На правах рукописи

Изюмов Павел Сергеевич

Разработка и исследование алгоритмов анализа сетевой инфраструктуры и интернет-трафика в условиях ограниченных вычислительных мощностей

Специальность: 2.3.1.
Системный анализ, управление и обработка информации, статистика

Автореферат
диссертации на соискание учёной степени
кандидата технических наук

Долгопрудный — 2025

Работа выполнена в «Московском физико-технический институте (национальный исследовательский университет)» (МФТИ, Физтех)..

Научный руководитель: кандидат технических наук
Ивченко Александр Владимирович

Ведущая организация: Федеральное государственное бюджетное образовательное учреждение высшего профессионального образования с длинным длинным длинным длинным названием

Защита состоится DD mmmmmmmm YYYY г. в XX часов на заседании диссертационного совета Д 123.456.78 при Название учреждения по адресу: Адрес.

С диссертацией можно ознакомиться в библиотеке Название библиотеки.

Отзывы на автореферат в двух экземплярах, заверенные печатью учреждения, просьба направлять по адресу: Адрес, ученому секретарю диссертационного совета Д 123.456.78.

Автореферат разослан DD mmmmmmmm2025 года.
Телефон для справок: +7 (0000) 00-00-00.

Ученый секретарь
диссертационного совета
Д 123.456.78,
д-р физ.-мат. наук Фамилия Имя Отчество

Общая характеристика работы

Актуальность темы. Современные интернет-сети становятся все более сложными, а также характеризуются высокой динамичностью изменений, что предъявляет высокие требования к методам их анализа и оценке производительности в режиме реального времени или близким к таковым по скорости. Эффективность работы сети напрямую зависит от множества факторов, включая нагрузку на каналы передачи данных, количество активных пользователей, процесс маршрутизации, а также используемые типы устройств. Все это влияет на качество обслуживания (QoS - Quality-of-Service) в условиях ограниченного вычислительного ресурса. С увеличением объема данных и числа подключенных устройств требуется разработка новых подходов для мониторинга и анализа работы сетей в реальном времени.

Однако в условиях ограниченного вычислительного ресурса, а также необходимости обеспечения быстродействия и масштабируемости, использование сложных моделей машинного обучения, таких как глубокие нейронные сети, оказывается не всегда оправданным. Эти модели требуют значительных вычислительных мощностей и памяти, что усложняет процесс внедрения в реальных системах с ограниченными ресурсами. Многие современные исследования указывают на необходимость в поиске альтернатив для задач сетевого мониторинга, в частности при применении мобильных и периферийных устройств (edge device) [1—3]. Также стоит подчеркнуть, что различные ресурсоемкие методы (в частности нейросетевые подходы) требуют больших серверных мощностей, а значит необходима передача данных с точки сбора до данного сервера, что может не проходить по ограничениям по времени, а также нарушать требования безопасности при транспортировке чувствительных данных. В связи со перечисленными проблемами требуется разработка и применение легковесных алгоритмов, которые позволяют на должном уровне точности анализировать производительность интернет-сетей при минимальных затратах ресурсов и таким образом не теряя свою эффективность с ростом сетевой инфраструктуры и увеличением количества пользователей.

Целью данной работы является исследование и разработка методов обработки и анализа интернет-трафика и данных интернет-сетей, в частности анализ доступности соединения, с упором на использование легковесных методов и алгоритмов, в частности относящихся к технологиям TinyML.

Для достижения поставленной цели были поставлены следующие **задачи**:

1. Провести исследование современных решений задачи анализа интернет-сетей, в частотности в условиях ограничений на вычислительные мощности.

2. Исследовать данные о трафике из открытых источников и провести апробацию и сравнение методов анализа соединений, применяемых с целью обеспечения необходимого уровня QoS.
3. Провести сбор и исследование данных о состоянии сетевой инфраструктуры. Разработать методы предобработки полученных данных.
4. Разработать методы анализа и оценки состояния компьютерных сетей, в том числе в режиме реального времени. При ведении разработки учитывать малую мощность применяемых устройств: отказ от использования глубоких нейронных сетей (DNN) и высоконагруженных ансамблевых методов машинного обучения.

Основные положения, выносимые на защиту:

1. Проведенный разведочный анализ, исследование и разработка методологии анализа данных сетевого домена позволили выявить ковариационный сдвиг в ряде датасетов и улучшить модели классификации (увеличение accuracy не менее чем на 0.2), а также выявить оптимальные методы предобработки.
2. Анализ и исследование особенностей программно-аппаратных платформ системы RIPE Atlas привел к разработке методов предобработки и калибровки получаемых данных, которые позволяют компенсировать временную задержку у 63% устройств, обеспечивая их функциональную эквивалентность 37% эталонных устройств и совместимость с едиными аналитическими методами.
3. Разработанные и исследованные алгоритмы динамической кластеризации узлов сети Интернет на основе данных об их производительности на территории РФ с применением методов машинного обучения позволили выделить 32% аномальных узлов, фильтрация которых привела к снижению ошибки (RMSE) оценки времени приема передачи до 50%.
4. Разработанные и исследованные алгоритмы оценки времени приема-передачи (round-trip-time, RTT) в детектируемых кластерах сети позволили использовать на 3 порядка меньше операций с плавающей запятой (FLOPs) при достижении сопоставимой точности в сравнении с передовыми нейростевыми подходами.

Достоверность полученных результатов обеспечивается использованием методов статистического анализа, машинного обучения, математическим моделированием, совпадением результатов исследования с экспериментальными данными, а также непосредственным участием автора в получении исходных данных и проведении экспериментов.

Апробация работы. Основные результаты работы докладывались на следующих научных конференциях:

1. VI международная конференция «Информационные технологии и технические средства управления» (ICCT-2022), Институт проблем управлений им. В.А.Трапезникова РАН совместно с Астраханским государственным техническим университетом, 2022 г.
2. X Международная конференция «Инжиниринг & Телекоммуникации — En&T-2023» МФТИ, Долгопрудный
3. 65-й Всероссийская научная конференция МФТИ в честь 115-летия Л.Д. Ландау 2023, МФТИ, Долгопрудный
4. XXVI Международная конференция «Цифровая обработка сигналов и ее применение — DSPA-2024», Институт проблем управления им. В.А.Трапезникова РАН, Москва, Россия.
5. XI Международная конференция «Инжиниринг & Телекоммуникации — En&T-2024» МФТИ, Москва.

Публикации. Основные результаты по теме диссертации изложены в 6 печатных работах, 1 из которых издана в журналах, рекомендованных ВАК, 2 – в рецензируемых изданиях, входящих в базу данных Scopus, 3 - в тезисах докладов. Получено 1 свидетельство о регистрации программы для ЭВМ.

Содержание работы

Во **введении** обосновывается актуальность исследования, проводимого в рамках данной диссертационной работы, приводится обзор научной литературы по изучаемой проблеме анализа сетей и интернет-трафика, формулируется цель, ставятся задачи для ее достижения, излагается научная новизна, практическая значимость, а также сведения об апробации описываемой работы. В последующих главах сначала представлен обзор существующих решений по тематике работы, затем описываются результаты анализа полученных сетевых данных и итоги применения на них разработанных методов и подходов.

Первая глава посвящена общей актуализации проблемы анализа сетей и обзору современных методов анализа интернет-трафика и сетевой инфраструктуры. Описываются основные факторы, обостряющие проблему мониторинга сетей как такового в настоящее время. Также представлен сравнительный анализ как классических методов анализа трафика и сетей, так и направления в разработке альтернативных подходов, в частности использование ML и DL подходов.

Раздел 1.1 посвящен общей актуализации проблемы анализа сетей и обзору современных методов анализа интернет-трафика и сетевой инфраструктуры. Описываются основные факторы, обостряющие

проблему мониторинга сетей и трафика как такового в настоящее время. В частности, такие как общий рост мировых объемов трафика (Рисунок 1), рост использования протоколов шифрования и рост разнообразия используемых устройств. Эти факторы приводят к уменьшению эффективности традиционных методов, на что указывают различные современные обзорные публикации [4; 5]. В связи с этим возникает необходимость в разработке альтернативных методов и инструменты, основанные на машинном обучении и статистическом анализе, являются перспективными и развивающимися.

Рисунок 1 — Средний и пиковый объемы трафика во Франкфурте в период с 2018 по 2022 год

В разделе 1.2 приведен обзор и сравнение классических методов анализа сетевого трафика и разрабатываемых подходов, основанных на машинном обучении и нейронных сетях. В последние годы наблюдается рост интереса к глубоким нейронным сетям (DNN) в различных сферах науки и анализ трафика не является исключением [6]. Их главное преимущество — автоматическое извлечение признаков, однако они требуют ресурсоёмкой настройки архитектуры и высоких вычислительных затрат. Несмотря на это, DNN часто обеспечивают высокую точность [7], особенно с учётом современных аппаратных улучшений.

Тем не менее, DNN не всегда применимы в условиях ограниченных вычислительных ресурсов, например, на промежуточных маршрутизаторах и одноплатных устройствах. Рост объёма и шифрования интернет-трафика делает традиционные методы анализа, такие как DPI, менее эффективными [8; 9]. В таких условиях классические алгоритмы машинного обучения, хоть и уступают в точности, могут быть предпочтительны из-за своей лёгкости [10; 11].

Deep Packet Inspection (DPI) остаётся важной технологией, применяемой для анализа содержимого пакетов и обнаружения угроз [12; 13]. DPI сопоставляет входящий трафик с базой сигнатур атак и может выявлять специфичные паттерны вредоносных программ [14] (Рисунок 2).

Однако DPI сталкивается с рядом ограничений:

- Широкое использование шифрования (например, TLS) затрудняет доступ к полезной нагрузке пакетов [15];
- Применение нестандартных портов и маскировка протоколов снижают точность анализа [16];
- Полиморфные вредоносные программы изменяют сигнатуры и обходят стандартные шаблоны [17];
- Высокая вычислительная нагрузка ограничивает применение DPI в реальном времени [18];
- Растущие объёмы сетевых данных усложняют эффективный анализ [19];
- DPI вызывает опасения по поводу конфиденциальности пользователей [20];
- Применение ИИ злоумышленниками требует более интеллектуальных методов защиты [21].

В ответ на эти вызовы возрастает интерес к использованию машинного обучения для анализа зашифрованного трафика без необходимости его расшифровки [22].

