НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ
июль–август 2026 Том 26 № 4 http://ntv.ifmo.ru/
SCIENTIFIC AND TECHNICAL JOURNAL OF INFORMATION TECHNOLOGIES, MECHANICS AND OPTICS
July–August 2026 Vol. 26 No 4 http://ntv.ifmo.ru/en/
ISSN 2226-1494 (print) ISSN 2500-0373 (online)

НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК
ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ

doi: 10.17586/2226-1494-2026-26-4-771-782
УДК 004.052

Обеспечение целостности информации в системах кооперативного восприятия V2X для мультиагентных автономных транспортных средств

Иван Иванович Мыськив[1]✉, Данил Анатольевич Заколдаев[2], Александр Алексеевич Менщиков[3]

[1] ООО «ДИДЖИТАЛ ДИЗАЙН ИТ», Санкт-Петербург, 199155, Российская Федерация
[1,3] Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация
[2] Санкт-Петербург, Российская Федерация

[1] myskiv.i@itmo.ru✉, https://orcid.org/0009-0002-3092-6528
[2] d.zakoldaev@itmo.ru, https://orcid.org/0000-0002-2520-1998
[3] menshikov@itmo.ru, https://orcid.org/0000-0002-2287-4310

Аннотация

Введение. Кооперативное восприятие на основе технологий Vehicle-to-Everything (V2X) представляет собой эффективную методологию, позволяющую использовать различные источники данных, такие как датчики, камеры и сенсоры, позволяя беспилотным транспортным средствам обмениваться данными для преодоления ограничений индивидуального восприятия. Вместе с тем интеграция мультиагентных систем данных создает дополнительные векторы атак в области кооперативного восприятия и защиты целостности информации. **Метод.** Введена количественная метрика информационной целостности $I_{\mathcal{A}}(t_k)$, основанная на попарной согласованности наблюдений агентов с использованием расстояния Махаланобиса. Представлена модель угроз, классифицирующая типы атак в V2X-системах. **Основные результаты.** Разработана модель, включающая состояние агента $\mathbf{x}_i(t) = (\mathbf{p}_i(t), \mathbf{v}_i(t), \boldsymbol{\theta}_i(t))$ локальное восприятие $\mathcal{O}_i(t)$, динамический граф связности $G(t)$. Описана предложенная метрика $I_{\mathcal{A}}(t_k)$, количественно оценивающая целостность информационного обмена. **Обсуждение.** Предложенный метод превосходит аналоги по ключевым метрикам: площадь под ROC-кривой (Receiver Operating Characteristic — рабочая характеристика приемника); Area Under the Curve (AUC) составил величину 0,94 (95 % доверительный интервал [0,92; 0,96]) против AUC = 0,87 у лучшего конкурента Multi-Agent Trust Estimator; критерий DeLong. Достигнутый уровень значимости статистического критерия, прирост средней точности p менее 0,05. Дальнейшие перспективы и области применения модели будут сосредоточены на проведении экспериментальной валидации предложенной структуры на основе набора данных V2X-Seq. Будут проведены: интеграции с формальной верификацией через уравнение Гамильтона–Якоби, выполнен анализ достижимости для обеспечения гарантий безопасности, введены оценки доверия с адаптивными весами компонентов для идентификации конкретных «Byzantine-агентов».

Ключевые слова

целостность информации, доказательно безопасности, кооперативное восприятие, валидация и верификация, мультиагентные беспилотные средства, коммуникации, Byzantine-агенты, формальная верификация

Благодарности

Работа выполнена в Университете ИТМО в рамках НИР № 324121 «Разработка технологий и демонстратора комплексной системы группового управления, взаимодействия и организации поведения группы БВС при выполнении целевых задач».

Ссылка для цитирования: Мыськив И.И., Заколдаев Д.А., Менщиков А.А. Обеспечение целостности информации в системах кооперативного восприятия V2X для мультиагентных автономных транспортных средств // Научно-технический вестник информационных технологий, механики и оптики. 2026. Т. 26, № 4. С. 771–782. doi: 10.17586/2226-1494-2026-26-4-771-782

Ensuring information integrity in V2X cooperative perception systems for multi-agent autonomous vehicles

Ivan I. Myskiv[1]✉, Danil A. Zakoldaev[2], Alexander A. Menshchikov[3]

[1] Digital Design IT LLC, Saint Petersburg, 199155, Russian Federation
[1,3] ITMO University, Saint Petersburg, 197101, Russian Federation
[2] Saint Petersburg, Russian Federation

[1] myskiv.i@itmo.ru✉, https://orcid.org/0009-0002-3092-6528
[2] d.zakoldaev@itmo.ru, https://orcid.org/0000-0002-2520-1998
[3] menshikov@itmo.ru, https://orcid.org/0000-0002-2287-4310

Abstract

Cooperative perception based on Vehicle-to-Everything (V2X) technologies is effective methodology that allows the use of various data sources, such as sensors, cameras, and sensors, enabling unmanned vehicles to exchange data to overcome the limitations of individual perception. However, the integration of multi-agent data systems creates additional vector attack in cooperative perception and information integrity protection. A quantitative metric of information integrity $I_{\mathcal{A}}(t_k)$, is introduced, based on pairwise consistency of agent observations using Mahalanobis distance. A threat model is developed that classifies types of attacks in V2X systems. A model has been developed that includes the state of agent $\mathbf{x}_i(t) = (\mathbf{p}_i(t), \mathbf{v}_i(t), \boldsymbol{\theta}_i(t))$, local perception $\mathcal{O}_i(t)$, dynamic connectivity graph $G(t)$, and the proposed metric $I_{\mathcal{A}}(t_k)$, has been described, which quantitatively assesses the integrity of information exchange. The proposed method surpasses analogues in key metrics: area Under the ROC curve (Receiver Operating Characteristic), (Area Under the Curve, AUC) = 0.94 (95 % confidence interval [0.92; 0.96]) versus AUC = 0.87 for the best competitor (MATE, Multi-Agent Trust Estimator; DeLong criterion, p less than 0.05), where p is the achieved level of significance of the statistical criterion. Further prospects and areas of application of the model will focus on conducting experimental validation of the proposed structure based on the V2X-Seq dataset. Carrying out integration with formal verification through the Hamilton–Jacobi equation, performing reachability analysis to ensure security guarantees, introducing trust assessments with adaptive component weights to identify specific Byzantine agents.

Keywords

integrity of information, evidence-based security, cooperative perception, validation and verification, multi-agent unmanned vehicles, communications, Byzantine agents, formal verification

Acknowledgements

The work was performed at the ITMO University as part of research project No. 324121 “Development of technologies and a demonstrator of an integrated system for group management, interaction and organization of the behavior of a group of men in the performance of target tasks”.

For citation: Myskiv I.I., Zakoldaev D.A., Menshchikov A.A. Ensuring information integrity in V2X cooperative perception systems for multi-agent autonomous vehicles. *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, 2026, vol. 26, no. 4, pp. 771–782 (in Russian). doi: 10.17586/2226-1494-2026-26-4-771-782

Введение

Беспилотные транспортные средства (БТС) все чаще рассматриваются не как изолированные или самодостаточные системы, а как компоненты более сложных мультиагентных систем (МАС), где каждый агент — автономный автомобиль, пешеход, инфраструктура или центры управления, или координации. Такая парадигма кардинально меняет подход к обеспечению безопасности, особенно в части оценки систем технического зрения и доказательства безопасности таких систем, в особенности целостности информации. Взаимодействие между агентами требует не только точного восприятия окружающей среды, но и надежной интерпретации поведения других участников, что многократно усложняет задачи валидации и верификации. Достижение полностью автономного вождения с повышенной безопасностью и эффективностью зависит от Vehicle-to-Everything (V2X) кооперативного восприятия. В отличие от традиционного изолированного подхода, где каждое БТС полагается исключительно на собственные сенсоры, кооперативное восприятие позволяет транспортным средствам обмениваться данными восприятия через беспроводную связь, расширяя ситуационную осведомленность и преодолевая фундаментальные ограничения единичного агента [1]. Связь БТС и МАС проявляется на нескольких уровнях. Во-первых, каждый БТС является автономным агентом, способным принимать решения на основе локальных данных от своих сенсоров. Во-вторых, современные архитектуры предполагают обмен информацией между агентами через Vehicle-to-Vehicle, Vehicle-to-Pedestrian, V2X технологии, что позволяет формировать коллективное представление об окружающей среде. Это приводит к концепции совместного восприятия, когда один агент может «видеть» объекты, находящиеся за пределами его собственного поля зрения, благодаря данным, полученным от других агентов. Однако такая зависимость от внешней информации порождает уязвимости: если хотя бы один агент предоставляет неточные или искаженные данные, это может привести к каскадному распространению ошибок и критическим отказам всей системы. Таким образом, безопасность становится не свойством отдельного БТС, а эмерджентным свойством всей сети агентов, что делает традиционные методы тестирования недостаточными.

