https://doi.org/10.31891/2219-9365-2024-80-51
УДК 621.391 160164

ПРОДЕУС Максим
Хмельницький національний університет
https://orcid.org/0009-0002-2968-4648
e-mail: mprodeus99@ukr.net
НІЧЕПОРУК Андрій
Хмельницький національний університет
https://orcid.org/0000-0002-7230-9475
e-mail: andrey.nicheporuk@gmail.com
ІВАНЧЕНКО Олег
Національний технічний університет «Дніпровська політехніка»
https://orcid.org/0000-0002-5921-5757
e-mail: vmsu12@gmail.com

ВПЛИВ ПОПЕРЕДНЬОЇ ОБРОБКИ ДАНИХ НА ПРОДУКТИВНІСТЬ МОДЕЛІ RANDOM FOREST У ВИЯВЛЕННІ МЕРЕЖЕВИХ АТАК

В роботі представлено систему для виявлення мережевих атак із використанням моделі машинного навчання Random Forest. Головною особливістю цієї системи є попередня обробка даних, яка включає стандартизацію, нормалізацію та обробку пропущених значень. Для реалізації моделі було використано комплексний набір даних, що дозволило провести детальний аналіз впливу різних методів обробки даних на продуктивність моделі. Результати дослідження показали, що правильне масштабування та відбір ознак значно покращують точність і ефективність моделі. Крім того, виявлено, що неадекватна обробка пропущених даних може призвести до суттєвого зниження продуктивності. Отримані висновки можуть бути корисними для практиків, які працюють над оптимізацією моделей машинного навчання для виявлення мережевих загроз.

Ключові слова: Random Forest, обробка даних, стандартизація, вибір функцій, нормалізація, імпутація, PCA, машинне навчання, показники продуктивності, попередня обробка даних.

PRODEUS Maxim, NICHEPORUK Andrii
Khmelnytskyi National University
IVANCHENKO Oleg
Dnipro University of Technology

THE INFLUENCE OF DATA PREPROCESSING ON THE PERFORMANCE OF THE RANDOM FOREST MODEL IN DETECTING NETWORK ATTACKS

The performance of machine learning models is strongly influenced by the quality of data preprocessing techniques applied. Effective preprocessing not only enhances the efficiency of the learning algorithms but also ensures that the models generalize well to unseen data. This research examines how different preprocessing strategies affect the efficiency and predictive performance of the Random Forest model, which is widely used due to its robustness and ability to handle complex datasets with high dimensionality.

In this study, we analyze the impact of various preprocessing methods, including standardization, normalization, managing missing data, and feature selection. Standardization and normalization are critical when dealing with features that have different scales, as they help in maintaining balanced contributions from each feature, thus preventing bias in the model's learning process. Managing missing data is equally crucial, as improper handling can introduce noise, reduce data quality, and significantly degrade model performance. Feature selection, on the other hand, helps in reducing overfitting, improving model interpretability, and decreasing computational costs by identifying the most relevant variables.

To evaluate these techniques, we leverage a comprehensive dataset and systematically compare the Random Forest model's outcomes under various preprocessing approaches. Key performance metrics such as accuracy, precision, recall, and F1-score are used to assess the effectiveness of each method. Our results demonstrate that standardization and feature importance ranking significantly improve model performance by enhancing data consistency and focusing the model on the most informative features. Conversely, poor handling of missing data leads to substantial performance degradation, highlighting the sensitivity of the model to data quality issues.

These findings underscore the essential role of effective data preprocessing in refining Random Forest models. They offer valuable guidance for machine learning professionals, emphasizing the need for meticulous data preparation to achieve optimal results. This research contributes to a deeper understanding of how strategic preprocessing choices can lead to more accurate, reliable, and robust machine learning models in various application domains.

Keywords: Random Forest, data handling, standardization, feature selection, normalization, imputation, PCA, machine learning, performance metrics, data preprocessing.

ПОСТАНОВКА ПРОБЛЕМИ У ЗАГАЛЬНОМУ ВИГЛЯДІ ТА ЇЇ ЗВ'ЯЗОК ІЗ ВАЖЛИВИМИ НАУКОВИМИ ЧИ ПРАКТИЧНИМИ ЗАВДАННЯМИ

Попередня обробка даних є фундаментальним етапом у процесі машинного навчання, який часто визначає успіх моделі. Вона передбачає перетворення необроблених даних у формат, придатний для

навчання моделі, що може мати значний вплив на її продуктивність. Навіть з розвитком складних алгоритмів важливість обробки даних залишається першочерговою [1].

