На правах рукописи

Мазурок Дмитрий Валерьевич

АЛГОРИТМЫ ГЛУБОКОГО МАШИННОГО ОБУЧЕНИЯ В СИСТЕМАХ АНАЛИЗА СЕТЕВОГО ТРАФИКА

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

АВТОРЕФЕРАТ
диссертации на соискание ученой степени
кандидата технических наук

Владимир – 2026

Работа выполнена на кафедре «Информатика и защита информации» в ФГБОУ ВО «Владимирский государственный университет имени Александра Григорьевича и Николая Григорьевича Столетовых» (ВлГУ).

Научный руководитель: **Монахов Михаил Юрьевич**
д.т.н., профессор, заведующий кафедрой «Информатика и защита информации» ФГБОУ ВО «Владимирский государственный университет имени Александра Григорьевича и Николая Григорьевича Столетовых», г. Владимир.

Официальные оппоненты: **Надеждин Евгений Николаевич**
доктор технических наук, профессор, профессор кафедры информационных технологий и систем ФГАОУ ВО «Российский государственный гуманитарный университет», г. Москва.

Пузанов Андрей Викторович
к.т.н., доцент, ведущий научный сотрудник АО «ВНИИ «Сигнал»», г. Ковров

Ведущая организация: Федеральное государственное автономное образовательное учреждение высшего образования «Омский государственный технический университет», г. Омск

Защита диссертации состоится «_13_» _мая_ 2026 года в 16 часов на заседании диссертационного совета 24.2.281.04 при ВлГУ имени А.Г. и Н.Г. Столетовых по адресу: 600000, г. Владимир, ул. Белоконская, д3, корп. 2, ауд.408-2 ВлГУ.

С диссертацией можно ознакомиться в библиотеке ВлГУ по адресу г.Владимир, ул. Горького, 87, корпус 1, ВлГУ и на сайте http://diss.vlsu.ru.

Отзывы на автореферат в двух экземплярах, заверенные печатью, просим направлять по адресу: 600000, г. Владимир, ул. Горького, 87, ВлГУ, ученому секретарю диссертационного совета 24.2.281.04. Тел. (4922) 47-97-46, E-mail: telnyy@vlsu.ru.

Автореферат разослан «_10_» _марта_ 2026 г.

Ученый секретарь диссертационного совета,
к.т.н., доцент А.В. Тельный

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность темы. В настоящее время сеть выполняет роль критически важной шины данных, обеспечивая взаимодействие множества устройств и сервисов, генерирующих значительный объем трафика. Необходимость поддержания таких характеристик сети, как доступность, производительность, стабильность обуславливает потребность в эффективных инструментах анализа сетевого трафика.

Несмотря на различие задач систем анализа сетевого трафика, все они являются неотъемлемой частью процесса принятия решений. Так, системы мониторинга производительности сети позволяют вовремя принять меры по увеличению вычислительной мощности оборудования, не справляющегося в пиковые нагрузки. Системы обнаружения утечек данных позволяют своевременно предотвращать нарушения законодательства, снижать репутационные риски и минимизировать финансовые потери, связанные с утечкой коммерческой тайны.

По мере увеличения сложности и объема информации, стандартные методы обработки и анализа данных, такие как корреляционный анализ, анализ временных рядов и другие, становятся недостаточно эффективными. Возникает потребность в более современных, высокотехнологичных и эффективных подходах. Одним из наиболее перспективных подходов является использование нейронных сетей.

Применение нейронных сетей позволяет, в большинстве случаев, добиться большей точности, производительности и улучшения других важных метрик, используемых при решении конкретных задач. Однако, повышая данные метрики, при использовании нейронных сетей приходится жертвовать такими свойствами, как прозрачность, или интерпретируемость решений нейросетевого классификатора, критически важным при принятии решений; робастность, или устойчивость к вариативности входных данных и состязательным атакам. Кроме того, при изменении распределения данных, например, ввиду изменения аппаратной конфигурации, требуется своевременно адаптировать модель.

В настоящее время, в большинстве случаев применения нейросетевых классификаторов в системах анализа сетевого трафика не уделяется должного внимания данным аспектам, что актуализирует тему настоящего исследования.

Степень разработанности темы исследования. Проблемами повышения эффективности принятия решений занимались такие ученые, как Бородачёв С.М., Романова Н.А., Свирина Л.Н., Михайлов Г.С., Абрахам Вальд, Даниэль Канеман, Амос Тверски. Проблемами интерпретации нейросетевых решений занимались такие ученые, как Маршаков Д.В., Гавриков М.М., Amirata Ghorbani, Abubakar Abid, James Y. Zou, David Alvarez-Melis, T. Jaakkola, Kopf L., Nakajima S., Kloft M., Höhne M.M., Barberan C.J., Balestriero R., Baraniuk R.G. и другие. Применение в сетях методов интеллектуального анализа данных, в частности, нейронных сетей, исследовали такие ученые, как Комашинский В.И., Смирнов Д.А., Амосов О.С., Амосова С.Г., Иванов Ю.С., Коцыняк М. А., Карпов М. А.,

Лаута О.С., Дементьев В. Е., Riyazahmed A.J., Resende P.A.A., Drummond A.C., Haripriya A.P., Kulothungan K., Tront, J. G., Marchany, R. и другие. Проблемой робастности и проблемой состязательных атак нейросетевых классификаторов занимались такие ученые, как Чехонина Е.А., Жуков В.Г., Колесниченко М.Д., Лапина М.А., Ржевская Н.В., Котляров Д.В., Дюдюн Г.Д., Jandrik Lana, Shuming Jiao, Z. Song, S. Xiang, Ian Goodfellow, W. Zhang, Quan Z. Sheng, A. Alhazmi, Chenliang Li, Shuyang Gu и др.

Объект исследования – процесс поддержки принятия решений в системах анализа сетевого трафика.

Предметом исследования – алгоритмы глубокого машинного обучения, используемые в системах анализа сетевого трафика.

Методология и методы исследования. Научные положения работы теоретически обосновываются при помощи аппарата теории вероятностей, математической статистики, математического анализа, технологии объектно-ориентированного программирования. Для экспериментальной проверки работоспособности предложенных алгоритмов использовалось разработанное программное обеспечение.

Цель исследования – решение научной задачи повышения эффективности принятия решений в системах анализа сетевого трафика, работающих на основе алгоритмов глубокого машинного обучения.

В связи с поставленной целью решались следующие **задачи**:

1. Проанализировать существующие модели и алгоритмы глубокого машинного обучения, используемые в задачах анализа сетевого трафика.
2. Разработать модели и алгоритмы повышения интерпретируемости нейросетевого классификатора анализатора сетевого трафика.
3. Разработать алгоритм оценки функциональной устойчивости нейросети анализатора сетевого трафика в условиях искажений данных.
4. Выполнить экспериментальное исследование разработанных моделей и алгоритмов.

Научная новизна проведенных исследований и полученных в ходе работы результатов заключается в следующем:

1. Разработан алгоритм оценки интерпретируемости результатов работы нейросети, отличающийся от существующих использованием положительных заключений модели и новой метрики точности, что позволяет численно сравнивать различные методы интерпретации решений нейросети анализатора сетевого трафика в задачах бинарной классификации редких событий.
2. Разработан алгоритм оценки устойчивости нейросети анализатора сетевого трафика, отличающийся от существующих использованием аугментации тестовой выборки при помощи вариационного автокодировщика с условием (cVAE) для проведения расчетов с учетом степени переобучения, позволяющий оценить пригодность использования классификатора в условиях зашумленности данных.