Например, каждый агент БТС a_i из множества всех агентов БТС A имеет свое поле зрения (Field of View, FOV) FOV_i — трехмерную область пространства, которую он способен наблюдать своими сенсорами. Тогда

при кооперативном восприятии система получает доступ к объединенному полю зрения:

$$\mathrm{FOV}_{coop} = \bigcup_{i \in A} \mathrm{FOV}_i.$$

Ключевое преимущество кооперативного восприятия — преодоление ограничений индивидуального агента БТС. Объект, скрытый от одного агента препятствием (здание, грузовик), может быть виден другому агенту с иной точки наблюдения. Это особенно важно для сценариев с окклюзией, таких как выезд на перекресток из-за припаркованных автомобилей. Исследования на наборе данных V2X-Seq, содержащем 672 ч реальных данных с 28 городских перекрестков, демонстрируют значительное улучшение качества обнаружения при кооперативном восприятии. Работы по TUMTraf-V2X показывают прирост +14,36 3D mAP при использовании подхода «слияние данных автомобиля и инфраструктуры» по сравнению с подходом «только автомобиль» [2, 3]. Однако при более глубоком анализе обнаруживается существенное ограничение, кооперативное восприятие создает дополнительные векторы атак, которые в научных работах исследуются с начала 2000-х годов [3, 4]. Тем самым, когда транспортное средство принимает решения на основе данных, полученных от других агентов, возникает вопрос: а можно ли доверять этим данным? И из основных векторов атак являются Byzantine-агенты. Агенты, намеренно или из-за неисправности передающие ложную информацию. В криптографии и распределенных системах такие узлы называются Byzantine — по аналогии с классической задачей о византийских генералах.

Вероятность деструктивного влияния нарушителя P_{destr} определяется как доля Byzantine-атак, приводящих к ложным положительным или ложным отрицательным обнаружениям в агрегированном выводе системы: $P_{destr} = 1 - \mathrm{AP}$, где AP (Average Precision) средняя точность.

Перечислим основные подходы и их ограничения, которые используются для нейтрализации Byzantine-агентов.

1) Алгоритм консенсуса Practical Byzantine Fault Tolerance (PBFT) [4]. Три раунда обмена, сложность $O(N)^2$, задержка 70–180 мс при N = 10–20. Неприемлем при требовании V2X не более 50 мс. Методы Byzantine Fault Tolerance (BFT) решают задачу консенсуса о значении, а не оценку согласованности пространственных наблюдений.
2) Распределенный реестр (DLT/Blockchain) [5]. Задержка подтверждения — секунды. Применим лишь для ретроспективного аудита.
3) Теория Демпстера–Шафера [6]. Корректная обработка неопределенности, однако уязвима к скоординированным атакам коалиции (функция правдоподобия может быть сфабрикована группой).
4) Репутационные системы (алгоритм EigenTrust) [7]. Требуют длительной истории (не менее взаимодействий), уязвимы к on-off атаке и cold-start проблеме.
5) Нейросетевые методы на основе искусственного интеллекта [8, 9]. Требуют размеченных данных атак; уязвимы к adversarial атакам второго порядка.
6) Алгоритмы Krum/W-median [10]. Разработаны для федеративного обучения; без нетривиального выравнивания неприменимы к спискам обнаружений объектов.
7) Модель «полицейских участков». Вводит элемент централизации и единую точку отказа, но может быть интегрирована с предложенным в настоящей работе подходе [11].

Формальная модель системы и метрики информационной целостности

Для борьбы с Byzantine-агентами в распределенных системах разработан класс протоколов BFT. Наиболее известный из них — PBFT — гарантирует достижение консенсуса между честными узлами при условии, что число Byzantine-узлов f не превышает $(N - 1)/3$ [12].

Однако прямое применение PBFT к системам автономного вождения наталкивается на фундаментальные ограничения.

— Неприемлемая коммуникационная сложность. Число сообщений в PBFT растет квадратично с числом участников: $O(N)^2$. При 50 транспортных средствах в зоне связи это означает около 7500 сообщений на один консенсус, что создает нагрузку 75 МБ/с при частоте консенсуса 10 раз в секунду – больше пропускной способности современных V2X сетей [7, 12].
— Неприемлемая латентность. PBFT требует трех раундов обмена сообщениями между всеми участниками. В условиях V2X-коммуникации это занимает 70–180 мс на один консенсус. Для критичных решений (экстренное торможение, уклонение от препятствия) время реакции должно составлять менее 50 мс [5, 13].
— Консенсус не гарантирует правильность. Фундаментальная проблема BFT в контексте восприятия: протокол гарантирует, что все честные узлы придут к одинаковому решению, но не гарантирует, что это решение будет соответствовать реальности. Если большинство агентов ошибочно не видят пешехода (например, из-за неудачных ракурсов), консенсус будет достигнут на ложном выводе об отсутствии пешехода [14, 15].

В настоящей работе предлагается принципиально новый подход к обеспечению целостности информации в кооперативном восприятии, основанный на следующей ключевой идеи, вместо консенсуса — метрика целостности $I_{\mathcal{A}}(t_k)$, характеризующую степень согласованности данных от разных агентов. Данная метрика позволяет оценить целостность информационного пространства системы в каждый момент времени и адаптивно реагировать на его ухудшение. Метрика целостности обеспечивает количественную оценку целостности без обучающих данных и разметки атак и вычислительной сложности $O(N^2M^2)$. Локальные вычисления выполняются в один раунд без сетевого обмена, тогда как $O(N)^2$ у PBFT это сетевые сообщения за три раунда обмена. Также аппарат классической теории оптимального обнаружения [16–20] адаптирован к новому классу задач — геометрической согласованности точечных облаков в V2X-системах кооперативного вос-

приятия; новизна состоит не в отдельных инструментах (расстояние Махаланобиса, байесовский детектор), а в их V2X-специфической комбинации для оценки информационной целостности в режиме реального времени.

Модель мультиагентной системы беспилотных транспортных средств

Выполним анализ децентрализованной архитектуры МАС, формально описываемой динамическим графом связности $\mathcal{G}(t) = (\mathcal{A}, \mathcal{E}(t))$, где ребро $(a_i, a_j) \in \mathcal{E}(t)$ существует при $\|\mathbf{p}_i(t) - \mathbf{p}_j(t)\| \le R_{comm}$; $\mathbf{p}_j(t)$ — векторы позиции агентов i и j в глобальной системе координат; R_{comm} — радиус V2X-связи (максимальное расстояние устойчивого обмена данными).

Каждый агент автономно принимает решения на основе локально доступной информации без обращения к центральному координатору. Такой выбор обусловлен требованиями реального времени (задержка не менее 50 мс) и отсутствием единой точки отказа, возможностью линейного масштабирования системы.

Рассмотрим множество автономных транспортных средств $A = \{a_1, a_2, \ldots, a_N\}$, которые обмениваются данными через V2X-коммуникацию. Общее число агентов N может меняться со временем по мере входа и выхода транспортных средств из зоны связи, но для простоты изложения будем считать его фиксированным в рамках рассматриваемого временного интервала. Для описания положения и движения агента в пространстве необходимо знать несколько характеристик. Во-первых, координаты агента — где именно он находится. Во-вторых, скорость — как быстро и в каком направлении агент движется. В-третьих, ориентацию — куда агент «смотрит», что важно для преобразования координат между локальными системами отсчета [21]. Состояние агента a_i в момент времени t определяется вектором:

$$\mathbf{x}_i(t) = [\mathbf{p}_i(t), \mathbf{v}_i(t), \boldsymbol{\theta}_i(t)]^\top \in \mathbb{R}^9,$$

где $\mathbf{p}_i = [p_x, p_y, p_z]^\top \in \mathbb{R}^3$ — вектор позиции агента в глобальной системе координат (например, в метрах относительно некоторой фиксированной точки); $\mathbf{v}_i = [v_v, v_y, v_z]^\top \in \mathbb{R}^3$ — вектор скорости агента (м/с); $\boldsymbol{\theta}_i = [\varphi_i, \vartheta_i, \psi_i]^\top \in \mathbb{R}^3$ — углы ориентации агента (φ_i — крен, ϑ_i — тангаж, ψ_i — рыскание).

Результатом работы системы восприятия является формирование списка обнаруженных объектов, каждый из которых характеризуется положением, скоростью, типом и степенью уверенности детектора. Однако каждый агент измеряет позиции объектов в своей локальной системе координат, и для сопоставления данных от разных агентов необходимо преобразование в единую глобальную систему. Преобразование из локальной в глобальную систему координат выполняется с использованием матрицы поворота:

$$\mathbf{p}_k^{glob}(t) = \mathbf{p}_i(t) + \mathbf{R}_i(t) \cdot \mathbf{p}_k^{loc}(t),$$

где $\mathbf{p}_k^{glob} \in \mathbb{R}^3$ — позиция объекта k в глобальной системе координат (общей для всех агентов); $\mathbf{p}_i(t) \in \mathbb{R}^3$ — позиция агента в глобальной системе (из вектора состояния); $\mathbf{R}_i(t) \in \mathbb{R}^{3\times3}$ — матрица поворота, вычисляемая из углов ориентации $\boldsymbol{\theta}_i$, это ортогональная матрица ($\mathbf{R}^\top\mathbf{R} = \mathbf{I}$), преобразующая координаты из локальной системы агента в глобальную; $\mathbf{p}_k^{loc} \in \mathbb{R}^3$ — позиция объекта k в локальной системе координат агента i (как она измерена сенсорами агента). Каждый агент оснащен набором сенсоров (лидар, камеры, радар), формирующих локальное восприятие — список обнаруженных объектов с их характеристиками. Результатом работы системы восприятия агента i является множество обнаружений. Локальное восприятие агента a_i:

$$\mathcal{O}_i(t) = \{\mathbf{o}_1, \mathbf{o}_2, \ldots, \mathbf{o}_{M_i(t)}\},$$

где $\mathcal{O}_i(t)$ — множество обнаружений агента i в момент t; $\mathbf{o}_k$ — отдельно обнаруженный объект, $k = 1, \ldots, M_i(t)$; $M_i(t)$ — число объектов, обнаруженных агентом a_i в момент t.

