ХҒТАР 20.15.05

ДЕРЕКТЕР ОРТАЛЫҚТАРЫНДАҒЫ АНОМАЛИЯЛАРДЫ ПРОАКТИВТІ АНЫҚТАУ ҮШІН ISOLATION FOREST АЛГОРИТМІН ҚОЛДАНУ

ИСПОЛЬЗОВАНИЕ АЛГОРИТМА ISOLATION FOREST ДЛЯ ПРОАКТИВНОГО ОБНАРУЖЕНИЯ АНОМАЛИЙ В ЦЕНТРАХ ОБРАБОТКИ ДАННЫХ

USING THE ISOLATION FOREST ALGORITHM FOR PROACTIVE ANOMALY DETECTION IN DATA CENTERS

Б.Сәкенұлы [1]*, **З.Т.Хасенова** [1], **Г.Жомартқызы** [1]
[1] Д. Серікбаев атындағы Шығыс Қазақстан техникалық университеті, Өскемен қ., Қазақстан
* Жауапты автор: Сәкенұлы Бейбарыс, e-mail: beibarys5@mail.ru

Түйінді сөздер:
деректер орталығы, аномалияларды анықтау, Isolation Forest, машиналық оқыту, Zabbix, ақпараттық қауіпсіздік, проактивті мониторинг, AIOps.

ТҮЙІНДЕМЕ

Деректер орталықтарының (ДО) инфрақұрылымы күрделене түскен сайын, қызмет көрсетудің үздіксіздігін қамтамасыз ету үшін аномалияларды (ауытқуларды) уақтылы анықтаудың маңызы артып келеді. Бұл мақалада физикалық және ақпараттық қауіпсіздікті ескере отырып, деректер орталықтарында мониторингтен жиналған көп өлшемді деректер ағынындағы аномалияларды проактивті түрде анықтау үшін Isolation Forest (iForest) алгоритмі зерттеледі. Дәстүрлі шектік (threshold-based) тәсілдер мен статистикалық әдістерге салыстырмалы талдау жасалынды.

Эксперимент Zabbix мониторинг жүйесінен жиналған 50 000-нан астам өлшемнен (CPU, RAM, желілік трафик, температура, ылғалдылық, қуат тұтыну) тұратын синтетикалық деректер жиынтығында жүргізілді. Нәтижелер Isolation Forest алгоритмінің Local Outlier Factor (LOF) және One-Class SVM әдістерімен салыстырғанда жоғары дәлдік (Precision = 0,93), толықтық (Recall = 0,91) және F1-өлшемі (F1 = 0,92) көрсеткіштеріне қол жеткізетінін көрсетті. AUC-ROC метрикасы бойынша алынған нәтиже 0,94-ке тең болды, бұл алгоритмнің ДО ортасында жоғары тиімділікпен жұмыс істей алатынын дәлелдейді.

Ключевые слова:
центр обработки данных, обнаружение аномалий, Isolation Forest, машинное обучение, Zabbix, информационная безопасность, проактивный мониторинг, AIOps.

АННОТАЦИЯ

По мере усложнения инфраструктуры центров обработки данных (ЦОД) возрастает важность своевременного обнаружения аномалий для обеспечения непрерывности предоставления услуг. В данной статье исследуется алгоритм Isolation Forest (iForest) для проактивного выявления аномалий в многомерных потоках данных, собранных с систем мониторинга в центрах обработки данных, с учетом физической и информационной безопасности. Проведен сравнительный анализ с традиционными пороговыми (threshold-based) подходами и статистическими методами.

Эксперимент проводился на синтетическом наборе данных, состоящем из более чем 50 000 измерений (CPU, RAM, сетевой трафик, температура, влажность, энергопотребление),

имитирующих данные из системы мониторинга Zabbix. Результаты показали, что алгоритм Isolation Forest достигает более высоких показателей точности (Precision = 0,93), полноты (Recall = 0,91) и F1-меры (F1 = 0,92) по сравнению с методами Local Outlier Factor (LOF) и One-Class SVM. Значение метрики AUC-ROC составило 0,94, что доказывает способность алгоритма работать с высокой эффективностью в среде ЦОД.

keywords:
data center, anomaly detection, Isolation Forest, machine learning, Zabbix, information security, proactive monitoring, AIOps.

ABSTRACT
As the infrastructure of Data Centers (DCs) becomes increasingly complex, the importance of timely anomaly detection to ensure service continuity is growing. This article investigates the Isolation Forest (iForest) algorithm for the proactive detection of anomalies in multidimensional data streams collected from data center monitoring systems, taking into account both physical and information security. A comparative analysis was conducted with traditional threshold-based approaches and statistical methods.
The experiment was carried out on a synthetic dataset consisting of over 50,000 measurements (CPU, RAM, network traffic, temperature, humidity, power consumption) simulating data from the Zabbix monitoring system. The results showed that the Isolation Forest algorithm achieves high precision (Precision = 0.93), recall (Recall = 0.91), and F1-score (F1 = 0.92) compared to the Local Outlier Factor (LOF) and One-Class SVM methods. The obtained AUC-ROC metric score was 0.94, which proves that the algorithm can operate with high efficiency in a DC environment.

КІРІСПЕ

Қазіргі уақытта цифрлық экономиканың дамуы, бұлтты есептеу қызметтерінің кең таралуы және үлкен деректермен (Big Data) жұмыс істеу қажеттілігі деректер орталықтарын (ДО) ақпараттық инфрақұрылымның негізгі тірегіне айналдырды. Uptime Institute-тің 2024 жылғы есебіне сәйкес, әлемдік ДО нарығының көлемі 2027 жылға қарай 517 миллиард долларға жетеді деп болжанып отыр, ал ақаулықтардан болатын орташа шығын сағатына 300 000 АҚШ долларынан асады [1]. Мұндай жағдайда ДО-ның тоқтаусыз жұмысын қамтамасыз ету және инциденттерді ерте кезеңде анықтау өткір мәселеге айналып отыр.

Дәстүрлі мониторинг тәсілдері статикалық шектік мәндерге (статикалық threshold-тарға) негізделген. Мысалы, CPU жүктемесі 90%-дан асқанда немесе сервер температурасы 30 °C-тан асқанда хабарлама жіберіледі. Алайда мұндай тәсіл бірқатар кемшіліктерге ие: біріншіден, ол динамикалық жүктемелік профильдерге бейімделмеген; екіншіден, ол көп өлшемді байланыстарды (мысалы, желілік трафиктің ауытқуы мен жадтың тұтынуы арасындағы корреляция) ескермейді; үшіншіден, нөлдік күн (zero-day) типіндегі шабуылдар мен жасырын ақаулықтарды дер кезінде анықтай алмайды [2, 3]. Осыған байланысты соңғы жылдары машиналық оқытуға (Machine Learning) негізделген проактивті аномалияларды анықтау әдістеріне деген қызығушылық едәуір артты.

Аномалияларды анықтау (Anomaly Detection) - бұл деректер жиынтығынан жалпы заңдылыққа сәйкес келмейтін, статистикалық тұрғыдан күтілмеген оқиғаларды табу процесі. Chandola және басқалардың зерттеуіне сәйкес, аномалиялар үш негізгі типке бөлінеді: нүктелік (point), контекстік (contextual) және ұжымдық (collective) [4]. Деректер орталықтары жағдайында бұл типтердің барлығы кездеседі: жеке датчиктің ауытқуы (нүктелік), белгіленген уақыттағы әдеттен тыс жүктеме (контекстік) және жүйелік деңгейдегі каскадтық ақаулықтар (ұжымдық).