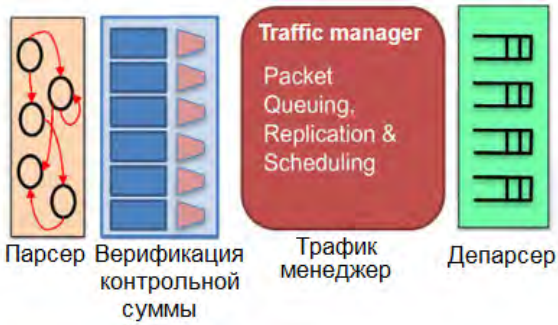


Рисунок 2 — Блок-схема этапов работы метода Deep Packet Inspection (DPI)

В разделе 1.3 рассматриваются проблемы анализа сетевой инфраструктуры, ее эффективности и поиска проблемных участков.

Анализ производительности современных компьютерных сетей сталкивается с рядом проблем, особенно при выявлении узких мест, так называемых «бутылочных горлышек» — компонентов, ограничивающих пропускную способность (см. Рисунок 3). Традиционные методы, включая протокол SNMP (Simple Network Management Protocol), пороговые правила

и сетевую томографию, становятся всё менее эффективными в условиях роста объёмов данных, усложнения архитектур и появления IoT-устройств [23; 24]. Они требуют ручной настройки, плохо адаптируются к изменяющемуся трафику и не обеспечивают полной видимости картины.

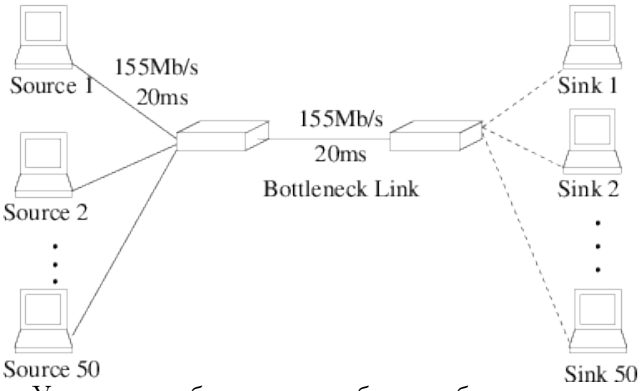


Рисунок 3 — Условное изображение проблемы «бутылочного горлышка» в структуре компьютерных сетей [25]

Методы машинного обучения предлагают более адаптивный подход. Классические модели, такие как SVM [26], деревья решений и нейронные сети, позволяют обнаруживать аномалии, прогнозировать загрузку и выявлять проблемные участки без вручную настраиваемых порогов. Они успешно применяются как в центрах обработки данных, так и в распределённых средах. Глубокое обучение (DL) с применением CNN, RNN и LSTM улучшают эту способность, извлекая сложные зависимости напрямую из трафика и временных рядов [27].

Тем не менее, методы DL имеют высокие вычислительные требования, что ограничивает их использование на маломощных устройствах, особенно в IoT-сетях [28]. Здесь необходимы лёгкие протоколы (например, LwM2M или CoMI) и энергоэффективная аналитика. Даже классические ML-модели требуют регулярного обновления и настройки, особенно при работе с изменяющейся нагрузкой.

Гибридные подходы, сочетающие лёгкий ML на периферии и более сложный анализ в центре, показывают хорошие результаты. Такие системы способны обеспечивать как предиктивное оповещение, так и выявление первопричин. Продолжающееся развитие графовых нейронных сетей и встроенной телеметрии делает перспективу полной автоматизации мониторинга всё более реальной [29].

Таким образом, традиционные подходы уже не справляются с задачами современного мониторинга. Машинное обучение предлагает более гибкие и точные инструменты, хотя их реализация требует учёта

ограничений устройств и сетевой инфраструктуры. Исследования в этой области направлены на создание масштабируемых, лёгких и адаптивных решений, пригодных для работы в реальном времени.

Итоги проведенного анализа приведены в таблице 1.

Таблица 1 — Сравнение подходов к способам мониторинга сетей

Подход	**Вычислительная сложность**	**Точность в комплексных сценариях**	**IoT и граничные вычисления**
Традиционный (SNMP, пороговые правила, выборка потоков)	Низкая нагрузка (простые опросы/ пороговые правила).	Приемлемая для известных проблем, но плохая для новых или сложных аномалий.	Хорошо подходит для высокоуровневого мониторинга (серверы, ЦОД), плохо для IoT (из-за стоимости опроса).
Классический ML(SVM, decision tree, k-NN, и т.д.)	Средняя (для обучения требуется ЦП; малая нагрузка инференса)	Хорошая во многих случаях - значительно лучше, чем статические правила [30].	Умеренно эффективно (может выполняться в облаке или на пограничных шлюзах).
Глубокое обучение(CNN, RNN, DNN и т.д.)	Высокая (крупные модели, рекомендуется использование GPU/TPU; высокая нагрузка на оборудование).	Отлично работает с большими и сложными наборами данных [27] (обнаруживает тонкие аномалии и нелинейные зависимости).	Плохо подходит для маломощных устройств Технологии TinyML могут помочь.

Во второй главе описываются результаты исследования данных интернет-трафика, предоставляемых исследователями из Канадского Института Кибербезопасности (Canadian Institute for Cybersecurity - CIC). Приведены основные результаты по применению методов генерации признаков с целью повышения метрик качества моделей машинного обучения.

В разделе 2.1 приведено описание использованных в работе наборов данных от CIC.

Одной из ключевых задач анализа трафика является его классификация по типу соединения и содержимому. Среди популярных наборов данных выделяются: ISCXVPN (2016) [31], ISCXTor2016 [32] и CCCS-CIC-AndMal (2020) [33], разработанные Канадским институтом кибербезопасности (CIC).

ISCXVPN (2016) позволяет проводить сравнительный анализ VPN и не-VPN трафика. Признаки включают:

- Среднее число пакетов в секунду
- Пропускная способность
- Длительность соединения
- Статистики по направлениям от и до адресата (среднее, максимум и т.д.)

Охватываются различные типы соединений: HTTP/HTTPS, общение в чатах, email, стриминг, FTP, VoIP, P2P.

ISCXTor2016 — содержит данные о трафике через сеть Tor и открытое соединение. Особенности:

- Объём: 22 ГБ
- Форматы: .pcap и CSV от ISCXFlowMeter

Ключевая проблема — сильный дисбаланс классов.

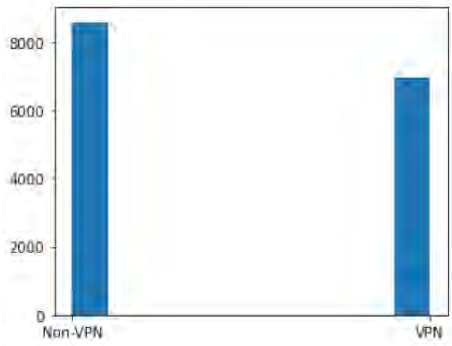


Рисунок 4 — Баланс VPN и не-VPN трафика в ISCXVPN (2016)

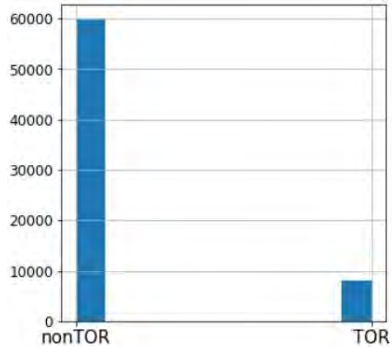


Рисунок 5 — Баланс Tor и не-Tor трафика в ISCXTor2016

CCCS-CIC-AndMal (2020) — классификация Android-программ: 200K вредоносных и 200K легитимных. Вредоносное ПО разбито на 14 категорий и 191 семейств.

Эти наборы используются для обучения и тестирования различных ML и DL моделей. Чаще всего используемые метрики: accuracy (точность), precision (точность), recall (полнота) и F1-метрика (F1-score).

Таблица 2 — Количественное описание изучаемых данных по категориям вредоносных программ

Категория	**Количество семейств**	**Количество образцов**
Рекламоносители (Adware)	48	47 210
Бэкдор (Backdoor)	11	1 538
Программы-вымогатели (Ransomware)	8	6 202
Троян (Trojan)	45	13 559
Троян-Банкир (Trojan-Banker)	11	887
Троян-Дроппер (Trojan-Dropper)	9	2 302
Троян-SMS (Trojan-SMS)	11	3 125
Тролян-Шпион (Trojan-Spy)	11	3 540
Вирус нулевого дня (Zero-Day)	-	13 340
Файловый вирус (FileInfector)	5	669
Потенциально нежелательное приложение (PUA)	8	6 202
ПО с потенциальным риском (Riskware)	21	97 349
Пугающее ПО (Scareware)	3	1 556
Без категории	-	2 296

Однако следует отметить, что метрика accuracy обладает существенным ограничением: её репрезентативность существенно снижается в условиях несбалансированности классов.

Анализ современных научных публикаций по теме анализа трафика (в том числе с применением описанных наборов данных) показывают что:

- Методы **DPI/DFI** — становятся все менее эффективны, особенно с ростом шифрования [34].
- Классическое машинное обучение, такие как **Наивный байес** [35], **SVM** [36] и **WKNN** — широко используются для анализа трафика и позволяют достичь метрики accuracy (точности) плоть до 95%.
- **Автоэнкодеры**, **CNN**, **DBN**, **LSTM** — позволяют автоматически извлекать признаки и достигать высоких показателей целевых метрик [37—40]
- **Эволюционные алгоритмы**, **LDA**, **PCA**, а также **методы ранжирования признаков** — значимо улучшают эффективность моделей и позволяют достичь accuracy до 98-99,8 % [37; 41].

Таким образом, можно подчеркнуть, что методы с использованием нейронных сетей являются довольно эффективными, но в то же время они обладают высокой вычислительной стоимостью. Классическое машинное обучение в купе с методами обработки признаков являются перспективным направлением в этом аспекте, а представленные датасеты и методы дают обширную основу для исследования в области сетевой безопасности и обнаружения вредоносных приложений.

Далее, в **разделе 2.2** описываются результаты применения методов генерации и обработки признаков, использованных на вышеописанных наборах данных.

В **подразделе 2.2.1** представлены результаты применения AGMV.

Одним из исследуемых методов генерации признаков был Accumulated Generalized Mean Value (AGMV, обобщенное накопленное среднее значение). Этот подход позволяет анализировать поведение временных рядов и выявлять переходы между классами данных. Формула AGMV определяется следующим образом:

$$AGMV(x_j(t_i), p, a, K, M) = \left\{ \frac{1}{M-K} \sum_{i=K, i=i+a}^{M-1} (|x_j(t_i)|^p) \right\}^{1/p},$$

где x_j — временной ряд на интервале [K, M], p — параметр чувствительности, a — шаг децимации, а $M - K$ — длина окна.