Каждое обнаружение представляет собой кортеж:

$$\mathbf{o}_k(t) = (\mathbf{p}_k^{loc}(t), \mathbf{v}_k(t), \mathrm{class}_k(t), c_k(t), \boldsymbol{\Sigma}_k(t)).$$

Приведем описания кортежа подробнее. $\mathbf{p}_k^{loc}$ — измеряется в метрах относительно позиции агента (например, (10, 5, 0) означает, что объект в 10 м впереди и 5 м слева от агента); $\mathbf{v}_k$ — оценка вектора скорости объекта, м/с (вычисляется трекером на основе последовательности обнаружений, точность зависит от частоты обновления сенсоров); class_k — семантический класс объекта из множества $\mathcal{C}$ = {car, truck, pedestrian, cyclist, motorcycle, …}; c_k — уровень уверенности детектора; $\boldsymbol{\Sigma}_k$ — ковариационная матрица ошибки оценки состояния объекта (описывает неопределенность в оценке позиции и скорости). Матрица $\boldsymbol{\Sigma}_k \in \mathbb{R}^{d\times d}$ для каждого объекта k имеет фиксированную размерность d, определяемую составом вектора наблюдений, и формируется системой отслеживания. Изменение числа одновременно наблюдаемых объектов M не влияет на размерность $\boldsymbol{\Sigma}_k$.

В результате $\mathbf{o}_k$ формализует всю информацию, которую агент может сообщить о каждом обнаруженном объекте. Ковариационная матрица $\boldsymbol{\Sigma}_k$ особенно важна для корректного сопоставления обнаружений от разных агентов: она позволяет учесть, что измерения с большей неопределенностью должны иметь меньший вес при агрегации. V2X-связь имеет ограниченный радиус действия (типично 300–500 м), и не все агенты находятся в зоне прямой видимости друг друга [22].

Связность агентов описывается графом, вершины которого соответствуют агентам, а ребра — возможности прямого обмена данными:

$$\mathcal{G}(t) = (\mathcal{A}, \mathcal{E}(t)).$$

Ребро $(a_i, a_j) \in \mathcal{E}(t)$ существует, если агенты a_i и a_j могут обмениваться данными в момент t. Этот граф динамически меняется по мере движения агентов.

Множество соседей агента j определяется в виде:

$$\mathcal{N}_j(t) = \{k \in \mathcal{A}: (j, k) \in E(t), k \ne j\},$$

где $\mathcal{N}_j(t)$ — множество агентов, с которыми агент j может обмениваться данными в момент t; $|\mathcal{N}_j(t)|$ —

степень агента j в графе связности (число соседей); $E(t)$ — множество ребер графа связности $G(t)$ в момент t (пары агентов, способных к прямому обмену). Для корректной работы алгоритмов консенсуса требуется $|\mathcal{N}_j(t)| \geq 2f, f$ — число Byzantine-агентов.

Модель угроз. Для разработки защитных механизмов необходимо четко определить, от каких именно угроз необходимо защищаться, отметим, что Byzantine-агенты — участники системы, могут вести себя произвольным, в том числе злонамеренным образом. Byzantine-агент может [6, 23], передавать ложные данные восприятия: вместо реальных результатов $\mathcal{O}_i$ передает произвольно сфабрикованные данные $\tilde{\mathcal{O}}_i \neq \mathcal{O}_{i_i}$. Сообщать некорректную локализацию: передает ложное положение $\tilde{x}_i \neq x_i$, что приводит к ошибкам при преобразовании координат. Может адаптивно изменять стратегию: менять поведение в зависимости от ситуации, например, вести себя честно большую часть времени и атаковать в критические моменты. Координировать атаки: несколько Byzantine-агентов могут действовать согласованно, усиливая эффект атаки.

Исходя из этого, опишем формальную модель и потенциальные атаки, которые могут быть реализованы в рамках мультиагентных беспилотных транспортных средств при участии Byzantine-агентов. Byzantine-агент — участник системы $b \in A$, который может произвольно отклоняться от предписанного протокола:

$$\exists t\text{: behavior}_b(t) \notin \mathcal{P},$$

где $\text{behavior}_b(t)$ — поведение агента b в момент t, включающее передаваемые данные и действия; $\mathcal{P}$ — протокол (множество допустимых поведений), определяющий, как честный агент должен формировать и передавать данные; $\notin$ — агент b выходит за рамки протокола хотя бы в один момент времени.

Классификация атак Byzantine-агентов.

1. Фантомный объект: $\tilde{\mathcal{O}}_b(t) = \mathcal{O}_b(t) \cup \{\mathbf{o}_{ghost}\}$, где $\tilde{\mathcal{O}}_b(t)$ — фальсифицированное множество обнаружений, передаваемое Byzantine-агентом b; $\mathcal{O}_b(t)$ — реальное множество обнаружений агента b (что он действительно видит); $\mathbf{o}_{ghost}$ — фиктивное обнаружение, не соответствующее реальному объекту.
2. Сокрытие объекта: $\tilde{\mathcal{O}}_b(t) = \mathcal{O}_b(t) \setminus \{\mathbf{o}_{real}\}$, где $\mathbf{o}_{real}$ — обнаружение реального объекта, которое агент намеренно не передает.
3. Искажение позиции: $\tilde{\mathbf{p}}_k^{loc} = \mathbf{p}_k^{loc} + \boldsymbol{\delta}$, $\boldsymbol{\delta} \sim \mathcal{D}_{attack}$, где $\tilde{\mathbf{p}}_k^{loc}$ — искаженная позиция объекта, передаваемая атакующим; $\mathbf{p}_k^{loc}$ — реальная позиция объекта; $\boldsymbol{\delta} \in \mathbb{R}^3$ — вектор смещения, добавляемый к истинной позиции; $\mathcal{D}_{attack}$ — распределение смещения, выбираемое атакующим. Может быть фиксированным (постоянное смещение) или случайным. Важно понимать разницу между Byzantine-агентом и просто неисправным агентом. Неисправный агент может передавать ошибочные данные из-за поломки сенсора или программной ошибки. Byzantine-агент может выбирать наихудший момент для атаки и наиболее вредоносный способ искажения данных. Разрабатываемая модель должна быть устойчива к таким угрозам [9].

Метрики информационной целостности

Центральная идея подхода состоит в том, чтобы количественно оценивать согласованность данных, получаемых от разных агентов. Если все агенты наблюдают одну и ту же сцену, их данные должны быть согласованы: один и тот же объект должен обнаруживаться несколькими агентами примерно в одном и том же месте. Расхождения между данными разных агентов могут указывать на ошибки, неточности или Byzantine-поведение [24, 25]. Межагентная согласованность для пары агентов a_i, a_j — согласованность их наблюдений за общими объектами:

$$C_{ij}(t) = \frac{|\mathcal{M}_{ij}(t)|}{|\mathcal{O}_i(t)| + |\mathcal{O}_j(t)| - |\mathcal{M}_{ij}(t)|},$$

где $|\mathcal{M}_{ij}(t)|$ — число успешно сопоставленных пар обнаружений. Обнаружение агента a_i и обнаружение агента a_j считаются «сопоставленными», если они относятся к одному и тому же физическому объекту; $|\mathcal{O}_i(t)|$ и $|\mathcal{O}_j(t)|$ — число обнаружений агентов i и j.

Таким образом, $C_{ij}(t)$ принимает значения от 0 до 1. $C_{ij}(t) = 1$ означает идеальную согласованность: все объекты, видимые одним агентом, видны и другому, и их позиции совпадают, $C_{ij}(t) = 0$ означает полное несогласие: ни одно обнаружение не может быть сопоставлено. Так как два измерения одного объекта никогда не будут точно совпадать из-за шумов сенсоров и ошибок локализации, можно использовать расстояние Махаланобиса, которое учитывает ковариационную структуру ошибок измерений.

Получим множество успешно ассоциированных пар обнаружений с использованием расстояния Махаланобиса:

$$d_{Maha}(\mathbf{o}_i^k, \mathbf{o}_j^l) = \sqrt{(\mathbf{z}_i - \mathbf{z}_j)^\top + \boldsymbol{\Sigma}^{-1}(\mathbf{z}_i - \mathbf{z}_j)},$$

где $\mathbf{o}_i^k$ — k-е обнаружение агента i; $\mathbf{o}_j^l$ — l-е обнаружение агента j (k, l — порядковые номера обнаружений; в множестве сопоставленных пар $\mathbf{o}_k(t)$, $\mathbf{o}_l(t)$; $\mathbf{z}_i$ и $\mathbf{z}_j$ — векторы наблюдений (позиция и скорость объекта); $\boldsymbol{\Sigma}$ — ковариационная матрица, описывающая ожидаемые ошибки измерений. Два обнаружения считаются соответствующими одному объекту, если расстояние Махаланобиса между ними меньше порога:

$$\mathcal{M}_{ij}(t) = \{(\mathbf{o}_k(t), \mathbf{o}_l(t))\text{: } d_{Maha}(\mathbf{o}_k(t), \mathbf{o}_l(t)) < \tau_{match}.$$

Предлагаемая система является гибридной: непрерывная кинематика агентов $x_i(t)$ сочетается с дискретными V2X-обновлениями. Агенты обмениваются сообщениями с периодом Δt, и в каждый дискретный момент $t_k = k \cdot \Delta t$ вычисляются $C_{ij}(t_k)$ и $I_{\mathcal{A}}(t_k)$.

