Bulletin Of High Technology N1(25) 2023.-pp. 59-66.

UDC - 004.725:004.852

RESEARCH OF ALGORITHM FOR EXPANDING THE DATABASE OF TRAINING DATASETS OF A GENERATIVE-ADVERSARIAL NETWORK

Robert G. Hakobyan
National Polytechnic University of Armenia
105 Teryan St. 0009, Yerevan
rob.hakonyan@polutechnic.am
ORCID iD: 0000-0002-6919-2106
Republic of Armenia

Timur V. Jamgharyan
National Polytechnic University of Armenia
105 Teryan St. 0009, Yerevan
t.jamgharyan@yandex.ru
ORCID iD: 0000-0002-9661-1468
Republic of Armenia

https://doi.org/10.56243/18294898-2023.1-59

Abstract

The paper presents the results of *calculations and tests of the* developed dataset expanding algorithm for training a generative-adversarial network. The research was conducted on two types of malicious software: mimikatz and cring. The boosting method was chosen as a method for expanding the database of datasets.

The process of expanding the database of datasets was carried out in a granular manner, *using timestamps.* Simulation of the algorithm operation at different iterations and visualization of the results have been carried out.

Key words: augmentation, boosting, training set, machine learning, weight coefficient, datasets, classification feature, mimikatz.

Introduction

An important place is occupied by the construction of a multi-level complementary security system to the operation of the network infrastructure. An important element of the network and infrastructure security architecture is an intrusion detection system (IDS).Various researchers and scientific communities are conducting research on the creation of an intrusion detection system based on generative-adversarial networks [1-4].

Generative-Adversarial Network (GAN) is an algorithm based on a combination of two neural networks one of which generates an object and the other tries to distinguish correct («real») objects from incorrect ones. The generating network G (generator) creates (generates) objects of a specified structure, the discriminating network D (discriminator) draws

conclusions about the similarity of the generated and true objects [5]. Concept of generative-adversarial networks was invented by Ian Goodfellow in 2014.

The problems of using the generative adversarial network and the whole concept of machine learning as a tool for detecting an attack on the Infrastructure are little explored.

It is necessary to distinguish that the generation of malicious software (software) using a GAN is a difficult task, due to the fact that both malicious and non-malicious software are implemented on the basis of a single software code base. Research is currently underway to generate «synthetic» training datasets for a GAN. Various researchers are conducting research on the research of methods and ways of preparing data, as well as creating methods for training a GAN to generate «synthetic» datasets of malware and detect it. In particular, ML researchers widely use the data augmentation method to create «synthetic» datasets (augmentation - is an increase in the data sample for training through the modification of existing data [6]).

The relevance of the research is due to the continuous improvement of the means and methods of attacks on the network infrastructure (NI), including the use of ML. The conditions for a successful attack on the Infrastructure using ML are considered in [7]. A substantive research was done on the introduction of certain data sets of malicious traffic into unencrypted VoIP traffic (Voice over IP, VoIP).

The choice of Internet telephony traffic as a transport for malware is due to several factors:

- the semantic content of telephone traffic is a priori unknown, which makes it difficult to analyze it even with «standard» IDS,
- traffic patterns allow an attacker to enter false data that is difficult to detect by IDS.

The scientific novelty of the research lies in the research of the possibility of creating malware traffic datasets with granular control of the augmentation process. The boosting method is used as a tool for increasing datasets.

Conflict Setting

It is necessary to carry out a quantitative change (expansion of the base) of training datasets for a generative-adversarial network without changing their quality.

Research Results

The importance of the ideas of "independence" and "freedom" is also evidenced by the

It is necessary to develop and programmatically implement an algorithm that will granularly expand training datasets for training GAN.

$$f_m(x) \cong f`_m(x) \tag{1}$$

where $f_m(x)$-original dataset, $f_m`(x)$-final dataset.

Border conditions: $N \geq \varepsilon$, (ε-epoch number), sign > 0, parameters and attributes of augmented datasets must be within the protocol.

For the correct operation of the algorithm and software, the following datasets are pre-formed.