Метод AGMV применяется к блокам данных отдельно и преобразует исходные признаки в новые, более информативные. Он может работать подобно среднему арифметическому, геометрическому или

гармоническому в зависимости от значения p. Подробности можно найти в [42].

Было проведено исследование эффективности AGMV на двух наборах данных: ISCXVPN (бинарная классификация трафика) и CCCS-CIC-AndMal-2020 (мультиклассовая классификация вредоносного ПО). Для обеих задач использовались четыре модели машинного обучения: Случайный лес (Random Forest), дерево решений (Decision Tree), градиентный бустинг (XGBoost) и метод опорных векторов (SVM). AGMV применялся с параметрами по умолчанию: $w = 400$, $a = 1$, $p = 1$.

На рисунке 6 представлены результаты бинарной классификации для лучших комбинаций гиперпараметров. Использование AGMV позволило достичь accuracy свыше 0,95 с моделью Random Forest. Аналогично, на рисунке 7 показаны результаты мультиклассовой классификации вредоносных программ. Здесь также достигнут высокий уровень accuracy, что демонстрирует эффективность подхода.

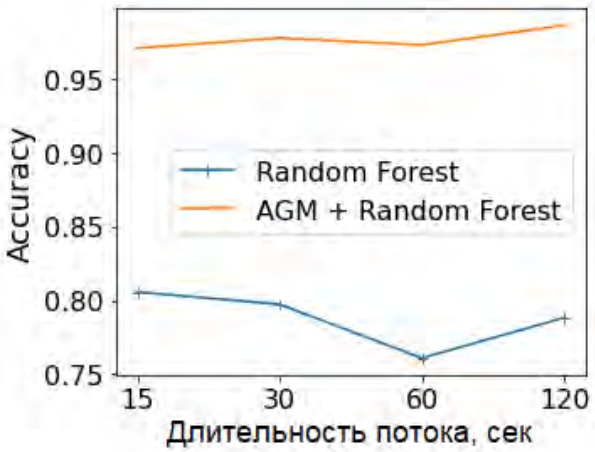


Рисунок 6 — Accuracy для лучших кандидатов (комбинации гиперпараметров) для бинарной классификации с и без использования AGMV на наборе данных ISCXVPN для алгоритма случайного леса (RandomForest)

Таким образом, метод AGMV оказался эффективным для улучшения качества классификации, как в задаче анализа сетевого трафика, так и в задаче обнаружения вредоносного ПО. В дальнейшем планируется исследовать влияние параметров AGMV на качество классификации и расширить применение метода на другие типы данных.

В **подразделе 2.2.2** представлены результаты применения CAPoNef.

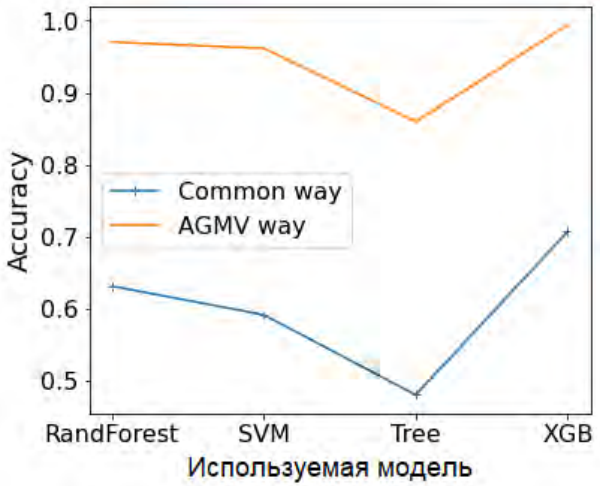


Рисунок 7 — Accuracy для лучших кандидатов (комбинации гиперпараметров) для мультиклассовой классификации с и без использования AGMV на наборе данных CCCS-CIC-AndMal (Before reboot сценарий) для моделей случайного леса (RandForest), дерева решений (Tree), машины опорных векторов (SVM) и градиентного бустинга (XGB)

Еще одним исследуемым методом генерации признаков был метод CAPoNef. Его принцип несколько отличается от рассматриваемого ранее AGMV. CAPoNeF (Comparative Analysis of the Positive and Negative Fluctuations) - это метод, которые производит анализ временных рядов, но его отличительной особенностью является то, что ряды в нем анализируются как бестрендовоые последовательности (trendless sequences (TLS)).

Суть метода заключается в извлечении признаков из временных последовательностей, при условии, что они являются бестрендовыми, то есть в них нет значимого роста или падения среднего значения наблюдаемой величины. Метод включает в себя среднее значение, но главное - ряд характеристик для отклонений значений ряда от среднего значения, таких как диапазон между положительными и отрицательными флуктуациями, баланс между ними, размах их значений и другие, связанные с ними характеристики. Метод применялся как к характеристикам исходящих, так и входящих сообщений и пакетов (рисунок 8). В итоге для каждого меняющегося во времени параметра было получено 7 новых признаков.

Метод был разработан в МФТИ и был впервые освящен в 2020 году. Данный метод уже показал свою эффективность в работах посвященных

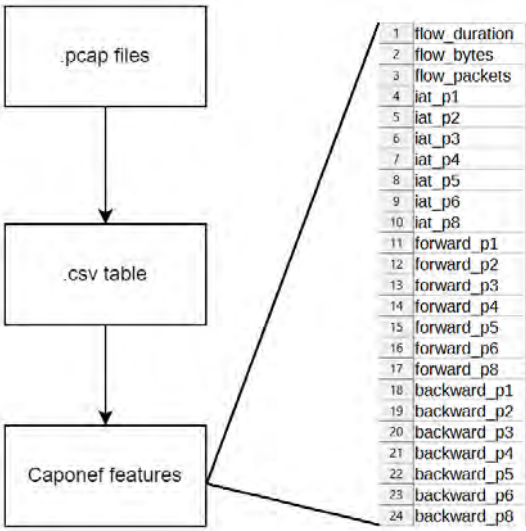


Рисунок 8 — Схема генерации признаков из исходных файлов

радиометрической идентификации, а также для задачи анализа уровня сахара в крови [43; 44].

В работе сравнивались два набора признаков: первый — базовые признаки, уже включенные в данные авторами данного датасета, а также признаки согласно методу CAPoNeF, полученные путем обработки исходных данных трафика в формате файлов .pcap. Также стоит отметить, что для верификации ранее полученных результатов признаки от авторов датасета были вычислены повторно из исходных файлов собранного трафика. Далее данные признаки обозначены как IAT (into arrival time). Они включают в себя среднее значение, среднеквадратичное отклонение, минимум и максимум времени между отправкой двух последовательных пакетов. Таким образом, можно заметить, что при применении CAPoNeF признаковое пространство значительно увеличилось (7 метрик против изначальных 4). На этих данных было обучено два алгоритма: случайный лес из библиотеки scikit-learn, а также градиентный бустинг из библиотеки XGB. В задаче бинарной классификации на данных VPN/non-VPN трафика при исследовании 15 секунд соединения удалось получить значение accuracy вплоть до 0,8, но, к сожалению, рассматриваемый метод не показал себя более эффективным по сравнению с базовым набором признаков. Результаты работы обоих алгоритмов представлены рисунке 9.

Исходя из вклада признаков в предсказывание категории трафика можно сделать вывод, что модель обобщилась хорошо, но, к сожалению, данный метод оказался не столь эффективным для рассматриваемой задачи. Исходя из этого можно сделать вывод, что для задачи

классификации VPN-трафика с данными, структура которых схожа с рассмотренной в работе, метод CAPoNeF оказывается неэффективным. Однако это не исключает потенциальную эффективность для других задач обработки трафика, а также эффективность для данных с иной структурой. Поэтому в дальнейшем исследовании планируется проделать более детальное исследование данного метода, тестирование для других задач и наборов данных с отличной от уже рассмотренной структурой.

Итак, был проведен анализ метода извлечения признаков для задачи анализа интернет-трафика. Хоть в целом метод показал себя работоспособным и полученная метрика показывает, что модели работают лучше, чем генератор случайных чисел (ГСЧ). Но тем не менее значимых приростов в показателях не наблюдалось при условии значительного увеличения количества используемых признаков (7 показателей соединения против изначальных 4). Поэтому следует провести тестирование для других задач и наборов данных с отличной от уже рассмотренной структурой. Тем не менее, результаты указывают на то, что поиск и тестирование стоит сместить в сторону других альтернативных методов, которые приведут к повышению точности в рассмотренных задачах.

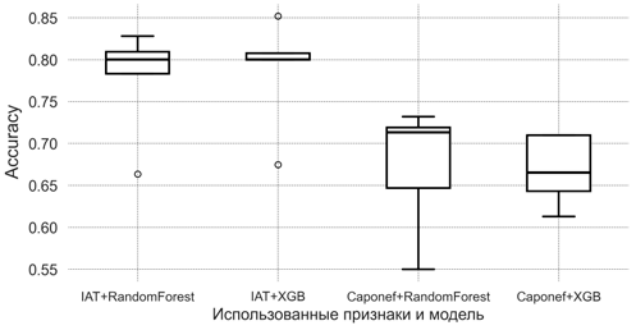


Рисунок 9 — Результаты применения генерации признаков для алгоритмов случайного леса(RandomForest) и градиентного бустинга(XGB)

В данной главе представлено исследование двух методов генерации и преобразования признаков — AGMV и CAPoNeF, ранее предложенных в научной литературе. Основное внимание было уделено оценке их эффективности в условиях, отличающихся от оригинальных: методы были адаптированы путём изменения ряда параметров и апробированы на расширенном спектре моделей машинного обучения и разнообразных наборах данных. Такая постановка задачи обеспечила более полную и независимую валидацию их применимости.

Полученные результаты выявили, что метод AGMV обеспечивает статистически значимое улучшение качества моделей, что подтверждает его универсальность и высокую адаптивность к различным структурам данных и архитектурам моделей. В отличие от него, применение метода CAPoNeF привело к систематическому ухудшению результатов, что может указывать на чувствительность метода к параметрам или ограничения его исходной концепции в условиях постановки задачи анализа трафика.

Основной результат заключается в углубленном анализе ранее предложенных решений и их эмпирической проверке за пределами оригинальных условий. Получилось расширить область применения методов AGMV и CAPoNeF, а также выявить границы их применимости на основе всестороннего экспериментального анализа. Полученные выводы могут быть использованы для более осознанного выбора стратегий генерации признаков в практических задачах анализа данных и развития автоматизированных систем построения моделей.

Третья глава посвящена теме анализа сетевой инфраструктуры и производительности узлов. В ней описывается существующие системы по сбору данных о работе интернет сетей, а также рассматриваются вопросы постобработки и калибровки получаемых данных.