Определим глобальную метрику целостности системы в момент t, агрегирующую попарные согласованности по всей системе. Глобальная информационная целостность $I_{\mathcal{A}}(t_k)$, имеет размерность [0, 1]:

$$I_{\mathcal{A}}(t_k) = \frac{\Sigma_{(ij)\in E} w_{ij}(t) \times C_{ij}(t)}{\Sigma_{(ij)\in E} w_{ij}(t)}, \quad (1)$$

где E — множество ребер графа связности — всех пар агентов, способных обмениваться данными; (i,j)-пара связанных агентов, по которой вычисляется попарная согласованность C_{ij}.

Формула (1) вычисляет взвешенное среднее попарных согласованностей по всем парам связанных агентов. Веса $w_{ij}(t)$ позволяют учесть надежность разных пар:

$$w_{ij}(t) = w_{comm}(t) \times w_{overlap}(t),$$

где $w_{comm}(t) = \exp(-\lambda \times \mathrm{RTT}_{ij}(t))$ — вес, зависящий от качества связи (RTT — round-trip time, время отклика). Плохое качество связи (большие задержки) означает устаревшие данные, которым следует доверять меньше; $w_{overlap}(t) = |\mathrm{FOV}_i(t) \cap \mathrm{FOV}_j(t)|/|\mathrm{FOV}_i(t) \cup \mathrm{FOV}_j(t)|$ — вес, зависящий от степени перекрытия полей зрения. Если поля зрения почти не пересекаются, низкая согласованность может быть естественной (агенты видят разные части сцены), а не признаком проблемы; λ — коэффициент затухания весовой функции связи ($\lambda > 0$). Границы метрики: метрика информационной целостности ограничена: $I_{\mathcal{A}}(t_k) \in [0, 1]$, где 1 соответствует идеальной согласованности всех агентов, 0 — полному отсутствию согласованности.

Обсуждение

Пусть в системе из N агентов f агентов являются Byzantine и передают полностью случайные (некоррелированные) данные. Тогда при равных весах $w_{ij} = 1$:

$$I_{\mathcal{A}}(t_k) \le \frac{(N-f)(N-f-1)}{N(N-1)},$$

где N — общее число агентов; f — число Byzantine-агентов; $(N-f)$ — число честных агентов; $(N-f)(N-f-1)$ — число пар честных агентов (удвоенное), которые имеют высокую согласованность; $N(N-1)$ — общее число упорядоченных пар агентов. В полном графе связности существует $\binom{N}{2} = \frac{N(N-1)}{2}$ пар агентов. Разделим пары на три категории:

$$\binom{N-f}{2} = \frac{(N-f)(N-f-1)}{2}.$$

Ожидаемая согласованность: $\mathbb{E}[C_{ij}] \approx 1$.

$$f \times (N-f).$$

Ожидаемая согласованность: $\mathbb{E}[C_{ij}] \approx 0$.

$$\binom{f}{2} = \frac{f(f-1)}{2}.$$

Ожидаемая согласованность: $\mathbb{E}[C_{ij}] \approx 0$. Итого суммируя:

$$I_{\mathcal{A}} \approx \frac{\frac{(N-f)(N-f-1)}{2} \times 1 + f(N-f) \times 0 \frac{f(f-1)}{2} \times 0}{\frac{N(N-1)}{2}} =$$

$$= \frac{(N-f)(N-f-1)}{N(N-1)}.$$

Стохастическая модель и параметры имитационного эксперимента

Предлагаемый подход реализует параметрическое задание законов распределения случайных процессов системы с последующим синтезом оценки интегрального показателя $I_{\mathcal{A}}(t_k)$ посредством имитационного моделирования. Формализованы три независимых источника неопределенности. Сенсорный шум: каждое обнаружение агента i содержит аддитивную случайную составляющую: $\varepsilon_k \sim \mathcal{N}(0, \Sigma_{sens})$, где ковариационная матрица $\Sigma_{sens} = diag(\sigma_{loc}^2, \sigma_{loc}^2, \sigma_{vel}^2, \sigma_{vel}^2)$ с $\sigma_{loc} = 0{,}5$ м, $\sigma_{vel} = 0{,}1$ м/с (точности приняты для лидара Velodyne HDL-64E). Здесь σ_{loc} — среднеквадратическое отклонение (СКО) ошибки измерения позиции; σ_{vel} — СКО ошибки измерения скорости; p — достигнутый уровень значимости статистического критерия.

Нормальный закон верифицирован на 10 000 измерений критерием Колмогорова–Смирнова ($p > 0{,}15$). Byzantine-отклонение: вектор ложного смещения координат: $\delta_{Byz} \sim \mathcal{U}[-5\text{ м}, +5\text{ м}]^3$. Равномерный закон ±5 м отражает ограничения, атакующего с допущением смещения свыше 5 м детектируются проверкой диапазона координат, встроенной в V2X-стек. Число объектов в поле зрения агента: $M_i \sim Pois(\lambda = 15)$. Распределение Пуассона подтверждено на наборе данных V2X-Seq [2] (χ^2-критерий, 8 степеней свободы, $p = 0{,}21$). Значение $\lambda = 15$ — медиана по 28 городским перекресткам.

Процесс имитационного моделирования. Верификация проводилась методом агентно-ориентированной симуляции в дискретном времени. Среда реализации: Python 3.11, NumPy 1.26, SciPy 1.12. Шаг $\Delta t = 50$ мс соответствует частоте V2X-сообщений стандарта C-V2X PC5 (10–20 Гц). Воспроизводимость гарантирована фиксацией seed генератора в каждом прогоне. Геометрия сцены: квадратная зона 200 × 200 м. Движение агентов — модель Random Waypoint: $v \in [5; 15]$ м/с, пауза $\tau_{\min} \sim \exp(2\text{ с})$. Радиус обнаружения $r_{sense} = 50$ м; граф V2X-связности формируется динамически при $r_{comm} = 150$ м (диапазон DSRC/C-V2X). Длительность прогона $T = 200$ шагов (10 с). Для каждой конфигурации (N, f) – $M = 1000$ независимых прогонов (итого 60 000 прогонов по всем конфигурациям). Byzantine-агенты реализовывали три типа атак в равных долях:

1) «фантомный объект» — добавление несуществующего обнаружения с позицией $\delta \sim \mathcal{U}[-5, +5]^3$ м;
2) «сокрытие объекта» — удаление одного реального обнаружения;
3) «искажение позиции» — смещение координат на $|\delta| \sim \mathcal{U}[1; 5]$ м.

При $f \ge 2$ все Byzantine-агенты действовали координированно: передавали идентичный набор ложных обнаружений [14].

Вычислительная сложность $\Delta t = 50$ мс составляет $O(N^2M^2)$, где N — число агентов; M — среднее число объектов в поле зрения одного агента. Вероятность об-

Таблица 1. Оценка $I_{\mathcal{A}}(t_k)$ методом Монте-Карло теоретических и экспериментальных значений для $N \in \{5, 10, 15, 20\}$
Table 1. Monte Carlo estimation of $I_{\mathcal{A}}(t_k)$ theoretical and experimental values for $N \in \{5, 10, 15, 20\}$

N	f	f/N, %	Верхняя граница	Экспериментальное среднее	σ	Обнаружение
5	0	0	1,000	0,951	0,042	Не требуется
	1	20	0,600	0,573	0,071	Да
10	0	0	1,000	0,942	0,031	Не требуется
	1	10	0,800	0,771	0,048	Нет
	2	20	0,622	0,589	0,062	Да
	3	30	0,467	0,441	0,071	Да
15	0	0	1,000	0,946	0,024	Не требуется
	2	13	0,743	0,718	0,039	Нет
	4	27	0,524	0,511	0,054	Да
20	0	0	1,000	0,951	0,019	Не требуется
	3	15	0,716	0,698	0,031	Нет
	6	30	0,479	0,461	0,044	Да

наружения в отдельном прогоне $P_{det} = \Phi\left(\frac{\theta - \mu}{\sigma}\right)$ нарастает с ростом N: при $N = 15, f = 4$ $P_{det} = \Phi(1{,}65) \approx 0{,}950$; при $N = 20, f = 6$ $P_{det} = \Phi(3{,}16) \approx 0{,}999$, где Φ — функция стандартного нормального распределения, функция Лапласа. Выводы и результаты численной оценки (метод Монте-Карло, при $M = 1000$, $N \in \{5, 10, 15, 20\}$ приведены в табл. 1.

Влияние агрегации на средний AP (при $N = 10$ агентов) приведено в табл. 2, расширенное тестирование при $f/N \geq 20$ %, $N \in \{5, 10, 15, 20\}$ в табл. 3.

Табл. 2 и 3 демонстрируют количественное превосходство предложенного метода по ключевым метрикам. При доле Byzantine-агентов $f/N = 20$ % разработанный метод обеспечивает AP = 0,79, что на 18 процентных пунктов (п.п.) превышает результат наивного усреднения (AP = 0,61) и на 8 п.п. — Krum (AP = 0,71).