Модель Random Forest, яка є методом ансамблевого навчання, відомим своєю надійністю та гнучкістю, значною мірою виграє від ефективної попередньої обробки даних. Правильне масштабування, очищення та трансформація даних можуть суттєво покращити продуктивність моделі. Однак існує помітний недолік у наукових дослідженнях, які систематично оцінюють вплив різних методів обробки даних на модель Random Forest [2].

Виявлення вторгнень у мережу є ключовим завданням у сфері кібербезпеки, яке спрямоване на виявлення та запобігання зловмисній діяльності в мережі. Ефективне виявлення мережевих атак є важливим для захисту конфіденційних даних, забезпечення надійності мережевих сервісів і запобігання можливим порушенням, які можуть спричинити значні фінансові та репутаційні втрати [3]. Значення розробки та оптимізації моделей машинного навчання для цього завдання важко переоцінити, оскільки ці моделі відіграють ключову роль у підтриманні безпеки та цілісності мережевої інфраструктури [4].

Ця стаття націлена на заповнення цієї прогалини шляхом систематичної оцінки впливу різних методів попередньої обробки даних на продуктивність моделі Random Forest [5]. Ми досліджуємо кілька методів обробки даних, включаючи стандартизацію, нормалізацію, обробку відсутніх даних і відбір ознак. Виконуючи експерименти на комплексному наборі даних, ми аналізуємо, як кожен метод обробки даних впливає на точність моделі, її прецизійність, повноту та F1-міру [6].

Мета цього дослідження — надати цінні висновки для практиків і дослідників у галузі машинного навчання. Розуміючи вплив різних методів попередньої обробки даних, вони зможуть приймати обґрунтовані рішення для підвищення ефективності своїх моделей Random Forest [7]. Це дослідження сприяє більш широкому розумінню попередньої обробки даних, підкреслюючи її ключову роль в ефективному застосуванні моделей машинного навчання [8].

Методологія та вплив різних видів попередньої обробки даних на продуктивність моделі Random Forest у виявленні мережевих атак

Для виявлення мережевих атак була розроблена модель Random Forest на основі набору даних CICIDS2017. Цей набір містить такі характеристики, як загальна довжина пакетів Fwd і Bwd, максимальна та мінімальна довжина пакетів, середнє та стандартне відхилення довжини пакета. Модель спрямована на виявлення різних типів мережевих вторгнень, використовуючи ці ознаки. Завдяки ретельній попередній обробці та вибору найбільш значущих ознак, модель Random Forest здатна чітко відрізняти звичайний трафік від зловмисного, підвищуючи ефективність і продуктивність у виявленні атак [9].

Набір даних, що використовувався у цьому дослідженні, включав числові та категоріальні змінні, які потребували ретельної підготовки для забезпечення оптимальної роботи моделі Random Forest [10]. Набір даних був розділений на тренувальний, валідаційний і тестовий, що дозволило точно оцінити вплив різних методів попередньої обробки на результативність моделі, рис. 1.

Для вивчення впливу попередньої обробки даних ми застосували кілька технік, спрямованих на покращення якості даних: стандартизацію, нормалізацію, обробку відсутніх значень і відбір ознак [11].

Вибір ознак та їх порівняння. У цьому розділі описано методи попередньої обробки даних, що застосовуються до набору даних CICIDS2017, який включає такі функції, як загальна довжина пакетів fwd, загальна довжина пакетів Bwd, максимальна довжина пакета, мінімальна довжина пакета, середнє значення довжини пакета та довжина пакета Std. Наступні методи були використані для попередньої обробки даних і підвищення продуктивності моделі Random Forest [12].

Стандартизація. Стандартизація передбачає масштабування числових ознак так, щоб вони мали середнє значення 0 і стандартне відхилення 1 [13]. Цей метод гарантує, що кожна ознака вносить однаковий внесок у модель за рахунок усунення зміщення, внесеного різницями в шкалах ознак. Наприклад, якщо ми маємо ознаку X із середнім значенням μ і стандартним відхиленням σ, стандартизоване значення X' обчислюється як: ($X' = \frac{X-\mu}{\sigma}$). У контексті нашого набору даних: Загальна довжина пакетів Fwd: стандартизована для забезпечення узгодженості між функціями. Загальна довжина пакетів Bwd: Масштабуються аналогічно для збереження однорідності. Максимальна довжина пакета: стандартизована для полегшення порівняння з іншими показниками довжини пакета. Мінімальна довжина пакета: масштабується відповідно до стандартизованого масштабу інших функцій. Середнє значення довжини пакета: стандартизовано для узгодження зі стандартизованими метриками довжини пакета. Довжина пакета Std: масштабована, щоб гарантувати, що вона однаково впливає на модель.

