ТЕХНИКАЛЫҚ ҒЫЛЫМДАР
ТЕХНИЧЕСКИЕ НАУКИ
TECHNICAL SCIENCE

CENTRAL ASIAN SCIENTIFIC JOURNAL
ISSN 3135-3061
2026. №4 (32), Том 2. С. 462–467
DOI: https://doi.org/10.5281/zenodo.24343170
Сайт журнала: https://cajournal.kz

УДК 004.056.5:004.8

AI-СИСТЕМА ОБНАРУЖЕНИЯ ФИШИНГОВЫХ ВЕБ-САЙТОВ НА ОСНОВЕ АНАЛИЗА URL, HTML И ВИЗУАЛЬНЫХ ПРИЗНАКОВ

С.Р. Зурдунова, Б.Қ. Зулпухарова, Б.К. Нуриддинов
Казахский национальный исследовательский технический университет имени К. И. Сатпаева, Алматы, Казахстан

Зурдунова Сабрина Рахимжановна — студент 4-го курса бакалавриата, Казахский национальный исследовательский технический университет имени К. И. Сатпаева, Алматы, Казахстан.

Зулпухарова Балымжан Қанатқызы — студент 4-го курса бакалавриата, Казахский национальный исследовательский технический университет имени К. И. Сатпаева, Алматы, Казахстан.

Нуриддинов Билолиддин Камалиддинович — студент 4-го курса бакалавриата, Казахский национальный исследовательский технический университет имени К. И. Сатпаева, Алматы, Казахстан.

© С.Р. Зурдунова, Б.Қ. Зулпухарова, Б.К. Нуриддинов, 2026

Аннотация: В статье рассматривается проблема обнаружения фишинговых веб-сайтов, которые представляют угрозу и опасность безопасности пользователей. Цель исследования заключается в анализе системы выявления фишинга на основе анализа веб-сайта, структуры страницы и визуальных характеристик. Методологическая основа исследования включает изучение научных статей и существующих способов обнаружения фишинга, а также обзор структуры системы с использованием машинного обучения и искусственного интеллекта. Предлагаемый подход предусматривает проверку URL адреса и углубленный анализ содержимого веб-страницы и внешнего вида сайта на признаки фишинга. Новизна работы заключается в последовательном выполнении нескольких проверок, при котором следующий этап выбирается в зависимости от результатов предыдущего. Для оценки эффективности подхода предполагается использовать показатели точности обнаружения фишинговых сайтов и количества ошибочных срабатываний. Практическая значимость заключается в том, что такой подход может помочь точнее выявлять фишинговые сайты. Особое внимание уделяется преимуществам и недостаткам существующих методов обнаружения фишинга. В работе рассматривается возможность совместного использования признаков URL-адреса и HTML-структуры для повышения информативности проверки веб-сайтов. Полученные результаты могут стать основой для дальнейшей разработки и экспериментального тестирования системы обнаружения фишинговых ресурсов.

Ключевые слова: Информационная безопасность, машинное обучение, фишинговые веб-сайты, обнаружение фишинга, анализ HTML, анализ URL, искусственный интеллект, обнаружение фишинга.

Аңдатпа: Бұл мақалада пайдаланушы қауіпсіздігіне қауіп төндіретін фишингтік веб-сайттарды анықтау мәселесі қарастырылады. Зерттеудің мақсаты - веб-сайтты талдау, бет құрылымы және визуалды сипаттамалары негізінде фишингті анықтау жүйесін талдау. Зерттеудің әдіснамалық негізі ғылыми мақалаларға және қолданыстағы фишингті анықтау әдістеріне шолуды, сондай-ақ машиналық оқыту мен жасанды интеллектті қолдана отырып, жүйе құрылымына шолуды қамтиды. Ұсынылған тәсіл URL мекенжайын тексеруді

және фишинг индикаторлары үшін веб-бет мазмұны мен сайттың сыртқы түрін терең талдауды жүргізуді қамтиды. Зерттеудің жаңалығы бірнеше тексерулерді дәйекті түрде орындауда жатыр, келесі кезең алдыңғысының нәтижелеріне негізделіп таңдалады. Тәсілдің тиімділігін бағалау үшін фишингтік сайттарды анықтаудың дәлдігі және жалған оң нәтижелер саны сияқты көрсеткіштерді пайдалану ұсынылады. Бұл тәсілдің практикалық маңыздылығы оның фишингтік сайттарды дәлірек анықтауға көмектесетіндігінде. Фишингті анықтаудың қолданыстағы әдістерінің артықшылықтары мен кемшіліктеріне ерекше назар аударылады. Мақалада веб-сайтты тексерудің ақпараттық деңгейін арттыру үшін URL және HTML құрылымының ерекшеліктерін біріктіру мүмкіндігі қарастырылады. Алынған нәтижелер фишингтік ресурстарды анықтау жүйесін одан әрі дамыту және эксперименттік сынақтан өткізу үшін негіз бола алады.

