УДК 004.56 **doi: 10.26906/SUNZ.2024.4.052**

Svitlana Gavrylenko, Vadim Poltoratskyi

National Technical University "Kharkiv Polytechnic Institute", Kharkiv, Ukraine

NETWORK INTRUSION DETECTION MODEL BASED ON CONVOLUTIONAL NEURAL NETWORKS AND TABULAR DATA CONVERTED INTO IMAGES

Abstract. The **object of the study** is the process of identifying the state of a computer systems and network. The **subject** of the study are the methods of identifying the state of computer systems and networks. The **purpose** of this paper is to improve the quality of detecting intrusions into computer networks. The UNSW-NB 15 set, which contains information about the normal functioning of the network and during synthetic intrusions, was used as input. Deep neural networks (DL), their advantages and problems in big data processing are considered. It was found that deep neural networks when processing tabular data require their transformation. Modern methods of tabular data transformation were studied. **The results obtained**. A method of converting tabular data into an image is proposed. The method converts each object of a separate class from a set of tabular data into an image by mapping the attribute values onto a two-dimensional plane. The method was implemented programmatically using the GOOGLE COLAB cloud service based on Jupyter Notebook. **Conclusions.** It was found that the use of the proposed conversion method of tabular data into an image made it possible to use a classification model based on the CNN neural network and increase the quality of detection of intrusions into computer networks up to 4%.

Keywords: intrusion detection systems, computer networks, machine learning, deep neural networks, tabular data conversion.

Introduction

Despite significant progress in the field of cyber security, today there is a need for continuous improvement of methods and technologies used to monitor and identify intrusions into computer networks [1].

Network traffic represents a complex set of data transmitted across a network. Its characteristics are determined by the interaction of numerous factors, including the properties of devices, software, the data itself, and the network. This forms the key features that characterize the traffic.

- **Hardware components:** The device type, technical specifications (processor, RAM), operating system, and network adapters significantly impact the nature of the generated traffic.
- **Software:** The use of different applications, their functionality, versions, and interaction protocols shape the characteristics of the traffic.
- **Data:** The type of data (text, images, video), its volume, and transmission frequency directly affect the network load.
- **Network infrastructure:** Bandwidth of channels, transmission delays, quality of service, and network topology determine the data transmission capabilities.

The set of features characterizing network traffic is flexible and can be adapted to specific analysis tasks. For example, when studying cyber threats, attributes related to attacks, such as type, source, and target are added to the traditional set of features. This approach allows for the creation of a detailed traffic profile but also increases its dimensionality. A large number of features complicates the process of training machine learning models, as it requires more computational resources and time. The presence of high-dimensional data introduces the problem of computational complexity, which can make training machine learning models slower and more expensive.

Today, models based on deep neural networks are the most popular for big data processing [2–4]. Neural networks are typically associated with image and text processing. However, they can also be successfully applied to tabular data, though this requires transforming the data first [5].

The purpose of this work is to develop a method of transforming tabular data into images and using deep neural networks to improve the quality of intrusion detection in computer networks.

1. Approaches and methods

Deep neural networks (DL) have a unique ability to autonomously discover and learn hidden patterns in data, forming internal representations of information. This feature makes them a powerful tool for solving complex problems such as image recognition, natural language processing, and speech signal analysis. Their ability to automatically identify abstract representations allows deep learning models to effectively adapt to new tasks and generalize acquired knowledge to new data. In many tasks, especially with large amounts of data, neural networks demonstrate higher accuracy compared to traditional machine learning methods like linear regression or decision trees. Additionally, deep neural networks can detect complex nonlinear relationships between data features, which is particularly useful for tabular data with many interactions between variables.

Due to these properties of deep neural networks, numerous attempts have been made to apply them to various types of tabular data [6].

Converting raw data into images is one of approach to applying DL to tabular data [7].