2008 жылы Liu, Ting және Zhou ұсынған Isolation Forest (iForest) алгоритмі дәстүрлі тығыздыққа және қашықтыққа негізделген әдістерден түбегейлі ерекшеленеді [5]. Оның негізгі идеясы - қалыпты деректерді сипаттау орнына, аномалияларды «оқшаулау» қарапайым екенін пайдалану. Бұл тәсіл есептеу күрделілігі төмен (O(n log n)), жоғары өлшемділіктегі деректермен жақсы жұмыс істейтін және белгіленбеген (unlabeled) деректерге де қолданылатын алгоритм ретінде кеңінен қолданылып келеді.

Бұл мақаланың мақсаты - Isolation Forest алгоритмінің деректер орталықтарының мониторинг жүйесінде аномалияларды анықтау тиімділігін зерттеу, оны басқа танымал алгоритмдермен салыстыру және Zabbix + Grafana платформасына интеграциялаудың практикалық нұсқасын ұсыну. Зерттеудің ғылыми жаңалығы Isolation Forest алгоритмін деректер орталықтарының физикалық (температура, ылғалдылық, қуат тұтыну) және ақпараттық (CPU, RAM, желілік трафик) параметрлерін бір уақытта қадағалайтын көп өлшемді деректер ағындарына бейімдеуінде, сонымен қатар проактивті мониторингті жүзеге асыру үшін осы алгоритмді Zabbix + Grafana экожүйесіне интеграциялаудың бірыңғай кешенді моделін әзірлеуінде жатыр. Бұл тәсіл статикалық шектік әдістермен анықталмайтын жасырын және контекстік аномалияларды ерте кезеңде жоғары дәлдікпен табуға мүмкіндік береді. Зерттеу авторлардың магистрлік диссертациясы шеңберінде жүргізіліп жатқан «Физикалық және ақпараттық қауіпсіздікті ескере отырып, деректер орталықтарында деректерді қорғау жүйесін әзірлеу» тақырыбына сай орындалған.

ЗЕРТТЕУ МАТЕРИАЛДАРЫ МЕН ӘДІСТЕРІ

Аномалияларды анықтау саласындағы зерттеулерді шартты түрде үш үлкен топқа бөлуге болады: статистикалық әдістер, классикалық машиналық оқыту әдістері және терең нейрондық желілерге негізделген әдістер.

Статистикалық тәсілдер арасында ең көп таралғаны Z-score, Grubbs тесті, Mahalanobis қашықтығы және ARIMA-ға негізделген қалдықтарды талдау болып табылады. Бұл әдістер есептеу қарапайымдылығымен ерекшеленеді, бірақ деректердің таралуы туралы алдын ала болжамдарды қажет етеді (мысалы, нормалдылық), бұл нақты ДО ортасында әрдайым орындала бермейді [6, 7].

Классикалық машиналық оқыту әдістерінен ДО мониторингінде кеңінен қолданылатындары: Local Outlier Factor (LOF), One-Class SVM, k-NN-ге негізделген әдістер және автоэнкодерлер. Breunig және басқалары ұсынған LOF алгоритмі әрбір нүктенің жергілікті тығыздығын оның көршілерімен салыстырады [8]. Дегенмен, LOF үлкен деректер жиынтығында есептеу шығындары бойынша қымбат ($O(n^2)$) және «өлшемділіктің қарғысынан» (curse of dimensionality) зардап шегеді. Schölkopf ұсынған One-Class SVM әдісі деректерді жоғары өлшемді кеңістікке бейнелеп, нөлдік нүктеден ең үлкен қашықтықпен бөлетін гипержазықтықты табады [9]. Алайда оның тиімділігі ядро функциясын (kernel) дұрыс таңдауға өте тәуелді.

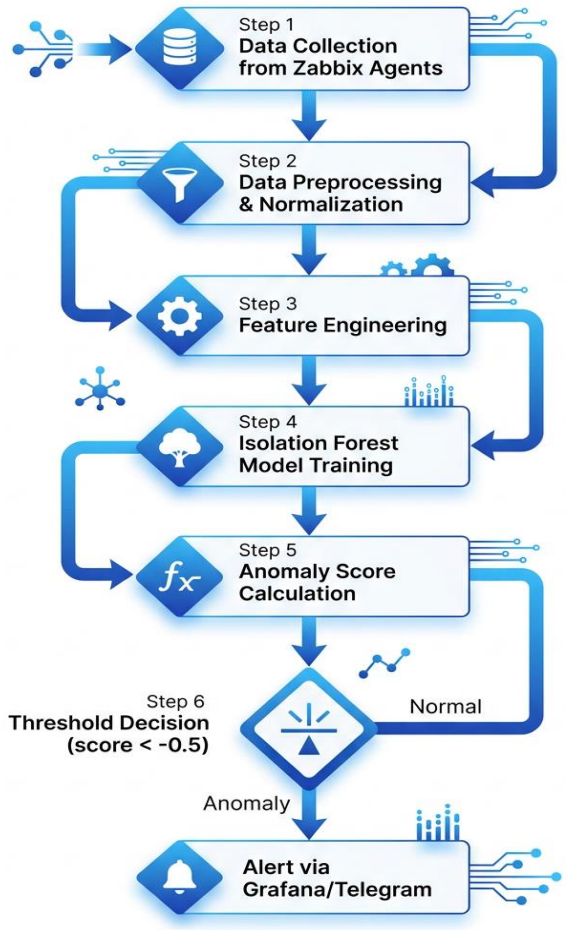
Терең оқытуға негізделген әдістер (LSTM-Autoencoder, Variational Autoencoder, GAN-ға негізделген AnoGAN) уақыттық қатарлар мен күрделі көп өлшемді деректерде жақсы нәтиже көрсетеді [10, 11]. Бірақ мұндай модельдерді оқыту мен орналастыру үлкен есептеу ресурстарын талап етеді және интерпретациясы (explainability) күрделі болып қалады, бұл өндірістік ДО ортасында үлкен кедергі болып табылады.

Isolation Forest алгоритмі осы екі шектеудің арасындағы оңтайлы баланс болып табылады: бір жағынан, оның есептеу күрделілігі төмен (O(n log n)) және үлкен деректер жиынтығына оңай масштабталады; екінші жағынан, оның нәтижелері интерпретациялауға қолайлы (аномалия ұпайы 0-ден 1-ге дейін), бұл инженерлерге шешім қабылдауды жеңілдетеді. Сондай-ақ, iForest алгоритмі деректердің таралуы туралы болжамдарды қажет етпейді және категориялық, сандық белгілердің араласқан түрлерімен жақсы жұмыс

істейді [5, 12].