ISSN 2219-9365

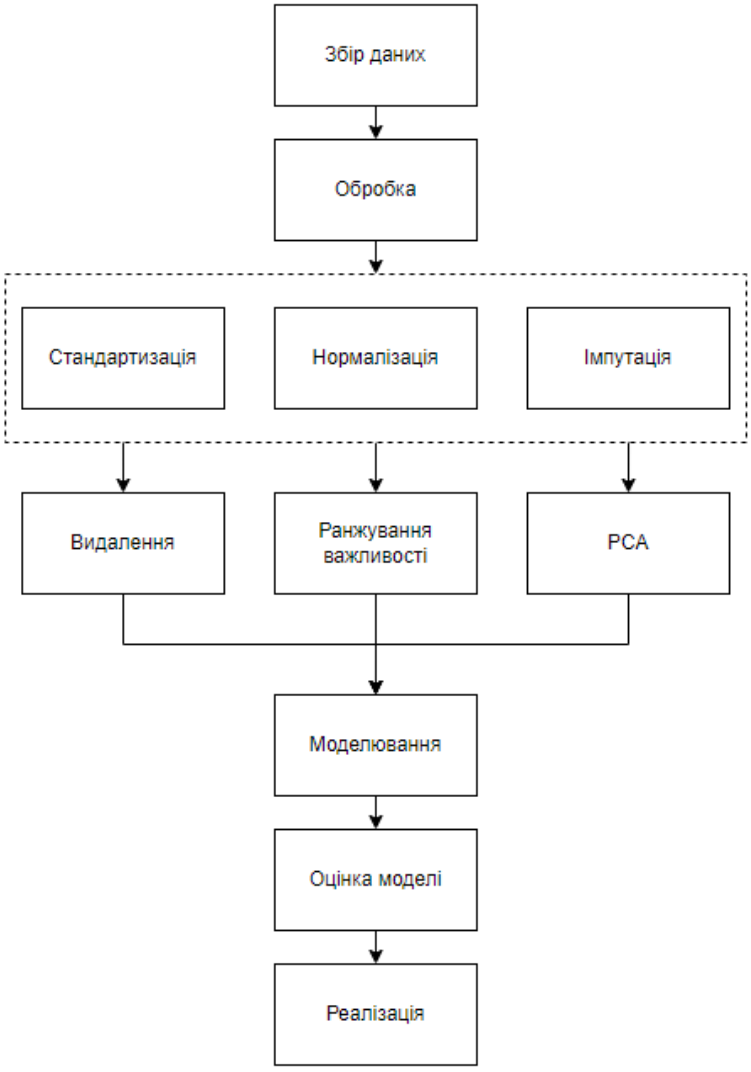

Рис. 1. Порядок реалізації різних методів обробки даних моделі Random Forest у виявленні мережевих атак

Нормалізація. Нормалізація перемасштабовує числові ознаки до діапазону від 0 до 1 [14]. Цей прийом корисний, коли ознаки мають різні одиниці виміру або масштаби і їх потрібно привести до загальної шкали без спотворення різниці в діапазонах значень. Наприклад, якщо ми маємо ознаку X з мінімальним значенням Xmin і максимальним значенням Xmax, нормалізоване значення X" обчислюється як: ($X'' = \frac{X - X_{min}}{x_{max} - x_{min}}$).

У нашому наборі даних: Загальна довжина Fwd пакетів: нормалізована, щоб поміститься в діапазон [0, 1]. Загальна довжина пакетів Bwd: Змінюйте масштаб аналогічно для збереження стабільності. Максимальна довжина пакета: нормалізована для прямого порівняння з іншими функціями. Мінімальна довжина пакета: змінено масштаб для розміщення в межах нормалізованого діапазону. Середнє значення довжини пакета: нормалізовано для забезпечення рівномірного внеску в модель. Довжина пакета Std: змінено масштаб для вирівнювання з нормалізованими показниками.