For example, Taehoon Kim et al. in his work [8] uses dimensionality reduction algorithms such as t-SNE, UMAP, PCA, and others to convert tabular data into grayscale images, which are then combined into a color image.

Sharma A. et al. [9] also proposes a method called DeepInsight, whi.h involves arranging similar or correlated features in neighboring areas of a two-dimensional feature map to facilitate learning their complex relationships and interactions. This method also involves using t-SNE or kPCA. The authors also note that the resolution of the image is significant; a lower resolution means that more features will overlap, which can lead to information loss.

Bazgir O. et al. [10] proposed a method called Representation of Features as Images with Neighborhood Dependencies (REFINED), which transforms tabular data into images using Bayesian Multidimensional Scaling (BMDS) and a hill climbing algorithm.

Finally, Zhu et al. [11] introduced a technique for converting tabular data into images called the "Image Generator for Tabular Data" (IGTD). This technique assigns one pixel in a grayscale image to each feature, positioning similar objects closer together. Research has shown that models based on recurrent neural networks and CNNs trained on images generated by IGTD perform are better than CNN models trained on images from other converters. However, the quality of such models still requires improvement, particularly in the context of multiclass classification.

2. Method development

In this work, the UNSW-NB15 dataset is used as the source data. This dataset was developed by the Cyber Range laboratory of the Australian Centre for Cyber Security (ACCS) and contains information on normal network operations as well as synthetic intrusions [12].

UNSW-NB15 represents nine main attack categories using the IXIA tool: PerfectStorm Analysis, Backdoors, DoS, Exploits, Fuzzers, Generic, Reconnaissance, Shellcode, and Worms. There are 45 features developed using the Argus and Bro-IDS tools, as well as twelve algorithms, which cover network packet characteristics.

The model was implemented programmatically using the GOOGLE COLAB cloud service based on Jupyter Notebook.

As part of the study, the following stages of data pre-processing were performed:

- Data cleaning. Removal of erroneous or invalid data such as empty values, invalid records.
- Removal of non-informative features (for example, id).
- Data balancing.
- Data Scaling. So that all features have the same range of values.

To simplify modeling for multiclass classification, four of the ten classes were selected based on the largest number of objects: Normal, Generic, Exploits, and Fuzzers.

Data analysis revealed that the data was imbalanced. To simplify modeling and reduce training time, the RandomUnderSampler algorithm [13, 14] was applied to remove excess objects.

To convert tabular data of an object into an image, the following algorithm was proposed:

1. Store the target features in a separate set.
2. Convert categorical data to numerical values using the “factorize()” method.
3. Normalize the data using the “MinMaxScaler()” method.
4. Iterate through each row and transpose it so that feature names become indices. This is necessary to position features along the x-axis and their values along the y-axis.
5. Generate a plot, convert it to a grayscale image, and save it in .png format.

Using the developed algorithm, all objects in the dataset were converted into image format. The images are created as two-dimensional line plots, where the x-axis represents the features and the y-axis displays their values (Fig. 1).