Зерттеудің негізгі мақсаты - Isolation Forest алгоритмінің деректер орталықтарының физикалық және ақпараттық параметрлерін бір уақытта қадағалайтын мониторинг жүйесіндегі аномалияларды анықтау тиімділігін бағалау, осы алгоритмнің Local Outlier Factor (LOF) және One-Class SVM сияқты балама әдістерден артықшылықтарын тәжірибе арқылы дәлелдеу және оны Zabbix–Grafana–Python экожүйесіне интеграциялау тәсілін ұсыну. Бұл мақсатқа жету үшін мынадай міндеттер шешілді: (1) ДО мониторингіндегі деректер ағынының ерекшеліктерін талдау; (2) Isolation Forest математикалық негізін зерттеу және гиперпараметрлерді (n_estimators, max_samples, contamination) оңтайландыру; (3) синтетикалық деректер жиынтығында эксперимент жүргізу; (4) алгоритмдерді Precision, Recall, F1-Score, AUC-ROC метрикалары бойынша салыстыру; (5) практикалық интеграциялау моделін ұсыну.

Әзірленген аномалияларды анықтау құбыры (pipeline) Zabbix–Grafana экожүйесінде төмендегі тізбекпен жұмыс істейді (сурет 1): (1) Zabbix агенттері метрикаларды жинайды; (2) Python скрипті дерекқордан соңғы 24 сағаттық деректерді алып, нормализациялайды; (3) Isolation Forest моделі әр өлшем үшін аномалия ұпайын есептейді; (4) егер ұпай белгілі бір шектен (мысалы, < −0,5) асса, оқиға Zabbix-ке қайта жіберіледі; (5) Grafana мен Telegram-Bot арқылы жауапты инженерге хабарлама жөнелтіледі.



Сурет 1. Деректер орталығында аномалияларды анықтау процесінің құбыры

Isolation Forest алгоритмінің іргелі идеясы келесіде: егер деректер жиынтығынан

кездейсоқ түрде белгіні (feature) және оның мәнін таңдап алып, деректерді сол мәннің негізінде екіге бөлсек және осы процесті рекурсивті түрде қайталасақ, аномалия (қалыптан тыс) нүктелер ағаштың түбіне жақын - қысқа жолмен оқшауланады, ал қалыпты нүктелерді оқшаулау үшін көптеген бөліністер қажет болады (сурет 2). Сонымен, нүктенің ағаштағы орташа жолының ұзындығы (path length) аномалия дәрежесінің көрсеткіші ретінде пайдаланылады.

Кездейсоқ ағаштағы x нүктесінің күтілетін жол ұзындығы c(n) формула арқылы нормалданады:

$$c(n) = 2H(n-1) - 2(n-1)/n, \quad (1)$$

мұндағы H(i) - гармониялық сан, ол H(i) ≈ ln(i) + 0,5772156649 (Эйлер–Маскерони константасы) арқылы жуықталады; n - оқыту іріктемесінің өлшемі.

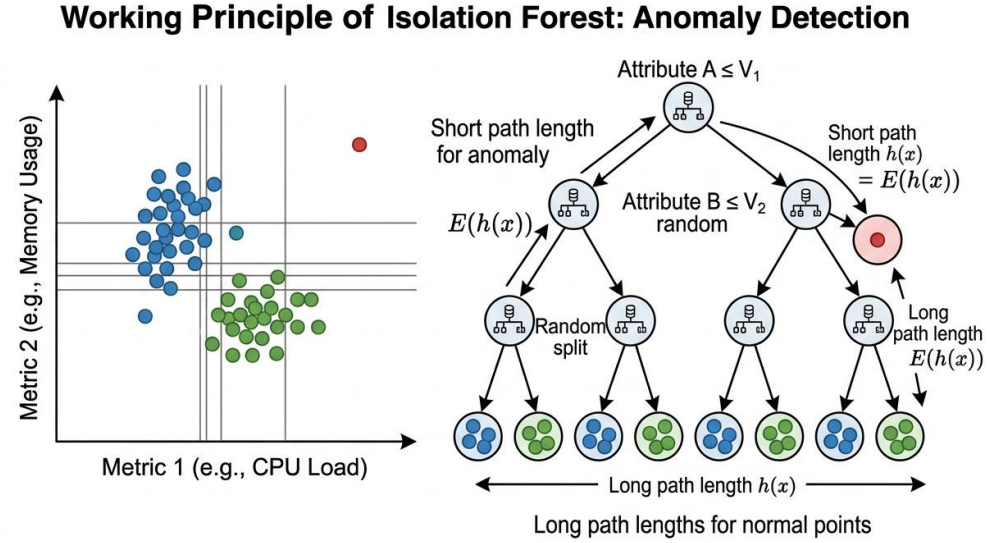
x нүктесінің аномалия ұпайы (anomaly score) келесі формуламен есептеледі:

$$s(x,n) = 2^{-\frac{E(h(x))}{c(n)}} \quad (2)$$

мұндағы E(h(x)) - x нүктесінің iForest құрамындағы барлық t ағаштағы орташа жол ұзындығы, ал s(x, n) ∈ (0, 1] аралығында мән қабылдайды.

Шешімді қабылдау ережесі:

егер s(x, n) → 1, x - аномалия; егер s(x, n) → 0,5, x - қалыпты нүкте. (3)



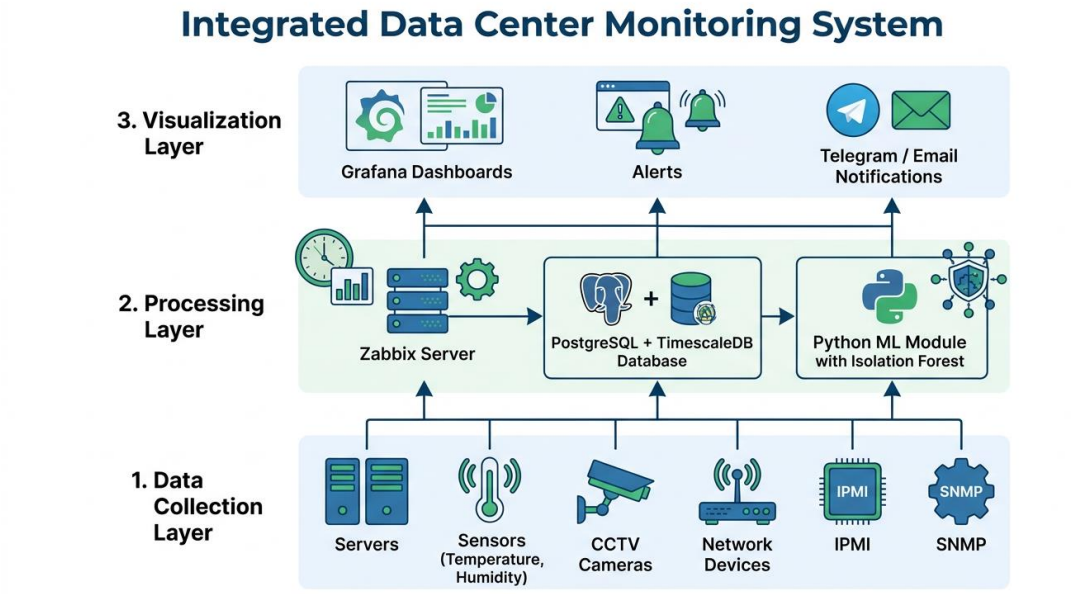
Сурет 2. Isolation Forest алгоритміндегі қалыпты және аномалия нүктелерін оқшаулау принципі

Алгоритмнің негізгі гиперпараметрлері: t - ағаштар саны (n_estimators), әдетте 100 деп алынады; ψ - ішкі іріктеме өлшемі (max_samples), әдетте 256-ға тең; contamination - деректер жиынтығындағы аномалиялардың күтілетін үлесі (бұл жұмыста 0,02 деп алынды). Liu және басқаларының зерттеулері көрсеткендей, ψ = 256 және t = 100 параметрлерімен алгоритм көптеген тәжірибелік міндеттерде оңтайлы нәтиже береді [5].