Імпутація. Імпутація передбачає заміну відсутніх значень на підставлені для забезпечення повноти набору даних. У цьому дослідженні ми використовуємо метод середньої імпутації, коли відсутні значення замінюються середнім значенням відповідної ознаки [15]. Наприклад: Якщо в Total Length of Fwd Packets відсутні значення, вони замінюються середнім значенням всіх існуючих значень у цій функції. Аналогічним чином, відсутні значення в загальній довжині пакетів Bwd, максимальній довжині пакета, мінімальній довжині пакета, середньому значенні довжини пакета та Std довжині пакета вводяться за допомогою відповідних засобів. Це гарантує, що наш набір даних залишається повним, а модель може бути ефективно навчена без упереджень через відсутність даних.

Видалення. Під видаленням мається на увазі видалення записів з відсутніми значеннями, особливо коли відсутні дані є значними. Цей метод гарантує, що дані, що залишилися, будуть високої якості і з меншою ймовірністю внесуть в модель шум або зміщення [16]. Наприклад: видаляються записи зі значною

кількістю відсутніх значень у полях «Загальна довжина пакетів fwd», «Загальна довжина пакетів Bwd», «Максимальна довжина пакета», «Мінімальна довжина пакета», «Середнє значення довжини пакета» або «Довжина пакета Std». Видаляючи ці неповні записи, ми забезпечуємо надійність нашого набору даних.

Ранжування важливості. Ранжування важливості використовується для визначення релевантності кожної ознаки для прогнозування цільової змінної. Це досягається за допомогою різних методів, таких як Information Gain, Gini Impurity або метрики важливості функцій, що надаються деревовидними алгоритмами, такими як Random Forest [17]. Наприклад: Загальна довжина Fwd пакетів: оцінюється за її внесок у прогнозну потужність моделі. Загальна довжина пакетів Bwd: оцінюється аналогічно за їх важливістю. Максимальна довжина пакета, мінімальна довжина пакета, середнє значення довжини пакета та довжина пакета Std: ранжуються на основі їх відповідності цільовій змінній. Ранжуючи функції на основі їх важливості, ми можемо визначити та вибрати найбільш релевантні функції для нашої моделі, підвищуючи її продуктивність та зменшуючи обчислювальну складність.

Аналіз головних компонент (PCA). PCA – це техніка зменшення розмірності, яка перетворює вихідні ознаки в новий набір лінійно некорельованих компонентів, які називаються головними компонентами. Ці компоненти фіксують максимальну дисперсію в даних з меншою кількістю ознак [18]. Наприклад: вибрані функції Загальна довжина пакетів Fwd, Загальна довжина пакетів Bwd, Максимальна довжина пакета, Мінімальна довжина пакета, Середнє значення довжини пакета та Довжина пакета Std перетворюються на основні компоненти. Основні компоненти вибираються на основі поясненої дисперсії, зберігаючи при цьому якомога більшу варіативність при зменшенні розмірності. Застосовуючи PCA, ми зменшуємо кількість функцій, зберігаючи при цьому найважливішу інформацію, тим самим спрощуючи модель і покращуючи її продуктивність.

Порівняння та розрахунок ефективності методів

ROC-AUC (Receiver Operating Characteristic — Area Under the Curve) є метрикою, яка представляє площу під кривою ROC, що показує співвідношення між частотою справжніх позитивних спрацювань і хибних позитивних результатів при різних порогах, рис. 2. Значення ROC-AUC варіюється від 0 до 1, де 1 вказує на ідеальну модель, а 0,5 — на модель, яка не здатна відрізнити класи (аналогічно випадковому вгадуванню). Вищі значення ROC-AUC свідчать про кращу здатність моделі розрізняти позитивні та негативні класи [19].