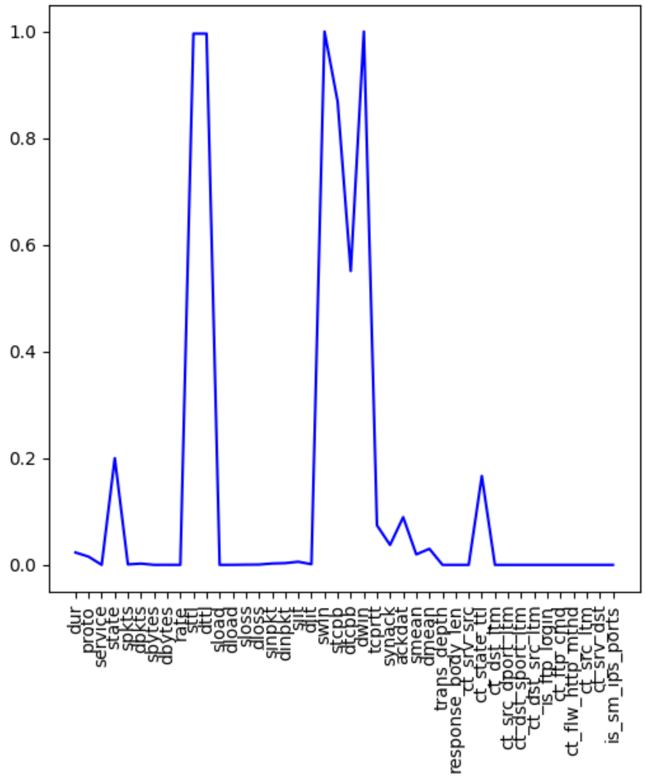

Fig. 1. An object of the Normal class with the standard set of features

Preserving the number of features eliminates the risk of information loss, thereby increasing the model's effectiveness. Moreover, converting data into images removes the need for a preprocessing step such as feature selection. Regardless of the number of features depicted in the image, the final volume of data fed into the model remains unchanged.

The next step in building the model is to create training and testing datasets for the classification model, which include these objects.

The following algorithm is proposed:

1. In a loop, load the current object's image using the `load_img()` method, convert it to a three-dimensional array using the `img_to_array()` method, and then add it to the list.
2. Serialize the list for storage. Serialization is needed for convenient storage of the dataset and quick loading.

3. Convert the list into a NumPy array, resulting in a four-dimensional array that can be fed into the neural network.

4. Also, load the class labels that were previously saved in a separate dataset.

5. Split the data into training and testing datasets.

The combination of the two previously developed algorithms forms a method for transforming tabular data into images, which are then used as input data for the model.

Since the IGTD method for converting tabular data into images showed the best results in previous studies, we will compare its results with those of the proposed method.

Fig. 2 shows an object of the Normal class transformed into an image using the IGTD method.

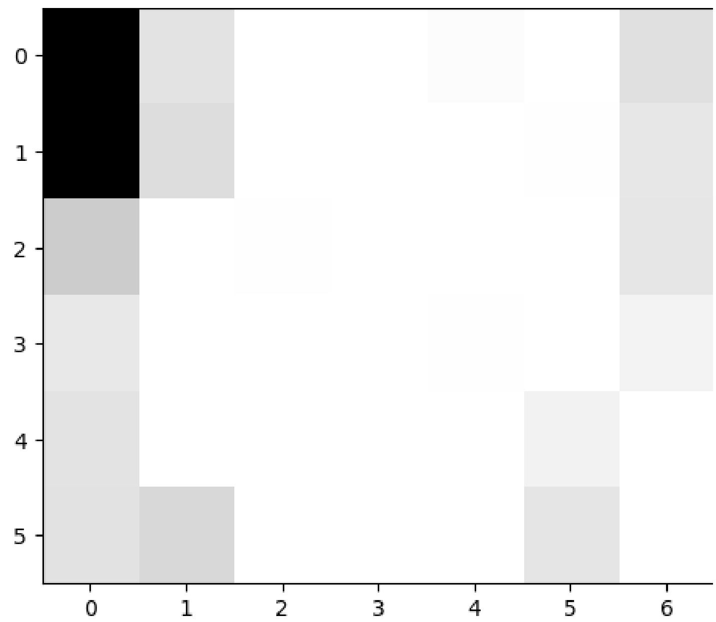

Fig. 2. An object of the Normal class transformed into an image using the IGTD method

For comparative analysis, a Convolutional Neural Network (CNN) model was constructed. The UNSW-NB15 dataset, which was converted from tabular data into images, was used as the input data. The model's performance was assessed using metrics such as Accuracy, F-1 Score, Precision, Recall, Training Time, and Recognition Time. The test results are presented in Table 1.