3. Предложен модифицированный в части представления и анализа данных алгоритм интерпретации решений классификатора анализатора сетевого трафика, отличающийся от существующих использованием большой языковой модели (LLM), что позволяет корректировать ошибочные решения модели.

Положения, выносимые на защиту:

1. Расширенный метод интерпретации позволяет снизить ошибку первого рода и повысить общую точность системы обнаружения сетевых вторжений на основе нейросетевого классификатора в среднем на 20% и более.
2. Использование LLM и мультиагентного подхода в совокупности с расширенным методом интерпретации позволяет повысить общую точность классификации нейросети NIDS без привлечения человека на 23% и более.
3. Методика оценки робастности нейросети IDS позволяет оценить стабильность точности работы модели в условиях зашумленности входных данных.

Теоретическая значимость работы заключается в предложенном подходе оценки интерпретируемости нейросетевого классификатора трафика, позволяющем производить численное сравнение разных методов интерпретации.

Предложенный в работе алгоритм оценки устойчивости нейросетевых классификаторов позволяет выбрать наиболее стабильную в реальных условиях эксплуатации модель.

В диссертационной работе предложено совершенствование существующего метода интерпретации, а также расширено применение больших языковых моделей, что в совокупности позволяет корректировать ошибки классификации.

Практическая значимость работы. Предложен модифицированный в части представления результатов алгоритм интерпретации решений нейросетевого классификатора сетевой IDS, использующий LLM, а также разработанное и зарегистрированное ПО для его применения, ПО для проведения эксперимента по оценке повышения интерпретируемости решений нейросетевого классификатора, модуль оценки значимости результатов эксперимента. Кроме того, были разработаны модуль оценки актуальности факторов защищенности детектора и классификатора и выдачи адаптивных рекомендаций по повышению защищенности, модуль тестирования скорости работы и точности классификаторов, модуль аудита нейронных сетей на подверженность состязательным атакам с графическим интерфейсом.

Предложенные в работе алгоритмы и средства позволяют повысить качество нейросетевого классификатора от 1,23 до 4,4 раз в сравнении с типовой реализацией классификатора, что подтверждается проведенным экспериментальным исследованием.

Достоверность и апробация. Степень достоверности результатов исследования подтверждается: рядом экспериментов, проводимых на исследуемых моделях с соблюдением требуемых условий случайности, выполненных на экспериментальных установках; положительным результатом

практического использования разработанных средств, а также апробацией в печати и на научных конференциях различного уровня. Полученные результаты были представлены на международной конференции 2020 Ural Symposium on Biomedical Engineering, Radioelectronics and Information Technology (USBEREIT), международных научно-практических конференциях: II «Наука, технологии и инновации: стратегии развития в современном мире», II «Исследования и инновации: синергия знаний и практики», IX и VIII «Наука и глобальные вызовы: перспективы развития», II «Европейские научные исследования», «Инновационное развитие современной науки: проблемы, закономерности, перспективы»; VIII Всероссийской научно-практической конференции «Информационные технологии и автоматизация управления». Научно-практическая значимость работы подтверждена рецензируемыми публикациями в журналах и в сборниках научных трудов, докладами на научных конференциях международного и российского уровня, свидетельствами о государственной регистрации программ для ЭВМ.

Практическая значимость работы подтверждена внедрением её результатов в ООО «ВОЙС КОММЬЮНИКЕЙШН» г. Москва, ООО «ЮКИТЕХ ЛАБ» г. Москва, ООО «НТЦ «СИСТЕМИНВЕСТ» г. Москва, ООО «АйТиАрт» г. Москва.

Публикации. По результатам диссертационной работы опубликовано 14 научных работ, в том числе в международных базах Scopus и Web of Science – 1, в изданиях, рекомендованных ВАК – 3, получено 7 свидетельств о государственной регистрации программы для ЭВМ.

Личный вклад. Все результаты, изложенные в научно-квалификационной работе, получены автором лично. Постановка цели и задач, обсуждение планов исследований и полученных результатов выполнены совместно с научным руководителем.

Соответствие паспорту специальности. Проблематика, исследованная в диссертации, соответствует пунктам 2, 3, 12 паспорта специальности 2.3.1 «Системный анализ, управление и обработка информации, статистика».

Структура и объем диссертационной работы. Диссертация состоит из введения, трех глав, заключения, списка обозначений и сокращений, списка использованных источников из 123 наименований, 9 приложений и содержит 116 страниц основного текста, иллюстрированного 27 рисунками, содержит 4 таблицы.

СОДЕРЖАНИЕ РАБОТЫ

Во введении обосновывается актуальность темы диссертации сформулированы цель, задачи, объект и предмет исследования, а также научная новизна, теоретическая и практическая значимость полученных результатов.

В первой главе анализируются сферы использования нейросетевых анализаторов в компьютерных сетях, изучаются проблемы, возникающие при использовании нейросетей, а также описываются составляющие качества нейросетевых классификаторов, используемых при принятии решений в

системах анализа сетевого трафика. Уточняются задачи исследования. Анализ сетевого трафика производится в следующих системах: системы обнаружения и предотвращения вторжений (IDS / IPS); системы мониторинга производительности сети и др. Применение нейронных сетей в данных системах призвано повысить точность работы, но в то же время это снижает интерпретируемость решений. Кроме того, нейросети часто не робастны, подвержены состязательным атакам и требуют периодической адаптации.

Эффективность принятия решений можно оценить через качество решения и скорость принятия решений. Со стороны системы поддержки принятия решений (СППР), на эффективность принятия решений влияют *факторы*: полнота информации, релевантность выданной информации, достоверность выданной информации, актуальность и скорость выдачи информации, доступность для понимания представленной информации.

Качество принятия решений можно оценить через метрику точности, т.е. отношение числа оптимальных решений к общему числу принятых решений. В работе используется устойчивая к дисбалансу классов метрика точности - F1-мера, учитывающая полноту и точность: $F = \frac{2*p*r}{p+r}$, где p – точность, r – полнота.

При расчете скорости принятия решений важно учитывать общее время принятия решения (D_{total_s}) от момента получения сигнала системой поддержки принятия решений (T_E) до фактического принятия решения (T_D): $D_{total_s} = T_D - T_E$.

Влияние СППР на эффективность принятие решения может быть оценено через прокси-метрики. Учитывая, что основу работы СППР составляет нейросетевая модель классификации, качество работы этой модели (P_c) напрямую влияет на представленные выше факторы. Взаимосвязь эффективности принятия решений и СППР представлена на рис. 1. При эксплуатации, системы анализа сетевого трафика постоянно классифицируют данные и получают определенный результат. Как одну из таких систем, возьмем за основу IDS.

