Scholar Publishing Group
Distributed Processing System
https://doi.org/10.38007/DPS.2021.020403
ISSN 2790-0916 Vol. 2, Issue 4: 18-27

Design and Implementation of Distributed System Based on Machine Learning Algorithm and Numerical Simulation

Sivakumar Shirley*

Democritus University of Thrace, Greece

**corresponding author*

Keywords: Network Traffic, Machine Learning, Data Simulation, Distributed Processing

Abstract: In recent years, the Internet industry has entered a period of rapid development, and smart life has brought people a lot of experiences they have never had before. Rich application scenarios not only bring convenience to people, but also expose more and more network security problems. Network traffic types are more diversified, network security monitoring is becoming more and more difficult, and network communication quality and user host security are constantly facing the threat of network intrusion. This paper studies the current network traffic anomaly detection methods. Aiming at the problems of low accuracy and difficult real-time monitoring caused by the limitations of data scale and processing capacity in the previous methods, combined with machine learning and data simulation technology, a multi model fusion streaming parallel anomaly detection method is proposed, which enables distributed processing of massive streaming data on the basis of ensuring algorithm accuracy, At the same time, a visualization system based on network traffic anomaly detection is developed. The system can monitor the flow, and the reliability experiment can also meet the daily needs.

1. Introduction

With the rapid update of network technology in recent years, the Internet industry has achieved rapid development, and more and more application services have been integrated into people's lives. By the end of last year, China's Internet penetration rate reached 70.4%, and the number of Internet users reached 989 million. The scale of Internet users is still expanding at a certain growth rate. China's information digitization plays an increasingly prominent role in the development of modern society.

In the 1980s, foreign scholars first proposed the concept of network intrusion detection. Its

purpose is to detect whether there is any behavior that damages the host system in the network environment. The key of anomaly detection is to find a suitable method to analyze the network traffic. There are many network traffic anomaly detection methods. According to the processing methods of different characteristics, the anomaly detection methods can be roughly divided into three categories. The first type is the method based on data clustering analysis, which regards the clusters with small data volume or the data that cannot be clustered as outliers, and K-means is the classical algorithm in this type of methods [1]. In previous years, a network hybrid anomaly detection method based on K-means algorithm was proposed. This method first uses K-means clustering algorithm to automatically generate abnormal data classes on the training data, and then matches the unknown network traffic data with these abnormal data classes to determine whether the data is abnormal. Experiments show that this method achieves high detection rate and low false alarm rate. The second type is based on data statistics, which relies on statistical analysis in mathematics to detect abnormal data [2]. On this basis, a parameterized method for detecting network anomalies using only aggregated traffic statistics is proposed. The method is compared and analyzed in three aspects: sequential probability ratio, traffic statistics, packet size statistics, and bit rate signal-to-noise ratio. The feasibility of the algorithm is verified. The third type is a method based on data proximity, which determines whether it is abnormal data by judging the relative distance between the point to be detected and the normal point. Due to the numerous dimensions of network traffic data, in the process of anomaly detection, we usually encounter problems such as long time-consuming and low utilization of CPU resources. Therefore, we must analyze the important characteristics of network traffic. A robust median nearest neighbor linear discriminant analysis method based on generalized mean can find relevant features by processing network connection data close to the median of each class. Through a large number of experiments, it is proved that this method is superior to many linear discriminant analysis [3]. With the increase of network traffic data fields and the rapid growth of data scale, a single anomaly detection algorithm can not meet the real-time detection needs of massive and high-dimensional data. The real-time streaming anomaly detection method based on model fusion and big data technology has become a research hotspot in this field.

Based on the above problems, this paper presents a streaming distributed anomaly detection method based on machine learning algorithm. By continuously accumulating abnormal data, extracting the common characteristics of abnormal network traffic data, establishing a black list of abnormal network traffic, and developing a distributed anomaly detection system based on network traffic. The system integrates an anomaly detection algorithm module, which can display the detected abnormal traffic data in real time and also visually present other information in the received network traffic data, thus improving the practical application value of the system.