В разделе 3.1 описываются две наиболее крупные системы мониторинга интернет сетей: RIPE Atlas и SamKnows. Рассмотрены их достоинства и недостатки, а также ключевые особенности доступности и распространенности по миру.

SamKnows — это платформа для мониторинга производительности интернет-сетей, предоставляющая пользователям и провайдерам возможность измерять качество соединения. Основанная в 2008 году компанией Cisco, система включила в себя решения на основе специализированного оборудования Whitebox (рисунок 10) и программного обеспечения, превращающего маршрутизаторы в измерительные зонды. Исследования подтверждают высокую точность данных SamKnows [45], включая замеры скорости загрузки, задержек и потерь пакетов в реальных условиях [46], [47]. Однако платформа коммерческая — часть функций доступна только по подписке, что ограничивает её использование в науке. Также остаются вопросы конфиденциальности, так как собирается информация о сетевом трафике пользователей.

Еще одной популярной системой является RIPE Atlas [48], насчитывающая более 13 тысяч измерительных зондов по всему миру, более 700 «Якорей» (более мощные устройства серверного типа), а также свыше 76 тысяч активных пользователей (на момент января 2025). Среди партнеров системы присутствуют такие крупные компании как Comcast и Console Connect. За 13 лет существования система значительно развилась: выпущено несколько версий аппаратных и программных зондов. На

данный момент используется пятая версия зонда (рисунок 11) на базе устройств Turris Mox [49].

Рисунок 10 — Измерительное оборудование SamKnows - «WhiteBox»

Рисунок 11 — Аппаратный зонд RIPE Atlas - Probe V5

RIPE Atlas была выбрана для дальнейших исследований благодаря возможности получения неагрегированных данных, в отличие от SamKnows, где доступны только усредненные метрики [50]. Кроме того, RIPE Atlas децентрализована и предоставляет гибкость в проведении измерений, что делает её более подходящей для научного анализа.

Ещё одним преимуществом таких систем является сбор данных из реальной сетевой инфраструктуры, а не смоделированного или лабораторного трафика. Примерами таковых могут служить многочленные наборы данных Канадского института кибербезопасности [51].

Раздел 3.2 более подробно описывает нюансы работы и особенности выбранной для исследований системы измерений - RIPE Atlas. В ходе работы с платформой RIPE Atlas было выявлено, что одним из факторов, ограничивающих исследования в области сетевого анализа, являются искажения, вносимые разными типами и версиями зондов. Например, первые две версии устройств добавляют значительные задержки. Различные версии прошивок также могут влиять на точность измерений. Таким образом, распределение метрик представляет собой смесь нескольких распределений, зависящих от типа и версии зонда.

Ещё одним фактором являются сетевые условия: RIPE Atlas не проверяет доступную пропускную способность перед измерениями, однако можно учитывать время суток для построения гипотез об искажениях. Также важно учитывать суточную активность некоторых устройств — например, они могут включаться и выключаться в определённое время.

Система RIPE Atlas использует балльную модель: пользователи накапливают баллы, устанавливая зонды, и тратят их на запуск измерений. Также доступ и баллы начисляют спонсорам и партнерам RIPE. Система поддерживает следующие типы измерений:

- PING - время круговой передачи (от исходного узла и обратно) до выставленного адреса
- TRACEROUTE - маршрут и время соединения между промежуточными адресами между зондом и выставленным адресом. Базируется на использовании параметра TTL (time-to-live)
- DNS - данные производительности DNS-серверов, в частности скорость процесса получения IP-адресов по запросу
- TLS - данные скорости работы сертификатов безопасности
- HTTP - данные об http-трафике, скорость обработки простых http-запросов
- NTP - отслеживание процесса синхронизации часов в сети

«Стоимость» измерений зависит от типа и параметров (например, таких как частота и длительность). Данные предоставляются в формате JSON через API или веб-интерфейс.

Таблица 3 — Стоимость в баллах системы RIPE Atlas различных видов сетевых измерений в расчете за один запрос к одному устройству

Тип измерения	**Стоимость**
TRACEROUTE	60
DNS (UDP)	20
DNS (TCP)	40
TLS	20
PING (ipv4/ipv6)	6
NTP	20
HTTP	20

Анализ данных показал, что устройства разных версий вносят различную дополнительную задержку. Например, при измерении TRACEROUTE до сервера gov.ru были задействованы 480 зондов с 8 версиями прошивок и ПО. Наибольшие задержки демонстрируют старые аппаратные версии V1 и V2. Современные версии V3, V4, «якоря» и программные зонды показывают задержки менее 1 мс. Результаты представлены на рисунке 12.

Устройства V1 и V2 существенно отличаются по характеристикам от V3 и V4. В связи с этим необходимо учитывать влияние аппаратной реализации при анализе данных. В новых выборках доля старых

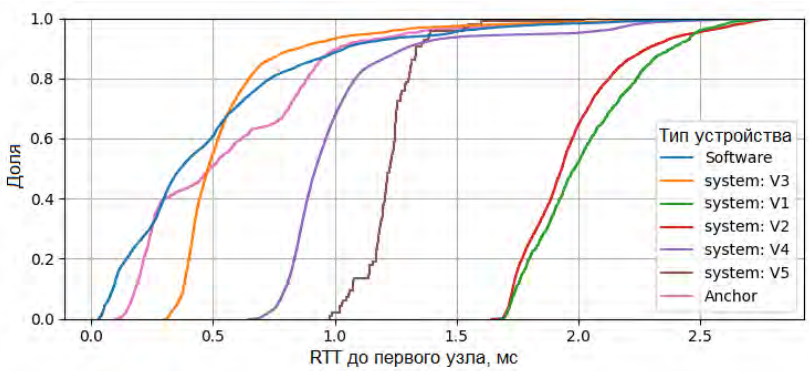


Рисунок 12 — Картина распределений времени отклика первого узла маршрута для различных типов и поколений устройств RIPE Atlas

версий прошивок невелика (<5%), поэтому основное внимание уделялось аномалиям, вызванных различиями между аппаратными версиями.

В разделе 3.3 представлены результаты калибровки данных путем оценки систематической ошибки.

Для устранения систематических ошибок в измерениях RTT были рассмотрены методы фильтрации зондов, обработки данных разных устройств разными методами и калибровка. Применение калибровки позволило снизить влияние систематической ошибки, однако статистические различия между распределениями остались значимыми (вероятность получить такие же или ещё большие отклонения при условии нулевой гипотезы (здесь и далее для краткости p-value) $p - value < 0{,}01$).

Программные зонды показали высокую дисперсию во времени отклика и требуют более сложной калибровки, что делает их использование на текущем этапе нецелесообразным.

Результаты калибровки приведены в таблице 4:

Таблица 4 — Результаты калибровки, выраженные в показаниях статистических тестов Манна–Уитни (U-test) и Колмогорова–Смирнова (KS-test). U/D - значение статистики в тестах

Стат. Тест	Этап	Тип калибруемого устройства					
		System V1		System V2		System V4	
		U/D	p-value	U/D	p-value	U/D	p-value
KS-test	До	1,0	0,0	1,0	0,0	0,89	0,0
KS-test	**После**	0,249	2,6e-42	0,23	5,7e-124	0,058	4,5e-19
U-test	До	13e6	0,0	53e6	0,0	168e6	0,0
U-test	**После**	6e6	0,0003	24e6	1,8e-14	85e6	0,0008

На основе тестов и визуального анализа сделан вывод о значительном отличии распределений от программных зондов. Было предложено разработать метод калибровки на основе преобразования распределений (Distribution Transforming).

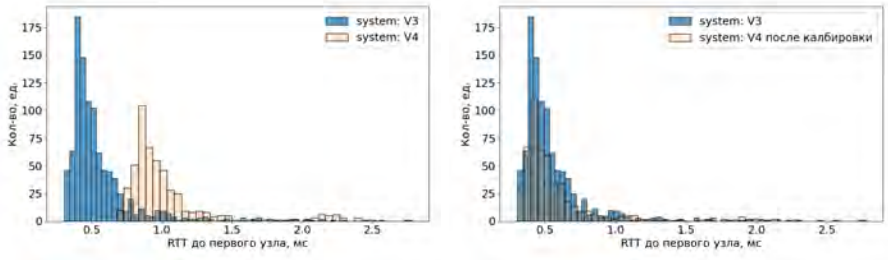
Раздел 3.4 описывает анализ одного из метода преобразования распределений - Квантиль Трансформер (Quantile Transformer).

Как было упомянуто ранее, для калибровки необходимо преобразовывать распределения. Учитывая наличие выбросов и скошенность данных, классические методы вроде Z-нормализации не подходят.

Одним из подходящих методов является квантильтрансформер (Quantile Transformer), который, например, реализован в библиотеке scikit-learn. Он позволяет переводить данные в равномерное или нормальное распределение и хорошо работает с ненормально распределенными данными [52; 53].

Для оценки эффективности метода проводились эксперименты на синтетических данных с гамма-распределением (1) и линейными деформациями. Оптимальные параметры — 24 и 43 квантиля для эталонной и калибруемой выборок соответственно — были применены к реальным данным. Визуальный и метрический анализ (рисунок 13) подтвердил улучшение соответствия эталонному устройству (устройства System: V3).

$$f(x) = x^{k-1} \frac{e^{-x/\theta}}{\theta^k \Gamma(k)} \tag{1}$$



а) б)

Рисунок 13 — Пример применения квантильтрансформера на данных RIPE Atlas. а) - Исходный вид выборки эталонного и калибруемого устройства, б) - Выборка после калибровки с параметрами, подобранными на синтетических данных

Несмотря на положительные результаты, метод требует значительных вычислительных затрат и тонкой настройки числа квантилей. С ростом разнообразия устройств его применимость усложняется, поэтому дальнейшее использование квантильтрансформера было отложено.

Раздел 3.5 посвящен анализу параметрических распределений и их применимости для калибровки полученных данных.

Визуальный анализ данных показал различия в формах распределений между типами устройств, несмотря на отсутствие статистически значимых отличий по критерию Манна–Уитни. Устройства размещены в разных городах и сетях, что указывает на влияние аппаратных особенностей зондов, нежели сетевой инфраструктуры.

Это обосновывает необходимость более точной калибровки, основанной не только на средних значениях, а на анализе всего распределения. Для этого использовался подход с параметрическими распределениями.

Были проверены гипотезы о нормальности и логнормальности с помощью тестов Шапиро–Уилка и χ^2, но они не подтвердились. Поэтому для оценки распределений применили ядерную оценку плотности (KDE) с гауссовым ядром и правилом Скотта (рисунок 14).