В то время как точность модели классификатора IDS высока (*в работе точность обозначается F*), остается шанс ошибки. Нейронная сеть не предоставляет обоснования своего решения. Принятие гипотезы об истинности решения влечет, потенциально, периодическую частичную или полную приостановку работы сети и, как следствие, снижает

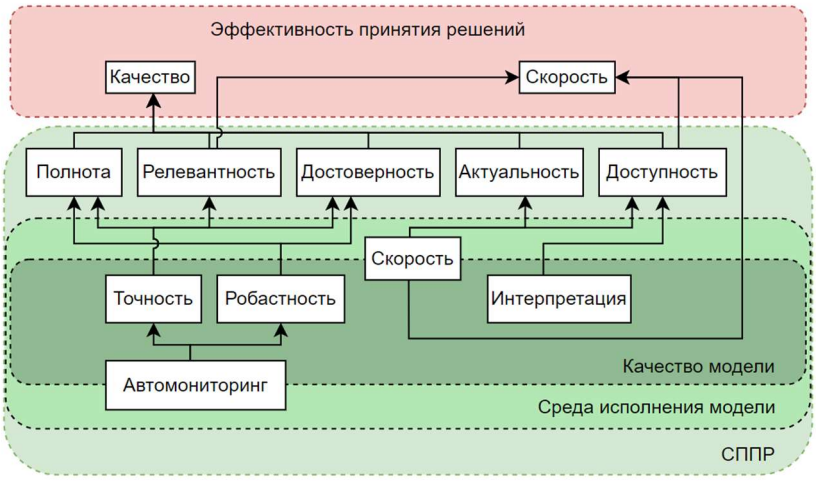

Рисунок 1 – Взаимосвязь эффективности принятия решений и СППР

доступность и ведет к материальным издержкам. Для принятия взвешенного решения требуется получить интерпретацию решения классификатора, что позволит сократить ошибку первого рода. *Оценка интерпретируемости в работе обозначается* I. Время, затрачиваемое на получение интерпретации конкретного решения модели, обозначается в работе T_I. Такая характеристика, как *робастность,* позволит учесть потенциальную подготовленность используемой модели классификатора к возможным искажениям в данных (R), а использование систем мониторинга сдвига (дрейфа) данных (*оценка адаптируемости* A) для определения потребности в адаптации модели позволит стабилизировать точность и робастность модели. Кроме того, важной характеристикой является скорость работы классификатора, зависящая как от используемого в промышленной среде оборудовании, так и конкретной модели и ее реализации. Если модель анализатора трафика обрабатывает данные ниже скорости реального времени, отставание накапливается, перечеркивая практическую пользу от системы в целом.

Таким образом, исходную точность F, оценку интерпретируемости I, оценку робастности R, оценку адаптируемости A, превышение скорости работы модели реальному времени C можно объединить в новое свойство модели – *качество***.** В данном случае, под моделью подразумевается не только физический файл модели, но и программно-аппаратная среда ее исполнения и мониторинга.

Под качественным нейросетевым классификатором в работе понимается такой классификатор, который позволяет получать интерпретацию решений, является в высокой степени робастным, в том числе на состязательных данных, с высокой точностью и достаточной скоростью классификации, при работе которого происходит мониторинг за распределением данных для своевременной адаптации.

Определим итоговую формулу оценки качества классификатора IDS (по аналогии, данный подход применим к оценке классификаторов других систем анализа сетевого трафика): $P_c = C \times \mathrm{F} \times (W_I \times I + W_R \times R + W_A \times A)$, где W_I, W_R, W_A – веса оценок интерпретации, робастности и адаптируемости соответственно. Общая сумма весов составляет 1; в рамках данного исследования все веса равны $\frac{1}{3}$. T_e определяет оценку повышения времени получения результатов. При этом важно, чтобы время работы модели не превышало реальное время, в то время как общее время работы СППР может его превышать ввиду редкости события.

Максимизация P_c напрямую влияет на качество и скорость принятия решений, и, в свою очередь, на общую эффективность принятия решений в системах анализа сетевого трафика, работающих на основе алгоритмов глубокого машинного обучения. Таким образом, задача исследования заключается в максимизации качества классификатора (P_c). Ограничения, используемые в работе: общее время принятия решения $\Delta D_{total_s} \leq 30$ секунд; Исходная точность $0{,}85 \leq F \leq 1$; Порог снижения точности при повышении робастности $0 \leq \delta \leq 0{,}05$.

Для повышения качества классификатора (P_c) предлагается:

1. Реализовать подход расширенной интерпретации решений модели классификатора, позволяющий оператору на основе анализа причин принятия конкретного решения снизить ошибку первого рода, максимизируя тем самым оценку интерпретируемости.
2. Разработать алгоритм оценки функциональной устойчивости нейросети анализатора сетевого трафика в условиях искажений входных данных, позволяющий отобрать наиболее робастную модель.

Результаты анализа проблем состязательных воздействий, робастности, интерпретации, мониторинга и адаптации нейросетей, определение качества классификатора и его составляющих получены автором лично и были частично опубликованы, в т. ч. в соавторстве (вклад автора более 80%) в [4, 7, 6, 9, 11].

Вторая глава содержит разработку алгоритма оценки интерпретируемости результатов работы нейросетевого анализатора трафика, алгоритма взаимодействия агентов большой языковой модели как части интерпретации, алгоритма оценки адаптируемости. В главе приводится разработанный в работе расширенный подход к интерпретации, описывается порядок и условия проведенных экспериментов и анализ результатов.

Рассмотрим типовой IDS, выступающей в роли СППР при принятии решений относительно факта атаки. IDS непрерывно анализирует трафик, при этом нейросетевая модель классифицирует сетевые соединения в режиме окна со смещением в 1 секунду. Абсолютное большинство примеров при классификации фактически являются не атакой, при этом IDS не выдает никаких сообщений оператору при корректной классификации. В случае, если модель определяет конкретное соединение как атаку, происходит выдача сообщения оператору – сетевому администратору. Оператор на основе имеющихся данных (в базовом случае – это только сообщение от IDS) должен принять решение, например, ничего не предпринимать (ложное срабатывание), разорвать соединение, временно отключить сеть и др. От характеристик поступившей информации оператору напрямую зависит эффективность принятого решения. Опираясь на это, в работе предлагается следующий алгоритм:

Алгоритм оценки интерпретируемости результатов работы нейросетевого анализатора трафика

Шаг 1. Обучить нейросетевую модель NIDS. Записать точность модели F.

Шаг 2. Отобрать данные с ложно и истинно положительными ответами модели.

Шаг 3.1 Сформировать выборку испытуемых с компетенциями, подобными оператору, принимающему решение на основе IDS (не менее 32 человек); провести инструктаж по работе с интерпретацией и действиями: выражением согласия или несогласия с ответом модели на основе имеющихся данных для оценки *I*.

Шаг 3.2 Выбрать LLM: GPT4, либо аналогичную, для оценки I_{LLM}.

Шаг 4. Используя исправленные ответы, рассчитать скорректированную точность модели $\hat{F}$ (усреднить для нескольких участников).

Шаг 5. Оценить значимость отличия точностей при помощи одновыборочного t-теста с альтернативной гипотезой о повышении точности при использовании метода интерпретации. В случае, если критическое t-значение выше рассчитанного, метод интерпретации принимается неэффективным в соответствии с выбранным уровнем значимости.

Шаг 6. Оценить прирост интерпретируемости по формуле:

$$I\ (I_{LLM}) = \max\ (0,\ \frac{\hat{F}\ - F}{1 - F})$$

Конец.

В базовом случае, когда изначально отсутствуют методы интерпретации, предполагается, что $I = 0,$ т. к. $F = \hat{F}$ ввиду невозможности принять взвешенное решение на основании лишь ответа модели, и рациональным подходом является полное согласие с ответами модели.

Предложенный алгоритм может быть использован для сравнения различных подходов интерпретируемости, в таком случае F может представлять собой скорректированную оценку точности, полученную в эксперименте с исходным методом интерпретации.

Для оценки адаптируемости предлагается следующий алгоритм.