1. **Initial dataset**. It is formed on the basis of VoIP telephony traffic operating over the SIP (Session Initiation Protocol, SIP) and RTP (Realtime Transport Protocol, RTP) protocols. As a software for capturing network traffic, a modified low-level library based was used the open source solution *tcpdump*.
2. **Training dataset.** Generated on the basis of the initial dataset by injecting malicious software obtained from open sources into it (was used *mimikatz* and *cring* malware from sources [8,9]).
3. **Test dataset**. Formed on the basis of a training dataset, but with a fixed value for both the type of malware and its percentage. Checked on online resources [10].Those datasets with embedded malware that are classified on the resources as malicious were not used in the training of the GAN (since they were detected by standard protection tools).
4. **Validation dataset.** It is formed on the basis of the initial and test, as well as on the basis of the initial and training data sets, XOR addition of these data types.

For each dataset, a signature calculation procedure was carried out (by the hashing method).

The developed algorithm is shown in Fig.1.

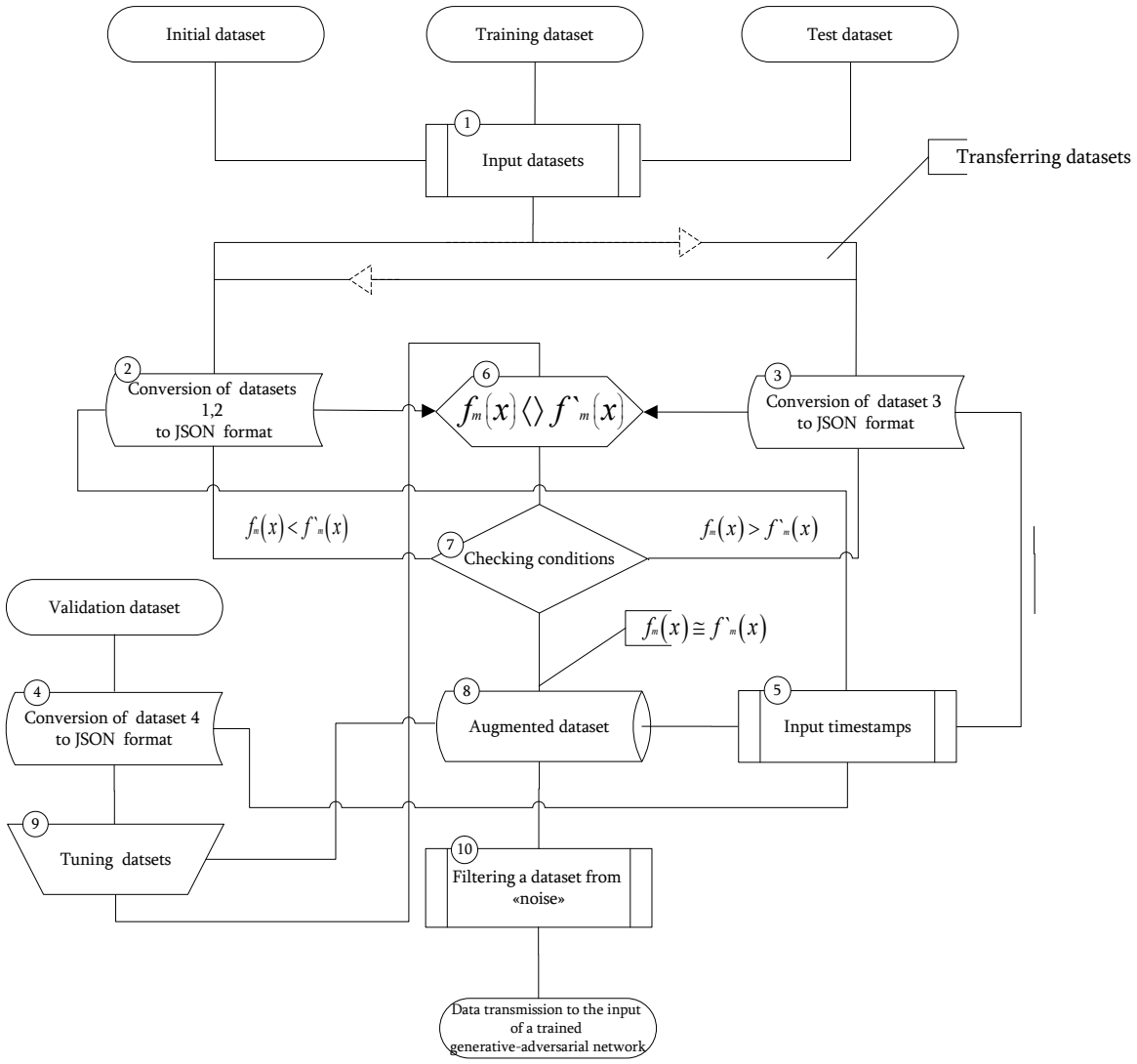

Fig. 1 The algorithm of the software for granularly expanding training datasets for GAN