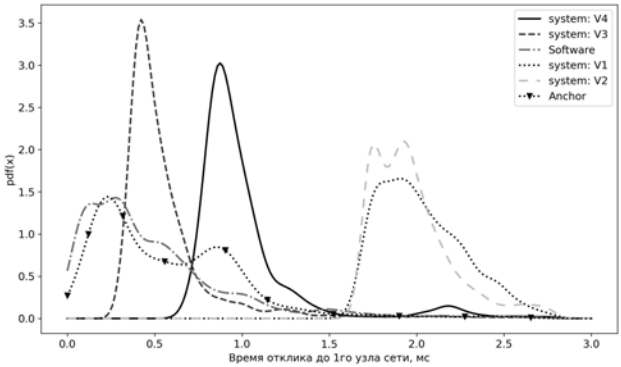


Рисунок 14 — KDE-оценка плотности вероятности для времени отклика

Далее оценили параметры для распространённых распределений: нормального, логнормального, смещённого нормального, обобщённого нормального и гамма-распределения. Пример приближения показан на рисунок 15.

Сравнение моделей выполнено с помощью дивергенции Джефриса (3) — симметричной версии KL-дивергенции (2).

$$D_{KL}(P||Q) = \int_X p \cdot \log \frac{p}{q} d\mu \tag{2}$$

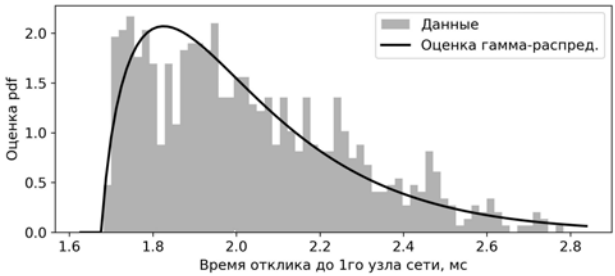


Рисунок 15 — Пример приближения pdf с использованием гамма-распределения

$$D_J(P||Q) = \frac{1}{2}D_{KL}(P||Q) + \frac{1}{2}D_{KL}(Q||P) \tag{3}$$

Результаты сравнения приведены в таблице 5 и на рисунке 16. Как можно отметить, лучшие приближения достигались с использованием логнормального и гамма-распределения. Поэтому в дальнейшем анализе использовались данные параметрические распределения.

Таблица 5 — Результаты сравнения наблюдаемых данных и рассматриваемых параметрических распределений, выраженные в расстоянии (дивергенции) Джефриса от табличного распределения до оценки по KDE. Обозначения распределений: lognormal - логнормальное, skewnorm - смещенное нормальное, gennorm - обобщенное нормальное

lognormal	skewnorm	gamma	gennorm	Тип устройства	Лучшее
0,101	0,271	0,153	0,823	system: V4	lognormal
0,050	0,411	0,081	0,458	system: V3	lognormal
0,026	0,531	0,018	0,064	Software	gamma
0,056	0,402	0,032	0,091	system: V1	gamma
0,066	0,167	0,022	0,603	system: V2	gamma
0,125	0,461	0,049	0,109	Anchor	gamma

Для калибровки было использовано следующее преобразование распределений:

$$x_{calib} = F_{ref}^{-1}(F_{obs}(x)), \tag{4}$$

где F_{ref} — функция распределения эталонного устройства (system: V3), а F_{obs} — калибруемого. Примеры результатов применения описанного метода для аппаратных устройств версии V3 и V4 приведены на рисунке 17.

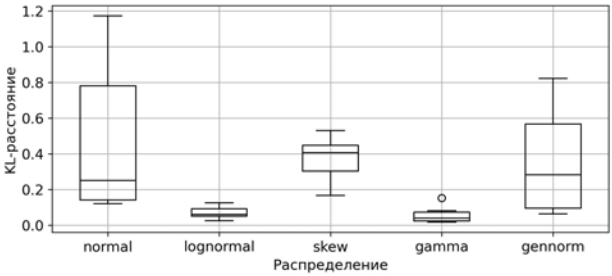
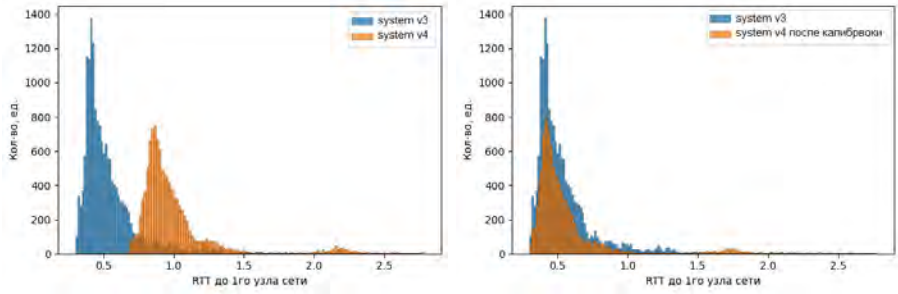


Рисунок 16 — Сравнение KDE и оценок параметрических распределений



а) б)

Рисунок 17 — Данные с устройств «system:V4» и «system:V3»: а) - до проведения калибровки, б) - после калибровки с использованием гамма-распределения

Сравнение методов выполнено по дивергенции Джефриса и расстоянию Вассерштейна (5):

$$W_1(P,Q) = \int_{-\infty}^{+\infty} |P(x) - Q(x)|dx, \tag{5}$$

где $P(x)$ и $Q(x)$ - это функции распределения сравниваемых распределений.

Особенно сильный эффект наблюдается для устройств Anchor и Software, где распределения имеют значительные различия. Преобразование распределений показало кратно лучшие результаты по сравнению с простой оценкой систематической ошибки (shift) (рисунок 18, 19).

В итоге, предложен новый метод калибровки, основанный на аппроксимации и трансформации распределений, который превосходит простую оценку сдвига, особенно при значительных отличиях формы

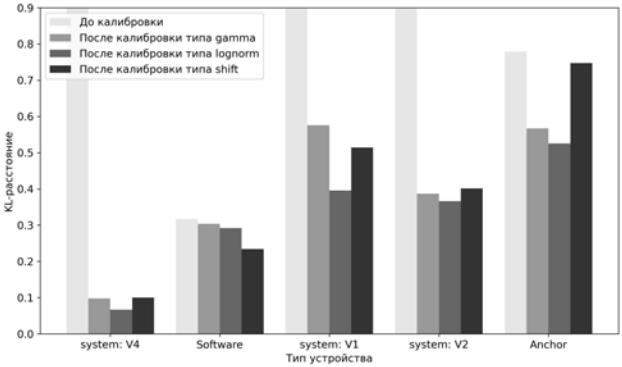


Рисунок 18 — Сравнение дивергенции Джефриса до и после применения калибровки (увеличенный масштаб).

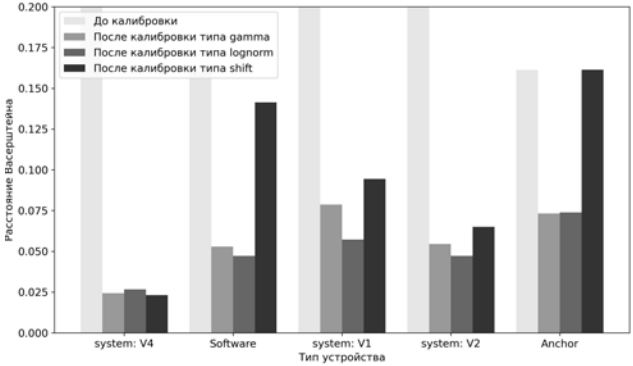


Рисунок 19 — Сравнение Расстояния Вассерштейна до и после применения калибровки (увеличенный масштаб).

распределений. Также стоит подчеркнуть его меньшую вычислительную стоимость в сравнении с Квантиль Трансформером. Тем не менее, для данного подхода видятся векторы его развития и улучшения: исследование других географических зон, анализ эффективности на новом типе устройств после роста его количества до сравнимого с предыдущими поколениями, а также исследование сходимости и оптимизации.

Четвертая глава содержит в себе результаты анализа данных о времени отклика, полученных в ходе измерений, проведенных с помощью системы RIPE Atlas, а также результаты применения методов оценки состояния узлов и загруженности сети.

Раздел 4.1 описывает характер проводимых измерений: период, частота и тип проводимых измерений сети интернет.

Для анализа доступности портала госуслуг gosuslugi.ru на территории РФ в период с 2024-05-26 по 2024-06-11 (16 суток) проводились измерения времени отклика с использованием 506 зондов RIPE Atlas. Сбор данных проводился с периодом в 20 минут (1200 секунд) путем отправки 4 пакетов и фиксированием среднего времени до пришедшего отклика. Всего было получено более 550 тысяч результатов измерений. Наибольшее количество устройств находилось в Центральном федеральном округе (246 шт.) (рисунок 20).

Рисунок 20 — Места размещения измерительных зондов системы RIPE Atlas на территории РФ, использованных в проведенных измерениях

Раздел 4.2 содержит анализ времени отклика, зафиксированного в разных регионах страны, с целью выявления как общенациональных, так и региональных особенностей функционирования сети.

Как уже упоминалось ранее, для замера времени круговой передачи (RTT) от различных узлов сети использовалась система RIPE Atlas, особенности функционирования которой описаны в работах [48] и [A1]. Эта платформа предоставляет специальный режим измерений под названием «PING», подходящий для целей данного исследования. В этом режиме выбранные узлы (станции) отправляют запросы на указанный IP-адрес или доменное имя, соблюдая заданные пользователем параметры — такие как интервал между запросами, количество и размер пакетов. По завершении каждого запроса фиксируется время, необходимое для доставки пакета до пункта назначения и получения ответа. Все полученные данные сохраняются в хранилище RIPE Atlas и могут быть выгружены после окончания тестирования. Несмотря на стратегическое значение инфраструктуры Интернета, в настоящее время отсутствуют

единые и масштабные системы его мониторинга. RIPE Atlas, являясь некоммерческим проектом, созданным международным сообществом энтузиастов, представляет собой ценный инструмент для исследователей благодаря открытому и свободному доступу к результатам измерений.

С целью более глубокого анализа характеристик полученных данных и повышения точности оценки RTT во времени, в дальнейшем был проведён кластерный анализ временных рядов. Для решения этой задачи применялся алгоритм K-средних (K-means), при этом в качестве меры сходства между рядами использовалась метрика, основанная на методе динамического преобразования временной шкалы (Dynamic Time Warping, DTW) [54].