Алгоритм оценки адаптируемости

Шаг 1. Разделить выборку на тренировочную и тестовую (при возможности - по времени) в соотношении 70/30.

Шаг 2. Для каждого признака из обучающего набора оценить изменение в распределении при помощи теста Колмогорова-Смирнова (Хи-квадрат для категориальных признаков).

Шаг 3. Для полученных значений рассчитать $p = \sup\left(\left[1 - p_{value_i}\right]\right)$ где i обозначает *i*-й признак.

Шаг 4. Оценить адаптируемость модели по формуле*:* $A = \max(\hat{A}, \frac{-log(p+e)}{-log(e)})$,

где $\hat{A}$ - флаг-индикатор наличия системы автомониторинга.

Конец.

В качестве тестируемого метода интерпретации был разработан и реализован в виде ПО следующий подход. За основу был выбран метод IntegratedGradients (IG) ввиду высокой скорости и точности работы.

Для сокращения времени, необходимой для понимания результатов интерпретации, в том числе, для большего понимания произошедшей ситуации, дополнительно к результатам IG были добавлены:

— окрашивание в оттенки красного / зеленого цветов в зависимости от значения важности признака и класса (атака / не атака);

— среднее значение данного признака в подобном случае при отсутствии атаки с окрашиванием значения в оттенки синего в случае сильного

отклонения (более двух стандартных отклонений) текущего значения от среднего;

— поле «Заметки», содержащее дополнительную информацию, зависящую от отличия текущего значения признака от среднего значения, рассчитанного для аналогичного случая.

При необходимости оператор может обратиться к большой языковой модели GPT4o, отвечающей на вопрос, действительно ли произошла атака, а также предоставляющей возможность контекстного взаимодействия.

Для повышения качества ответа большой языковой модели был применен мультиагентный подход, взаимодействие которого описывает алгоритм:

Алгоритм взаимодействия агентов LLM для повышения интерпретации

Дано 3 агента: Агент 1 (DLExpert), Агент 2 (NetSecExpert), Агент 3 (Judger).

Шаг 1. Подобрать промпт 1 для Агента 1 таким образом, чтобы Агент 1 решал задачу изучения существующего набора признаков, анализа интерпретации, определения наиболее характерных комбинаций признаков и важности. Агент должен учитывать возможность модели ошибаться на редковстречаемых комбинациях данных и др.

Шаг 2. Подобрать промпт 2 для Агента 2 таким образом, чтобы Агент 2 решал задачу изучения комбинаций признаков с точки зрения возможных сетевых атак, определял характерность подобного набора для каких-либо потенциальных атак.

Шаг 3. Подобрать промпт 3 для Агента 3 таким образом, чтобы Агент 3 решал задачу анализа представленных заключений агентов 1 и 2 и генерации окончательного обоснованного заключение. Формат ответа предполагает удобство извлечения итогового решения большой языковой модели автоматически.

Шаг 4. Передать Агенту 1 промпт 1, данные, значения интерпретаций, ответ модели. Записать ответ.

Шаг 5. Передать Агенту 2 промпт 2, данные, ответ модели. Записать ответ.

Шаг 6. Передать Агенту 3 промпт 3, ответы агентов с Шага 4 и Шага 5.

Шаг 7. Передать ответ Агента 3 оператору; извлечь согласие/несогласие из ответа для расчета метрик.

Конец.

Для экспериментального исследования было собрано 3 стенда:

1. Стенд обучения модели: Ubuntu 18.04, видеокартой NVIDIA RTX 2080TI, ЦП Intel Core I7 8700, 32 Гб оперативной памяти DDR3.
2. Стенд подготовки данных и произведения статистических расчетов: ЦП AMD Ryzen 7950X, 64GB DDR5, графическим ускорителем NVIDIA RTX 3090.
3. Стенд проведения экспериментов: выделенный виртуальный сервер провайдера SebekVPS. Сервер использует 4 CPU ядра AMD Ryzen 3900 с производительностью 4,3 Ггц, 8 Гб DDR4, операционной системой Ubuntu.

Для проверки разработанных алгоритмов и методов оценки использовался набор данных CSE-CICIDS2018. Итоговый набор, содержащий 16232943 строк данных, включает 45 признаков. Обучающий, тестовый и валидационный наборы содержат 12986354, 1623295, 1623294 строк соответственно. Используются классы «нет атаки» (benign) и «атака» (attack). В качестве архитектуры модели выступает полносвязная многослойная модель прямого распространения (DNN) с 4 слоями размерностью 45-128-512-128-2 и пакетной нормализацией после каждого скрытого слоя. Исходный F1-score – 0,9875.

Перед проведением эксперимента установлены ограничения: $\Delta D_{total_s} \leq 30$ секунд; $0{,}9 \leq F \leq 1$. Уровень значимости выбран равным 0,05.

Для проведения эксперимента были приглашены 60 студентов ВлГУ кафедры «Информатика и защита информации», обладающие компетенциями, схожими с компетенциями типового сетевого инженера. Было отобрано 60 вопросов – 50% ложноположительные, 50% истинно положительные. Точность модели на данной выборке F1-score 0,667.

В результате эксперимента были собраны данные и проведен t-тест. Среднее значение F1-score среди испытуемых составило 0,801 (повышение на 20,09% по сравнению с базовым случаем), медианное 0,813. При этом точность ответов LLM составила 0,8065. Значение t-статистики составило 30,358, что указывает на статистически значимое отличие средней точности при использовании предложенного подхода интерпретации (рис. 2) и свидетельствует о повышении интерпретируемости ответов нейросети. Далее было проведено тестирование предложенного мультиагентного подхода при анализе результатов интерпретации, что позволило получить итоговую оценку точности F1=0,826 (на 23,8% выше исходной) при среднем времени на ответ в 5,47 секунд. Полученные в ходе эксперимента оценки интерпретируемости: I=0,403, I_{LLM}=0,419 и I_{LLM}=0,478 при применении только мультиагентного подхода LLM без человека.