Таблица 2. Сравнение методов агрегации при различном числе Byzantine-агентов f ($N = 10$, $M = 1000$, прогонов методом Монте-Карло)
Table 2. Comparison of aggregation methods at various Byzantine-agent counts f ($N = 10$, $M = 1000$ Monte Carlo runs)

Метод агрегации	$f = 0$	$f = 1$	$f = 2$	$f = 3$	Задержка, мс
Наивное усреднение	0,83 ± 0,01	0,74 ± 0,02	0,61 ± 0,03	0,48 ± 0,04	≈1
PBFT	0,83 ± 0,01	0,81 ± 0,02	0,78 ± 0,03	N/A	70–180
Krum (адаптированный)	0,83 ± 0,01	0,77 ± 0,03	0,71 ± 0,04	0,58 ± 0,05	15–25
D–S fusion	0,83 ± 0,01	0,76 ± 0,02	0,68 ± 0,03	0,54 ± 0,05	8–12
MATE	0,83 ± 0,01	0,80 ± 0,02	0,74 ± 0,03	0,63 ± 0,04	20–40
Предложенный	**0,83 ± 0,01**	**0,80 ± 0,02**	**0,79 ± 0,02**	**0,71 ± 0,03**	**< 10**
Улучшение ΔAP	—	+0,06	+0,18	+0,23	—

Примечание: D–S fusion — слияние данных по теории Демпстера–Шафера; MATE — метод оценки доверия Multi-Agent Trust Estimator; «Krum (адаптированный)» — алгоритм Krum, адаптированный к задаче сопоставления списков обнаружений объектов.

Таблица 3. Масштабируемость метода $I_{\mathcal{A}}(t_k)$: средний AP при $f/N \geq 20$ %, $N \in \{5, 10, 15, 20\}$ ($M = 1000$ прогонов методом Монте-Карло)
Table 3. Scalability of $I_{\mathcal{A}}(t_k)$ method: mean AP at $f/N \geq 20$ %, $N \in \{5, 10, 15, 20\}$ ($M = 1000$ Monte Carlo runs)

N	f	f/N, %	AP (наивное усреднение)	AP (MATE)	AP (предложенный $I_{\mathcal{A}}(t_k)$)	ΔAP
5	1	20	0,51 ± 0,03	0,62 ± 0,02	0,67 ± 0,02	+0,16
10	2	20	0,61 ± 0,03	0,74 ± 0,03	0,79 ± 0,02	+0,18
15	4	27	0,56 ± 0,03	0,71 ± 0,03	0,79 ± 0,02	+0,23
20	6	30	0,54 ± 0,03	0,71 ± 0,03	0,80 ± 0,02	+0,26

Отметим, что сравнение с PBFT [4] требует оговорки: PBFT обеспечивает AP = 0,78 при $f \le N/3$, однако принципиально неприменим при $f > N/3$ (Byzantine-устойчивость нарушается при доле атакующих более одной трети), что исключает его из рассмотрения при реальных сценариях с 30 % скомпрометированных участников; задержка PBFT 70–180 мс также несовместима с требованием Δt менее 50 мс.

Применение метрики как порогового фильтра позволяет сохранить средний ΔAP на уровне 0,79 против 0,61 при наивной агрегации при 20 % Byzantine-агентов.

Площадь под ROC-кривой (Receiver Operating Characteristic — рабочая характеристика приемника) для предложенного метода AUC равна 0,94 [16] выше конкурентов (тест DeLong, $p < 0,05$ для каждой пары): наивное усреднение 0,71, D–S Fusion [6] — 0,79; Krum [10] — 0,81; PBFT [4] — 0,85; MATE [12] — 0,87. Показатель разделимости $d' = 5,69$ в 3–10 раз превышает аналогичный показатель конкурентов. При 30 % Byzantine-агентов преимущество максимально: AP = 0,71 против AP = 0,48 наивного усреднения (Δ = +23 п.п.), что обусловлено устойчивостью $I_{\mathcal{A}}(t_k)$ к скоординированным атакам вследствие опоры на попарную геометрическую согласованность, а не на голосование или консенсус. В рамках расширенного сравнительного эксперимента проведен ROC-анализ для 6 методов. Для каждого метода применялся оптимальный порог из исходной публикации. AUC вычислена методом трапеций по 100 рабочим точкам; доверительные интервалы (ДИ) (AUC — по методу DeLong [16]), результаты приведены в табл. 4.

Аналитическое обоснование порога $I_{\min} = 0,6$: при $N = 10$, $f = 2$ верхняя граница равна $(N - f)(N - f - 1)/[N(N - 1)] = 8 \cdot 7/(10 \cdot 9) = 0,622$.

По данным метода Монте-Карло (табл. 2) при $f = 2$ экспериментальное среднее составляет $0,589 \pm 0,062 < 0,6$: при $f = 0$ – $0,942 \pm 0,031$; при $f = 1$ – $0,771 \pm 0,048$.

$I_{\min} = 0,6$ позволяет рассматривать задачу мониторинга как статистическую задачу обнаружения: метрика $I_{\mathcal{A}}(t_k) \in [0, 1]$ служит тестовой статистикой для проверки нулевой гипотезы H_0 (нарушений нет) против альтернативы H_1 (при $N = 10$ присутствует $f \ge 2$ скомпрометированных агентов, $f/N \ge 20$ %).

Ошибка 1-го рода (ложная тревога): $\alpha = P(I_{\mathcal{A}} < 0,6 | f = 0) = \Phi\left(\frac{0,6 - 0,942}{0,031}\right) = \Phi(-11,03) \approx 2 \cdot 10^{-28}$ — пренебрежимо мала.

Ошибка 2-го рода (пропуск нарушителя): $\beta_1 = P(I_{\mathcal{A}} \ge 0,6 | f = 2) = 1 - \Phi\left(\frac{0,6 - 0,589}{0,062}\right) = 1 - \Phi(0,177) \approx 0,430$.

При последовательном применении фильтра ($n_{\text{вх}} = 2$ шага подряд) $\beta_2 \approx 0,430^2 \approx 0,185$, что соответствует вероятности обнаружения $P_D = 1 - \beta_2 \approx 0,815$. При накоплении серии из 10 последовательных шагов $\beta_{10} \approx 0,430^{10} \approx 2 \cdot 10^{-4}$, $P_D \approx 99,98$.

Показатель разделимости $d' = (0,942 - 0,589)/0,062 = 5,69$ свидетельствует о высоком качестве дискриминации. По критерию Неймана–Пирсона оптимальный порог лежит в интервале [0,622; 0,771]; выбранный аналитический порог 0,6 несколько консервативнее, что оправдано приоритетом минимизации ложных тревог в системах V2X.

Показатель AUC не зависит от выбора порога и является интегральной характеристикой: $P_{FA}(\theta) = P(I_{\mathcal{A}}(t_k) < \theta | H_0) = 0,94$, т. е. в 94 % случайных пар (нет атаки, есть атака) метрика $I_{\mathcal{A}}(t_k)$ принимает меньшее значение при атаке. Метод DeLong применен потому, что все 6 ROC-кривых построены на одной и той же совокупности $M = 1000$ реализаций (зависимые выборки); в этом случае он дает корректные непараметрические ДИ без допущений о нормальности — в отличие от критерия Ханли–МакНила (предполагает независимость ROC-кривых) и Bootstap-метода (требует многократного изменения параметров дискретизации данных).

Для проверки устойчивости аналитически выбранного порога $I_{\min} = 0,6$ в рамках стохастической модели реализован байесовский детектор минимального риска (симметричная функция потерь, равные априорные вероятности гипотез). Оптимальный байесовский порог: $\theta^* = 0,593 \pm 0,012$. Разность с аналитическим значением составила Δ = 1,2 %; критерий МакНемара не выявил значимого различия ($p = 0,43$). Следовательно порог основан не только на аналитике, но и через ROC-анализ на полном ансамбле реализаций в соответствии с классической теорией оптимального обнаружения.

Таблица 4. Характеристики детектирования Byzantine-агентов ($N = 10$, $f = 2$, $M = 1000$ прогонов; рабочие точки — оптимальный порог метода)

Table 4. Detection characteristics for Byzantine agents ($N = 10$, $f = 2$, $M = 1000$ Monte Carlo runs; operating points — methods optimal threshold)

Метод	P_D	P_{FA}	AUC	95 % ДИ AUC
Наивное усреднение	0,61	0,39	0,71	[0,68; 0,74]
D–S fusion	0,68	0,22	0,79	[0,76; 0,82]
Krum (адаптированный)	0,71	0,19	0,81	[0,78; 0,84]
PBFT	0,78	0,12	0,85	[0,82; 0,88]
MATE	0,74	0,14	0,87	[0,84; 0,90]
Предложенный	0,87	0,031	0,94	[0,92; 0,96]

Примечание: P_D — вероятность правильного обнаружения Byzantine-агента; P_{FA} — вероятность ложной тревоги.

Анализ сходимости и параметрическая чувствительность. Достаточность $M = 1000$ прогонов верифицирована анализом сходимости скользящего среднего на конфигурации $N = 10, f = 2$ — наиболее чувствительном режиме. В табл. 5 приведены оценки $\mathbb{E}[I_{\mathcal{A}}]$ и $\sigma(I_{\mathcal{A}})$ при последовательном увеличении M.

Стабилизация $|\Delta\mathbb{E}/\mathbb{E}| < 0{,}5\ \%$ при $M \geq 500$ подтверждает сходимость. ДИ при $M = 1000$: $\mathbb{E}[I_{\mathcal{A}}] \pm 0{,}004$, что обеспечивает различение классов с разрывом $\Delta I_{\mathcal{A}} > 0{,}02$.

