) (ім'я та прізвище)

АНОТАЦІЯ

Андрій Петриковський «Прогнозування, виявлення та обробка аномалій даних журналу серверів методами машинного навчання». Кваліфікаційна робота за спеціальністю 122 “Комп’ютерні науки” складається з текстової частини, що містить 3 розділи, 67 с., 17 рис., 2 табл., 16 джерел.

Об’єкт дослідження – процес аналізу та виявлення аномалій в лог-файлах інформаційних систем.

Предмет дослідження – методи, моделі та інформаційні технології, що використовуються для аналізу аномалій в лог-файлах з використанням машинного навчання.

Мета роботи – аналіз і порівняння ефективності різних моделей машинного навчання для виявлення аномалій в лог-файлах, що дозволить підвищити рівень кібербезпеки інформаційних систем.

Магістерська кваліфікаційна робота спрямована на дослідження та вибір оптимальних моделей машинного навчання для аналізу аномалій в лог-файлах, оцінку їх ефективності та можливостей інтеграції у системи кібербезпеки. Проведено детальний огляд існуючих моделей машинного навчання, їх переваг та недоліків у контексті аналізу лог-файлів. Визначено критерії ефективності та впроваджено експериментальне дослідження для оцінки роботи обраних моделей. Розроблено комплексний підхід до виявлення та аналізу аномалій у лог-файлах інформаційних систем, що включає попередню обробку даних, вибір моделі, тренування, тестування та оцінку ефективності. Запропоновано рекомендації щодо впровадження виявлених оптимальних моделей машинного навчання в практику роботи спеціалістів з кібербезпеки, що дозволить підвищити захист інформаційних систем від потенційних загроз.

МЕТОДИ МАШИННОГО НАВЧАННЯ, АНАЛІЗ АНОМАЛІЙ, ЛОГ-ФАЙЛИ, КІБЕРБЕЗПЕКА, ІНФОРМАЦІЙНІ СИСТЕМИ.

ABSTRACT

Andriy Petrykovskyi "Forecasting, Detection, and Processing of Anomalies in Server Log Data Using Machine Learning Methods". The qualification work for the specialty 122 "Computer Science" consists of a textual part, which contains 3 sections, 67 pages, 17 figures, 2 tables, 16 sources.

The object of research is the process of analysing and detecting anomalies in log files of information systems. The subject of the study is methods, models, and information technologies used for analysing anomalies in log files using machine learning.

The purpose of the work is to analyse and compare the effectiveness of different machine learning models for detecting anomalies in log files, which will increase the level of cybersecurity of information systems.

The master's qualification work is aimed at researching and selecting optimal machine learning models for analysing anomalies in log files, evaluating their effectiveness and integration possibilities into cybersecurity systems. A detailed review of existing machine learning models, their advantages, and disadvantages in the context of log file analysis has been conducted. Criteria for effectiveness have been defined, and experimental research has been introduced to evaluate the performance of selected models. A comprehensive approach to detecting and analysing anomalies in log files of information systems has been developed, which includes data preprocessing, model selection, training, testing, and effectiveness evaluation. Recommendations have been proposed for implementing the identified optimal machine learning models into the practice of cybersecurity specialists, which will enhance the protection of information systems against potential threats.

MACHINE LEARNING METHODS, ANOMALY ANALYSIS, LOG FILES, CYBERSECURITY, INFORMATION SYSTEMS.

ЗМІСТ

ПЕРЕЛІК УМОВНИХ СКОРОЧЕНЬ

ML – Machine Learning (машинне навчання)

DL – Deep Learning (глибинне навчання)

NN – Neural Networks (нейронні мережі)

LSTM – Long short-term memory (модель довгої короткочасної пам’яті)

CNN – Convolutional neural network (згорткова нейронна мережа)

RNN – Recurrent neural networks (рекурентні нейронні мережі)

ВСТУП

Актуальність теми. Збір та обробка даних запуску та роботи різноманітного програмного забезпечення, сервісів та комплексних серверних рішень завжди була основним та надійним ресурсом інформації про роботу та цілісність цих систем. Даний підхід завжди забезпечував інженерів інформацією про різноманітні події та загальну роботу інформаційних систем. Водночас, зі збільшенням та розростаннями комплексних програмних та інших рішень, об’єм інформації, що надходить від них для аналізу інженеру, збільшується в досить великих масштабах. Дане явище свідчить, що традиційні підходи до виявлення аномалій у такому об'ємі даних стають непрактичними та часто займають багато часу для аналізу класичними статистичними методами, що може призвести до серйозних кіберінцидентів різноманітного характеру, чи то апаратний збій, чи то дії зловмисників.