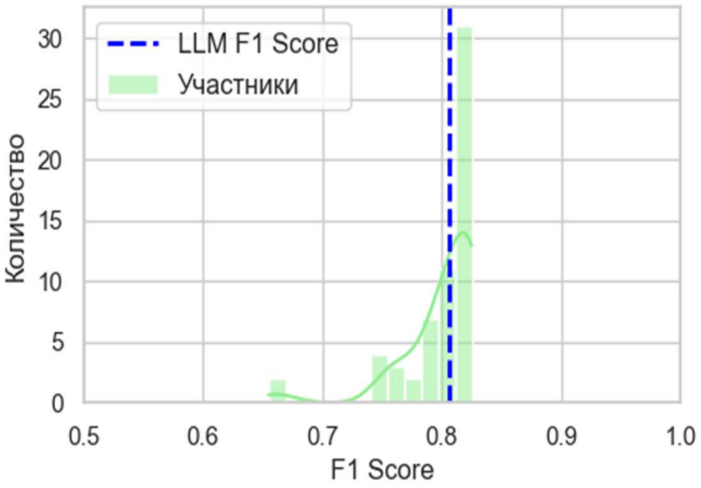

Рисунок 2 – F1-score участников

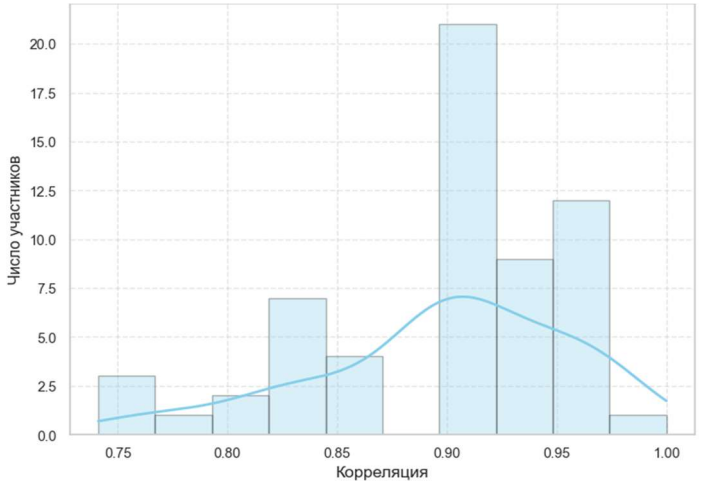

Рисунок 3 – Корреляция с LLM

Таким образом, в ходе эксперимента применение разработанного подхода интерпретации значительно повысило интерпретируемость с 0 до 0,419 (и до 0,478 в случае мультиагентного подхода). Дополнительным фактом, установленном в ходе эксперимента, стал корреляционный анализ ответов участников и ответов LLM: большинство участников практически полностью повторили ответы большой языковой модели. Средняя корреляция Мэтьюса составила 0,898 (рис. 3), что является статистически значимым. Это в совокупности с более высокой точностью и скоростью работы свидетельствует о возможности исключения из цепочки принятия решений человека. Среднее время ответа участников составило 24,3 секунд, что удовлетворяет начальным требованиям (рис. 4).

Данные результаты получены автором лично и были частично опубликованы в соавторстве (вклад автора более 80%) в [7, 10].

В третьей главе предлагается алгоритм оценки функциональной устойчивости нейросети анализатора сетевого трафика в условиях искажений входных данных. Описывается порядок и условия проведенных экспериментов и анализ результатов. Природу искажений данных можно разделить на случайные и злонамеренные. Актуальными искажениями являются случайные, причиной которых является изменение распределения данных, помехи в работе источников данных, изменения среды эксплуатации модели.

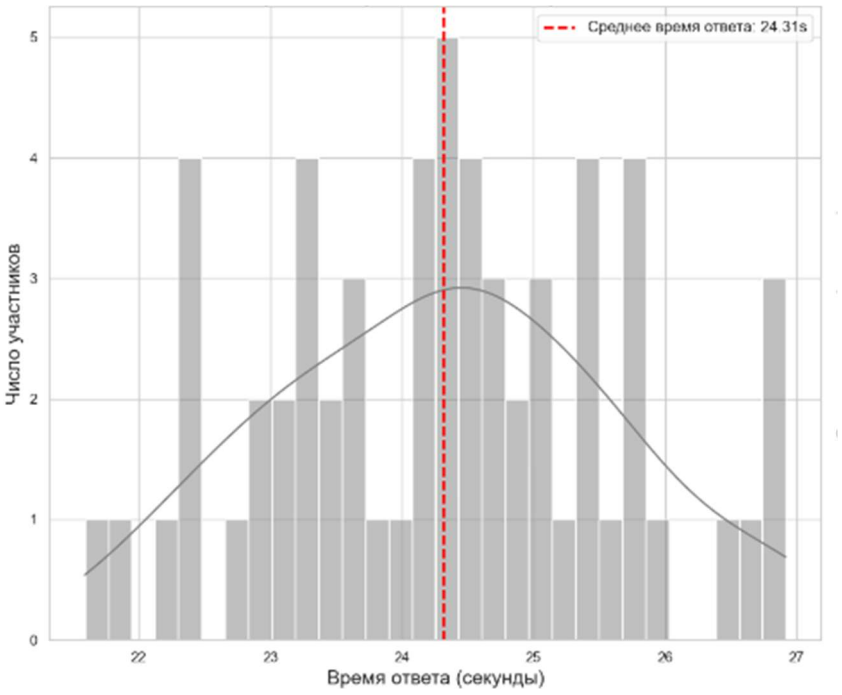

Рисунок 4 – Распределение времени на ответ

К злонамеренным искажениям относятся, прежде всего, состязательные атаки. Проведение классических состязательных атак значительно затруднено ввиду природы сетевых данных. Так, например, число пакетов, используемый протокол и другие – физически получаются из сетевого оборудования, часто имеют ограниченный набор значений. В то же время актуальной остается состязательная атака, заключающаяся в «отравлении» обучающей выборки, однако данный тип состязательной атаки не вносит шум в классифицируемые данные и находится за пределами исследований настоящей работы.

Робастностью, или функциональной устойчивостью модели нейросети анализатора сетевого трафика называется способность сохранять достигнутые в ходе обучения и тестирования точностные характеристики моделью при

внедрении ее в промышленную эксплуатацию при возможных искажениях в данных, вызванных изменением в распределении данных, помехами сетевого оборудования, изменениями в среде эксплуатации модели. Робастность бывает локальная и глобальная (относится к одному / всем примерам данных).

Алгоритм оценки глобальной робастности модели - базовый:

Шаг 1. Получить тестовую выборку размера не менее M элементов, не использованную при обучении классификатора.

Шаг 2. Выбрать подходящий под тип данных метод аугментации.

Шаг 3. Произвести классификацию тестовых данных и оценить точность *S*.

Шаг 4. Для элементов тестовой выборки сформировать N аугментированных примеров с выбранной силой аугментации - магнитудой (при возможности).

Шаг 5. Провести классификацию полученных на шаге 4 данных (без учета исходных данных) и оценить точность классификации по выбранной метрике S.

Шаг 6. Произвести расчет оценки робастности для модели.

Конец.

Значение для N выбирается 1 и более. Значение для M обычно равно размеру валидационной или тестовой выборки.

При создании искажений в данных для оценки робастности важно учитывать природу данных. Так, недопустимо, чтобы генератор шума создавал примеры с вещественным числом пакетов, либо использовал несовместимые с определенным протоколом флаги. Учитывая это, в настоящей работе предлагается в качестве генератора зашумленных синтетических данных использовать cVAE - вариационный автокодировщик с условием. Для оценки робастности предлагается следующая формула: $R = \frac{\min(S,\hat{s})}{S} * (1 - R_{overfit})$, где где S – исходная точность, $\hat{s}$ - точность на аугментированных данных.

$R_{overfit} = \frac{\max(0;F_{train}-F_{test})}{F_{train}}$, где F – точность на обучающем или тестовом наборе.

В экспериментах используется метрика точности F1. Для аугментации данных в исследовании предлагается использовать следующий алгоритм:

Алгоритм аугментации данных

1. Обучение модели cVAE высокой точности
2. Генерация N данных:
    a. Кодирование данных для получения векторов среднего (mu) и логарифма дисперсии, к которому добавляется шум из z-распределения с выбранной магнитудой. Категориальные данные, включая метку класса – используются как условия.
    b. Вектор z сэмплируется из нормального распределения с параметрами mu и std, при помощи повторной параметризации
    c. Денормализация данных.

d. Расчет зависимых признаков, округление с учетом типов данных.
e. Нормализация данных.

Конец.