2. Overview of relevant concepts

2.1. Machine Learning

In recent years, the concept of machine learning has been proposed, and many fields have developed and designed machine learning algorithms. Make the machine thinking more human learning thinking, so as to optimize its own functions and make the machine closer to the learning system. The scholar T. Michel has analyzed the machine learning thinking in his research. It relies on computer programs to study the experience similar to human learning, complete a series of tasks T, and then evaluate and measure the learning effect. Through the continuous learning process, it

can strengthen its own functions and set the performance e to evaluate the machine itself [4]. From the analysis of real life, if people are engaged in chess games, then t represents the result of winning chess; P represents chessboard evaluation analysis; E means experience or chess score.

In the process of machine learning, a large number of data will be processed, and the calculation process is cumbersome, but it can be used in many practical operations. The machine itself has the learning ability to repeatedly build models for the system. With the mature development of computing technology, many machine learning can also be put into the equipment.

There are many classifications of machine learning methods in different practice fields. There are two common types of learning: supervised learning and unsupervised learning. The difference between the two is: learning whether the input characteristics and output labels correspond; Whether the input data learning process corresponds. If the computer learning content and learning method are told, the process is called supervised learning; If the computer is not clear about the learning content, combined with the input characteristics, the computer can find the learning rules and complete the learning content. This process is unsupervised learning. However, from its learning results, supervised learning results are better than unsupervised learning results, but this does not mean that unsupervised learning is not as good as supervised learning process. Therefore, it is still necessary to select appropriate learning methods according to specific practical activities. For example, in the intuitive learning activities, if the training samples are available, the supervised learning method can be directly selected. If there is no shaped sample in daily learning and the cost is high, it is suggested to choose unsupervised learning for such a situation [5].

Supervised learning process can also be divided into two ways: regression problem and classification problem. The classification method is to implement classification management on some rules according to input characteristics, which is a learning method of discrete data [6]. Fundamentally, machine learning presents computer programs. The important task for this program is learning. Through experience learning to optimize the system, the actual fault can be solved. The machine learning framework is roughly divided into four parts: learning, performance evaluation, knowledge base and execution system [7].

The regression algorithm of machine learning is to use the regression analysis method formed by mathematical statistics to predict and analyze future data, find the quantitative relationship between different variables, and then refer to other system data to build a model to find the development law of future data [8]. Such an algorithm is widely used, and the calculation process has been very mature. It has become a common algorithm for supervised learning. The principle of regression learning is to implement the model hypothesis on the data, then analyze the output data and the actual data, find the difference between them and adjust the model parameters to ensure that the data error can be controlled within a reasonable range [9].

Since the development of machine learning algorithms, various mature algorithms have been applied to various scenarios. In order to find the algorithm suitable for this paper, this paper analyzes the results according to the classification of machine learning. Machine learning can be divided into different categories according to different dimensions, such as classification according to learning strategies, classification according to learning methods, classification according to learning methods, classification according to data forms, and classification according to learning purposes. Different classifications have different characteristics. This paper will refer to the classification of learning objectives, and select the applicable algorithm according to the final learning objectives and the results achieved [10-11].

Concept, rule, function and category learning are four categories of machine learning objectives. The research content of this paper is server fault monitoring. The goal is to realize fault prediction

and early warning and real-time fault alarm and analysis based on data analysis and learning. Therefore, algorithms that can learn data and predict and algorithms that can discover laws based on historical data and existing data are required [12].