Algorithm operation

step 1 splitting into blocks of input datasets,

step 2,3,4 conversions to JSON format (Java Script Object Notation, JSON), initial (dataset 1), training (dataset 2), test (dataset 3) and validation (dataset 4) datasets,

step 5 setting the boosting time interval,

step 6 dataset augmentation,

step 7 checking the conditions for executing the augmentation algorithm,

step 8,9 verification of the augmented data set and adjustment based on the validation dataset. Implemented the ability to «track» the state of datasets at a given point in time based on timestamps,

step 10 filtering the augmented data set from «noise».

Processing, transformation and augmentation of input datasets (block 6 of the algorithm) is implemented on the basis of the «recurrent neural network with attention» (RNN) mechanism. The scheme of augmentation of datasets based on a RNN is shown in Fig. 2. Visualized results of generating datasets are is shown in Fig. 3-4.

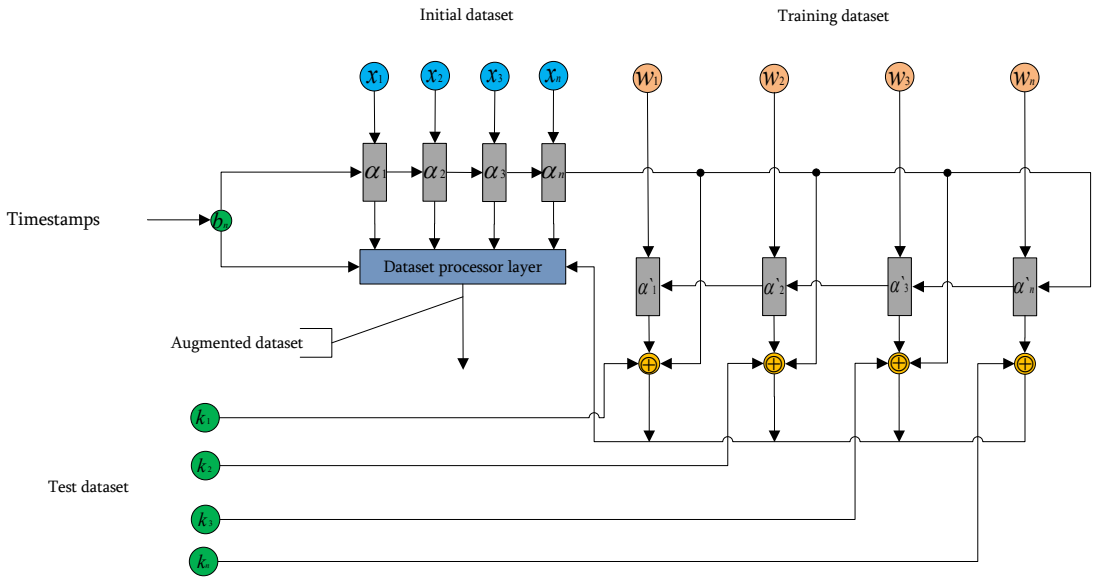

Fig. 2 ***Dataset augmentation mechanism based RNN***

where: $k_1....k_n$ test dataset, $x_1....x_n$ initial dataset, $w_1....w_n$ training dataset, $b_1....b_n$ timestamp hash values, $\alpha_1....\alpha_n$ coefficients (weights) of the primary boosting algorithm, $\alpha`_1....\alpha`_n$ biased coefficients (weights) of the boosting algorithm.

«Quality datasets» - are augmented datasets with embedded malware that have been tested on the virus total resource and not detected by the resource monitoring systems as malicious software.

Recurrent neural network with attention (attention mechanism) - is a technique used in recurrent neural networks and convolutional neural networks to search for relationships between different parts of the input and output data.