Для проверки разработанного алгоритма требуется определить подходы, доказанно повышающие робастность нейросетевой модели. Среди них: использование регуляризации, дроп-аут, пакетная нормализация, балансировка классов, в т.ч. через распределение весов. Предполагается, добавление в исходную (базовую) архитектуру модели перечисленных методов, оценка робастности (при ее адекватности) должна статистически значимо повышаться. Кроме того, ожидается, что модели с более высокой оценкой робастности будут демонстрировать более высокую точность на состязательных данных.

Для проверки выдвинутых гипотез был подготовлен стенд с ЦПУ AMD Ryzen 7950X, оперативной памятью 64GB DDR5, графическим ускорителем NVIDIA RTX 3090. Была обучена модель cVAE, содержащая 9 слоев и использующая комбинированную функцию потерь: среднеквадратичную ошибку и дивергенцию Кульбака-Лейблера. По завершении обучения итоговое значение функции потерь составило 8,42. В качестве базовой архитектуры выступает полносвязная многослойная модель прямого распространения с 4 слоями.

Было проведено 3 эксперимента, в ходе которых было обучено 153 модели, разбитых на группы: 1 группа состоит из 32 базовых моделей с различными комбинациями методов повышения робастности; 2 группа состоит из 57 базовых моделей, отличающиеся друг от друга используемой инициализацией генератора случайных значений. 3 группа включает 64 модели: по 32 модели с различным числом и шириной слоев, с различными комбинациями методов повышения робастности (группа 1) и без методов повышение робастности (группа 2).

В экспериментах применялись выборки из исходного набора AWS IDS-2018. Модели обучались на наборе 10^6 строк, тестировались на валидационном и тестовом наборах размером 600000 строк. Использовался уровень значимости 0,05.

Односторонний тест Манна-Уитни показал значимо большее значение R в группе 1 ($p = 0{,}00068$). Интересным заключением является наличие среди группы 2 моделей с высокой оценкой робастности R (рис. 5, слева). Еще более

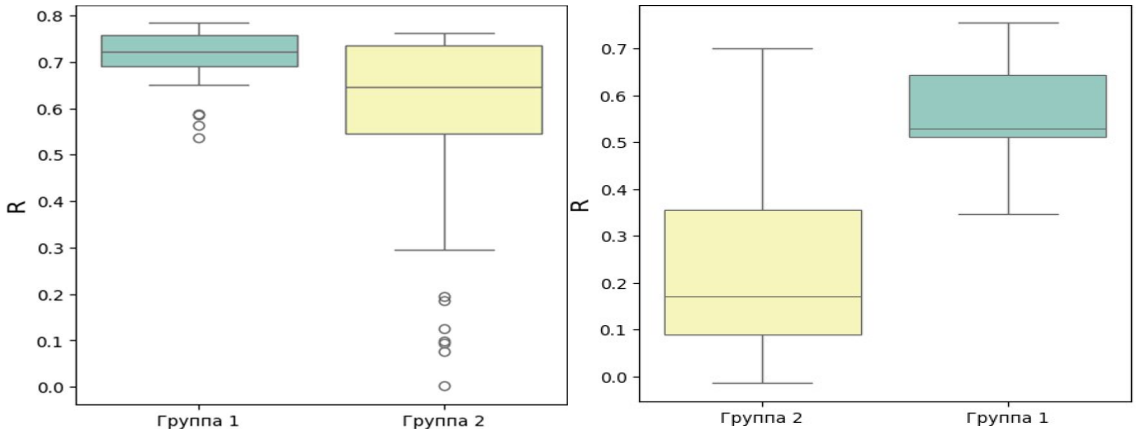

Рисунок 5 – R в группах: одна архитектура (слева) и различные (справа)

выраженное отличие (рис. 5, справа) было получено при использовании моделей с различной архитектурой - значение R значимо больше в группе 1 ($p = 5{,}26 \times 10^{-6}$).

Следующей эксперимент проверял падение уверенности модели в предсказаниях на состязательных данных. Для этого был использован Быстрый Метод Знака Градиента (FGSM) с epsilon=0,04. Ввиду ненормальности данных в выборках была рассчитана корреляция Спирмена, составившая -0,288 в случае одной (рис. 6, слева) и -0,704 в случае различных архитектур (рис. 6, справа), что является значимым ($p = 0{,}005$ и $p = 8{,}46 \times 10^{-11}$ соответственно).

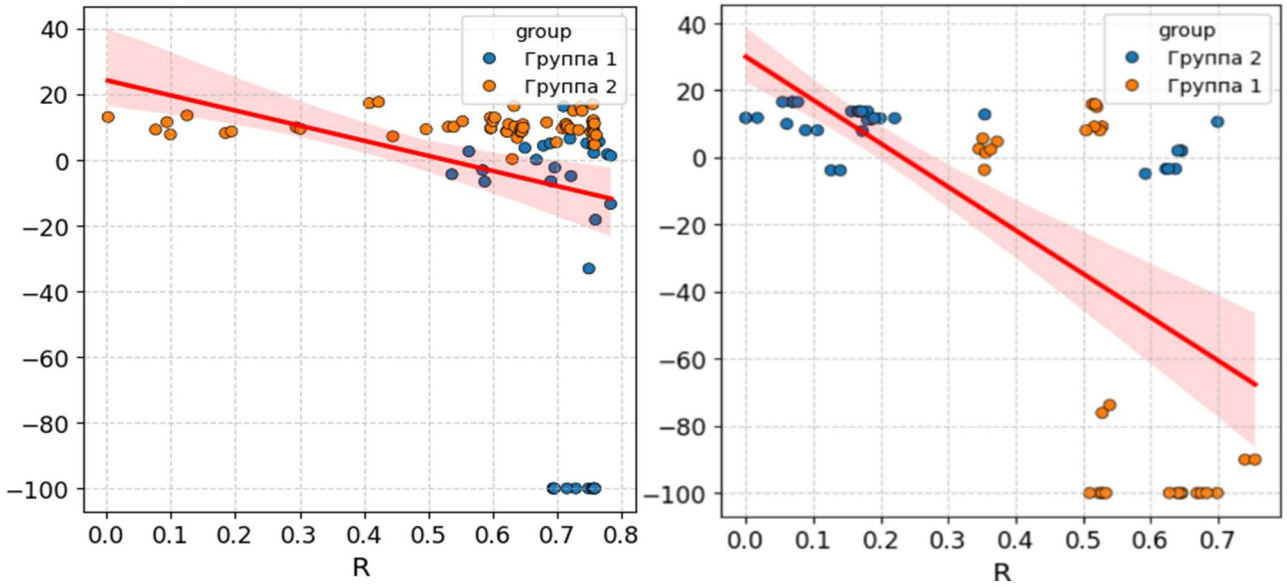

Рисунок 6 – Падение уверенности FGSM: одна архитектура (слева) и различные (справа)

В следующем эксперименте проверялась корреляция между требуемым количеством шагов для смены предсказания на противоположное. Применялся итеративный метод генерации состязательных данных – DeepFool с шагом 0,02. Корреляция Спирмена составила 0,518 в случае одной (рис. 7, слева) и 0,674 в