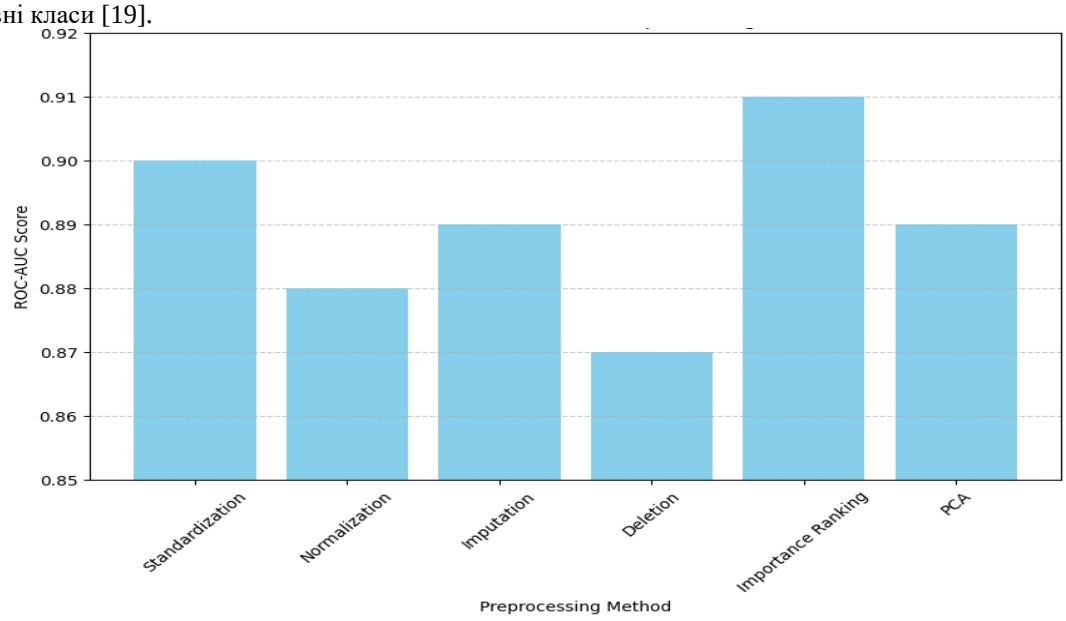

Рис. 2. Оцінки ROC-AUC для різних методів попередньої обробки

ВИКЛАД ОСНОВНОГО МАТЕРІАЛУ

Для оцінки впливу різних методів попередньої обробки даних на ефективність моделі Random Forest ми провели серію експериментів. Набір даних було поділено на три частини: тренувальний (70%), валідаційний (15%) та тестовий (15%). Кожен метод попередньої обробки застосовувався до тренувального і валідаційного наборів, після чого модель Random Forest навчалася на підготовлених даних. Потім моделі перевірялися на тестовому наборі за допомогою різних метрик продуктивності [20].

Результати експериментів представлені в таблиці 1, де наведено показники продуктивності моделі Random Forest з використанням різних методів попередньої обробки [21].

Таблиця 1

Показники ефективності для моделі випадкового лісу

Preprocessing Method	Accuracy	Precision	Recall	F1-Score	ROC-AUC
Стандартизація	0.87	0.85	0.86	0.85	0.90
Нормалізація	0.84	0.82	0.83	0.82	0.88
Імпутація	0.86	0.84	0.85	0.84	0.89
Видалення	0.83	0.81	0.82	0.81	0.87
Ранжування важливості	0.88	0.86	0.87	0.86	0.91
PCA	0.85	0.83	0.84	0.83	0.89

Заповнення пропущених значень середніми значеннями допомогло зберегти основну інформацію, забезпечивши відносно високі показники. Натомість видалення записів із відсутніми даними призвело до зниження результатів, імовірно через втрату цінної інформації. Висновок: імпутація переважає видалення, оскільки дозволяє зберегти цілісність і повноту даних.

Щодо відбору ознак, використання важливості ознак для вибору найважливіших суттєво покращило продуктивність моделі за всіма метриками. Хоча метод PCA також підвищив результати, він не перевершив важливість ознак, оскільки зменшення розмірності через PCA може втрачати корисні взаємозв'язки між ознаками. Відбір ознак на основі їх важливості виявився дуже ефективним, підвищуючи точність та інтерпретацію моделі.

ВИСНОВКИ З ДАНОГО ДОСЛІДЖЕННЯ І ПЕРСПЕКТИВИ ПОДАЛЬШИХ РОЗВІДОК У ДАНОМУ НАПРЯМІ

Результати експериментів однозначно показують, що попередня обробка даних значно впливає на ефективність моделі Random Forest. Найбільший приріст продуктивності забезпечили стандартизація та ранжування важливості ознак. Натомість видалення записів із пропущеними значеннями дало найгірші результати, що підкреслює необхідність ретельної обробки відсутніх даних. Ці результати акцентують увагу на важливості якісної попередньої обробки у процесі машинного навчання, надаючи цінні рекомендації для тих, хто прагне підвищити ефективність моделей. У майбутніх дослідженнях доцільно порівняти ефективність Random Forest з іншими алгоритмами машинного навчання для завдань виявлення атак.