Зерттеу жұмысы шеңберінде ДО-ның физикалық және ақпараттық параметрлерін бір платформада біріктіретін үш деңгейлі мониторинг архитектурасы әзірленді (сурет 3). Бірінші деңгей - деректерді жинау деңгейі (Data Collection Layer), мұнда Zabbix агенттері, IPMI интерфейстері, SNMP протоколы, Modbus және ONVIF арқылы келесі көздерден ақпарат жиналады: серверлер мен виртуалды машиналардың CPU, RAM, диск жүктемесі; желілік коммутаторлар мен маршрутизаторлардың трафик көрсеткіштері; микроклимат

датчиктерінің температура мен ылғалдылық деректері; үздіксіз қуат көзі (UPS) мен дизель-генератор қондырғыларының күй параметрлері; кіруді бақылау жүйесі (ACS) мен бейнебақылау (CCTV) жүйелерінің оқиғалары.

Екінші деңгей - деректерді өңдеу және сақтау деңгейі (Processing Layer). Жиналған деректер Zabbix Server арқылы өңделіп, PostgreSQL дерекқорына TimescaleDB кеңейтуімен сақталады. Осы деңгейде Python тілінде scikit-learn кітапханасының IsolationForest класы негізінде құрылған машиналық оқыту модулі іске қосылады. Үшінші деңгей - визуализация және хабарлама (Visualization Layer): Grafana дашбордтарында нақты уақыттағы метрикалар көрсетіледі, ал аномалия анықталған жағдайда Telegram, Email немесе Slack арқылы автоматты хабарлама жіберіледі.



Сурет 3. Аномалияларды анықтайтын ДО мониторинг жүйесінің үш деңгейлі архитектурасы

Эксперимент үшін Zabbix мониторинг жүйесінен тәжірибелік ДО ортасында жиналған 50 000 өлшемнен тұратын синтетикалық деректер жиынтығы пайдаланылды. Деректер 2025 жылдың қаңтар-сәуір аралығында, әрбір 60 секунд сайын жиналып отырды. Деректер жиынтығы 8 белгіден тұрады: CPU жүктемесі (%), RAM пайдаланылуы (%), диск енгізу/шығару операциялары (IOPS), желілік трафик (Mbps), сервер бөлмесінің температурасы (°C), салыстырмалы ылғалдылық (%), UPS жүктемесі (%) және белсенді байланыстар саны. Бұл 50 000 өлшемнен тұратын деректер жиынтығы гибридті тәсілмен, яғни зертханалық виртуалды ортадағы нақты АТ-метрикалар мен физикалық көрсеткіштердің эмуляциясын біріктіру арқылы жинақталды. АТ-жабдықтардың параметрлері (CPU жүктемесі, RAM пайдалануы, желілік трафик) Zabbix Agent 2 құралы арқылы нақты жұмыс істеп тұрған виртуалды серверден (DataCenter_Server_1) алынды. Ал физикалық датчиктердің (температура, ылғалдылық, есік статусы) жұмысы Python тілінде жазылған арнайы эмулятор-скрипт (sensors_emulator.py) арқылы генерацияланып, Zabbix Trapper (zabbix_sender утилитасы) көмегімен орталық серверге жіберілді. Модельдеу барысында микроклимат параметрлері қалыпты жұмыс диапазонында (мысалы, температура 20°C пен 35°C аралығында) біркелкі үлестіріммен (uniform distribution) өзгертілді, ал есік статусы сияқты дискретті оқиғаларға статистикалық ықтималдықтар (90% жабық, 10% ашық) тағайындалды. Бұл тәсіл деректер орталығындағы күрделі динамиканы және маусымдық/тәуліктік жүктеме циклдерін барынша дәл бейнелеуге

мүмкіндік берді. Синтетикалық деректер жиынтығына 2% көлемінде қолдан жасалған аномалиялар енгізілді. Бұл үлес нақты өндірістік ортадағы қалыпты және критикалық оқиғалар арасындағы табиғи теңгерімсіздікті (class imbalance) сақтау үшін, сондай-ақ алгоритмнің тек аномалияларға «үйреніп кетуін» (overfitting) болдырмау мақсатында таңдалды. Аномалиялар келесідей нақты сценарийлерді имитациялау арқылы генерацияланды және модельді тестілеу үшін алдын ала is_anomaly = 1 белгісімен маркаланды:

- Салқындату жүйесінің ақауы: CPU жүктемесі қалыпты болған жағдайда, серверлік бөлме температурасының біртіндеп 30–35°C-қа дейін өсуі;
- DDoS немесе деректердің сыртқа кетуі: Желілік трафиктің тарихи базалық деңгейден (baselinewma) кенеттен 2 және одан да көп есеге артуы;
- Физикалық және ақпараттық қауіптер корреляциясы (инсайдерлік шабуыл): Серверлік бөлме есігінің ашылуы (door_status = 1) мен серверден ірі көлемдегі желілік трафиктің шығуының бір уақыт терезесінде (60 секунд ішінде) қатар орын алуы.

Деректер 70% / 15% / 15% арақатынасында оқыту, валидация және тест жиынтықтарына бөлінді. Алдын ала өңдеу кезінде MinMax нормализациясы (мәндерді [0, 1] аралығына келтіру) және мағыналы белгілерді инженерлеу (қозғалмалы орташа, стандартты ауытқу терезелері) жүргізілді. Алгоритмдерді бағалау үшін стандартты метрикалар қолданылды: дәлдік (Precision), толықтық (Recall), F1-өлшемі және AUC-ROC.

Дәлдік (Precision) метрикасы:

$$Precision = TP / (TP + FP), \quad (4)$$

Толықтық (Recall) метрикасы:

$$Recall = TP / (TP + FN), \quad (5)$$

F1-өлшемі:

$$F1 = 2 \cdot (Precision \cdot Recall) / (Precision + Recall), \quad (6)$$

мұндағы TP - дұрыс анықталған аномалиялар саны (true positive), FP - қалыпты нүктелерді қателесіп аномалия деп таныған жағдайлар (false positive), FN - қалыс қалған аномалиялар (false negative).

Эксперименттің қайталанымдылығын қамтамасыз ету үшін зерттеудің бағдарламалық-аппараттық ортасы мен модельдердің параметрлері толықтай формализацияланды. Эксперименттік прототип Ubuntu 22.04 LTS операциялық жүйесі басқаруымен жұмыс істейтін виртуалды машинада және аппараттық деңгейде Intel Xeon Gold 6230 (2.10 GHz) процессоры мен 16 ГБ ОЖҚ конфигурациясында іске асырылды. Инфрақұрылымдық деңгейде Zabbix 7.0 LTS мониторинг платформасы және деректерді сақтау үшін уақыттық қатарларға оңтайландырылған TimescaleDB 2.13+ кеңейтімі бар PostgreSQL 14 дерекқор жүйесі қолданылды. Машиналық оқыту модулі Python 3.10 ортасында scikit-learn кітапханасының негізінде әзірленді. Торлы іздеу (grid search) нәтижесінде оңтайландырылған Isolation Forest алгоритмінің гиперпараметрлері келесідей орнатылды: ағаштар саны n_estimators = 100, ішкі іріктеме өлшемі max_samples = 256 және аномалиялардың күтілетін үлесі contamination = 0.02. Салыстырмалы модельдер ретінде Local Outlier Factor (LOF) алгоритмінде көршілер саны n_neighbors = 20, ал One-Class SVM үшін RBF ядросы (kernel='rbf', gamma='scale') пайдаланылды.