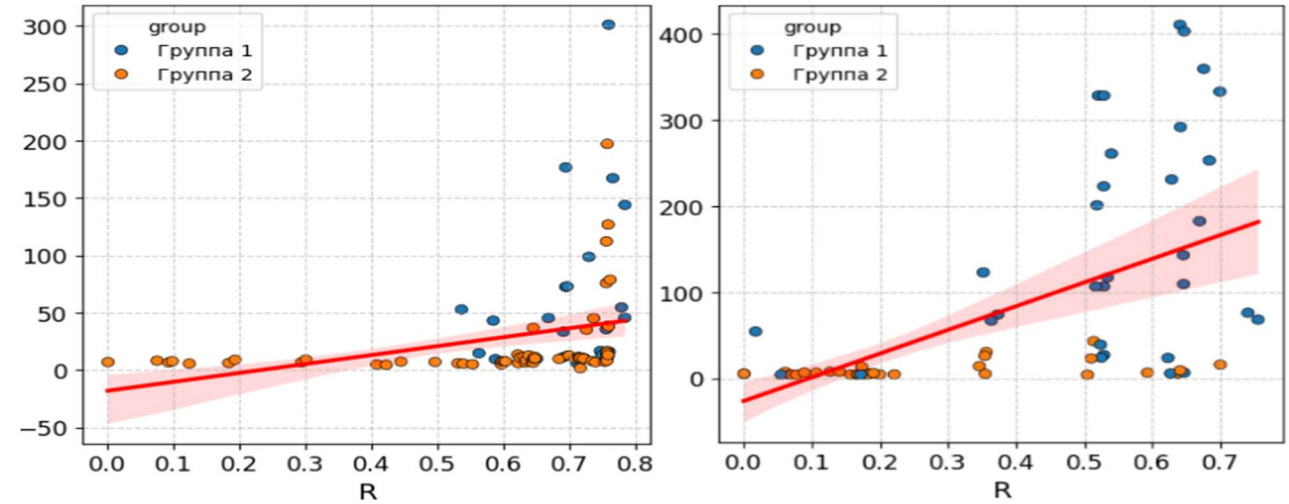

Рисунок 7 – Шаги до смены предсказания, DeepFool: одна архитектура (слева) и различные (справа)

случае различных архитектур (рис. 7, справа), что является статистически значимым ($p = 8{,}46 \times 10^{-11}$ и $p = 1{,}4 \times 10^{-7}$, соответственно).

Таким образом, в ходе экспериментов подтвердилась адекватность предполагаемой формулы оценки функциональной устойчивости нейросети анализатора сетевого трафика.

Оценим повышение качества классификатора, достигнутое в ходе настоящего исследования (табл. 1). В качестве *Базовой модели 1* выступает отобранная перебором параметров в ходе экспериментов полносвязная модель с 4 слоями с размерностями 45-128-512-128-2, пакетной нормализацией после каждого скрытого слоя, имеющая точность на тестовой выборке F1-score 0,9875. Оценка робастности (R) данной модели составляет 0,5105. *Базовая модель 2*– та же базовая модель 1, но при эксплуатации которой применяются системы автомониторинга. *Базовая модель 3* представляет собой случай, когда была случайно выбрана наиболее робастная модель, а также применяется система автомониторинга. *Конечной* моделью является наиболее робастная модель из моделей с высокой точностью, при использовании которой применяется разработанный алгоритм повышения интерпретации решений и применяется мониторинг потребности в адаптации.

Таблица 1 – Расчет качества классификаторов

	Базовая 1	Базовая 2	Базовая 3	Конечная	Конечная LLM	Конечная $LLM_{м.аг.}$
C	Да	Да	Да	Да	Да	Да
F	0,988	0,988	0,984	0,984	0,984	0,984
I	Нет	Нет	Нет	0,403	0,419	0,478
R	0,511	0,511	0,758	0,758	0,758	0,758
A	0,002	Есть сист. автомон.	Есть сист. автомон.	Есть сист. автомон.	Есть сист. автомон.	Есть сист. автомон.
P	0,169	0,497	0,577	0,709	0,714	0,733

Таким образом, в ходе работы удалось повысить качество модели в 4,2 раза и 4,4 относительно базовой модели *1*, на 43% и 48% относительно базовой 2 и на 23% и 27% относительно базовой модели 3 при использовании людей в экспериментах и с применением только большой языковой модели с мультиагентным подходом соответственно.

Все результаты получены автором лично и частично опубликованы в [8], а также в соавторстве (доля автора более 80%) в [11].

В заключении приведены основные результаты и выводы, полученные автором.

В приложениях представлены листинги программ, акты внедрения и свидетельства о регистрации программ для ЭВМ.

Основные результаты диссертационного исследования

1. Проанализированы существующие модели и алгоритмы глубокого машинного обучения, используемые в задачах анализа сетевого трафика. Выделены ключевые проблемы, связанные с использованием нейросетей: отсутствие интерпретации решений, низкая функциональная устойчивость на зашумленных данных, подверженность состязательным атакам, потребность в адаптации. Обоснована связь качества нейросетевого классификатора с качеством СППР и эффективностью принятия решений. Сформулирована задача оптимизации, заключающаяся в максимизации показателя качества классификатора P_c. Результаты анализа проблем состязательных воздействий, робастности, интерпретации, мониторинга и адаптации нейросетей, определение качества классификатора и его составляющих получены автором лично и были частично опубликованы, в т. ч. в соавторстве (вклад автора более 80%) в [4,7,6,9,11].

2. Разработан алгоритм оценки интерпретируемости результатов работы нейросетевого анализатора трафика и алгоритм взаимодействия агентов большой языковой модели. Реализован модифицированный в части представления и анализа данных алгоритм интерпретации решений классификатора анализатора сетевого трафика, использующий большую языковую модель для повышения интерпретации. Созданы экспериментальные стенды и проведены эксперименты, подтверждающие повышение интерпретируемости результатов модели до 0,419 (и до 0,478 в случае мультиагентного подхода). Предложенный метод интерпретации позволяет повысить итоговую точность системы более чем на 20%. Все результаты получены автором лично и были частично опубликованы в соавторстве (вклад автора более 80%) в [7, 10].

3. Разработан алгоритм оценки функциональной устойчивости нейросети анализатора сетевого трафика в условиях искажений входных данных. Создан экспериментальный стенд и проведены эксперименты, подтверждающие адекватность предлагаемого алгоритма и формулы робастности: существует значимая разница средних значений R в группах с использованием повышающих робастность подходов обучения и без, в различных архитектурах. Кроме того, R имеет значимую корреляцию в экспериментах на состязательных данных. Данные результаты получены автором лично и частично опубликованы в [8], а также в соавторстве (доля автора более 80%) в [11].

4. Разработанные в ходе настоящего исследования модели и алгоритмы позволяют повысить качество модели в 4,2 раза и 4,4 относительно базовой модели 1, на 43% и 48% относительно базовой модели 2 и на 23% и 27% относительно базовой модели 3 при использовании людей в экспериментах и с применением только большой языковой модели с мультиагентным подходом соответственно.

Перспективы дальнейшего развития темы. В рамках дальнейших исследований планируется доработка процесса обучения классификаторов трафика с проверкой робастности для автоматизации отбора оптимальной

модели, а также продолжение исследования использования LLM для повышения точности моделей.

СПИСОК РАБОТ, ОПУБЛИКОВАННЫХ ПО ТЕМЕ ДИССЕРТАЦИИ

Научные статьи, опубликованные в журналах из перечня ВАК

1. Определение контекстной тональности сообщений в задачах информационного влияния. Часть 1 / Д. В. Мазурок, М. Ю. Монахов, Ю. М. Монахов, Е. А. Матвеева // Проектирование И Технология Электронных Средств. – 2023. – № 2. – С. 7-11.