В данном методе кластеризации пользователь сам выставляет предполагаемое число кластеров. Поэтому в целях выбора оптимального количества ядер результаты кластеризации с различными заданными значениями (вплоть до 9 кластеров) сравнивались между собой по среднему коэффициенту силуэта. Это метрика, которая базируется на численной оценке баланса расстояний между кластерами и внутри кластеров (6).

$$s = \frac{b - a}{max(a, b)}, \tag{6}$$

где a – это среднее расстояние от рассматриваемого временного ряда до других того же кластера, а b – среднее расстояние до рядов ближайшего кластера. Согласно этому критерию оптимальным значением для количества кластеров является $n = 3$ (рисунок 21).

В ходе применения кластеризации с найденными оптимальными параметрами были получены кластеры, в которых можно выделить три различных паттерна поведения в производительности узлов сети (постоянные флуктуации вокруг среднего значения и два паттерна ступенчатого вида) (рисунок 22). Баланс кластеров получился следующий: кластер №1 – 68%, кластер №2 – 13%, кластер №3 – 19%.Наличие в совокупности собранных значений RTT различных семантически и статистически отличающихся семейств временных рядов указывает на целесообразность их сегментации при смене динамических паттернов. Такая предварительная декомпозиция временных рядов позволяет реализовать подход к построению специализированных моделей прогнозирования, адаптированных к характеристикам отдельных кластеров или сегментов. Это, в свою очередь, способствует повышению точности предсказания и снижению вычислительной нагрузки за счёт уменьшения структурной и параметрической сложности применяемых моделей.

В **разделе 4.3** рассматриваются вопросы анализа и оценки производительности интернет-сети на территории страны. Предлагается

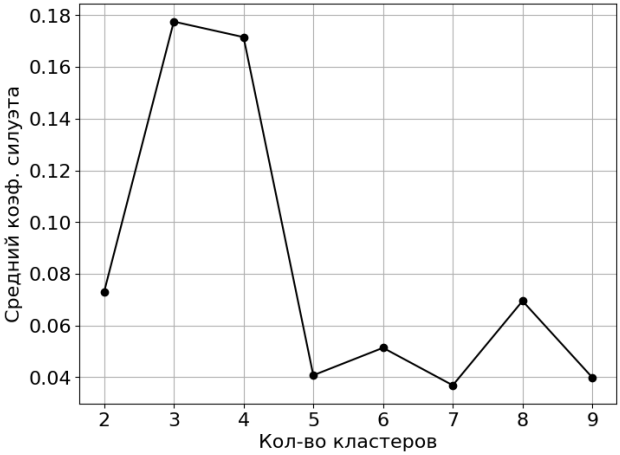


Рисунок 21 — Значения коэфициента силуэта при применении алгоритма кластеризации для различного выставленного количества кластеров

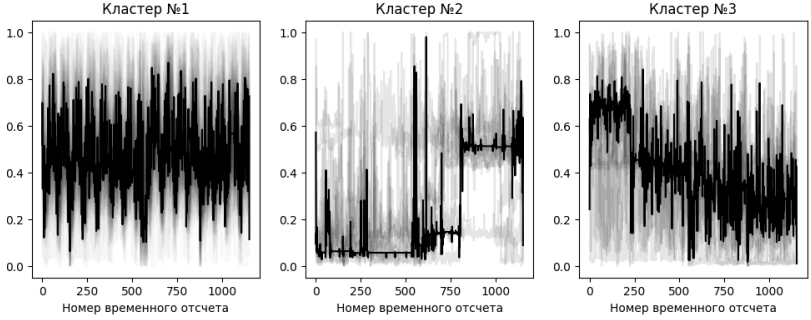


Рисунок 22 — Визуальное отображение применения кластеризации к временным рядам времени отклика. Черным отображены полученные ядра кластеров, а серым - множество временных рядов, отнесенных к этому кластеру.

комплекс методологических подходов к оценке быстродействия сети, включающих обработку аномальных наблюдений, сравнительный анализ территориальных зон, а также компонентный анализ временных рядов, полученных в ходе мониторинга. Особое внимание уделяется исследованию факторов, оказывающих влияние на скорость интернет-соединения, таких как расстояние до серверов, суточные и временные колебания активности, а также особенности работы различных интернет-провайдеров. Проведённый анализ позволяет выявить ключевые факторы, определяющие производительность сетевой инфраструктуры.

Также было исследовано изменение времени отклика в зависимости от взаимной удаленности станции и заданного сервера. Для этого были применены две модели: линейная регрессия (с и без применения регуляризации), а также метод опорных векторов (SVR). В качестве признаков выступали широта, долгота и расстояние до сервера. Лучшие результаты показала линейная регрессия с Lasso-регуляризацией (рисунок 23), с медианным значением R^2 более 0,84 на кроссвалидации. При этом веса для широты и долготы оказались равными нулю, что указывает на преобладание удаленности, нежели региональных особенностей.

Было выявлено 17 аномальных точек измерений (рисунок 24а), чьи значения времени отклика значительно превышали ожидаемые. Это может указывать на необходимость оптимизации оборудования или сетей в этих локациях.

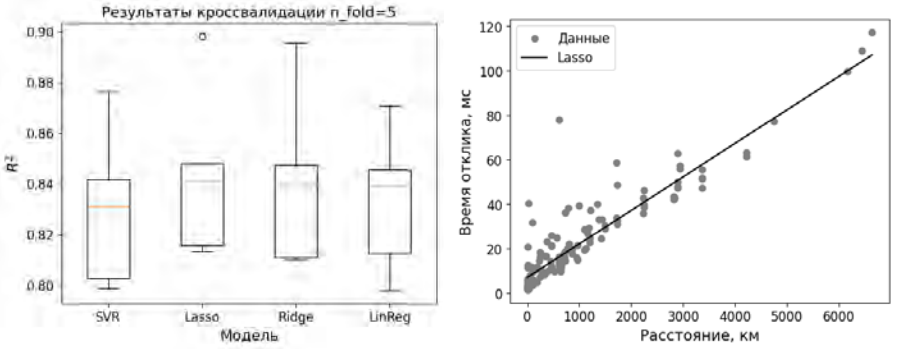
На следующем этапе был проведён анализ временных рядов: декомпозиция, изучение автокорреляции и проверка стационарности. Перед этим данные были очищены от выбросов (рисунок 24б), которые были заменены с помощью линейной интерполяции.

Анализ частичной автокорреляции (PACF) показал отсутствие нестационарности, подтверждённое тестом Филлпса-Перрона. Однако данные содержали высокочастотный шум, который был снижен с помощью фильтра Савицкого-Голея (рисунок 25).

После фильтрации и однократного дифференцирования обнаружена сезонная компонента с периодом T = 20 (6 часов 40 минут). Эти результаты могут быть использованы при настройке параметров STL-разложения и ARIMA-моделей.

Спектральный анализ также подтвердил наличие суточной сезонной компоненты (период 24 часа), а сравнительный фазовый анализ с применением преобразования Гильберта показал синхронность изменений нагрузки на сеть по всей стране, независимо от часового пояса (рисунки 26, 27).

Анализ временных рядов показал важность сегментирования данных перед прогнозированием. Результаты исследования могут быть использованы интернет-провайдерами и операторами связи для оптимизации сетей и улучшения качества услуг.



а) б)

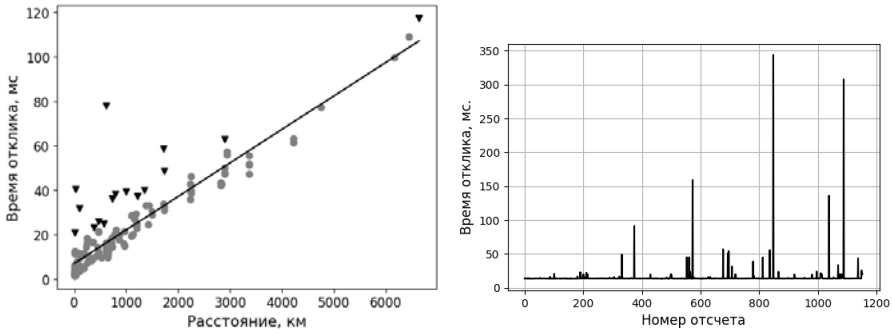
Рисунок 23 — Результаты анализа зависимости времени отклика от расстояния. На рисунке а) отображены диаграммы размаха для коэффициента R^2 для обученных регрессионных моделей зависимости медианного времени отклика от географического расположения станций. SVR – регрессия методом опорных векторов, Lasso и Ridge – линейные регрессии с регуляризацией соответствующего типа, LinReg – линейная регрессия без регуляризации. На рисунке б) - визуализация результатов применения линейной регрессии с регуляризацией типа Lasso

В финальном **подразделе 4.4** представлены результаты сравнительного анализа вычислительной сложности разработанного подхода и передовых методов, решающих схожую задачу анализа сети. В рассматриваемых работах [55] и [56] использовались три передовых нейросетевых архитектуры: Графовая нейронная сеть (GNN), Трансформер(Transformer) и долгая краткосрочная память (LSTM). Результаты проведенного анализа приведены в таблице 6.

Таблица 6 — Сравнение метрик рассматриваемых подходов

Метрика	GNN[55]	Transformer[55]	LSTM[56]	ARIMA
MAE	0,0023	2,0	2,2	0,1178
RMSE	0,0040	-	-	0,1493
MSE	-	-	-	0,0362
MAPE	-	-	-	2,4573
Примечание: прочерки означают, что авторы не предоставляют значения указанных метрик в своих работах.				

Таким образом, сравнительный анализ предложенного подхода с существующими методами решения задачи прогнозирования времени отклика показал его существенное преимущество в вычислительной эффективности: предлагаемая модель требует на 3–4 порядка меньше



а) б)

Рисунок 24 — а) - наблюдаемые данные и оценка модели (прямая). Треугольными метками изображены «аномальные» измерения. б) - исходный вид полученных данных до предобработки на примере устройства prb_id=11337. Временной шаг между отчетами 20 мин

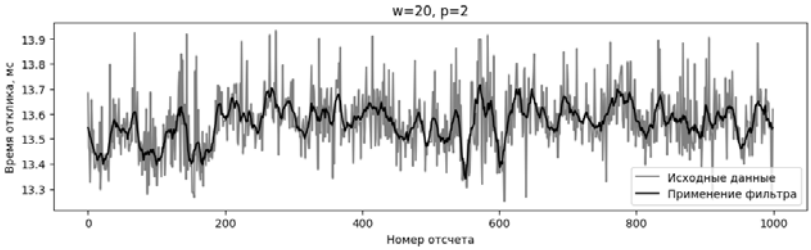


Рисунок 25 — График изменения времени отклика в течение периода проведения измерений для устройства prb_id = 12794 до и после фильтрации Савицкого-Голея. Время между отчетами $\Delta t = 20$ мин.

операций с плавающей запятой (FLOPs), что делает её пригодной для масштабирования и применения в условиях ограниченных вычислительных ресурсов.