Table 1 – **Results of multiclass classification of the dataset converted into images using the IGTD method and the proposed method**

Transformation method	Model of classifier	Accuracy, %	F1-score, %	Precision, %	Recall, %	Trainin g time, s	Recognition time, s
IGTD	CNN	83.21	83.35	84.9	81.66	680.2	10.45
Proposed method	CNN	86.89	87.06	88.02	85.59	669.3	10.62

As shown in Table 2, using the proposed method for converting tabular data into images enabled the construction of a classification model based on a CNN and improved classification quality by up to 4%.

Conclusions

This work investigates the effectiveness of using modern deep neural network models for intrusion detection in computer networks.

Deep neural networks are among the most popular methods for analyzing big data. Typically, neural network models are used for processing images and texts. For handling tabular data, these models require a transformation of the input data.

The research analyzed various data conversion approaches. It was found that the most effective technique for converting tabular data into images is the "Image Generator for Tabular Data" (IGTD). However, the quality of the model remains insufficient.

To improve model quality, a method for converting tabular data into images has been proposed. Initially, images are created for each object in the form of two-dimensional line plots, where features are represented on the x-axis and their values on the y-axis. The images are then converted into a three-dimensional array, and a list of objects is formed. This list is converted into a NumPy array, resulting in a four-dimensional array that can be fed into the neural network.

The proposed method was implemented using the GOOGLE COLAB cloud service based on Jupyter Notebook. An intrusion detection model for computer networks based on a Convolutional Neural Network (CNN) was developed. The model also includes a block for converting the input data into images.

In this work, the UNSW-NB15 dataset is used as the source data. This dataset was developed by the Cyber Range laboratory of the Australian Centre for Cyber Security (ACCS) and contains information about normal network operations as well as synthetic intrusions.

For comparative analysis, the input data was converted into images using both the proposed method and the IGTD method.

Using the proposed method for converting tabular data into images enabled the application of a CNN-based classification model and improved classification quality by up to 4%.

Future research will focus on augmenting the input data with synthetic features and evaluating their impact on model performance.