2.2. Machine Learning Algorithm

Although logistic regression is called regression, it is actually a classification algorithm, which is often used for the task of classifying samples. For a given dichotomous task dataset:

$$D = \{(x_1, y_1), (x_2, y_2), ..., (x_N, y_N), x_i \in R^n, y_i \in 0,1), i = 1,2,..., N\} \quad (1)$$

Because the value range of linear regression is continuous, it cannot be used to fit the discrete variable Yi, but it can be used to fit the conditional probability. Therefore, the logistic function is used to map the output range to (0,1) to obtain the probability of each category in the case of two categories, so there is a probability prediction function constructed by the formula [13].

$$g(x) = \frac{1}{1+e^{-x}} \quad (2)$$

$$h_w(x) = g(w^T x) = \frac{1}{1+e^{-w^T x}} \quad (3)$$

Such a model is called logistic regression or log probability regression. In the process of model construction, the meaning of regression is to use linear regression to fit and approximate a decision boundary. The meaning of logic is to use logistic function to map the discrete value to the probability between 0 and 1 on the basis of linear regression, i.e. [14]:

$$P(y = 1 | x, w) = h_w(x) \quad (4)$$

$$P(y = 0 | x, w) = 1 - h_w(x) \quad (5)$$

2.3. Distributed Stream Processing (Kafka)

In big data processing, Apache Kafka is usually used as an advanced message queue, which can realize real-time streaming collection and processing of data messages. Its high throughput and strong real-time characteristics make it have a wide range of application scenarios. Apache Kafka is a data forwarding system based on the Hadoop ecosystem and capable of cluster deployment. Kafka has three advantages: (1) strong real-time. Kafka can receive and send different topics message data at the same time (topic can be understood as the classification of data), without causing data blocking, loss and redundancy; (2) Disk based data storage. The message queue and the consumer group are associated through topic. Since the message queue obtained by Kafka cluster is stored on the disk of the server node, the disk based storage method can be permanently saved. Even if the data sent by the consumer group of the Kafka cluster is suddenly interrupted, the data can continue to be processed from the interrupted position; (3) Strong expandability. The number of brokers that send or receive data can be expanded horizontally according to the size of the data. There will be multiple brokers in the Kafka cluster, so large-scale data can be easily processed [15-16].

2.4. Algorithm Model Fusion Theory

Because there are many attribute value categories of network traffic data, and there are certain associations between attributes, PCA algorithm can extract key features from many records and reduce the dimension of attribute features. Therefore, PCA algorithm can meet the processing requirements of feature extraction and dimension reduction of network traffic data. In addition, abnormal network traffic and normal network traffic will show differences in global characteristics and local density characteristics, thus forming outlier data. Iforest algorithm and lof algorithm are currently popular global and local anomaly detection algorithms, which are suitable for outlier detection of abnormal network traffic data. At the same time, anomaly detection needs to be verified again manually in the actual application environment, and the threshold value is set according to the actual situation, so that the actual application can meet the double requirements of accuracy and efficiency. LR algorithm can map the discrimination result to the probability estimation of anomaly detection, which is convenient for the threshold adjustment of detection results and the selection and ranking of feature importance [17-18].

3. System design and Implementation

3.1. Experimental Environment

The overall experimental test was completed on 5 servers, and the basic environment configuration of each server was the same. A CDH version of Hadoop cluster was also built. Table 3.3 lists the installation version information of the cluster platform, and table 1 lists the hardware parameters of a single server.

Table 1. Cluster platform component version

Software	Version type
Operating system	Centos6.5
JDK version	1.8.0
Scala version	2.10.6
Hadoop version	2.6.0
Spark version	1.6.0
Kylin version	2.2.0
Kafka version	3.0.0
Number of CPU cores / core	8
CPU main frequency / GHz	2.10
Memory / GB	Master node 32GB, slave node 16GB
Hard disk capacity / GB	Master node 600gb slave node 300gb

The system adopts a development architecture that separates the front end from the back end. The back end interacts with the database, uses SQL statements to analyze, obtains data resources, and writes logic processing codes to provide JSON interfaces for the front end; The front-end visualization is connected with the back-end through JSON format data to realize data display and function page Jump; Use HTTP / HTTPS protocol to display the front-end page in the browser. This architecture mode solves the problem of difficult business communication between front-end and back-end engineers, making everyone's business more simple and efficient, and later maintenance more convenient and simple.