2. Определение контекстной тональности сообщений в задачах информационного влияния. Часть 2 / Д. В. Мазурок, М. Ю. Монахов, Ю. М. Монахов, В. А. Вилкова // Проектирование И Технология Электронных Средств. – 2023. – № 3. – С. 49-54.

3. Автоматизированная система контроля целостности политики информационной безопасности сетевого оборудования / Д. В. Мазурок, М. М. Монахова, Г. В. Путинцев, С. Д. Лучинкин // Перспективы Науки. – 2015. – № 8 (71). – С. 80-83.

Научные публикации, индексируемые в международных базах Scopus и/или Web of Science

4. Improving Security of Neural Networks in the Identification Module of Decision Support Systems / Y. Monakhov, M. Monakhov, A. Telny et al. – Text: electronic // 2020 Ural Symposium on Biomedical Engineering, Radioelectronics and Information Technology (USBEREIT) 2020 Ural Symposium on Biomedical Engineering, Radioelectronics and Information Technology (USBEREIT). – Yekaterinburg, Russia: IEEE, 2020. – P. 571-574

Публикации в прочих изданиях

5. Мазурок, Д. В. Метод автоматического ранжирования заявок в системах поддержки пользователя корпоротивных телекоммуникационных сетей / Д. В. Мазурок // Инновационное развитие современной науки: проблемы, закономерности, перспективы: сборник статей победителей международной научно-практической конференции, Пенза, 23 февраля 2017 года. – Пенза: "Наука и Просвещение" (ИП Гуляев Г.Ю.), 2017. – С. 30-32.

6. Мазурок, Д. В. Мониторинг нейросетевых моделей: подходы и инструменты / Д. В. Мазурок // Наука и глобальные вызовы: перспективы развития: сборник статей IX Международной научно-практической конференции, Саратов, 20 июня 2024 года. – Саратов: Научно-образовательная платформа «Цифровая Наука», 2024. – С. 43-50.

7. Мазурок, Д. В. Повышение информативности нейросетевых анализаторов сетевого трафика при принятии решений / Д. В. Мазурок // Научный аспект. – 2024. – Т. 17, № 8. – С. 2084-2088.

8. Мазурок, Д. В. Составляющие оценки качества модели нейросетевого классификатора / Д. В. Мазурок // Наука и глобальные вызовы:

перспективы развития: сборник статей VIII Международной научно-практической конференции, Саратов, 30 мая 2024 года. – Саратов: Научно-образовательная платформа «Цифровая Наука», 2024. – С. 62-67.

9. Мазурок, Д. В. Мониторинг нейросетевых моделей: подходы и инструменты / Д. В. Мазурок // Исследования и инновации: синергия знаний и практики: сборник статей II Международной научно-практической конференции, Самара, 20 февраля 2025 года. – Москва: Издательство "Доброе слово и Ко", 2025. – С. 102-109.

10. Мазурок, Д. В. Мультиагентный подход анализа результатов интерпретации с применением больших языковых моделей / Д. В. Мазурок, М. Ю. Монахов // Наука, технологии и инновации: стратегии развития в современном мире : сборник статей II Международной научно-практической конференции, Москва, 23 апреля 2024 года. – Москва: Издательство "Доброе слово и Ко", 2024. – С. 130-136.

11. Мазурок, Д. В. Подход к оценке робастности модели нейросетевого классификатора / Д. В. Мазурок, А. В. Рунов // Наука, технологии и инновации: стратегии развития в современном мире: сборник статей II Международной научно-практической конференции, Москва, 23 апреля 2024 года. – Москва: Издательство "Доброе слово и Ко", 2024. – С. 137-146.

12. Мазурок, Д. В. Образ системы поддержки пользователей по устранению сбоев в работе оборудования корпоративной телекоммуникационной сети / Д. В. Мазурок, И. И. Семенова // Информационные технологии и автоматизация управления : Материалы VIII Всероссийской научно-практической конференции студентов, аспирантов, работников образования и промышленности, Омск, 26–28 апреля 2016 года. – Омск: Омский государственный технический университет, 2016. – С. 92-97.

13. Алгоритм локализации участка корпоративной информационно-телекоммуникационной сети предприятий с аномальным поведением / М. М. Монахова, Г. В. Путинцев, Д. В. Мазурок, А. А. Порфирьев // Вопросы безопасности. – 2017. – № 2. – С. 13-24.

14. Мазурок, Д. В. Методы определения появления неполадки в корпоративных телекоммуникационных сетях / Д. В. Мазурок // Европейские научные исследования: сборник статей победителей II международной научно-практической конференции, Пенза, 20 февраля 2017 года. – Пенза: Наука и Просвещение, 2017. – С. 35-36.

Свидетельства о государственной регистрации программы для ЭВМ

15. Свидетельство о государственной регистрации программы для ЭВМ № 2021618155 Российская Федерация. Программная реализация графического интерфейса для модуля аудита нейронных сетей на подверженность состязательным атакам: № 2021617158: заявл. 12.05.2021: зарег. 24.05.2021 / Д. В. Мазурок, Ю. М. Монахов, М. М. Агафонова; заявитель ВлГУ.

16. Свидетельство о государственной регистрации программы для ЭВМ № 2021618197 Российская Федерация. Программный модуль для тестирования скорости работы и точности классификаторов: № 2021617359: заявл. 12.05.2021:

зарег. 24.05.2021 / Д. В. Мазурок, М. М. Агафонова, Ю. М. Монахов; заявитель ВлГУ. – EDN NYHHQE.

17. Свидетельство о государственной регистрации программы для ЭВМ № 2021618487 Российская Федерация. Программный модуль оценки актуальности факторов защищенности детектора и классификатора и выдачи адаптивных рекомендаций по повышению защищенности: № 2021617213: заявл. 12.05.2021 : зарег. 27.05.2021 / Д. В. Мазурок, Ю. М. Монахов, М. М. Агафонова; заявитель ВлГУ.

18. Свидетельство о государственной регистрации программы для ЭВМ № 2021619081 Российская Федерация. Программная реализация модуля аудита нейронных сетей на подверженность состязательным атакам: № 2021617231: заявл. 12.05.2021: зарег. 03.06.2021 / Д. В. Мазурок, Ю. М. Монахов, М. М. Агафонова; заявитель ВлГУ.

19. Свидетельство о государственной регистрации программы для ЭВМ № 2024682642 Российская Федерация. Модуль оценки значимости по результатам эксперимента: № 2024681285: заявл. 14.09.2024: зарег. 25.09.2024 / Д. В. Мазурок, В. С. Полховский.

20. Свидетельство о государственной регистрации программы для ЭВМ № 2024682393 Российская Федерация. Программа получения расширенной интерпретации результатов нейросетевого классификатора трафика: № 2024681474: заявл. 15.09.2024: зарег. 23.09.2024 / Д. В. Мазурок.

21. Свидетельство о государственной регистрации программы для ЭВМ № 2024682916 Российская Федерация. Программа-API для проведения эксперимента по оценке повышения интерпретируемости решений нейросетевого классификатора: № 2024681297: заявл. 15.09.2024: зарег. 30.09.2024 / Д. В. Мазурок, Т. Д. Басов, А. Д. Масляных, А. Д. Гришин.

Автореферат
диссертации на соискание ученой степени
кандидата технических наук

Подписано в печать 04.03.2026 г.
Формат 60×84/16. Усл. печ. л. 1,0. Тираж 100 экз.

Издательство Владимирского государственного университета
имени Александра Григорьевича и Николая Григорьевича Столетовых 600000,
Владимир, ул. Горького, 87