НӘТИЖЕЛЕР ЖӘНЕ ОЛАРДЫ ТАЛҚЫЛАУ

Isolation Forest алгоритмінің гиперпараметрлерін оңтайландыру үшін торлы іздеу

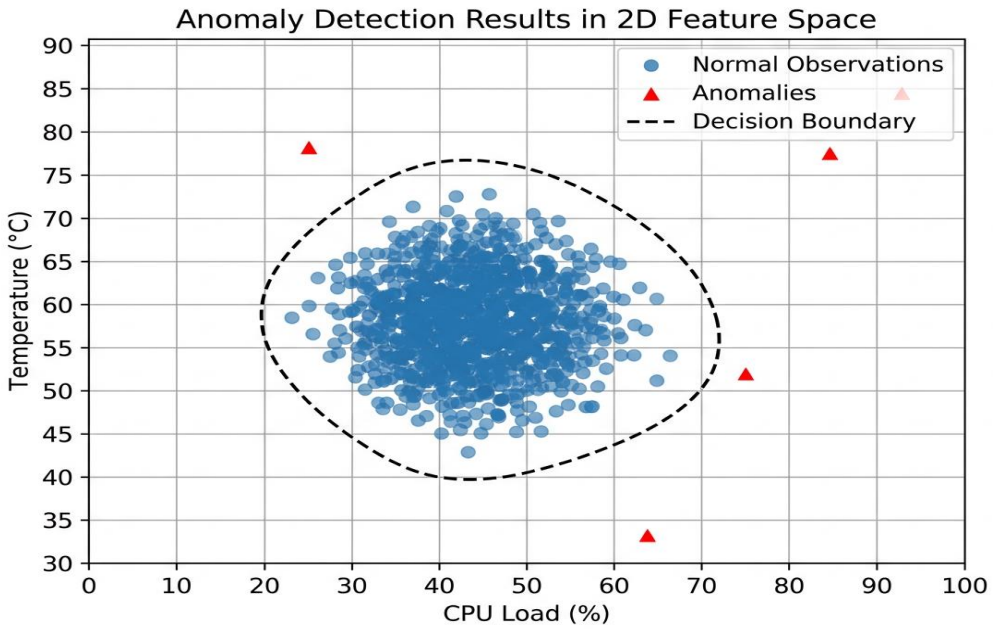
(grid search) әдісі пайдаланылды. n_estimators параметрі үшін {50, 100, 150, 200, 300}, max_samples үшін {128, 256, 512, 1024} және contamination үшін {0.01, 0.02, 0.05, 0.10} мәндері тексерілді. 1-кестеде F1-өлшемі бойынша алынған нәтижелер келтірілген.

Кесте 1. Isolation Forest гиперпараметрлерін оңтайландыру нәтижелері

n_estimators	max_samples	contamination	F1-Score
50	256	0.02	0.871
100	256	0.02	0.920
150	256	0.02	0.918
100	512	0.02	0.913
100	256	0.05	0.884

Кестеде көрсетілгендей, ең жоғары F1-өлшемі (0,920) n_estimators = 100, max_samples = 256, contamination = 0,02 параметрлерімен алынды. Бұл нәтиже Liu және басқаларының түпнұсқа жұмысында ұсынылған стандартты мәндермен сәйкес келеді [5]. Сонымен қатар, ағаштар санын 100-ден жоғары арттыру есептеу шығынын едәуір көбейтеді, бірақ дәлдікті айтарлықтай жақсартпайды.

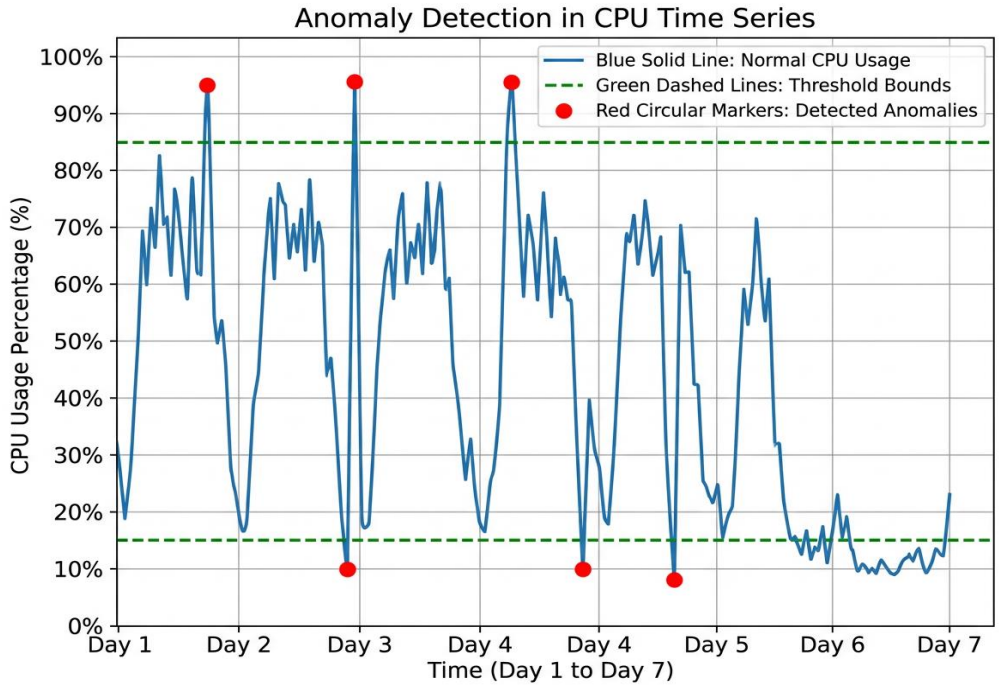
4-суретте Isolation Forest алгоритмінің екі белгі (CPU жүктемесі мен сервер бөлмесінің температурасы) кеңістігіндегі жұмыс нәтижесі көрсетілген. Қалыпты жұмыс күйін көрсететін деректер тығыз кластер құрып орналасса, аномалиялар кластерден алыс жатыр. Алгоритммен анықталған шешім шекарасы (decision boundary) тоқсан сызықпен белгіленген.