REFERENCES

1. Chen, Ying. "Big data technology for computer intrusion detection" Open Computer Science, vol. 13, no. 1, 2023, 20220267. https://doi.org/10.1515/comp-2022-0267
2. Gavrylenko , S., Poltoratskyi , V., & Nechyporenko , A. “Intrusion detection model based on improved transformer”. Advanced Information Systems, 2024, 8(1), 94–99. https://doi.org/10.20998/2522-9052.2024.1.12
3. Dusan Nedeljkovic, Zivana Jakovljevic. “CNN based method for the development of cyber-attacks detection algorithms in industrial control systems” ,Computers & Security, 2022, vol.114, 102585, https://doi.org/10.1016/j.cose.2021.102585
4. A. El-Rady, H. Osama, R. Sadik and H. El Badwy, "Network Intrusion Detection CNN Model for Realistic Network Attacks Based on Network Traffic Classification," 2023 40th National Radio Science Conference (NRSC), Giza, Egypt, 2023, pp. 167-178, https://doi.org/10.1109/NRSC58893.2023.10152872
5. S. Golubev and E. Novikova, "Transformation of Network Flow Data into Images for Intrusion Detection Using Convolutional Neural Networks," 2023 International Russian Automation Conference (RusAutoCon), Sochi, Russian Federation, 2023, pp. 948-952, https://doi.org/10.1109/RusAutoCon58002.2023.10272890
6. Borisov V, Leemann T, Seßler K, Haug J, Pawelczyk M, Kasneci G. “Deep neural networks and tabular data: A survey.” IEEE Transactions on Neural Networks and Learning Systems. 2022 https://doi.org/ 10.1109/TNNLS.2022.3229161
7. Medeiros Neto L, Rogerio da Silva Neto S, Endo PT, “A comparative analysis of converters of tabular data into image for Kim the classification of Arboviruses using Convolutional Neural Networks.” PLoS ONE 2023? 18(12): e0295598. https://doi.org/10.1371/journal.pone.0295598
8. Taehoon & Pak, Wooguil. (2023). “Deep Learning-Based Network Intrusion Detection Using Multiple Image Transformers.” Applied Sciences. 13. 2754. https://doi.org/10.3390/app13052754
9. Sharma, A., Vans, E., Shigemizu, D. et al. “DeepInsight: A methodology to transform a non-image data to an image for convolution neural network architecture”. *Sci Rep* **9**, 11399 (2019). https://doi.org/10.1038/s41598-019-47765-6
10. Bazgir O, Zhang R, Dhruba SR, Rahman R, Ghosh S, Pal R. “Representation of features as images with neighborhood dependencies for compatibility with convolutional neural networks.” Nature communications. 2020;11(1):4391. https://doi.org/10.1038/s41467-020-18197-y
11. Zhu Y, Brettin T, Xia F, Partin A, Shukla M, Yoo H, Evrard YA, Doroshow JH, Stevens RL. “Converting tabular data into images for deep learning with convolutional neural networks.” Sci Rep. 2021 Jul 1;11(1):14036. https://doi.org/10.1038/s41598-021-93376-5
12. N. Moustafa and J. Slay, "UNSW-NB15: a comprehensive data set for network intrusion detection systems (UNSW-NB15 network data set)," 2015 Military Communications and Information Systems Conference (MilCIS), Canberra, ACT, Australia, 2015, pp. 1-6, https://doi.org/ 10.1109/MilCIS.2015.7348942
13. Abdi L, Sattar H. “To combat multi-class imbalanced problems by means of over-sampling techniques”. IEEE Trans Knowl Data Eng. 2016;28(1):238–251. https://doi.org/10.1109/TKDE.2015.2458858
14. S. Gavrylenko, V. Zozulia, and N. Khatsko. “Methods for Improving the Quality of Classification on Imbalanced Data”, Proceedings of the IEEE 4th KhPI Week on Advanced Technology (KhPIWeek), Kharkiv, Ukraine, 2023, pp. 1-5, https://doi.org/ 10.1109/KhPIWeek61412.2023.10312879

Received (Надійшла) 15.07.2024
(Accepted for publication) Прийнята до друку 23.10.2024

Модель виявлення вторгнень у комп’ютерну мережу
на основі згорткових нейронних мереж та табличних даних, перетворених на зображення

С. Ю. Гавриленко, В. О. Полторацький

Анотація. Об’єктом дослідження є процес ідентифікації стану комп’ютерної системи та мережі. **Предметом дослідження** є методи ідентифікації стану комп'ютерних систем і мереж. **Метою даної роботи** є підвищення якості виявлення вторгнень в комп'ютерні мережі. В якості вхідних даних використовувався набір UNSW-NB 15, який містить інформацію про нормальне функціонування мережі та під час синтетичних вторгнень. Розглянуто глибокі нейронні мережі (DL), їх переваги та проблеми в обробці великих даних. Виявлено, що глибокі нейронні мережі при обробці табличних даних потребують їх трансформації. Досліджено сучасні методи трансформації табличних даних. **Отримано такі результати**. Запропоновано метод перетворення табличних даних в зображення. Метод перетворює кожен об’єкт окремого класу з набору табличних даних на зображення шляхом відображення значень атрибутів на двовимірну площину. Метод реалізовано програмно за допомогою хмарного сервісу GOOGLE COLAB на базі Jupyter Notebook. **Висновки**. Встановлено, що використання запропонованого методу перетворення табличних даних в зображення дозволило використати модель класифікації на основі нейронної мережі CNN та підвищити якість виявлення вторгнень у комп’ютерні мережі до 4%.

Ключові слова: системи виявлення вторгнень, комп'ютерні мережі, машинне навчання, глибокі нейронні мережі, перетворення табличних даних.