3.2. System Realization

The whole system realizes the basic functions through the cooperation and association of data access layer, business processing layer and page application layer. The system in this paper is a front-end and back-end separation system, which is divided into two parts: the front-end page of the system and the back-end logic of the system. These two parts communicate with each other by configuring address routing and forwarding. The front-end page of the system is mainly used to visualize the system functions, and the back-end logic of the system is responsible for configuring the system security and completing the interaction with the database. The front-end page of the system mainly includes the visualization interface of traffic data, the data analysis interface, the result display interface of traffic anomaly detection and the user information management interface. The back-end logic of the system mainly includes user login verification module, user page authority verification module, personal information management module, exception detection algorithm module, and traffic data storage and query display module. Each web page interface has its corresponding business logic module. After processing the data in the database, the logic module uploads it to the page application layer, rendering and typesetting through visualization technology, and finally realizes the interactive operation between the front end, the back end and the database.

4. System Test and Analysis

4.1. User Management and Personal Information Center Performance Test

The user management interface is a part of the system management module. Only the super management user can enter this page. This page is responsible for the management and maintenance of the system user information. The administrator can query the user through the search area. At the same time, the user can be managed by deleting or disabling / enabling buttons to ensure whether the user has the right to log in to the system. As shown in Figure 1, after 200 performance tests on user management and personal information center, there were 7 times of bugs, which have been repaired by technical means. The normal rate reached 96.5%, which can meet the needs of normal use.

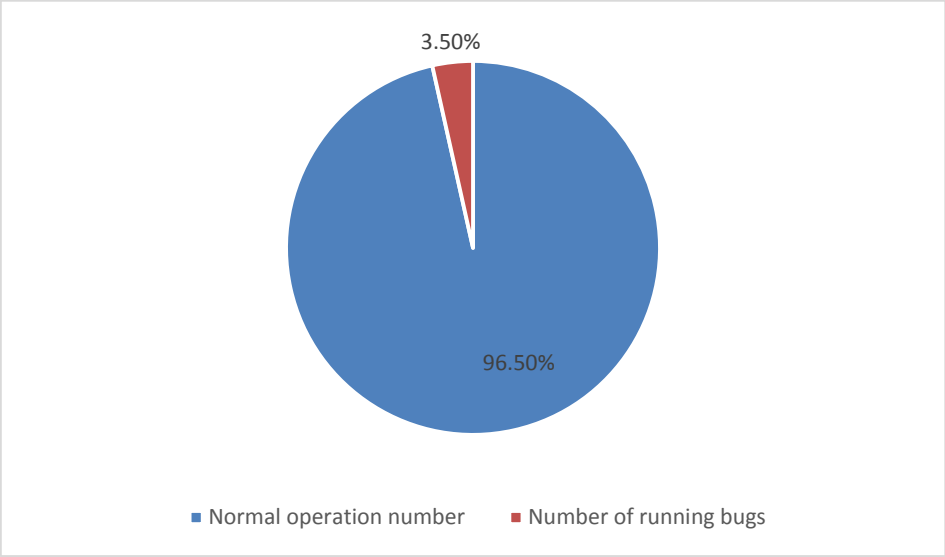

Figure 1. User management interface bug test

4.2. Flow Data Analysis

As shown in Fig. 2, Fig. 3 and Fig. 4, the analysis and display of traffic data are realized by means of the ecarts visualization tool. The data analysis page is used to analyze and count the existing network traffic, and can deeply understand the content of the data. The whole page is divided into three parts. The first part is to make statistics on all the stored traffic data, show the data types of network traffic through the pie chart, and understand the distribution of various types of data; The second part is to compare the data sets of traffic data.; The third part is to analyze the abnormal network traffic data, extract the key characteristic fields of the abnormal traffic data, and make statistics on the data content of these characteristics, and show the support of the abnormal data in the form of horizontal histogram.