Түйінді сөздер: Ақпараттық қауіпсіздік, машиналық оқыту, фишингтік веб-сайттар, фишингті анықтау, HTML талдауы, URL талдауы, жасанды интеллект, фишингті анықтау.

ABSTRACT: This article addresses the problem of detecting phishing websites that pose a threat to user security. The study aims to analyze a phishing detection system based on the examination of website content, page structure, and visual characteristics. The research methodology involves reviewing academic literature and existing phishing detection methods, as well as examining a system architecture that utilizes machine learning and artificial intelligence. The proposed approach entails URL verification combined with an in-depth analysis of webpage content and visual appearance to identify signs of phishing. The study’s novelty lies in a sequential verification process, where the subsequent stage is determined by the outcome of the preceding one. The approach's effectiveness is evaluated using metrics such as detection accuracy and the rate of false positives. The practical significance of this work lies in the potential for more accurate identification of phishing websites. Particular attention is paid to the advantages and limitations of current detection methods. The study also explores the combined use of URL features and HTML structure to enhance the informativeness of the website analysis. The findings provide a foundation for the further development and experimental testing of a phishing detection system.

Keywords: Information security, machine learning, phishing websites, phishing detection, HTML analysis, URL analysis, artificial intelligence, phishing detection.

Введение

В настоящее время фишинговые веб-сайты остаются одной из актуальных угроз информационной безопасности. Они создаются для обмана пользователей, получения конфиденциальной информации и доступа к учетным записям. Такие сайты могут имитировать страницы банков, интернет-магазинов и других известных организаций. Развитие технологий искусственного интеллекта также облегчает создание поддельных сайтов, поэтому обнаружение фишинга требует постоянного улучшения методов защиты.

Для выявления фишинговых веб-сайтов применяются различные методы, включая анализ URL-адресов, HTML-кода, текстового содержимого и внешнего вида страниц. Методы машинного обучения позволяют классифицировать веб-сайты на основе их признаков. Однако каждый подход имеет свои ограничения. Например, анализ URL-адреса позволяет быстро проверить подозрительные признаки, но не всегда дает достаточно информации. Анализ HTML-кода и визуальных характеристик помогает получить дополнительные сведения о странице, однако требует более сложной обработки данных. Использование только одного типа признаков может снижать эффективность обнаружения сайтов, которые используют новые способы маскировки.

В связи с этим актуальной задачей является разработка системы, которая объединяет несколько методов анализа и последовательно проверяет подозрительные веб-сайты. Такой подход предполагает первоначальную проверку URL-адреса с последующим анализом HTML-структуры и визуальных характеристик страницы, если первоначальных данных недостаточно для принятия решения. Дополнительное внимание необходимо уделить

обнаружению сайтов на новых доменах и фишинговых страниц, созданных с использованием искусственного интеллекта.

Цель исследования – исследовать AI – систему обнаружения фишинговых веб-сайтов на основе анализа URL-адресов, HTML-структуры и визуальных характеристик страниц с использованием методов искусственного интеллекта.

Для достижения поставленной цели необходимо изучить существующие методы обнаружения фишинга и определить их проблемы; определить показатели оценки эффективности системы; предусмотреть возможность объяснения результатов классификации.

Объектом исследования являются процессы обнаружения фишинговых веб-сайтов в системах информационной безопасности. Предметом исследования выступают методы и модели искусственного интеллекта для анализа URL-адресов, HTML-структуры и визуальных характеристик веб-страниц.

Научная новизна предлагаемого подхода заключается в организации этапов анализа, при которой необходимость перехода к следующему уровню определяется результатами предыдущей проверки. Предполагается, что объединение различных типов признаков позволит лучше оценивать потенциально опасные веб-сайты. Практическая значимость работы заключается в возможности применения предложенного подхода в средствах защиты веб-доступа для повышения надежности обнаружения фишинговых сайтов.