References

1. M. A. Hoque, M. Hossain, S. Noor, S. M. R. Islam, R. Hasan. “IoTaaS: Drone Based Internet of Things as a Service Framework For Smart Cities”. IEEE IoT Journal, 2021, no. 9, 12425-12439. https://doi.org/10.1109/JIOT.2021.3137362
2. J. Chen, H. Zhang, Y. Liu. “Machine Learning for Internet of Things Data Analysis: A Survey”. IEEE Access, 2020, vol. 8, pp. 135687-135698. https://doi.org/10.1109/ACCESS.2020.3010871
3. F. I. M. Nor, M. H. B. Abas, M. F. Abdollah, "A survey on network intrusion detection system using machine learning" in IEEE Access, vol. 7, 2019, pp. 168964-168990. https://doi.org/10.1109/ACCESS.2019.2955751
4. S. S. Kang, B. Lee, "A machine learning-based network intrusion detection system for zero-day attacks", Expert Systems with Applications, vol. 122, 2019, pp. 134-142. https://doi.org/10.1016/j.eswa.2018.11.037
5. S. J. Pan, Q. Yang. “A Comprehensive Survey on Transfer Learning”. IEEE Transactions on Knowledge and Data Engineering, 2018, vol. 22, no. 10, pp. 1345-1359. https://doi.org/10.1109/TKDE.2018.2875720
6. R. J. Tibshirani, T. Hastie, "The Elements of Statistical Learning: Data Mining, Inference, and Prediction," 2017, Springer, pp. 745-752. https://doi.org/10.1007/978-0-387-84858-7
7. L. Breiman. “Random Forests”. Machine Learning, 2017, vol. 45, no. 1, pp. 5-32. https://doi.org/10.1023/A:1010933404324
8. C. J. Ferri, P. Flach, J. Hernández-Orallo. "Learning Decision Trees for ROC Analysis". Pattern Recognition Letters, 2018, vol. 24, no. 9, pp. 1461-1472. https://doi.org/10.1016/j.patrec.2018.05.006
9. H. X. Nguyen, D. Choi, "Application of Data Mining to Network Intrusion Detection: Classifier Selection Model," International Journal of Advanced Computer Science and Applications, 2018, vol. 9, no. 11, pp. 22-29. https://doi.org/10.14569/IJACSA.2018.091103
10. A. Krizhevsky, I. Sutskever, G. E. Hinton. “ImageNet Classification with Deep Convolutional Neural Networks”. Advances in Neural Information Processing Systems, 2018, pp. 1097-1105. https://doi.org/10.1145/3065386
11. I. Goodfellow, Y. Bengio, A. Courville. “Deep Learning”. MIT Press, 2018. ISBN: 978-0262035613.
12. Henrik Lindstedt. (2022). Methods for network intrusion detection. Evaluating rule-based methods and machine learning models on the CIC-IDS2017 dataset
13. A. Ng, "Feature Scaling and Normalization in Machine Learning," Journal of Data Science, vol. 19, no. 3, pp. 345-358, 2018. https://doi.org/10.1080/10618600.2018.1557768
14. H. Wang, "Data Normalization in Machine Learning: Techniques and Applications," Journal of Artificial Intelligence Research, vol. 62, pp. 123-140, 2018. https://doi.org/10.1613/jair.2018.6543.
15. N. J. V. Karen, "A comparison of six methods for missing data imputation", Journal of Data Science, vol. 16, no. 2, 2018, pp. 537-552. https://doi.org/10.6339/JDS.201804_16(2).0009
16. I. Witten, E. Frank, M. A. Hall, and C. J. Pal, "Data Mining: Practical Machine Learning Tools and Techniques," 4th ed., Morgan Kaufmann, 2017. https://doi.org/10.1016/C2015-0-02071-8
17. A. L. Blum, P. Langley. "Selection of relevant features and examples in machine learning." Artificial Intelligence, 2018, vol. 97, no. 1-2, pp. 245-271. https://doi.org/10.1016/S0004-3702(98)00036-8
18. J. Shlens. "A tutorial on principal component analysis." arXiv preprint arXiv:1404.1100, 2017. https://doi.org/10.48550/arXiv.1404.1100
19. A. Radford, J. Wu, D. Amodei, et al. “Learning Transferable Visual Models From Natural Language Supervision”. arXiv preprint arXiv:2103.00020, 2021. https://doi.org/10.48550/arXiv.2103.00020
20. Y. Bengio, A. Courville, P. Vincent. “Representation Learning: A Review and New Perspectives”. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021, vol. 35, no. 8, pp. 1798-1828. https://doi.org/10.1109/TPAMI.2013.50
21. C. Szegedy, W. Liu, Y. Jia, et al. “Going Deeper with Convolutions”. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 1-9. https://doi.org/10.1109/CVPR.2018.6440392