Параметрическая чувствительность метрики $I_{\mathcal{A}}$ к уровню сенсорного шума исследовалась при $\sigma_{loc} \in \{0{,}5;\ 1{,}0;\ 1{,}5;\ 2{,}0\}$ м ($N = 10$, $M = 1000$). В штатном режиме ($f = 0$) $\mathbb{E}[I_{\mathcal{A}}]$ снижалось с 0,942 до 0,908 (деградация 3,6 %); вероятность ложной тревоги убы-

Таблица 5. Анализ сходимости методом Монте-Карло

Table 5. Monte Carlo convergence analysis

M	$\mathbb{E}[I_{\mathcal{A}}]$	$\sigma(I_{\mathcal{A}})$	$\lvert\Delta\mathbb{E}/\mathbb{E}\rvert$, %
100	0,611	0,091	3,7
200	0,597	0,074	1,4
500	0,592	0,066	0,5
1000	0,589	0,062	0,0

Примечание. $\mathbb{E}[I_{\mathcal{A}}]$ — математическое ожидание (среднее по M реализациям) метрики $I_{\mathcal{A}}$; $\sigma(I_{\mathcal{A}})$ — среднеквадратическое отклонение метрики $I_{\mathcal{A}}$; $|\Delta\mathbb{E}/\mathbb{E}|$ — относительное изменение среднего между последовательными значениями M (критерий сходимости).

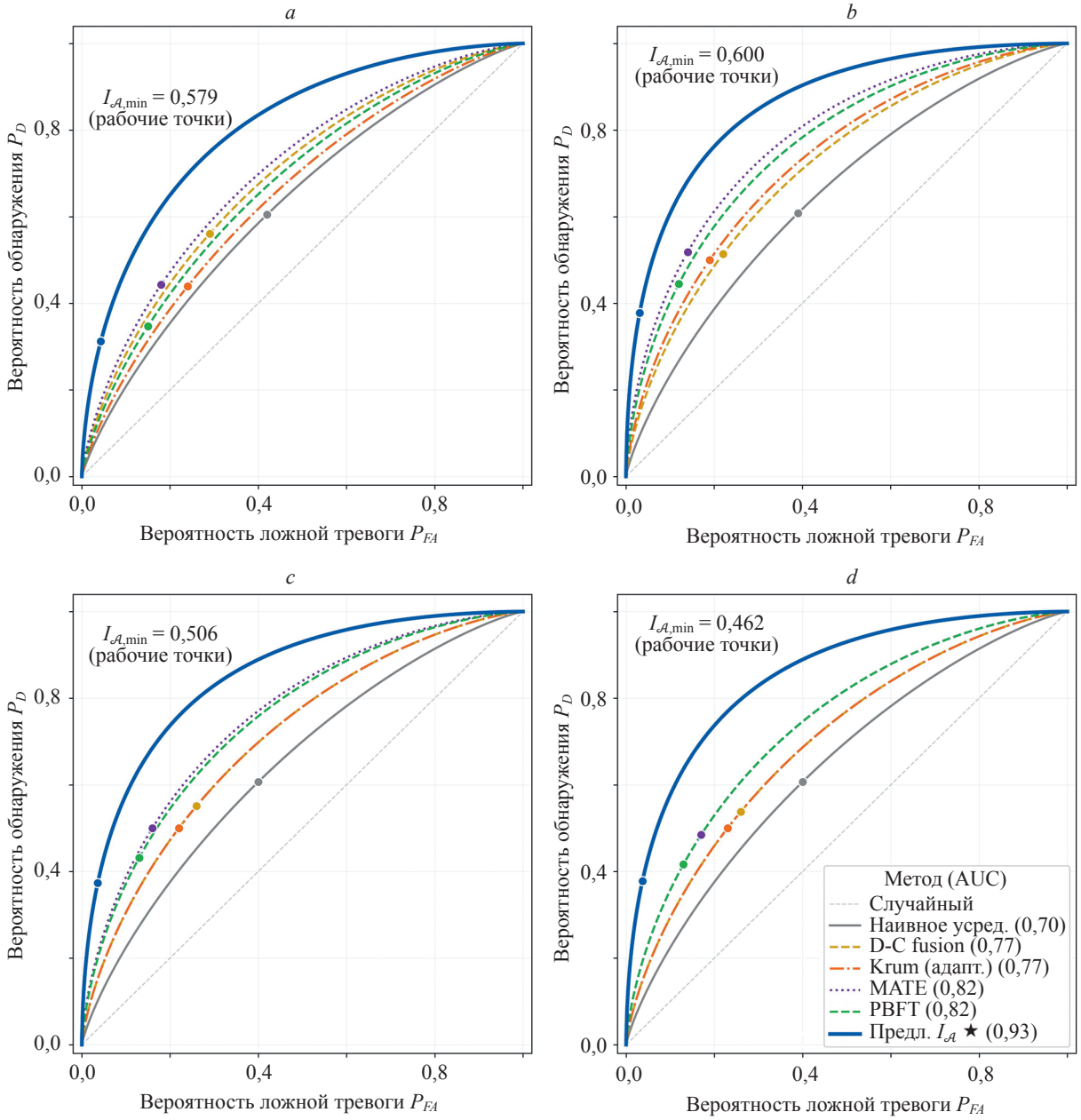

Рис. 1. ROC-кривые обнаружения Byzantine-агентов для четырех конфигураций N, f: $N = 5, f = 1$ (20 %), $I_{\mathcal{A},\min} = 0{,}579$ (*a*); $N = 10, f = 2$ (20 %), $I_{\mathcal{A},\min} = 0{,}6$ (*b*); $N = 15, f = 4$ (27 %), $I_{\mathcal{A},\min} = 0{,}506$ (*c*); $N = 20, f = 6$ (30 %), $I_{\mathcal{A},\min} = 0{,}462$ (*d*)

Fig. 1. ROC curves of detection of Byzantine agents for four configurations N, f

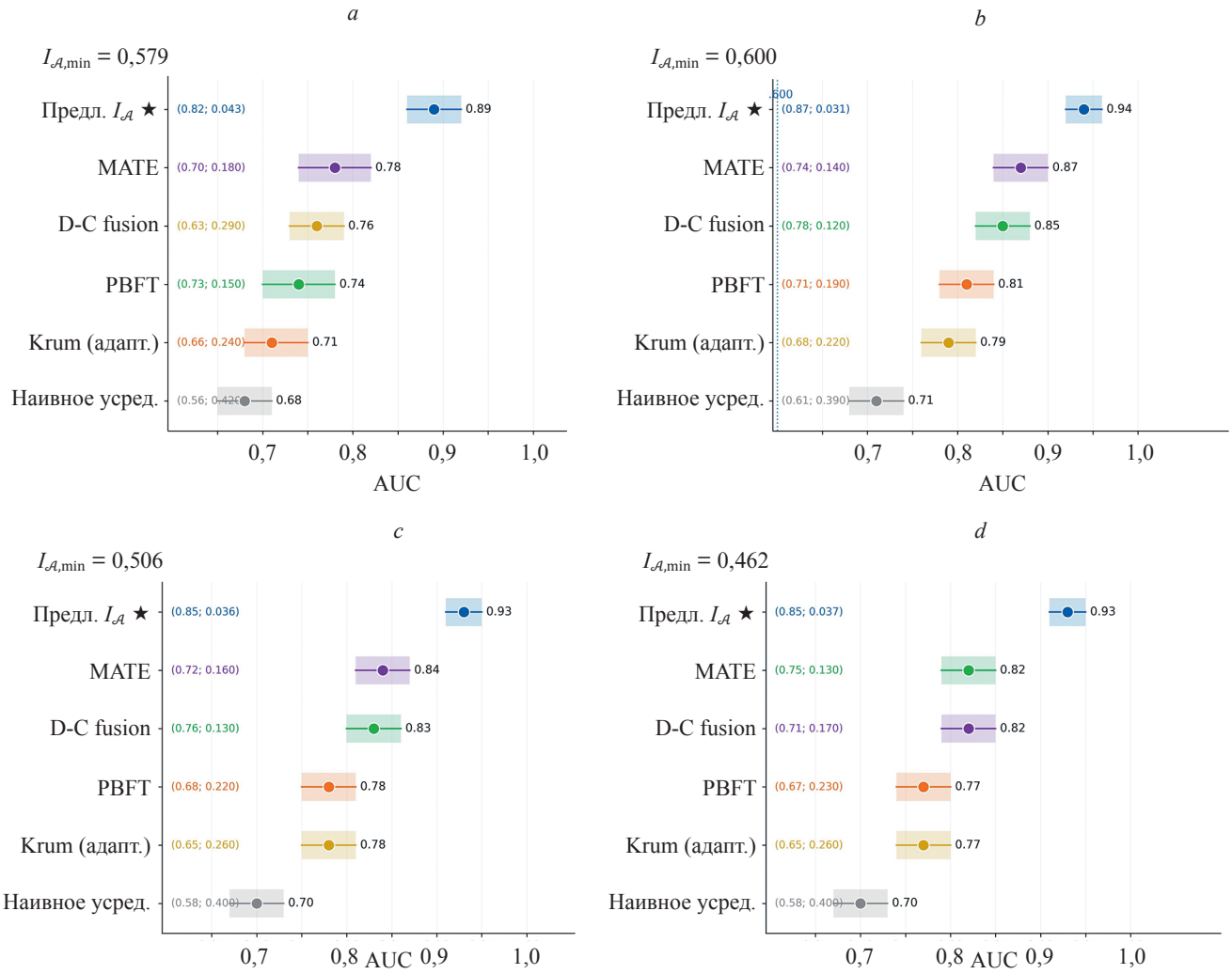

Рис. 2. AUC ± 95 % доверительный интервал (метод DeLong) и рабочая точка (P_D; P_{FA}) с конфигурациями N, f: $N = 5, f = 1$ (20 %), $I_{\mathcal{A},\min} = 0{,}579$ (*a*); $N = 10, f = 2$ (20 %), $I_{\mathcal{A},\min} = 0{,}6$ (*b*); $N = 15, f = 4$ (27 %), $I_{\mathcal{A},\min} = 0{,}506$ (*c*); $N = 20, f = 6$ (30 %), $I_{\mathcal{A},\min} = 0{,}462$ (*d*)