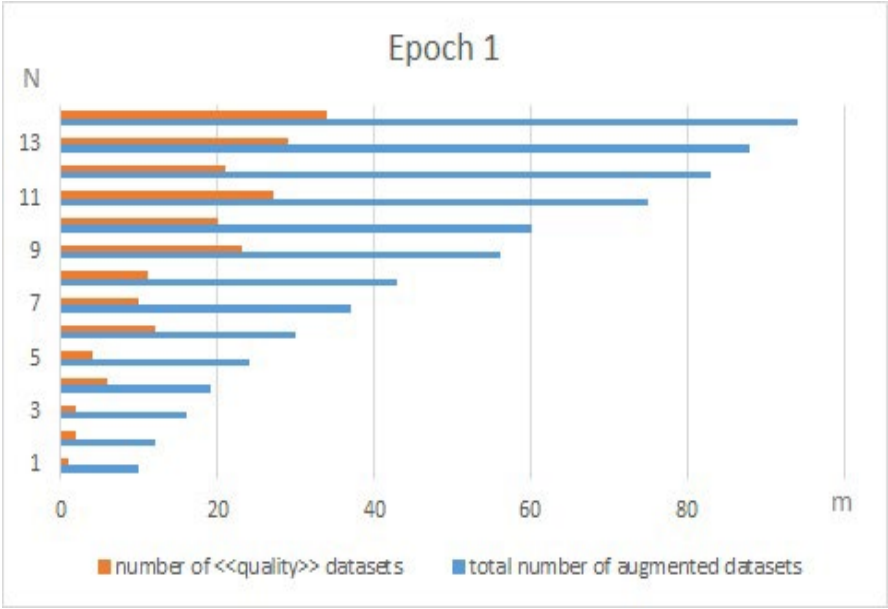

Fig. 3 Number of generated datasets and number of «quality» datasets (epoch 1)

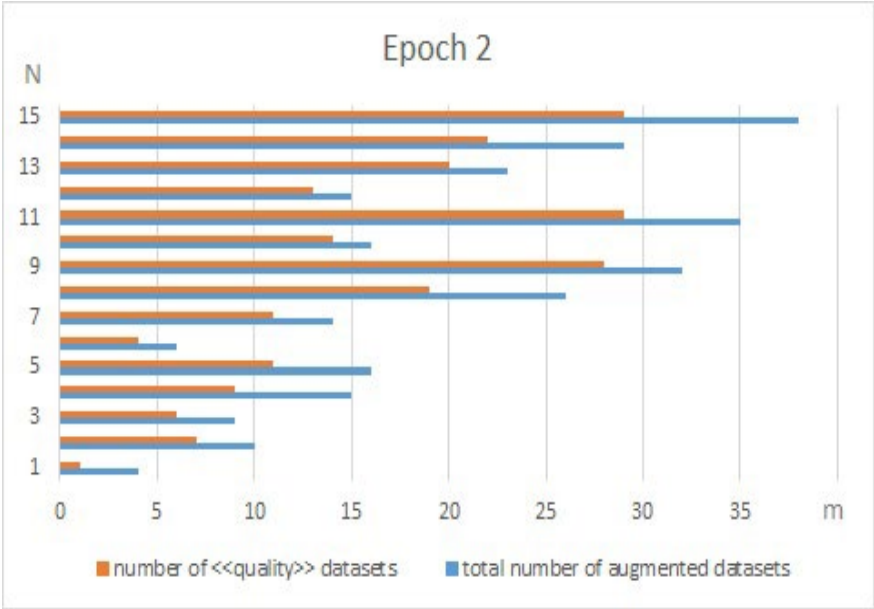

Fig. 4 Number of generated datasets and number of «quality» datasets (epoch 2)

Used software and hardware

1. The hash function was SHA-1 (Secure Hash Algorithm, SHA),
2. The calculations were performed on the Dell Power Edge T-330 server,
3. The open source platform asterisk was used as a VoIP telephony server,
4. The research was conducted in a virtual environment based on the Windows Server 2016 operating system with the preinstalled Hyper-V role,
5. The developed software is implemented in the python programming language in the PyCharm development environment.

An increase in boosting iterations and a timestamp value leads to an increase in the number of generated data sets, but these datasets are of little use for training a generative

adversarial network, as they are detected even by «standard» intrusion detection systems as malware.