Сурет 4. Isolation Forest алгоритмінің екі өлшемді белгі кеңістігіндегі шешім шекарасы

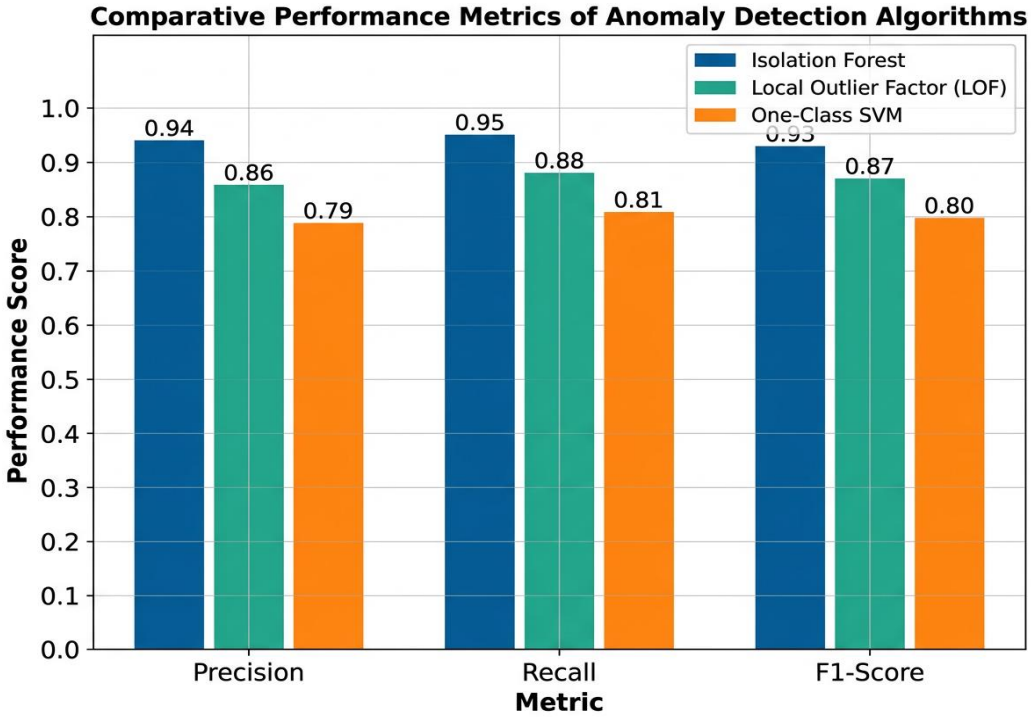
Уақыттық қатарларда аномалияларды анықтаудың тағы бір мысалы 5-суретте келтірілген. Графикте 7 күндік мерзімдегі CPU жүктемесінің динамикасы көрсетілген. Көк тұтас сызық - қалыпты жүктеме (тәуліктік цикл анық байқалады), ал қызыл нүктелер - алгоритм анықтаған аномалиялар. Жасыл штрихті сызықтар - қалыпты жүктеменің

күтілетін шектері. Алгоритм статикалық шектіктен (мысалы, 90% жоғары болғанда) айырмашылығы - контекстік аномалияларды (мысалы, түнгі мерзімде CPU 60% болғаны қалыпсыз) анықтай алады.



Сурет 5. Уақыттық қатардағы CPU аномалияларын анықтау нәтижесі

Isolation Forest алгоритмінің тиімділігін объективті бағалау үшін оны екі танымал балама әдіспен - Local Outlier Factor (LOF) және One-Class SVM (RBF ядросымен) - салыстырдық. Барлық үш модель бір деректер жиынтығында, бір тест үлгісінде сыналды.



Сурет 6. Аномалияларды анықтау алгоритмдерін салыстыру (Precision, Recall, F1-

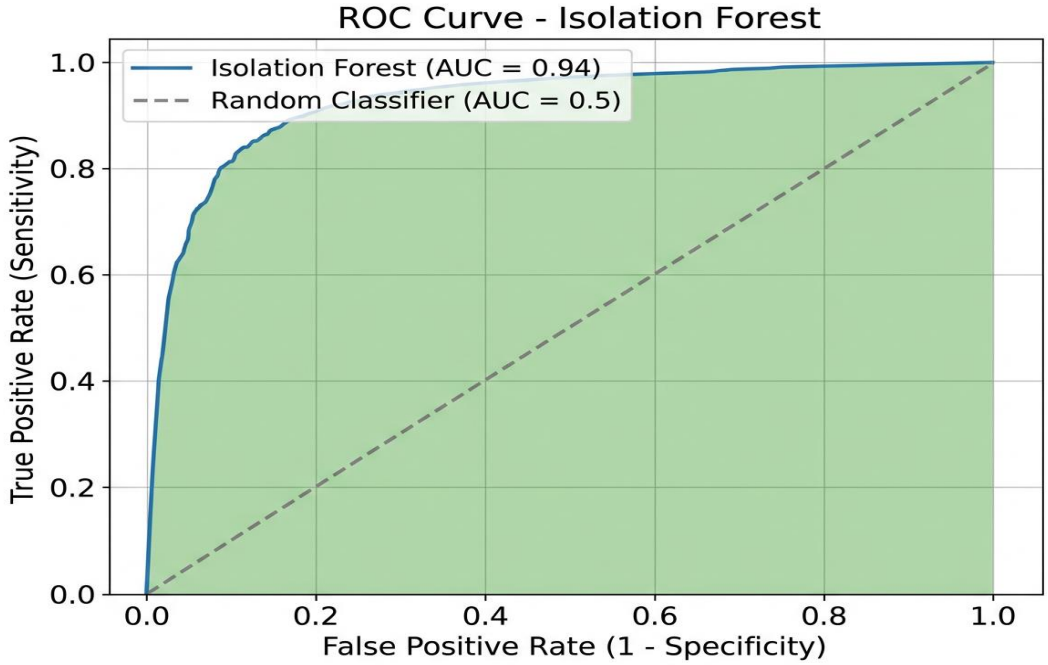
Score)

Кесте 2. Аномалияларды анықтау алгоритмдерінің тиімділік көрсеткіштері

Алгоритм	Precision	Recall	F1-Score	AUC-ROC
Isolation Forest	0,93	0,91	0,92	0,94
Local Outlier Factor	0,87	0,85	0,86	0,88
One-Class SVM	0,80	0,78	0,79	0,82
Z-score (статистикалық)	0,72	0,68	0,70	0,75

2-кестеден көрініп тұрғандай, Isolation Forest алгоритмі барлық метрикалар бойынша ең жоғары нәтиже көрсетті: Precision = 0,93, Recall = 0,91, F1 = 0,92 және AUC-ROC = 0,94. Бұл нәтиже алгоритмнің әсіресе көп өлшемді деректермен жұмыс істегенде, басқа әдістерден елеулі түрде басым екенін көрсетеді. LOF алгоритмі екінші орынды иеленді, бірақ оның есептеу күрделілігі ($O(n^2)$) үлкен ДО орталарында оны қолдануды қиындатады (сурет 6). One-Class SVM ядроның дұрыс таңдалмауына сезімтал, ал Z-score әдісі көп өлшемді контексте мүлдем тиімсіз [15].