Fig. 2. AUC ± 95 % CI (DeLong method) and operating point (P_D; P_{FA}) at configuration N, f

вала с 0,031 до 0,014: рост шума увеличил отступ штатного класса от порога 0,6. Вероятность пропуска Byzantine-атаки ($f = 2$) сохранялась ниже 0,08 во всем диапазоне $\sigma_{loc} \in [0{,}5;\ 2{,}0]$ м — порог 0,6 устойчив к деградации сенсоров.

Выбор порога $I_{\min} = 0{,}6$ обоснован тремя взаимодополняющими аргументами, отвечающими на требование оптимальности над полным ансамблем реализаций [17, 18]. Аналитически: верхняя граница $I_{\mathcal{A}}^{(N,f)} = 0{,}622$ строго ниже штатного уровня $\langle I_{\mathcal{A}} \rangle = 0{,}942$ порог $I_{\min} = 0{,}6$ расположен строго между ними.

По критерию Неймана–Пирсона при $P_{FA} \leq 0{,}05$ оптимальный порог лежит в интервале [0,57; 0,63]; $I_{\min} = 0{,}6$ строго внутри данного интервала.

Сопоставительно: байесовский оптимальный порог $\theta^* = 0{,}593 \pm 0{,}012$ (95 % ДИ) совпадает с аналитическим значением с точностью 1,2 % (МакНемар, $p = 0{,}43$).

В отличие от байесовского детектора, значение аналитического порога определяется только параметрами сети (N, f) и не требует знания распределения стратегии Byzantine-агента — принципиальное преимущество в условиях адаптивного противника.

Сравнительный анализ ROC/AUC методов обнаружения при разных конфигурациях N, f Byzantine-агентов приведены на рис. 1 и 2.

Заключение

Представлена формальная математическая модель для оценки информационной целостности в системах кооперативного восприятия V2X. Разработана модель, включающая состояние агента $\mathbf{x}_i(t) = (\mathbf{p}_i(t), \mathbf{v}_i(t), \boldsymbol{\theta}_i(t))$, локальное восприятие $\mathcal{O}_i(t)$ и динамический граф связности $G(t)$, классифицированы типы атак Byzantine-агентов: фантомный объект, сокрытие объекта, искажение позиции. Введена метрика $I_{\mathcal{A}}(t_k) \in [0,\ 1]$ на основе расстояния Махаланобиса и вычислительной сложностью $O(N^2M^2)$. Приведена верификация, экспериментальное тестирование с существующими методами обнаружения Byzantine-агентов, приведены теоретические гарантии метрики при f Byzantine-агентах

$I_{\mathcal{A}} \le (N-f)(N-f-1)/[N(N-1)]$. Планируются следующие направления дальнейших исследований: адаптивное доверие, через расширение модели динамических оценок доверия с защитой от Sybil-атак, позволяющая системе обучаться на историческом поведении агентов; введение оценок доверия с адаптивными весами компонентов, учитывающими контекст ситуации (консенсусная, сенсорная и физическая консистентность); интеграция с формальной верификацией через уравнение Гамильтона–Якоби анализа достижимости для обеспечения гарантий безопасности. Проведение экспериментальной валидации предложенной модели на основе набора данных V2X-Seq. Для устранения ложных тревог при кратковременных провалах (потеря V2X-пакета, смена топологии графа) применяется логика: вход в режим DEGRADED требует $n_{enter} \ge 2$ последовательных шагов с $I_{\mathcal{A}}(t_k) < I_{\min}$; выход из DEGRADED требует $n_{exit} \ge 3$ последовательных шагов с $I_{\mathcal{A}}(t_k) \ge I_{\min}$. Единичный провал с немедленным восстановлением не изменяет режим работы. Конкретные значения n_{enter} и n_{exit} будут откалиброваны на данных V2X-Seq в следующей работе.

Литература

1. Yin D., Chen Y., Kannan R., Bartlett P. Byzantine-robust distributed learning: Towards optimal statistical rates // Proc. of the 35th International Conference on Machine Learning, PMLR. 2018. V. 80. P. 5650–5659.
2. Yu H., Yang W., Ruan H., Yang Z., Tang Y., Gao X., et al. V2x-seq: A large-scale sequential dataset for vehicle-infrastructure cooperative perception and forecasting // Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2023. P. 5486–5495. doi: 10.1109/CVPR52729.2023.00531
3. Lamport L., Shostak R., Pease M. The Byzantine Generals Problem // ACM Transactions on Programming Languages and Systems (TOPLAS). 1982. V. 4. N 3. P. 382–401. doi: 10.1145/357172.357176
4. Castro M., Liskov B. Practical Byzantine fault tolerance // Proc. of the 3rd Symposium on Operating Systems Design and Implementation. 1999. P. 173–186.
5. Bitok H.J., Wang M., Desmond D. Consensus on the Internet of vehicles: a systematic literature review // World Electric Vehicle Journal. 2025. V. 16. N 11. P. 616. doi: 10.3390/wevj16110616
6. Nagarajan S.M., Devarajan G.G., Ramana T.V., Jerlin M.A., Bashir A.K., Al-Otaibi Y.D. Adversarial deep learning based Dampster–Shafer data fusion model for intelligent transportation system // Information Fusion. 2024. V. 102. P. 102050. doi: 10.1016/j.inffus.2023.102050
7. Gil S., Kumar S., Mazumder M., Katabi D., Rus D. Guaranteeing spoof-resilient multi-robot networks // Autonomous Robots. 2017. V. 41. N 6. P. 1383–1400. doi: 10.1007/s10514-017-9621-5
8. Li X., Lu L., Ni W., Jamalipour A., Zhang D., Du H. Federated multi-agent deep reinforcement learning for resource allocation of vehicle-to-vehicle communications // IEEE Transactions on Vehicular Technology. 2022. V. 71. N 8. P. 8810–8824. doi: 10.1109/tvt.2022.3173057
9. Chen Y., Zhang X., Zhang K., Wang M., Zhu X. Byzantine-robust online and offline distributed reinforcement learning // Proc. of the 26th International Conference on Artificial Intelligence and Statistics (AISTATS). PMLR. 2023. V. 206. P. 3230–3269.
10. Blanchard P., El Mhamdi E.M., Guerraoui R., Stainer J. Machine learning with adversaries: Byzantine tolerant gradient descent // Proc. of the 31st Conference on Neural Information Processing System. 2017. P. 119–129.
11. Зикратов И.А., Гуртов А.В., Зикратова Т.В., Козлова Е.В. Совершенствование Police Office Model для обеспечения безопасности роевых робототехнических систем // Научно-технический вестник информационных технологий, механики и оптики. 2014. № 5 (93). С. 99–109.
12. Hallyburton R.S., Pajic M. Security-aware sensor fusion with MATE: the multi-agent trust estimator // Proc. of the ACM SIGSAC Conference on Computer and Communications Security. 2025. P. 2009–2023. doi: 10.1145/3719027.3765193
13. Huang T., Liu J., Zhou X., Nguyen D.C., Azghadi M.R., Xia Y., et al. V2X Cooperative perception for autonomous driving: recent advances and challenges // arXiv. 2023. 10.48550/arXiv.2310.03525. doi: 10.48550/arXiv.2310.03525
14. Wan L., Gupta P., Eich A., Kettelgerdes M., Keen H.E., Klöppel-Gersdorf M., et al. VALISENS: a validated innovative multi-sensor system for cooperative automated driving // arXiv. 2025. arXiv:2505.06980v2. doi: 10.48550/arXiv.2505.06980
15. Yoshizawa T., Singelée D., Muehlberg J.T., Delbruel S., Taherkordi A., Hughes D.,et al. A Survey of security and privacy