An increase in the number of training epochs has a positive effect on the «quality» of the generated data sets, since with an increase in the training epoch, the number of generated and «quality» data sets practically coincides. As a disadvantage, it should be noted that an increase in the number of epochs, increasing the «quality» of the generated datasets, increases their «noisiness».

An important requirement for training a neural network is to ensure that the training set is balanced.

The use and input of a timestamp into the set of training datasets made it possible to granularly detect the moment of «retraining» of the generative-adversarial network.

Only datasets that have been filtered from «noise» are suitable for practical use.

The developed algorithm and software allows, in the presence of the source code of various malicious software, to create training data for an intrusion detection system based on ML, increasing the protection of NI.

The initial data of the dataset 1, part of the source code of the developed software and third-party libraries are presented in the repository to [11].

Conclusion

An increase in the number of training epochs has a positive effect on the «quality» of augmented datasets, since with an increase in the training epoch, the number of generated and «quality» datasets practically coincides. As a disadvantage, it should be noted that increasing the number of epochs and increasing the «quality» of augmented data sets increase their «noisiness». The use and input of a timestamp into the set of training datasets made it possible to granularly detect the moment of «overfitting» of the generative-adversarial network. The developed algorithm and software allow, in the presence of the source code of various malicious software, to create training data for an intrusion detection system based on machine learning by increasing the protection of the network infrastructure.

References

1. J.Chan, Y.Zhao, Q.Li, X.Feng, K Xu, FeDDeF: Defence Against Gradient Leakage in Federated Learning-based Network Intrusion Systems. [Online].Available: https://arxiv.org/abs/2210.04052
2. S.Das, FGAN: Federated Generative Adversarial Networks for Anomally Detection in Network Traffic
[Online].Available: https://arxiv.org/abs/2203.11106
3. Y. Cui, W.Shen, J.Zhang, W.Lu, L.Sun, S.Chen, Usung EBGAN for Anomally Intrusion Detection
[Online].Available: https://arxiv.org/abs/2206.10400
4. Z.Zhi lin, T.D. Pike, A Hypergraph-Based Mashine Learning Ensemble Network Intrusion Detection System. [Online].Available: https://arxiv.org/abs/2211.03933
5. Ian J.Goodfellow,J.Pouget-Abadie, M. Mirza, B.Xu, D.Warde-Farley,S.Ozair, A. Courville,Y.Bengio. Generative Adversarial Networks. [Online]. Available: https://arxiv.org/abs/1406.2661

6. Z.Yuan, Z.Zhao, Y/Liu, X.Zhang, X.Hou. RPN: A Word Vector Level Data Augmentataion Algorithm in Deep Learning for Language Understanding.
[Online].Available: https://arxiv.org/abs/2212.05961
7. T.V. Jamgharyan, V. H. Ispiryan. Model of Generative Network Attack, CSIT Conference 2021, Yerevan, Armenia,// (2021, September 27-October 1),
[Online].Available: https://csit.am/2021/proceedings/IS/IS_3.pdf
8. Mimikatz software
[Online].Available: https://github.com/search?q=mimikatz
9. Internet resource dedicated to encryption viruses
https://id-ransomware.blogspot.com/2021/01/cring-ransomware.html
10. Internet resource for checking various types of malicious files.
[Online].Available: https://www.virustotal.com/gui/home/upload
11. Programmatic learning source code and full research results
[Online].Available: https://github.com/T-JN/Research-of-Algorithm-for-Expanding-the-Database-of-Training-Datasets-of-a-GAN