Материалы и методы исследования

Материалами исследования являются научные публикации, посвященные обнаружению фишинговых веб-сайтов, применению искусственного интеллекта в информационной безопасности и анализу признаков веб-страниц. В работе рассматриваются существующие подходы к анализу URL-адресов и HTML-структуры сайтов, а также их преимущества и недостатки. Экспериментальные данные в рамках настоящего исследования не собирались, обучение моделей и тестирование программного обеспечения не проводились.

В исследовании применяются методы анализа научной литературы, сравнения и систематизации существующих подходов. Анализ литературы используется для определения основных признаков фишинговых веб-сайтов и выявления ограничений существующих методов. Сравнение позволяет рассмотреть возможности URL-анализа и HTML-анализа, а систематизация – определить последовательность их применения в предлагаемой системе.

Первый уровень предлагаемой архитектуры предусматривает анализ URL-адреса веб-сайта. На этом этапе рассматриваются такие признаки, как длина адреса, структура доменного имени, наличие специальных символов и использование IP-адреса вместо доменного имени. Эти признаки могут указывать на потенциальный риск, однако сами по себе не доказывают, что сайт является фишинговым.

Второй уровень предусматривает анализ HTML-структуры страницы. При необходимости дополнительной проверки рассматриваются формы ввода данных, ссылки, адреса переходов и другие элементы веб-страницы. Цель данного этапа заключается в получении дополнительной информации для оценки потенциальной опасности сайта.

На основе рассмотренных методов предлагается многоуровневая архитектура, в которой сначала выполняется анализ URL-адреса, а затем при наличии подозрительных признаков или недостаточной определенности результата проводится HTML-анализ. Такой принцип позволяет разделить проверку на последовательные этапы и учитывать результаты предыдущего анализа при выборе следующего действия.

Результаты и их обсуждение

В результате анализа научной литературы были рассмотрены основные подходы к обнаружению фишинговых веб-сайтов, основанные на анализе URL-адресов и HTML-структуры страниц. Рассмотрение этих подходов позволило определить их основные

особенности и ограничения, а также предложить свою многоуровневую систему обнаружения фишинга.

В результате аналитического обзора научных работ, книг, статей, диссертаций и патентов были рассмотрены основные подходы к обнаружению фишинговых веб-сайтов. Анализ показал, что методы обнаружения фишинга развиваются от использования черных списков и эвристических правил к машинному обучению, глубокому обучению и системам, объединяющим несколько типов признаков.

Черные списки позволяют быстро блокировать известные фишинговые адреса, но не обеспечивают эффективного обнаружения новых доменов, которые еще не были внесены в базу. Эвристические методы позволяют проверять URL по заранее определенным признакам, но их эффективность зависит от выбранных правил. Машинное обучение таких моделей зависит от обучающих данных и их соответствия реальному миру.

Современные исследования также рассматривают анализ HTML-структуры, визуальных характеристик и текстового содержимого веб-страниц. Например, в работе по обнаружению фишинговых сайтов с использованием моделей глубокого обучения рассматривается автоматическая классификация веб-ресурсов. В диссертации Khalid Alrokhaimi изучается обнаружение фишинговых URL на основе лексических признаков, при этом визуальные характеристики страниц не анализируются. В патенте «Выявление фишинговых сайтов с помощью машинного обучения» рассматривается классификация ресурсов на основе признаков.

Таблица 1. Сравнение методов обнаружения фишинговых веб-сайтов

Метод обнаружения	Преимущества	Ограничения
Черные списки URL-адресов	Быстрая проверка адресов, простота реализации, возможность блокировки ранее выявленных фишинговых ресурсов.	Не всегда обнаруживают новые фишинговые сайты, отсутствующие в базе данных.
Анализ URL и эвристические методы	Позволяют выявлять подозрительные признаки адреса сайта, включая необычную структуру домена, специальные символы и чрезмерную длину URL.	Отдельные признаки не доказывают наличие фишинга; злоумышленники могут изменять адреса для обхода проверки.
Машинное обучение	Позволяет выявлять закономерности в характеристиках легитимных и фишинговых сайтов.	Эффективность зависит от соответствия данных реальным угрозам
Глубокое обучение	Позволяет автоматически извлекать признаки.	Может требовать значительных вычислительных ресурсов и большого объема обучающих данных.
Анализ HTML-структуры	Позволяет исследовать формы ввода данных и другие элементы веб-страницы.	Требует получения и обработки содержимого страницы.
Многоуровневый анализ URL и HTML	Объединяет признаки адреса и структуры страницы.	Требует разработки механизма перехода между этапами и проведение экспериментов.