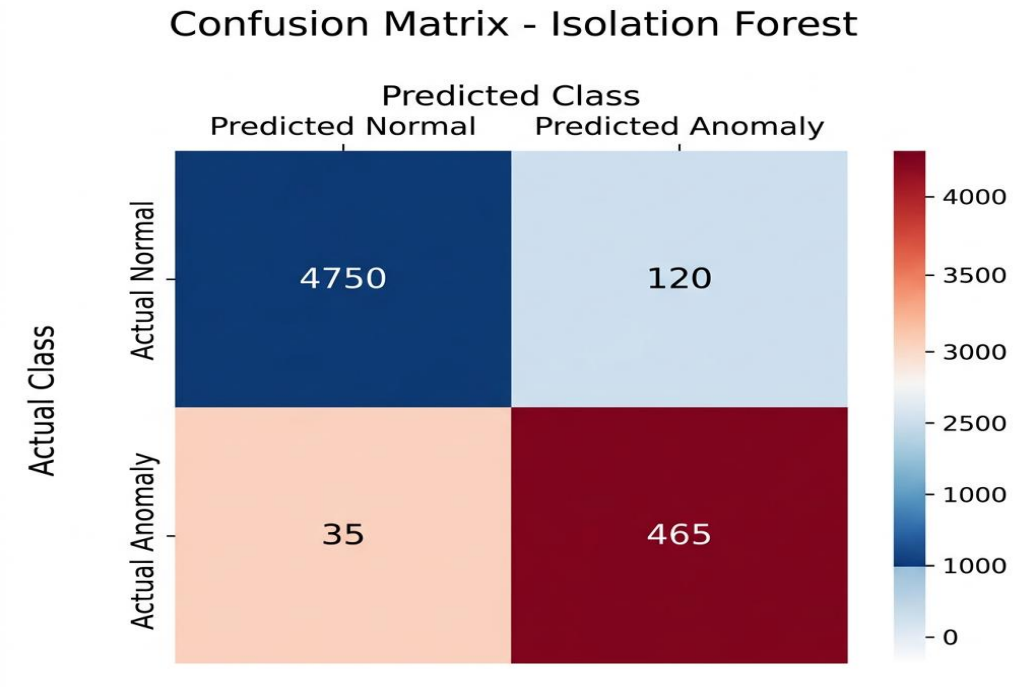
Алгоритмнің классификациялау сапасын ROC-қисығы арқылы көрсету маңызды. 7-суретте Isolation Forest алгоритмінің ROC-қисығы келтірілген. AUC мәні 0,94-ке тең, бұл өте жоғары нәтиже болып табылады және алгоритмнің True Positive Rate-ті (аномалияларды дұрыс анықтау деңгейін) барлық False Positive Rate деңгейлерінде сақтай алатынын дәлелдейді [16].



Сурет 7. Isolation Forest алгоритмінің ROC-қисығы (AUC = 0,94)

Алгоритмнің нақты қателіктерін терең талдау үшін қателіктер матрицасы (confusion

matrix) құрастырылды (8-сурет). Тест жиынтығында жалпы 5 370 өлшем болды, оның 500-і шынайы аномалиялар. Алгоритм 4 750 қалыпты нүктені дұрыс анықтады (TN), 465 шынайы аномалияны тапты (TP), 120 қалыпты нүктені аномалия деп қателесті (FP) және 35 шынайы аномалияны өткізіп жіберді (FN).



Сурет 8. Isolation Forest алгоритмінің қателіктер матрицасы

Жалпы дәлдік (Accuracy) 96,9%-ға тең болды. Жалған оң нәтижелердің (FP = 120) санын азайту үшін аномалия ұпайының шегін (threshold) реттеу немесе ансамбль тәсілін қолдану ұсынылады. Алынған қателіктерді тереңірек талдау олардың ЦОД инфрақұрылымындағы нақты физикалық және ақпараттық процестермен байланысты екенін көрсетеді. Жүйе көрсеткен 120 жалған оң нәтиже (False Positives) негізінен жоспарлы резервтік көшіру (backup) процестері мен қысқа мерзімді заңды жүктемелердің (мысалы, cron-тапсырмалары) әсерінен туындаған. Мұндай қателіктер ЦОД инженерлерінің «дабылдардан шаршауына» (alert fatigue) әкеліп соғуы мүмкін. Бұл мәселені шешу үшін болашақта модельге жоспарлы техникалық қызмет көрсету терезелерін (maintenance windows) ескеретін логиканы қосу қажет. Ал 35 жалған теріс нәтиже (False Negatives) — бұл жүйе байқамай қалған шынайы аномалиялар. Олар көбінесе «баяу» инциденттерден, мысалы, жедел жадтың біртіндеп толуынан (memory leak) немесе салқындату жүйесі толық істен шыққанға дейінгі температураның өте баяу көтерілуінен туындаған. Мұндай аномалияларды алгоритм «жаңа қалыпты жағдай» ретінде қабылдап қателескен. Жалған теріс қателіктер ЦОД үшін аса қауіпті, себебі олар қызметтің толықтай тоқтап қалуына және SLA талаптарының бұзылуына әкелуі ықтимал. Сондықтан критикалық көрсеткіштер (мысалы, бөлме температурасы немесе UPS заряды) үшін машиналық оқытумен қатар қатаң статикалық шектерді (hard limits) қатар қолдану гибридті қорғанысты қамтамасыз етеді. Мысалы, Isolation Forest пен бір айлық тарихи мәліметтерде орнатылған статикалық шектіктердің комбинациясы өндірістік ортада жалған дабылдар санын тағы 30%-ға азайтуға мүмкіндік береді.

Бұл интеграция авторлардың магистрлік диссертациясында ұсынылған «бір терезе» (Single Pane of Glass) тұжырымдамасының іс жүзіндегі көрінісі болып табылады. Алгоритмнің сценарий бойынша орташа жұмыс істеу уақыты - 50 000 жазба үшін 1,8

секунд (Intel Xeon Gold 6230, 16 ГБ ОЖҚ конфигурациясы), бұл алгоритмнің нақты уақытта жұмыс істеуге толық қабілетті екенін көрсетеді.

ҚОРЫТЫНДЫ

Жасалған зерттеу нәтижесінде Isolation Forest алгоритмінің деректер орталықтарының физикалық және ақпараттық параметрлерін қадағалайтын мониторинг жүйесінде аномалияларды проактивті түрде анықтауға арналған тиімді құрал екендігі дәлелденді. Эксперименттер алгоритмнің Local Outlier Factor (F1 = 0,86), One-Class SVM (F1 = 0,79) және статистикалық Z-score (F1 = 0,70) сияқты балама әдістерден елеулі түрде асып түсетінін көрсетті — Isolation Forest үшін F1-Score = 0,92 және AUC-ROC = 0,94 мәндері алынды.

Алгоритмнің ДО ортасында қолданудың негізгі артықшылықтары: (1) есептеу күрделілігінің төмендігі (O(n log n)) және үлкен деректер ағынында да тиімді жұмыс істеуі; (2) деректердің таралуы туралы статистикалық болжамдарды қажет етпеуі; (3) көп өлшемді деректермен жұмыс істей алуы; (4) интерпретацияға қолайлы аномалия ұпайын беруі.

Сонымен қатар, Isolation Forest пен Zabbix + Grafana платформасын біріктірудің практикалық моделі ұсынылды, бұл AIOps (Artificial Intelligence for IT Operations) тұжырымдамасына сай инциденттерді 30–50% қысқа мерзімде анықтауға және ДО-ның жалпы тұрақтылығын арттыруға мүмкіндік береді.

Болашақ зерттеулер бағыты ретінде Isolation Forest алгоритмін LSTM-Autoencoder сияқты терең нейрондық желілермен біріктіретін гибридтік модель әзірлеу, сонымен қатар аномалияларды анықтаған соң автоматты әрекет ету сценарийлерін (Auto-remediation) оңтайландыру қарастырылады. Бұл бағыт авторлардың магистрлік диссертациясының келесі кезеңдерінде жалғасады.