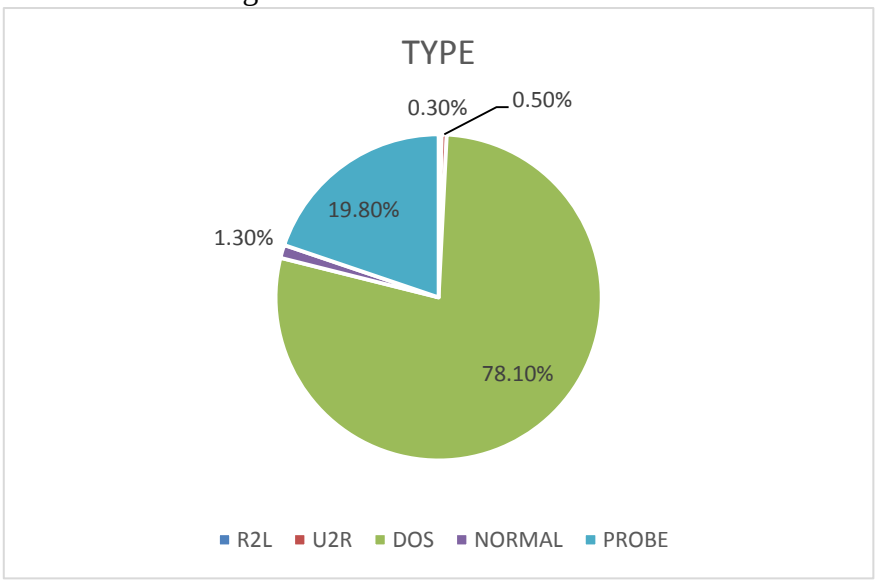

Figure 2. Comparison of network traffic data types

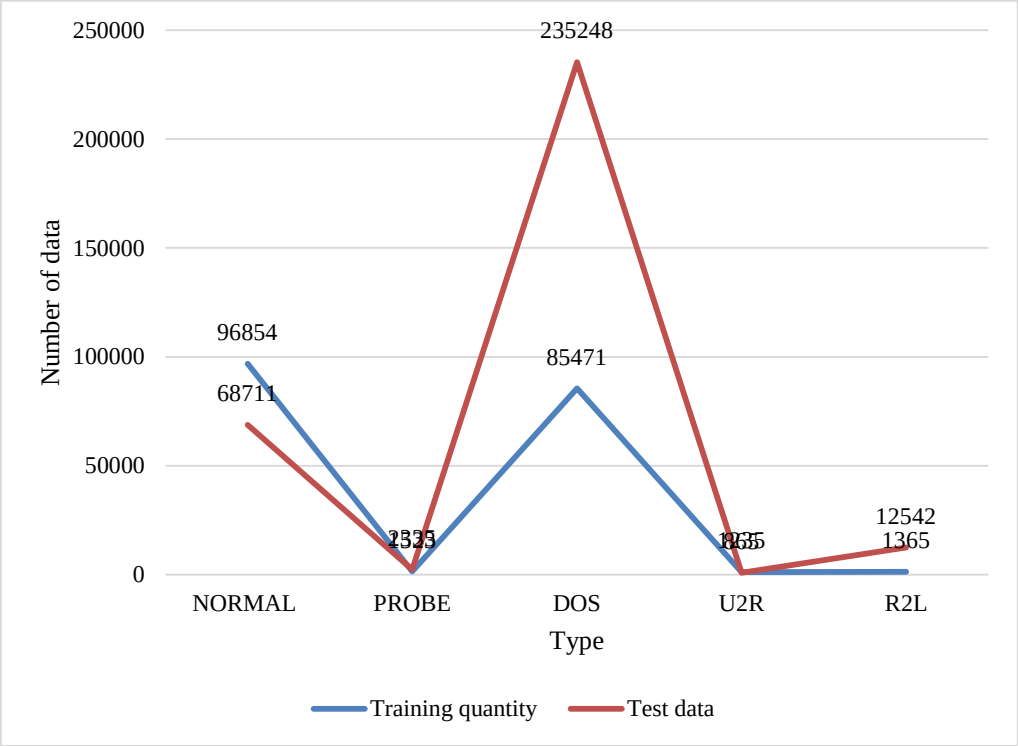

Figure 3. Comparison of experimental data sets

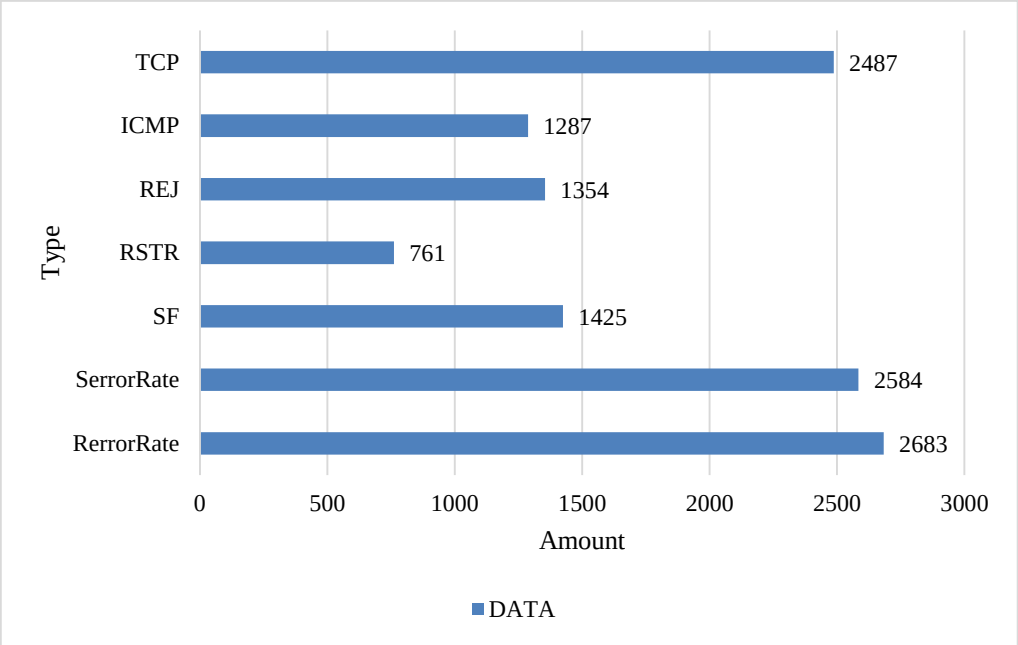

Figure 4. Abnormal traffic data field

5. Conclusion

Under the background of the big data era, how to detect anomalies efficiently and accurately

from the massive network traffic and ensure network security has become an important topic that scholars continue to study. Based on the network traffic data, this paper studies and implements the distributed anomaly detection and analysis system. Firstly, the purpose and significance of the research are analyzed, and the current research status at home and abroad is introduced in detail. Then, the relevant anomaly detection algorithms and the applied big data technology are introduced, and the feasibility and necessity of the research and implementation of this method are pointed out; Secondly, according to the characteristics of network traffic, the implementation of multi model fusion distributed anomaly detection method is introduced in detail, and the performance of this method is verified and analyzed by using network data; Finally, in order to display and detect the network traffic data efficiently, this paper develops a set of network traffic anomaly detection and analysis system, and integrates the anomaly detection model into the system to further verify the model. After verification, the system can meet the requirements mentioned in the article, and basically everything is normal in the network data processing.

Funding

This article is not supported by any foundation.

Data Availability

Data sharing is not applicable to this article as no new data were created or analysed in this study.

Conflict of Interest

The author states that this article has no conflict of interest.

References

[1] Ali R, Lee S, Chung T C. Accurate multi-criteria decision making methodology for recommending machine learning algorithm. Expert Systems with Applications, 2017, 71:257-278. https://doi.org/10.1016/j.eswa.2016.11.034

[2] Rindal O, Seeberg T M, Tjnns J, et al. Automatic Classification of Sub-Techniques in Classical Cross-Country Skiing Using a Machine Learning Algorithm on Micro-Sensor Data. Sensors, 2018, 18(1):75. https://doi.org/10.3390/s18010075

[3] Boland M R, Polubriaginof F, Tatonetti N P. Development of A Machine Learning Algorithm to Classify Drugs Of Unknown Fetal Effect. Scientific Reports, 2017, 7(1):12839.