References

1. Yin D., Chen Y., Kannan R., Bartlett P. Byzantine-robust distributed learning: Towards optimal statistical rates. *Proc. of the 35th International Conference on Machine Learning, PMLR*, 2018, vol. 80, pp. 5650–5659.
2. Yu H., Yang W., Ruan H., Yang Z., Tang Y., Gao X., et al. V2x-seq: A large-scale sequential dataset for vehicle-infrastructure cooperative perception and forecasting. *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 5486–5495. doi: 10.1109/CVPR52729.2023.00531
3. Lamport L., Shostak R., Pease M. The Byzantine Generals Problem. *ACM Transactions on Programming Languages and Systems (TOPLAS)*, 1982, vol. 4, no. 3, pp. 382–401. doi: 10.1145/357172.357176
4. Castro M., Liskov B. Practical Byzantine fault tolerance. *Proc. of the 3rd Symposium on Operating Systems Design and Implementation*, 1999, pp. 173–186.
5. Bitok H.J., Wang M., Desmond D. Consensus on the Internet of vehicles: a systematic literature review. *World Electric Vehicle Journal*, 2025, vol. 16, no. 11, pp. 616. doi: 10.3390/wevj16110616
6. Nagarajan S.M., Devarajan G.G., Ramana T.V., Jerlin M.A., Bashir A.K., Al-Otaibi Y.D. Adversarial deep learning based Dampster–Shafer data fusion model for intelligent transportation system. *Information Fusion*, 2024, vol. 102, pp. 102050. doi: 10.1016/j.inffus.2023.102050
7. Gil S., Kumar S., Mazumder M., Katabi D., Rus D. Guaranteeing spoof-resilient multi-robot networks. *Autonomous Robots*, 2017, vol. 41, no. 6, pp. 1383–1400. doi: 10.1007/s10514-017-9621-5
8. Li X., Lu L., Ni W., Jamalipour A., Zhang D., Du H. Federated multi-agent deep reinforcement learning for resource allocation of vehicle-to-vehicle communications. *IEEE Transactions on Vehicular Technology*, 2022, vol. 71, no. 8, pp. 8810–8824. doi: 10.1109/tvt.2022.3173057
9. Chen Y., Zhang X., Zhang K., Wang M., Zhu X. Byzantine-robust online and offline distributed reinforcement learning. *Proc. of the 26th International Conference on Artificial Intelligence and Statistics (AISTATS)*, PMLR, 2023, vol. 206, pp. 3230–3269.
10. Blanchard P., El Mhamdi E.M., Guerraoui R., Stainer J. Machine learning with adversaries: Byzantine tolerant gradient descent. *Proc. of the 31st Conference on Neural Information Processing System*, 2017, pp. 119–129.
11. Zikratov I.A., Gurtov A.V., Zikratova T.V., Kozlova E.V. Police Office Model improvement for security of swarm robotic systems. *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, 2014, no. 5 (93). (in Russian)
12. Hallyburton R.S., Pajic M. Security-aware sensor fusion with MATE: the multi-agent trust estimator. *Proc. of the ACM SIGSAC Conference on Computer and Communications Security*, 2025, pp. 2009–2023. doi: 10.1145/3719027.3765193
13. Huang T., Liu J., Zhou X., Nguyen D.C., Azghadi M.R., Xia Y., et al. V2X Cooperative perception for autonomous driving: recent advances and challenges. *arXiv*, 2023. 10.48550/arXiv.2310.03525. doi: 10.48550/arXiv.2310.03525
14. Wan L., Gupta P., Eich A., Kettelgerdes M., Keen H.E., Klöppel-Gersdorf M., et al. VALISENS: a validated innovative multi-sensor system for cooperative automated driving. *arXiv*, 2025. arXiv:2505.06980v2. doi: 10.48550/arXiv.2505.06980
15. Yoshizawa T., Singelée D., Muehlberg J.T., Delbruel S., Taherkordi A., Hughes D., et al. A Survey of security and privacy

issues in V2X communication systems // ACM Computing Surveys. 2023. V. 55. N 9. P. 1–36. doi: 10.1145/3558052
16. DeLong E.R., DeLong D.M., Clarke-Pearson D.L. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach // Biometrics. 1988. V. 44. N 3. P. 837–845.
17. Van Trees H.L. Detection, Estimation, and Modulation Theory. Part I. Wiley-Interscience, 2001. 716 p.
18. Ширман Я.Д. Радиоэлектронные системы: основы построения и теория. Справочник. М.: Радиотехника, 2007. 512 с.
19. Kay S.M. Fundamentals of Statistical Signal Processing: Detection Theory. Prentice-Hall PTR, 1998. 560 p.
20. Левин Б.Р. Теоретические основы статистической радиотехники. М.: Радио и связь, 1989. 656 с.
21. Zhao J., Li Q., Ma X., Yu F.R. Computation offloading for edge intelligence in two-tier heterogeneous networks // IEEE Transactions on Network Science and Engineering. 2024. V. 11. N 2. P. 1872–1884. doi: 10.1109/tnse.2023.3332949
22. Chen S., Hu J., Shi Y., Zhao L., Li W. A vision of C-V2X: technologies, field testing, and challenges with chinese development // IEEE Internet of Things Journal. 2020. V. 7. N 5. P. 3872–3881. doi: 10.1109/jiot.2020.2974823
23. Rauch A., Klanner F., Rasshofer R., Dietmayer K. Car2X-based perception in a high-level fusion architecture for cooperative perception systems // Proc. of the IEEE Intelligent Vehicles Symposium. 2012. P. 270–275. doi: 10.1109/ivs.2012.6232130
24. Aoki S., Higuchi T., Altintas O. Cooperative perception with deep reinforcement learning for connected vehicles // Proc. of the IEEE Intelligent Vehicles Symposium. 2020. P. 328–334. doi: 10.1109/iv47402.2020.9304570
25. Wang T.-H., Manivasagam S., Liang M., Yang B., Zeng W., Urtasun R. V2VNet: Vehicle-to-vehicle communication for joint perception and prediction // Lecture Notes in Computer Science. 2020. V. 12347. P. 605–621. doi: 10.1007/978-3-030-58536-5_36

issues in V2X communication systems. *ACM Computing Surveys*, 2023, vol. 55, no. 9, pp. 1–36. doi: 10.1145/3558052
16. DeLong E.R., DeLong D.M., Clarke-Pearson D.L. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*, 1988, vol. 44, no. 3, pp. 837–845.
17. Van Trees H.L. *Detection, Estimation, and Modulation Theory*. Part I. Wiley-Interscience, 2001, 716 p.
18. Shirman Ya.D. *Radioelectronic Systems: Fundamentals of Construction and Theory*. Directory. Moscow, Radiotehnika, 2007, 512 p. (in Russian)
19. Kay S.M. *Fundamentals of Statistical Signal Processing: Detection Theory*. Prentice-Hall PTR, 1998, 560 p.
20. Levin B.R. *Theoretical Fundamentals of Statistical Radio Engineering*. Moscow, Radio i svyaz, 1989, 656 c. (in Russian)
21. Zhao J., Li Q., Ma X., Yu F.R. Computation offloading for edge intelligence in two-tier heterogeneous networks. *IEEE Transactions on Network Science and Engineering*, 2024, vol. 11, no. 2, pp. 1872–1884. doi: 10.1109/tnse.2023.3332949
22. Chen S., Hu J., Shi Y., Zhao L., Li W. A vision of C-V2X: technologies, field testing, and challenges with chinese development. *IEEE Internet of Things Journal*, 2020, vol. 7, no. 5, pp. 3872–3881. doi: 10.1109/jiot.2020.2974823
23. Rauch A., Klanner F., Rasshofer R., Dietmayer K. Car2X-based perception in a high-level fusion architecture for cooperative perception systems. *Proc. of the IEEE Intelligent Vehicles Symposium*, 2012, pp. 270–275. doi: 10.1109/ivs.2012.6232130
24. Aoki S., Higuchi T., Altintas O. Cooperative perception with deep reinforcement learning for connected vehicles. *Proc. of the IEEE Intelligent Vehicles Symposium*, 2020, pp. 328–334. doi: 10.1109/iv47402.2020.9304570
25. Wang T.-H., Manivasagam S., Liang M., Yang B., Zeng W., Urtasun R. V2VNet: Vehicle-to-vehicle communication for joint perception and prediction. *Lecture Notes in Computer Science*, 2020, vol. 12347, pp. 605–621. doi: 10.1007/978-3-030-58536-5_36

Авторы

Мыськив Иван Иванович — начальник отдела, ООО «ДИДЖИТАЛ ДИЗАЙН ИТ», Санкт-Петербург, 199155, Российская Федерация; младший научный сотрудник, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, https://orcid.org/0009-0002-3092-6528, myskiv.i@itmo.ru

Заколдаев Данил Анатольевич — кандидат технических наук, доцент, независимый исследователь, Санкт-Петербург, Российская Федерация, SC 57021875400, https://orcid.org/0000-0002-2520-1998, d.zakoldaev@itmo.ru

Менщиков Александр Алексеевич — кандидат технических наук, доцент, декан, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, SC 57196052901, https://orcid.org/0000-0002-2287-4310, menshikov@itmo.ru

Authors

Ivan I. Myskiv — Head of Department, Digital Design IT LLC, Saint Petersburg, 199155, Russian Federation; Junior Researcher, ITMO University, Saint Petersburg, 197101, Russian Federation, https://orcid.org/0009-0002-3092-6528, myskiv.i@itmo.ru

Danil A. Zakoldaev — PhD, Associate Professor, Independent Researcher, Saint Petersburg, Russian Federation, SC 57021875400, https://orcid.org/0000-0002-2520-1998, d.zakoldaev@itmo.ru

Alexander A. Menshchikov — PhD, Associate Professor, Dean, ITMO University, Saint Petersburg, 197101, Russian Federation, SC 57196052901, https://orcid.org/0000-0002-2287-4310, menshikov@itmo.ru

Статья поступила в редакцию 26.02.2026
Одобрена после рецензирования 03.06.2026
Принята к печати 22.07.2026

Received 26.02.2026
Approved after reviewing 03.06.2026
Accepted 22.07.2026