Застосування методів машинного навчання, у тому числі глибинного навчання (Deep Learning), у роботі, може допомогти з аналізом, передбаченням та запобіганням критичним подіям інформаційних систем. Брак застосування даних підходів у сьогоднішніх рішеннях комплексних інформаційних систем свідчить про актуальність теми дослідження та розвиває тему практичного застосування машинного навчання у практиці.

Метою дослідження є подолання розриву між передовими дослідженнями машинного навчання у виявленні аномалій на основі журналів роботи інформаційних систем та їх фактичним впровадженням у індустрії.

Для досягнення поставленої мети були визначені наступні завдання:

- підбір даних та вибір програмного рішення для їх обробки;
- класифікація методів машинного навчання;
- порівняння класичних методів машинного навчання і методів та моделей глибинного навчання;
- комплексний аналіз та оцінка якості обраних моделей;

Об'єктом дослідження є методи машинного навчання, де на основі отриманих даних, модель аналізує та виявляє аномалії.

Предметом дослідження є аномалії серед даних журналу роботи інформаційних систем.

Методи дослідження. Для досягнення поставлених завдань та вирішення проблем в рамках дослідження використовувалися різноманітні наукові методи, спрямовані на аналіз та виявлення аномалій в роботі інформаційних систем, зокрема комплексних серверних рішень та програмного забезпечення. Зазначені методи були вибрані з метою забезпечення достовірності та ефективності отриманих результатів, а також з урахуванням великого обсягу інформації, що підлягала аналізу.

1. Аналіз та обробка даних:
   - Для збору та первинної обробки даних були використані методи аналізу журналів роботи інформаційних систем.
   - Методи обробки даних включали в себе стандартні процедури фільтрації, сортування та агрегації, спрямовані на виявлення ключових параметрів та взаємозв'язків.
2. Класифікація методів машинного навчання:
   - Здійснено аналіз та класифікацію різних методів машинного навчання, враховуючи їх ефективність та відповідність до задач виявлення аномалій в журналах роботи інформаційних систем.
3. Порівняння класичних методів та методів глибинного навчання:
   - Проведено порівняльний аналіз класичних методів машинного навчання і методів глибинного навчання з метою визначення їхньої ефективності в контексті виявлення аномалій.
4. Комплексний аналіз та оцінка якості моделей:
   - Застосовано комплексний підхід до аналізу та оцінки якості моделей машинного навчання, враховуючи показники точності, стійкості та ефективності в контексті виявлення аномалій в інформаційних системах.

Наукова новизна одержаних результатів полягає в виявленні та прогнозуванні аномалій в роботі інформаційних систем, в тому числі комплексних серверних рішень з об'ємними автоматизованими рішеннями, взаємозалежних між собою. Підвищено стійкість та надійність роботи, передбачення та запобігання критичних збоїв в роботі методами машинного навчання.

Апробація результатів роботи. Основні теоретичні положення та результати магістерського дослідження були презентовані у доповіді на CXXXVII Міжнародній науково-практичній конференції «Розвиток науки і технологій: Виклики 2024 року», м. Чернівці, 19 січня 2024 р., та на Всеукраїнській науково-практичній конференції «Вітчизняна наука на зламі епох: проблеми та перспективи розвитку» у збірнику матеріалів конференції (№ 94, 2024 р.), м. Переяслав, 23 січня 2024 р.

Особистий внесок здобувача. Кваліфікаційна робота є узагальненням результатів теоретичних і експериментальних досліджень, проведених автором самостійно. У роботі виконавцю належать налаштування, збір та формування датасету із журналу роботи різних інформаційних систем, методами формування потоку, виокремлення та парсингу даних. Порівняння класичних моделей машинного навчання з моделями глибинного навчання у контексті точності результатів, швидкодії та стійкості. Створення підґрунтя для майбутніх досліджень та імплементації в робочих рішення комплексних інформаційних систем.

ВИСНОВКИ

Магістерська робота присвячена аналізу ефективності моделей машинного навчання, зокрема моделі Трансформера, у виявленні аномалій в даних, що має важливе значення для забезпечення кібербезпеки в сучасному цифровому світі. Дослідження охоплює широкий спектр аспектів, від теоретичних основ машинного навчання до практичного застосування отриманих моделей для аналізу реальних даних.