Также стоит отметить, что для оценки точности использовались данные только «неаномального» кластера. В силу случайных изменений в аномальных кластерах ошибка оценки времени отклика может достигать $RMSE = 0{,}5$. Таким образом, исходя из баланса кластеров можно оценить ошибку на всех устройства оценивается как $RMSE_{all} \approx \sqrt{(}0{,}68 \times 0{,}15^2 + 0{,}32 \times 0{,}5^2) \approx 0{,}3$. Таким образом, фильтрация аномальных устройств обеспечивает снижение RMSE до 50%.

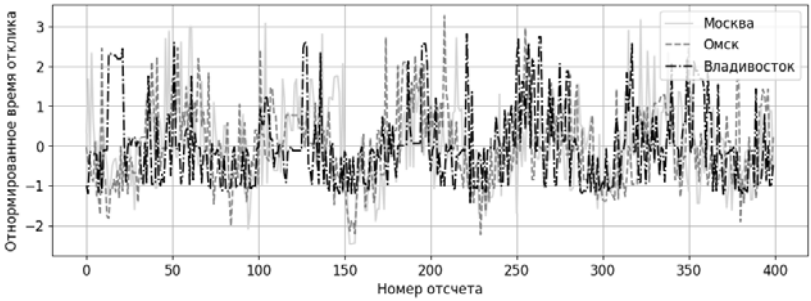


Рисунок 26 — Времена отклика для трех случайных устройств из трех городов после нормировки. Время между отчетами $\Delta t = 20$ мин.

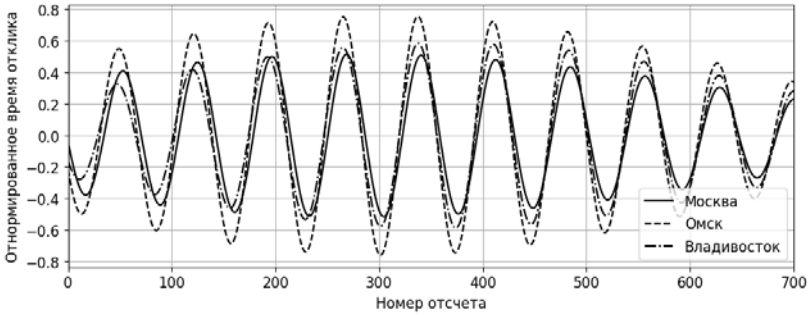


Рисунок 27 — Время отклика после использования полосового фильтра Баттерворта на данных трех случайных устройствах из разных городов после нормировки.

В **заключении** приведены основные результаты работы, которые заключаются в следующем:

1. На основе анализа данных интернет-трафика от Канадского Института Кибербезопасности был проведен углубленный анализ ранее предложенных решений за пределами оригинальных условий. Получилось расширить область применения метода AGMV и получить прирост метрики accuracy не менее чем на 0,2, а для метода CAPoNeF выявить границы его применимости на основе экспериментального анализа. Полученные выводы могут быть использованы для более осознанного выбора стратегий генерации признаков.
2. На основе данных RIPE Atlas разработаны методы предобработки и калибровки показателей времени отклика, учитывающие особенности устройств и временные флуктуации. Это позволило

повысить точность оценки производительности узлов путем устранения искажений у 63% используемых устройств.

3. На основе анализа данных о RTT, собранных с измерительных узлов, расположенных на территории Российской Федерации, был разработан набор методов выявления аномальных устройств, демонстрирующих отклонения от типового сетевого поведения. Применение разработанного метода позволило выделить 32% аномальных узлов, фильтрация которых привела к снижению ошибки (RMSE) оценки времени приема передачи до 50%.
4. На основе обработанных данных был разработан подход к оценке быстродействия сетевых узлов. Предложенный метод требует на 3–4 порядка меньше операций с плавающей запятой (FLOPs), что делает его пригодным для масштабирования и применения в условиях ограниченных вычислительных ресурсов.

Публикации автора по теме диссертации

A1. *Izyumov, P. S.* Analysis of Network State by RIPE Atlas Distributed Measurement System [Text] / P. S. Izyumov, A. V. Ivchenko // 2024 26th International Conference on Digital Signal Processing and its Applications (DSPA). IEEE. — 2024.

A2. *Izyumov, P. S.* AGMV Approach for Reduce Complexity of Classification Tasks [Text] / P. S. Izyumov, A. V. Ivchenko // 2022 6th International Scientific Conference on Information, Control, and Communication Technologies (ICCT). IEEE. — 2022.

A3. *Изюмов, П. С.* Анализ быстродействия сети интернет на территории РФ с использованием данных системы распределенных измерений RIPE Atlas [Текст] / П. С. Изюмов, А. В. Ивченко // Труды МФТИ. — 2025. — Т. 1.

A4. *Свидетельство о гос. регистрации программы для ЭВМ.* Инструмент для сбора, предобработки и калибровки данных о работе сети [Текст] / П. С. Изюмов, А. В. Ивченко ; П. Изюмов. — № 2025663192 ; заявл. 21.05.2025 ; опубл. 04.06.2025, 2025664540 (Рос. Федерация).

A5. *Изюмов, П. С.* Применение AGMV для уменьшения вычислительной сложности в задачах классификации [Текст] / П. С. Изюмов, А. В. Ивченко // Технические средства систем управления и связи, International Scientific Forum on Control and Engineering. Материалы Международного научного форума. Материалы VI Международной конференции. Материалы 15-й Международной конференции. — 2022.

А6. *Изюмов, П. С.* Исследование метода CAPoNeF к задаче анализа интернет-трафика [Текст] / П. С. Изюмов, А. В. Ивченко // Труды 65-й Всероссийской научной конференции МФТИ. — 2023.

А7. *Изюмов, П. С.* Анализ сетевой инфраструктуры с помощью распределенной платформы RIPE Atlas [Текст] / П. С. Изюмов, А. В. Ивченко // Инжиниринг и Телекоммуникации - EN&T – 2023, сборник тезисов X Международной конференции. — 2023.

Список литературы

1. LightweightNet: Toward fast and lightweight convolutional neural networks via architecture distillation [Текст] / T.-B. Xu [и др.] // Pattern Recognit. — 2019. — Т. 88. — С. 272—284.
2. Lightweight Deep Learning: An Overview [Текст] / C.-H. Wang [и др.] // IEEE Consumer Electronics Magazine. — 2024. — Т. 13. — С. 51—64.
3. *Liu, X.* Analysis on Lightweight Network Methods and Technologies [Text] / X. Liu // Highlights in Science, Engineering and Technology. — 2022.
4. *El-Maghraby, R. T.* A survey on deep packet inspection [Текст] / R. T. El-Maghraby, N. M. Abd Elazim, A. M. Bahaa-Eldin // 2017 12th International Conference on Computer Engineering and Systems (ICCES). — 2017. — С. 188—197.
5. *Kumar, K.* Network Traffic Classification Techniques: A Survey [Текст] / K. Kumar, M. Punia, Vandana // 2023 Seventh International Conference on Image Information Processing (ICIIP). — 2023. — С. 205—211.
6. *Rezaei, S.* Deep Learning for Encrypted Traffic Classification: An Overview [Текст] / S. Rezaei, X. Liu // IEEE Communications Magazine. — 2019. — Т. 57, № 5. — С. 76—81.
7. *Liu, H.* Machine Learning and Deep Learning Methods for Intrusion Detection Systems: A Survey [Текст] / H. Liu, B. Lang // Applied Sciences. — 2019. — Т. 9, № 20. — URL: https://www.mdpi.com/2076-3417/9/20/4396.
8. BlindBox: Deep Packet Inspection over Encrypted Traffic [Текст] / J. Sherry [и др.] // SIGCOMM Comput. Commun. Rev. — New York, NY, USA, 2015. — Авг. — Т. 45, № 4. — С. 213—226. — URL: https://doi.org/10.1145/2829988.2787502.
9. DE CIX Company, Frankfurt statistics [Электронный ресурс]. — URL: https://www.de-cix.net/en/locations/frankfurt/statistics (дата обр. 22.07.2022).

10. *Theofilatos, A.* Comparing Machine Learning and Deep Learning Methods for Real-Time Crash Prediction [Текст] / A. Theofilatos, C. Chen, C. Antoniou // Transportation Research Record. — 2019. — Т. 2673, № 8. — С. 169—178. — eprint: https : / / doi . org / 10 . 1177 / 0361198119841571. — URL: https://doi.org/10.1177/0361198119841571.

11. A nonlinear adaptive noise canceller with multiple reference channels for speech enhancement using both bone-and air-conducted measurements [Текст] / Y. Xiao [и др.] // 2018 International Symposium on Intelligent Signal Processing and Communication Systems (ISPACS). — 2018. — С. 258—263.

12. *Chaudhary, A.* Software Based Implementation Methodologies for Deep Packet Inspection [Текст] / A. Chaudhary, A. Sardana // 2011 International Conference on Information Science and Applications. — IEEE, 2011. — С. 1—10.

13. *Shubbar, R.* Fast 2D filter with low false positive for network packet inspection [Текст] / R. Shubbar, M. Ahmadi // IET Networks. — 2017. — Т. 6, № 6. — С. 224—231. — URL: http : / / dx . doi . org / 10 . 1049 / iet-net.2017.0055.

14. Guest Editorial Deep Packet Inspection: Algorithms, Hardware and Applications [Текст] / Y.-D. Lin [и др.] // IEEE Journal on Selected Areas in Communications. — 2014. — Т. 32, № 10. — С. 1781—1783.

15. Artificial Intelligence-Based Anomaly Detection Technology over Encrypted Traffic: A Systematic Literature Review [Текст] / I. H. Ji [и др.] // Sensors. — 2024. — Т. 24, № 3. — С. 898. — URL: http : //dx.doi.org/10.3390/s24030898.

16. Joint Analysis of Port and Protocol via Endpoint Measurement: An Empirical Study [Текст] / C. Hou [и др.] // 2020 21st Asia-Pacific Network Operations and Management Symposium (APNOMS). — IEEE, 2020. — С. 231—234.

17. A Novel Deep Packet Inspection Method for Polymorphic Network [Text] / M. Xue [et al.] // 2024 Sixth International Conference on Next Generation Data-driven Networks (NGDN). — IEEE, 2024. — P. 268—271. — URL: http : / / dx . doi . org / 10 . 1109 / ngdn61651 . 2024 . 10744179.

18. Reconfigurable regular expression matching architecture for real-time pattern update and payload inspection [Текст] / J. Nam [и др.] // Journal of Network and Computer Applications. — 2022. — Т. 208. — С. 103507. — URL: http://dx.doi.org/10.1016/j.jnca.2022.103507.