Сравнение показывает, что ни один из рассмотренных подходов не следует считать универсальным. Черные списки эффективны при проверке уже известных угроз, однако ограничены полнотой и актуальностью содержащихся в них данных. Эвристический анализ позволяет оценивать URL-адреса без сложной обработки содержимого страницы, но отдельные признаки не всегда позволяют достоверно определить характер сайта. Методы машинного и глубокого обучения расширяют возможности классификации, однако требуют подходящих обучающих данных и проверки устойчивости к изменению характеристик фишинговых ресурсов.

На основании проведённого обзора предлагается концептуальная многоуровневая система обнаружения фишинговых веб-сайтов, объединяющая анализ URL-адреса и HTML-структуры страницы. На первом этапе выполняется анализ URL, включающий исследование длины адреса, структуры доменного имени, специальных символов и других лексических

признаков. Если полученных данных недостаточно для принятия решения или выявлены подозрительные признаки, система переходит ко второму этапу — анализу HTML-структуры страницы.

На втором этапе могут исследоваться формы ввода пользовательских данных, ссылки, адреса переходов и другие элементы HTML-документа. Результаты обоих этапов используются для формирования итоговой оценки риска. При этом отдельный признак не рассматривается как достаточное доказательство фишинга: решение должно основываться на совокупности характеристик и результатах комплексного анализа.

Предлагаемая последовательность позволяет разделить процесс обнаружения на предварительную и углубленную проверки. Предполагается, что такой подход может сократить необходимость проведения полного анализа каждой страницы и тем самым уменьшить вычислительную нагрузку. Однако это преимущество требует экспериментальной проверки и не может считаться доказанным результатом настоящей работы.

Таким образом, аналитический обзор позволил выявить ограничения существующих методов и обосновать необходимость последовательного применения анализа URL и HTML.

Научно-практическая значимость предложенного подхода заключается в разработке концептуальной архитектуры, предусматривающей использование разных типов признаков на последовательных этапах проверки. Поскольку программная реализация и экспериментальное тестирование системы не проводились, ее фактическая точность и эффективность пока не установлены.

В дальнейшем предполагается реализовать прототип системы и оценить его на размеченном наборе легитимных и фишинговых веб-сайтов. Для количественной оценки могут использоваться метрики Precision, Recall, F1-score и ROC-AUC. Дополнительно целесообразно исследовать частоту ложноположительных и ложноотрицательных классификаций, время обработки страницы и устойчивость системы к появлению новых фишинговых доменов.

Заключение

В ходе исследования были рассмотрены современные методы обнаружения фишинговых веб-сайтов и выявлены их основные преимущества и ограничения. На основе аналитического обзора предложена концептуальная многоуровневая система обнаружения фишинга, предусматривающая последовательный анализ URL-адреса и HTML-структуры веб-страницы.

Особенность предложенного подхода заключается в переходе от предварительной проверки URL к углубленному анализу HTML при выявлении подозрительных признаков или недостаточной определенности результата предыдущего этапа. Такой принцип позволяет учитывать различные характеристики веб-ресурса при оценке потенциального риска. Научная новизна работы состоит в концептуальной организации последовательного анализа двух типов признаков в рамках единой архитектуры обнаружения фишинга.

Практическая значимость исследования заключается в возможности дальнейшего применения предложенной архитектуры при разработке средств защиты пользователей от фишинговых веб-сайтов. Вместе с тем, поскольку программная реализация и экспериментальная оценка не проводились, эффективность предложенной системы требует дополнительной проверки.

В дальнейшем планируется разработать программный прототип, обучить и протестировать модели машинного обучения на размеченных наборах легитимных и фишинговых веб-сайтов, а также оценить качество классификации с использованием метрик Precision, Recall, F1-score и ROC-AUC. Полученные результаты позволят определить эффективность предложенного подхода и возможности его практического применения.