Основні результати дослідження підкреслюють, що використання моделі Трансформера дозволяє значно підвищити точність ідентифікації аномалій в даних завдяки його здатності аналізувати великі обсяги інформації та враховувати контекстуальні залежності між різними частинами датасету. Це відкриває нові можливості для вдосконалення систем кібербезпеки, забезпечуючи виявлення та нейтралізацію потенційних загроз набагато ефективніше, ніж це було можливо раніше.

Дослідження вносить важливий вклад у розвиток теоретичних засад машинного навчання та їх застосування в області кібербезпеки. Практичне значення роботи полягає у розробці рекомендацій для створення та впровадження ефективних систем виявлення аномалій, заснованих на моделях машинного навчання, що можуть бути використані у різних сферах діяльності для забезпечення захисту інформаційних ресурсів.

Рекомендації для подальших досліджень включають необхідність розширення експериментальної бази даних для тестування моделей, що дозволить оцінити їх ефективність у ширшому спектрі умов. Важливим аспектом є також розвиток нових методів оптимізації моделей машинного навчання, що може підвищити їх точність і швидкодію. Окрім того, рекомендується розглянути можливість інтеграції моделей машинного навчання з іншими технологіями, для розробки комплексних рішень у сфері кібербезпеки.

Загальний висновок полягає в тому, що використання моделей машинного навчання, особливо моделі Трансформера, відіграє ключову роль у сучасних системах кібербезпеки, забезпечуючи високу ефективність виявлення та аналізу аномалій. Результати дослідження демонструють великий потенціал цих технологій для покращення захисту інформаційних систем і можуть слугувати основою для подальших наукових розробок у цій галузі.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. M. Mohri, A. Rostamizadeh, A. Talwalkar. (2018). “Foundations of Machine Learning, second edition”, The MIT Press, pp. 1-5.
2. I. Goodfellow, Y. Bengio, A. Courville. (2016). “Deep Learning”, The MIT Press, pp. 152-160, 2016.
3. OpenVPN: Business VPN For Secure Networking. https://openvpn.net/ [Електронний ресурс]
4. Elasticsearch: The Official Distributed Search & Analytics Engine. https://www.elastic.co/ [Електронний ресурс]
5. Kibana is a source-available data visualization dashboard software for Elasticsearch. https://www.elastic.co/kibana [Електронний ресурс]
6. H. Mi, H. Wang, Y. Zhou, R. Lyu, and H. Cai. (2013). “Toward fine-grained, unsupervised, scalable performance diagnosis for production cloud computing systems.” IEEE Transactions on Parallel and Distributed Systems, 24:1245–1255, 2013.
7. What is unsupervised learning? https://cloud.google.com/discover/what-is-unsupervised-learning [Електронний ресурс]
8. X. Fu, R. Ren, J. Zhan, W. Zhou, Z. Jia, and G. Lu. (2012). “Logmaster: mining
   event correlations in logs of large-scale cluster systems.” In SRDS’12:
   Proc. of the 31st IEEE Symposium on Reliable Distributed Systems,
   pages 71–80. IEEE.
9. M. Chen, A. X. Zheng, J. Lloyd, M. I. Jordan, and E. Brewer. (2004). “Failure diagnosis using decision trees.” In ICAC’04: Proc. of the 1st International Conference on Autonomic Computing, pages 36–43. IEEE.
10. Q. Lin, H. Zhang, J.G. Lou, Y. Zhang, and X. Chen. (2016). “Log clustering based problem identification for online service systems.” In ICSE’16:
    Proc. of the 38th International Conference on Software Engineering.

11. Min Du, Feifei Li, Guineng Zheng, and Vivek Srikumar. (2017). “Deeplog: Anomaly detection and diagnosis from system logs through deep learning.” In Proc. of CCS’17. pp. 1285–1298.
12. Shilin He, Qingwei Lin, Jian-Guang Lou, Hongyu Zhang, Michael R Lyu, and Dongmei Zhang. (2018). “Identifying impactful service system problems via log analysis”. In Proc. of ESEC/FSE’18, pp. 60–70
13. Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. (2017). “Attention is All
you Need.” In Proc. of NIPS’17.
14. Sasho Nedelkoski, Jasmin Bogatinovski, Alexander Acker, Jorge Cardoso, and Odej Kao. (2020). Self-attentive classification-based anomaly detection in unstructured logs. arXiv.
15. Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. (1998). Gradient-based learning applied to document recognition. Proc. of the IEEE 86, 11, pp.2278–2324.
16. Min Du, Feifei Li, Guineng Zheng, and Vivek Srikumar. (2017). “Deeplog: Anomaly detection and diagnosis from system logs through deep learning.” In Proc. of CCS’17, pp. 1285–1298.