19. Advanced Network Representation Learning for Container Shipping Network Analysis [Текст] / L. Jiang [и др.] // IEEE Network. — 2021. — Т. 35, № 2. — С. 182—187. — URL: http://dx.doi.org/10.1109/mnet.011.2000444.

20. R1DIT: Privacy-Preserving Malware Traffic Classification With Attention-Based Neural Networks [Текст] / O. Barut [и др.] // IEEE Transactions on Network and Service Management. — 2023. — Т. 20, № 2. — С. 2071—2085. — URL: http://dx.doi.org/10.1109/tnsm.2022.3211254.

21. *Roshan, K.* Untargeted White-box Adversarial Attack with Heuristic Defence Methods in Real-time Deep Learning based Network Intrusion Detection System [Текст] / K. Roshan, A. Zafar, S. B. U. Haque. — 2023. — URL: https://arxiv.org/abs/2310.03334.

22. Machine Learning-Powered Encrypted Network Traffic Analysis: A Comprehensive Survey [Text] / M. Shen [et al.] // IEEE Communications Surveys; Tutorials. — 2023. — Vol. 25, no. 1. — P. 791—824. — URL: http://dx.doi.org/10.1109/comst.2022.3208196.

23. *Kentik.* The Evolution of Network Monitoring: From SNMP to Modern Network Observability [Text] / Kentik. — URL: https://www.kentik.com/kentipedia/evolution-of-network-monitoring-snmp-to-network-observability/ (visited on 04/15/2025).

24. *LOGIC, N.* The Future of Network Monitoring [Text] / N. LOGIC. — URL: https://www.netflowlogic.com/the-future-of-network-monitoring-how-ai-and-machine-learning-are-changing-the-game/ (visited on 04/15/2025).

25. Locating internet bottlenecks: algorithms, measurements, and implications [Текст] / N. Hu [и др.] // Proceedings of the 2004 Conference on Applications, Technologies, Architectures, and Protocols for Computer Communications. — Portland, Oregon, USA : Association for Computing Machinery, 2004. — С. 41—54. — (SIGCOMM '04). — URL: https://doi.org/10.1145/1015467.1015474.

26. Anomaly Detection and Bottleneck Identification of The Distributed Application in Cloud Data Center using Software–Defined Networking [Текст] / A. El-shamy [и др.] // Egyptian Informatics Journal. — 2021.

27. *Abbasi, M.* Deep Learning for Network Traffic Monitoring and Analysis (NTMA): A Survey [Текст] / M. Abbasi, A. Shahraki, A. Taherkordi // Computer Communications. — 2021. — Март. — Т. 170. — С. 19—41. — URL: http://dx.doi.org/10.1016/j.comcom.2021.01.021.

28. *Qasim Jebur Al-Zaidawi, M.* Advanced Deep Learning Models for Improved IoT Network Monitoring Using Hybrid Optimization and MCDM Techniques [Текст] / M. Qasim Jebur Al-Zaidawi, M. Çevik // Symmetry. — 2025. — Т. 17, № 3. — URL: https://www.mdpi.com/2073-8994/17/3/388.

29. Empirical analysis of performance bottlenecks in graph neural network training and inference with GPUs [Текст] / Z. Wang [и др.] // Neurocomputing. — 2021. — Т. 446. — С. 165—191.

30. *Mirza, M.* A Machine Learning Approach to Problems in Computer Network Performance Analysis [Текст] : дис. ... канд. / Mirza Mariyam. — Madison, WI, USA : University of Wisconsin–Madison, 2012. — Date of final oral examination: 05/25/12.

31. Characterization of Encrypted and VPN Traffic using Time-related Features [Текст] / G. Draper-Gil [и др.] // Proceedings of the 2nd International Conference on Information Systems Security and Privacy - ICISSP. — INSTICC. SciTePress, 2016. — С. 407—414.

32. *Abu Al-Haija, Q.* Machine-Learning-Based Darknet Traffic Detection System for IoT Applications [Текст] / Q. Abu Al-Haija, M. Krichen, W. Abu Elhaija // Electronics. — 2022. — Т. 11, № 4. — URL: https://www.mdpi.com/2079-9292/11/4/556.

33. DIDroid: Android Malware Classification and Characterization Using Deep Image Learning [Текст] / A. Rahali [и др.] // Proceedings of the 2020 10th International Conference on Communication and Network Security. — Tokyo, Japan : Association for Computing Machinery, 2021. — С. 70—82. — (ICCNS '20). — URL: https://doi.org/10.1145/3442520.3442522.

34. Using per-Source measurements to improve performance of Internet traffic classification [Текст] / S. Bregni [и др.] // 2010 IEEE Latin-American Conference on Communications. — 2010. — С. 1—5.

35. *Moore, A. W.* Internet traffic classification using bayesian analysis techniques [Текст] / A. W. Moore, D. Zuev // SIGMETRICS Perform. Eval. Rev. — New York, NY, USA, 2005. — Июнь. — Т. 33, № 1. — С. 50—60. — URL: https://doi.org/10.1145/1071690.1064220.

36. Semantics-Aware Android Malware Classification Using Weighted Contextual API Dependency Graphs [Текст] / M. Zhang [и др.] // Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security. — Scottsdale, Arizona, USA : Association for Computing Machinery, 2014. — С. 1105—1116. — (CCS '14). — URL: https://doi.org/10.1145/2660267.2660359.

37. *Jamil, H. A.* Feature Selection and Machine Learning Classification for Live P2P Traffic [Текст] / H. A. Jamil // Proceedings of the International Conference on Industrial Engineering and Operations Management (IEOM). — 2019. — С. 1—9.

38. Droid-Sec: deep learning in android malware detection [Текст] / Z. Yuan [и др.] // Proceedings of the 2014 ACM Conference on SIGCOMM. — Chicago, Illinois, USA : Association for Computing Machinery, 2014. — С. 371—372. — (SIGCOMM '14). — URL: https://doi.org/10.1145/2619239.2631434.

39. DroidDelver: An Android Malware Detection System Using Deep Belief Network Based on API Call Blocks [Текст] / S. Hou [и др.] //. Т. 9998. — 06.2016. — С. 54—66.

40. *Nix, R.* Classification of Android apps and malware using deep neural networks [Текст] / R. Nix, J. Zhang // 2017 International Joint Conference on Neural Networks (IJCNN). — 2017. — С. 1871—1878.

41. *Saber, A.* Encrypted Network Traffic Identification: LDA-KNN Approach [Текст] / A. Saber, B. Fergani, M. Abbas // Proceedings of the 9 ème édition du colloque Tendances dans les Applications Mathématiques en Tunisie Algérie et Maroc. — 02.2019. — С. 1—3.

42. *Nigmatullin, R.* Accumulated Generalized Mean Value - a New Approach to Flow-Based Feature Generation for Encrypted Traffic Characterization [Текст] / R. Nigmatullin, A. Ivchenko, S. Dorokhin //. — 01.2021. — С. 165—169.

43. *Nigmatullin, R.* A Novel Approach to Radiometric Identification [Текст] / R. Nigmatullin, S. Dorokhin, A. Ivchenko // Machine Learning and Artificial Intelligence. — IOS Press, 12.2020. — URL: http://dx.doi.org/10.3233/FAIA200806.

44. Differentiation of Different Sorts of Sugars by the CAPoNeF Method [Текст] / R. R. Nigmatullin [и др.] // Electroanalysis. — 2021. — Сент. — Т. 33, № 12. — С. 2508—2515. — URL: http://dx.doi.org/10.1002/elan.202100291.

45. Building a standard measurement platform [Текст] / M. Bagnulo [и др.] // IEEE Communications Magazine. — 2014. — Т. 52. — С. 165—173.

46. *Bajpai, V.* A Survey on Internet Performance Measurement Platforms and Related Standardization Efforts [Text] / V. Bajpai, J. Schonwalder // IEEE Communications Surveys. — 2015. — Vol. 17, no. 3. — P. 1313—1341.

47. Measuring home broadband performance [Текст] / S. Sundaresan [и др.] // Communications of the ACM. — 2012. — Нояб. — Т. 55, № 11. — С. 100—109. — URL: http://dx.doi.org/10.1145/2366316.2366337.

48. RIPE Atlas: A Global Internet Measurement Network [Text] / M. Candela [et al.] // The Internet Protocol Journal. — 2015. — Jan. — Vol. 18.

49. Turrs Devices [Electronic Resource]. — URL: www.turris.com (visited on 08/15/2024).

50. *Bajpai, V.* Managing SamKnows probes using NETCONF [Text] / V. Bajpai, R. Krejčí // 2014 IEEE Network Operations and Management Symposium (NOMS). — 2014. — P. 1—2.

51. Developing Realistic Distributed Denial of Service (DDoS) Attack Dataset and Taxonomy [Text] / I. Sharafaldin [et al.] // 2019 International Carnahan Conference on Security Technology (ICCST). — 2019. — P. 1—8.

52. Permeability prediction and uncertainty quantification base on Bayesian neural network and data distribution domain transformation [Текст] / L. MingXuan [и др.]. — 01.2023.

53. Scikit-learn: Machine Learning Without Learning the Machinery [Text] / G. Varoquaux [et al.] // GetMobile: Mobile Computing and Communications. — 2015. — June. — Vol. 19, no. 1. — P. 29—33. — URL: http://dx.doi.org/10.1145/2786984.2786995.

54. *Sakoe, H.* Dynamic programming algorithm optimization for spoken word recognition [Текст] / H. Sakoe, S. Chiba // IEEE Transactions on Acoustics, Speech, and Signal Processing. — 1978. — Т. 26, № 1. — С. 43—49.

55. *Ge, Z.* GNN-based End-to-end Delay Prediction in Software Defined Networking [Текст] : Master's thesis / Ge Zhun. — Ottawa, Canada : University of Ottawa, 2022. — Thesis submitted in partial fulfillment of the requirements for the Master of Applied Science Electrical and Computer Engineering degree, School of Electrical Engineering and Computer Science, Faculty of Engineering.

56. Probabilistic Delay Forecasting in 5G Using Recurrent and Attention-Based Architectures [Текст] / S. Mostafavi [и др.]. — 2025. — URL: https://arxiv.org/abs/2503.15297.

Изюмов Павел Сергеевич

Разработка и исследование алгоритмов анализа сетевой инфраструктуры и интернет-трафика в условиях ограниченных вычислительных мощностей

Автореф. дис. на соискание ученой степени к. т. н.

Подписано в печать ___.___._____. Заказ № _____
Формат 60×90/16. Усл. печ. л. 1. Тираж 100 экз.
Типография _____