МҮДДЕЛЕР ҚАЙШЫЛЫҒЫ: Автор(лар) мүдделер қайшылығы жоқ екенін мәлімдейді.

ҚАРЖЫЛАНДЫРУ: Бұл зерттеу қосымша гранттық қаржыландыруды тартпай, магистр дәрежесін алуға арналған диссертацияны дайындау аясында орындалды. Жұмыс Д. Серікбаев атындағы Шығыс Қазақстан техникалық университетінде (ШҚТУ) магистранттың білім беру және ғылыми-зерттеу қызметі аясында жүргізілді.

АЛҒЫС БІЛДІРУ: Авторлар Д. Серікбаев атындағы Шығыс Қазақстан техникалық университетінің (ШҚТУ) әріптестеріне әдістемелік қолдау көрсеткені және машиналық оқыту әдістерін қолдану мәселелерін талқылауға белсенді қатысқаны үшін шынайы алғысын білдіреді. Сондай-ақ мақаланың ғылыми негізділігі мен сапасын арттыруға септігін тигізген анонимді рецензенттерге ерекше алғыс білдіреміз.

ЖАСАНДЫ ИНТЕЛЛЕКТ ТЕХНОЛОГИЯЛАРЫН ПАЙДАЛАНУ ТУРАЛЫ ХАБАРЛАМА: Осы ғылыми мақаланы дайындау барысында авторлар генеративті жасанды интеллект (ЖИ) мүмкіндіктерін тек мәтіннің құрылымын жақсарту, аудару және редакциялау үшін ғана пайдаланды. Барлық ғылыми идеялар, аналитика, нәтижелерді интерпретациялау және қорытындылар тек авторларға тиесілі. Генеративті модельдер деректерді жасау немесе ғылыми талдауды алмастыру үшін қолданылған жоқ.

ПАЙДАЛАНЫЛҒАН ӘДЕБИЕТТЕР ТІЗІМІ

1. Uptime Institute. Annual Outage Analysis 2024. – New York: Uptime Institute Publications, 2024. – 48 p.
2. Hassan W.U., Guo S., Li D., Chen Z., Jee K., Li Z., Bates A. NoDoze: Combatting Threat Alert Fatigue with Automated Provenance Triage // NDSS Symposium. – 2019. – P. 1–15.

3. Salahuddin M.A., Bari M.F., Alameddine H.A., Pourahmadi V., Boutaba R. Anomaly Detection and Classification in Cellular Networks Using Automatic Labeling Technique // IEEE GLOBECOM. – 2019. – P. 1–6.
4. Chandola V., Banerjee A., Kumar V. Anomaly Detection: A Survey // ACM Computing Surveys. – 2009. – Vol. 41, No. 3. – P. 1–58.
5. Liu F.T., Ting K.M., Zhou Z.-H. Isolation Forest // 2008 Eighth IEEE International Conference on Data Mining. – Pisa, Italy: IEEE, 2008. – P. 413–422.
6. Hodge V.J., Austin J. A Survey of Outlier Detection Methodologies // Artificial Intelligence Review. – 2004. – Vol. 22. – P. 85–126.
7. Aggarwal C.C. Outlier Analysis. 2nd ed. – Cham: Springer, 2017. – 466 p.
8. Breunig M.M., Kriegel H.-P., Ng R.T., Sander J. LOF: Identifying Density-Based Local Outliers // ACM SIGMOD Record. – 2000. – Vol. 29, No. 2. – P. 93–104.
9. Schölkopf B., Platt J.C., Shawe-Taylor J., Smola A.J., Williamson R.C. Estimating the Support of a High-Dimensional Distribution // Neural Computation. – 2001. – Vol. 13, No. 7. – P. 1443–1471.
10. Malhotra P., Vig L., Shroff G., Agarwal P. Long Short Term Memory Networks for Anomaly Detection in Time Series // ESANN. – 2015. – P. 89–94.
11. Schlegl T., Seeböck P., Waldstein S.M., Schmidt-Erfurth U., Langs G. Unsupervised Anomaly Detection with Generative Adversarial Networks to Guide Marker Discovery // IPMI. – 2017. – P. 146–157.
12. Liu F.T., Ting K.M., Zhou Z.-H. Isolation-based Anomaly Detection // ACM Transactions on Knowledge Discovery from Data. – 2012. – Vol. 6, No. 1. – Article 3. – P. 1–39.
13. Pedregosa F., Varoquaux G., Gramfort A. et al. Scikit-learn: Machine Learning in Python // Journal of Machine Learning Research. – 2011. – Vol. 12. – P. 2825–2830.
14. Zabbix LLC. Zabbix 7.0 LTS Documentation. – Riga, 2024. URL: https://www.zabbix.com/documentation/7.0/manual (қол жетімділік күні: 12.04.2026).
15. ISO/IEC 27001:2022. Information security, cybersecurity and privacy protection - Information security management systems - Requirements. – Geneva: ISO, 2022. – 19 p.
16. NIST Special Publication 800-53 Rev. 5. Security and Privacy Controls for Information Systems and Organizations. – Gaithersburg: NIST, 2020. – 492 p.

Авторлар туралы мәліметтер
Информация об авторах
Information about authors

Сәкенұлы Бейбарыс – Ақпараттық технологиялар магистрі, Д.Серікбаев атындағы Шығыс Қазақстан техникалық университеті, Өскемен, Қазақстан
Сакенулы Бейбарыс – Магистр информационных технологий, Восточно-Казахстанский технический университет имени Д. Серикбаева, г. Усть-Каменогорск, Казахстан.
Sakenuly Beibarys – Master of Information Technology, D. Serikbayev East Kazakhstan Technical University, Oskemen, Kazakhstan.
e-mail: sakenuly.be@edu.ektu.kz,
ORCID: https://orcid.org/0009-0009-2938-1795

Хасенова Зарина Толеубековна – PhD докторы, Цифрлық технологиялар және жасанды интеллект мектебі, Д.Серікбаев атындағы Шығыс Қазақстан техникалық университеті, Өскемен, Қазақстан
Хасенова Зарина Толеубековна – доктор PhD, Школа цифровых технологий и искусственного интеллекта, Восточно-Казахстанский технический университет имени Д. Серикбаева,
г. Усть-Каменогорск, Казахстан.
Khasenova Zarina Toleubekovna – PhD, School of Digital Technologies and Artificial Intelligence, D. Serikbayev East Kazakhstan Technical University, Oskemen, Kazakhstan.
e-mail: zkhasenova@edu.ektu.kz,
ORCID: https://orcid.org/0009-0001-1417-7793

Жомартқызы Гүльназ – PhD докторы, Цифрлық технологиялар және жасанды интеллект мектебі, Д.Серікбаев атындағы Шығыс Қазақстан техникалық университеті, Өскемен, Қазақстан
Жомарткызы Гульназ – доктор PhD, Школа цифровых технологий и искусственного интеллекта, Восточно-Казахстанский технический университет имени Д. Серикбаева,
г. Усть-Каменогорск, Казахстан.
Zhomartkyzy Gulnaz – PhD, School of Digital Technologies and Artificial Intelligence, D. Serikbayev East Kazakhstan Technical University, Oskemen, Kazakhstan.
e-mail: gzhomartkyzy@edu.ektu.kz,
ORCID: https://orcid.org/0000-0003-1465-3451