References

1. J.Chan, Y.Zhao, Q.Li, X.Feng, K Xu, FeDDeF: Defence Against Gradient Leakage in Federated Learning-based Network Intrusion Systems. [Online].Available: https://arxiv.org/abs/2210.04052
2. S.Das, FGAN: Federated Generative Adversarial Networks for Anomally Detection in Network Traffic,
[Online].Available: https://arxiv.org/abs/2203.11106
3. Y. Cui, W.Shen, J.Zhang, W.Lu, L.Sun, S.Chen, Usung EBGAN for Anomally Intrusion Detection
[Online].Available: https://arxiv.org/abs/2206.10400
4. Z.Zhi lin, T.D. Pike, A Hypergraph-Based Mashine Learning Ensemble Network Intrusion Detection System. [Online].Available: https://arxiv.org/abs/2211.03933
5. Ian J.Goodfellow,J.Pouget-Abadie, M. Mirza, B.Xu, D.Warde-Farley,S.Ozair, A. Courville,Y.Bengio. Generative Adversarial Networks. [Online]. Available: https://arxiv.org/abs/1406.2661
6. Z.Yuan, Z.Zhao, Y/Liu, X.Zhang, X.Hou. RPN: A Word Vector Level Data Augmentataion Algorithm in Deep Learning for Language Understanding.
[Online].Available: https://arxiv.org/abs/2212.05961
7. Timur V. Jamgharyan, Vahe H. Ispiryan. Model of Generative Network Attack, CSIT Conference 2021, Yerevan, Armenia,// (2021, September 27-October 1),
[Online].Available: https://csit.am/2021/proceedings/IS/IS_3.pdf
8. Программное обеспечение mimikatz
[Online].Available: https://github.com/search?q=mimikatz
9. Интернет ресурс посвященный вирусам-шифровальщикам
https://id-ransomware.blogspot.com/2021/01/cring-ransomware.html
10. Интернет ресурс проверки различного типа вредоносных файлов.
https://www.virustotal.com/gui/home/upload
11. Исходный код программного обучения и полные результаты исследования
https://github.com/T-JN/Research-of-Algorithm-for-Expanding-the-Database-of-Training-Datasets-of-a-GAN

Bulletin Of High Technology N1(25) 2023.-pp. 59-66.

R.G. Hakobyan, T.V. Jamgharyan

ԳԵՆԵՐԱՏԻՎ-ՄՐՑԱԿՑԱՅԻՆ ՑԱՆՑԻ ՈՒՍՈՒՑՄԱՆ ՀԱՄԱՐ ՏՎՅԱԼՆԵՐԻ ՀԱՎԱՔԱԾՈՒՆԵՐԻ ՀԵՆՔԻ ԸՆԴԼԱՅՆՄԱՆ ՀԱՇՎԵԿԱՐԳԻ ՀԵՏԱԶՈՏՈՒՄ

Հակոբյան Ռ.Գ., Ջամղարյան Թ.Վ.
Հայաստանի ազգային պոլիտեխնիկական համալսարան

Հոդվածում ներկայացված են գեներատիվ-մրցակցային ցանցի ուսուցման համար տվյալների հավաքածուների հենքի ընդլայնման մշակված հաշվկարգի հետազոտության արդյունքները: Հետազոտությունն իրականացվել է երկու տեսակի վնասաբեր ծրագրային ապահովման ելակետային կոդի հիման վրա՝ mimikatz-ի և cring-ի: Որպես հավաքածուների հենքի ընդլայնման մեթոդ ընտրվել է բուսթինգը: Տվյալների հավաքածուների ընդլայնման գործընթացն իրականացվել է հատիկավոր եղանակով՝ օգտագործելով ժամանակի պիտակները: Իրականացվել է ալգորիթմի գործողության մոդելավորում տարբեր կրկնություններում և արդյունքների արտացոլում:

Բանալի բառեր. աուգմենտացիա, բուսթինգ, գեներատիվ-մրցակցային ցանց, ուսուցման հավաքածու, տվյալների հավաքածու, մեքենայական ուսուցում:

ИССЛЕДОВАНИЕ АЛГОРИТМА РАСШИРЕНИЯ БАЗЫ ОБУЧАЮЩИХ НАБОРОВ ДАННЫХ ДЛЯ ГЕНЕРАТИВНО-СОСТЯЗАТЕЛЬНОЙ СЕТИ

Акопян Р.Г., Джамгарян Т.В.
Национальный политехнический университет Армении

В статье представлены результаты исследований алгоритма расширения базы наборов данных для обучения генеративно-состязательной сети. Исследование проводилось на двух типах зловредного программного обеспечения mimikatz и cring. В качестве метода расширения базы наборов данных выбран метод бустинга *(процедура последовательного построения композиции алгоритмов машинного обучения)*.

Процесс расширения наборов данных был выполнен гранулярным способом с использованием меток времени. Проведено моделирование работы алгоритма при разных итерациях и визуализация результатов.

Ключевые слова: аугментация, бустинг, генеративно-состязательная сеть, обучающая выборка, набор данных, машинное обучение.

Submitted on 25.04.2022
Sent for review on 26.04.2022
Guaranteed for printing on 28.04.2023