[4] Yu J, Hall J J, Yu T M. Automatic Equatorial GPS Amplitude Scintillation Detection Using a Machine Learning Algorithm. IEEE Transactions on Aerospace & Electronic Systems, 2017, PP(1):1-1.

[5] Park B J, Kang M S, Lee M, et al. A Study on Efficient Memory Management Using Machine Learning Algorithm. International journal of advanced smart convergence, 2017, 6(1):39-43.

[6] Comertpay B, Gov E. Identification of molecular signatures and pathways of obese breast cancer gene expression data by a machine learning algorithm. Journal of Translational Genetics and Genomics, 2021, 6(1):84-94. https://doi.org/10.20517/jtgg.2021.44

[7] Teluguntla P, Thenkabail P S, Oliphant A, et al. A 30-m Landsat-derived Cropland Extent Product of Australia and China using Random Forest Machine Learning Algorithm on Google

Earth Engine Cloud Computing Platform. ISPRS Journal of Photogrammetry and Remote Sensing, 2018, 144(OCT.):325-340.

[8] Gao J, Nuyttens D, Lootens P, et al. Recognising weeds in a maize crop using a random forest machine-learning algorithm and near-infrared snapshot mosaic hyperspectral imagery. Biosystems Engineering, 2018, 170:39-50.

[9] Cho M J, Hallac R R, Effendi M, et al. Comparison of an unsupervised machine learning algorithm and surgeon diagnosis in the clinical differentiation of metopic craniosynostosis and benign metopic ridge. scientific reports, 2018, 8(1):6312. https://doi.org/10.1038/s41598-018-24756-7

[10] Alhudhaif A, Cmert Z, Polat K. Otitis media detection using tympanic membrane images with a novel multi-class machine learning algorithm. PeerJ Computer Science, 2021, 7(7):e405. https://doi.org/10.7717/peerj-cs.405

[11] Richter C, Petushek E, Grindem H, et al. Cross-validation of a machine learning algorithm that determines anterior cruciate ligament rehabilitation status and evaluation of its ability to predict future injury. Sports biomechanics, 2021:1-11.

[12] Allen A, Ektefaie Y, Garikipati A, et al. Sa102 A Machine Learning Algorithm To Predict Gastrointestinal Bleeding Requiring Intervention. Gastroenterology, 2021, 160(6):S-422.

[13] Turnquist M, Lewis P, Lau T, et al. Adaptive Focused Ion Beam Milling through Machine Learning Algorithm Integration. Microscopy and Microanalysis, 2021, 27(S1):1624-1624. https://doi.org/10.1017/S1431927621005985

[14] Zoss B M, Mateo D, Kuan Y K, et al. Distributed system of autonomous buoys for scalable deployment and monitoring of large waterbodies. Autonomous Robots, 2018(11):1669-1689.

[15] Hillah L M, Maesano A P, Rosa F D, et al. Automation and intelligent scheduling of distributed system functional testing. International Journal on Software Tools for Technology Transfer, 2017, 19(3):281-308. https://doi.org/10.1007/s10009-016-0440-3

[16] Siami M, Skaf J. Structural Analysis and Optimal Design of Distributed System Throttlers. IEEE Transactions on Automatic Control, 2017, PP(99):1-1.

[17] Srivastava A K, Kumar S. Dynamic Reconfiguration of robot software component in real time distributed system using clustering techniques. Procedia Computer Science, 2018, 125:754-761. https://doi.org/10.1016/j.procs.2017.12.097

[18] Nesterov R A, Lomazova I A. Using Interface Patterns for Compositional Discovery of Distributed System Models. Proceedings of the Institute for System Programming of RAS, 2017, 29(4):21-38. https://doi.org/10.15514/ISPRAS-2017-29(4)-2