© Central Asian Scientific Journal, 2026

Список литературы:

1. Bhanu Prasad, Y., Dondeti, V., & Dondeti, R. (2026). Phishing URL Detection in Real-Time: Machine Learning and Deep Learning Approaches on a Custom Threat Intelligence Dataset. Procedia Computer Science. https://www.sciencedirect.com/science/article/pii/S1877050926021745

2. Fares, H., Kilani, J., Fagroud, F., Toumi, H., Lakrami, F., Baddi, Y., & Aknin, N. (2024). Machine Learning Approach for Email Phishing Detection. https://www.sciencedirect.com/science/article/pii/S1877050924034136

3. Zara, U., Ayyub, K., Khan, H. U., Daud, A., Alsahfi, T., & Ahmad, S. G. (2024). Phishing Website Detection Using Deep Learning Models. IEEE. https://ieeexplore.ieee.org/document/10735206

4. Zhou, L., Gaurav, A., Arya, V., Attar, R. W., Bansal, S., & Alhomoud, A. (2024). Enhancing Phishing Detection in Semantic Web Systems Using Optimized Deep Learning Models. https://www.sciencedirect.com/org/science/article/pii/S1552628324000929

5. Bustio-Martínez, L., Herrera-Semenets, V., González-Ordiano, J. A., Pérez-Guadarrama, Y., Zúñiga-Morales, L. N., Montoya-Godínez, D., Álvarez-Carmona, M. A., & van den Berg, J. (2025). Enhanced Phishing Detection Using Multimodal Data. Knowledge-Based Systems. https://www.sciencedirect.com/science/article/abs/pii/S0950705125021434

6. Khandelwal, S., & Das, R. (2022). Phishing Detection Using Content-Based Image Classification. CRC Press. https://books.google.kz/books?id=HKkkzwEACAAJ

7. Sonowal, G. (2022). Phishing and Communication Channels: A Guide to Identifying and Mitigating Phishing Attacks. Apress. https://doi.org/10.1007/978-1-4842-7744-7

8. Bertino, E., Bhardwaj, S., Cicala, F., Gong, S., Karim, I., Katsis, C., Lee, H., Li, A. S., et al. (2023). Machine Learning Techniques for Cybersecurity. Springer. https://doi.org/10.1007/978-3-031-28259-1

9. Rochester Institute of Technology. (б. д.). [Название диссертации и автор требуют уточнения]. RIT Digital Institutional Repository. https://repository.rit.edu/cgi/viewcontent.cgi?article=13539&context=theses

10. Rashid, F. (2025). Robust Phishing URL Detection Through Deep Learning and Domain Shift Mitigation [Докторская диссертация]. University of Sydney. https://rahul.gopinath.org/phdcommittee/2025/05/26/fariza-rashid/

11. Kuo, J. H. C., Lee, C. P., & Lee, J. K. C. (2019). Phishing detection with machine learning. Патент США US10313352B2. https://patents.google.com/patent/US10313352B2/en

12. Akhter, S., Pandey, S., Rizvi, S., et al. (2023). Detection of phishing websites using machine learning. Патентная заявка США US20230065787A1. https://patents.google.com/patent/US20230065787A1/en

13. International Organization for Standardization, & International Electrotechnical Commission. (2022). ISO/IEC 27001:2022. Information security, cybersecurity and privacy protection — Information security management systems — Requirements. https://www.iso.org/standard/27001?archive=all

14. Alam, M., Rahman, M. L., Paul, S. K., Hays, A. W., Hussain, A., Huq, M. I., & Saxena, N. (2026). SoK: PHILTER: Uncovering Security and Functional Gaps in AI-based Phishing Website Detection Literature via an LLM-based Reasoning Framework. 35th USENIX Security Symposium (USENIX Security 26), 4981–5000. https://www.usenix.org/conference/usenixsecurity26/presentation/alam

15. Li, W., Manickam, S., Chong, Y.-W., et al. (2026). PhishIntentionLLM: Uncovering Phishing Website Intentions Through Multi-agent Retrieval-Augmented Generation. Digital Forensics and Cyber Crime (ICDF2C 2025), 88–106. Springer. https://doi.org/10.1007/978-3-032-22545-0_6

16. Czybik, S., Kouam, A. J., Heubl, P., Nold, J. M., & Rieck, K. (2026). A Large-Scale Study of Personalized Phishing using Large Language Models. 35th USENIX Security Symposium (USENIX Security 26), 1687–1706. https://www.usenix.org/conference/usenixsecurity26/presentation/czybik

17. Anti-Phishing Working Group. (б. д.). Phishing Activity Trends Reports. https://apwg.org/trendreports

18. Microsoft. (2025). Microsoft Digital Defense Report 2025. https://www.microsoft.com/en-us/corporate-responsibility/topics/cybersecurity/reports/microsoft-digital-defense-report-2025/

19. Cisco. (б. д.). Cyber Threat Trends Report. https://www.cisco.com/c/en/us/products/security/cyber-threat-trends-report.html