МРНТИ 81.93.29

10.51889/2959-5894.2025.89.1.022

Д.А. Султанова

Казахский национальный университет им. аль-Фараби, г.Алматы, Казахстан
e-mail: dan1ssimo0320@gmail.com

СРАВНИТЕЛЬНОЕ ИССЛЕДОВАНИЕ АЛГОРИТМОВ МАШИННОГО ОБУЧЕНИЯ ДЛЯ ОБНАРУЖЕНИЯ ФИШИНГОВЫХ САЙТОВ

Аннотация

Фишинговые атаки на веб-сайты – это тип кибератаки, в ходе которой злоумышленники создают мошеннические веб-сайты, имитирующие законные платформы, такие как социальные сети, с целью обмануть ничего не подозревающих пользователей и заставить их разгласить конфиденциальную информацию. Это включает в себя пароли, данные кредитных карт, имена пользователей и другие персональные данные. Эти фишинговые веб-сайты выглядят аутентично и часто используют различные методы, такие как подмена URL-адресов, социальная инженерия и фишинг по электронной почте или текстовым сообщениям, чтобы заманить жертв раскрыть свою конфиденциальную информацию. Веб-приложения становятся все более сложными, и их трудно идентифицировать с первого взгляда, особенно когда они используют методы шифрования и обфускации. Чтобы эффективно обнаруживать и предотвращать загрузку фишинговых веб-приложений на сервер в режиме реального времени, необходимо разработать машинное обучение. В дополнение к анализу алгоритмов машинного обучения для выявления атак на основе веб-приложений, исследование калибрует свежие анализы путем выполнения алгоритмов машинного обучения и подтверждения результатов. В исследовании используются уникальные и категоризированные результаты из набора данных машинного обучения. Согласно результатам, полученным в результате экспериментального и сравнительного анализа применяемых алгоритмов классификации, метод случайного леса продемонстрировала высочайшую точность, достигнув показателя 97%, за ней следуют модель дерева решений с показателем 94% и экстремальный градиентный бустинг (XGBoost).

Ключевые слова: дерево решений, логистическая регрессия, фишинг, случайный лес, XGBoost.

Д.А.Султанова
Әл-Фараби атындағы Қазақ ұлттық университеті, Алматы қ., Қазақстан

ФИШИНГТІК САЙТТАРДЫ АНЫҚТАУҒА АРНАЛҒАН МАШИНАЛЫҚ ОҚЫТУ АЛГОРИТМДЕРІНІҢ САЛЫСТЫРМАЛЫ ЗЕРТТЕУІ

Аңдатпа

Веб-сайттарға фишингтік шабуылдар – бұл кибершабуылдың бір түрі, онда шабуылдаушылар бейхабар пайдаланушыларды алдау және құпия ақпаратты жария ету мақсатында әлеуметтік медиа сияқты заңды платформаларға еліктейтін алаяқтық веб-сайттар жасайды. Бұған Құпия сөздер, несие картасы деректері, пайдаланушы имен және басқа да жеке деректер кіреді. Бұл фишингтік веб-сайттар шынайы көрінеді және құрбандарды өздерінің құпия ақпаратын ашуға тарту үшін URL мекенжайларын ауыстыру, әлеуметтік инженерия және электрондық пошта немесе мәтіндік фишинг сияқты әртүрлі әдістерді жиі пайдаланады. Веб-қосымшалар барған сайын күрделене түседі және оларды бір қарағанда анықтау қиын, әсіресе шифрлау және обфускация әдістерін қолданған кезде. Фишингтік веб-қосымшалардың нақты уақыт режимінде серверге жүктелуін тиімді анықтау және болдырмау үшін машиналық оқытуды дамыту қажет. Веб-қосымшаларға негізделген шабуылдарды анықтау үшін машиналық оқыту алгоритмдерін талдаудан басқа, зерттеу Машиналық оқыту алгоритмдерін орындау және нәтижелерді растау арқылы жаңа талдауларды калибрлейді. Зерттеу Машиналық оқыту деректер жиынтығынан бірегей және санатталған нәтижелерді пайдаланады. Қолданылатын жіктеу алгоритмдерін эксперименттік және салыстырмалы талдау нәтижесінде алынған нәтижелерге сәйкес, кездейсоқ орман моделі ең жоғары дәлдікті көрсетті, әсерлі 97%, одан кейін 94% шешім ағашының моделі және экстремалды градиентті күшейту (XGBoost).

Түйін сөздер: шешім ағашы, логистикалық регрессия, фишинг, кездейсоқ орман, XGBoost

D.A. Sultanova
Al-Farabi Kazakh National University, Almaty, Kazakhstan

A COMPARATIVE STUDY OF MACHINE LEARNING ALGORITHMS FOR PHISHING SITE DETECTION

Abstract

Phishing attacks on websites are a type of cyberattack in which attackers create fraudulent websites that imitate legitimate platforms such as social media in order to deceive unsuspecting users and disclose confidential information. This includes passwords, credit card data, user ID, and other personal data. These phishing websites look real and often use various methods such as URL swapping, social engineering, and email or text phishing to attract victims to reveal their sensitive information. Web applications are becoming more and more complex and difficult to detect at first glance, especially when using encryption and obfuscation methods. Machine learning needs to be developed to effectively detect and prevent phishing web applications from being uploaded to the server in real time. In addition to analyzing machine learning algorithms to detect attacks based on web applications, the study calibrates new analyses by executing machine learning algorithms and validating the results. The study uses unique and categorized results from machine learning data sets. According to the results obtained from the experimental and comparative analysis of the classification algorithms used, the random forest model showed the highest accuracy, with an impressive 97%, followed by the decision tree model of 94% and extreme gradient amplification (XGBoost).

Keywords: Decision Tree, Logistic Regression, Phishing, Random Forest, XGBoost.

Основные положения

В исследовании используются уникальные и категоризированные результаты из набора данных машинного обучения. Согласно результатам, полученным в результате экспериментального и сравнительного анализа применяемых алгоритмов классификации, метод случайного леса продемонстрировала высочайшую точность, достигнув показателя 97%, за ней следуют модель дерева решений с показателем 94% и экстремальный градиентный бустинг (XGBoost).

Введение

Фишинговые атаки – это мошеннические попытки, в которых киберпреступники создают вводящие в заблуждение сообщения, такие как электронные письма, сообщения или веб-сайты, которые выглядят как исходящие из надежных источников. Эти атаки стали серьезной проблемой для компаний, убытки от которых составляют около 100 миллиардов долларов в год. Кроме того, они находятся на подъеме, увеличившись на 200% по сравнению с предыдущими годами. Имеющиеся в настоящее время решения для борьбы с этими атаками неэффективны, и существует острая потребность в новых и инновационных методах защиты как компаний, так и частных лиц. В связи с растущей зависимостью от компьютеризированной финансовой деятельности и уменьшением количества операций с наличными киберпреступники используют эту тенденцию, используя методы фишинга для мошеннического получения конфиденциальной финансовой информации от ничего не подозревающих жертв. Преступные организации перешли от эксплуатации технических уязвимостей систем к эксплуатации человеческих уязвимостей, таких как неспособность отличить подлинные онлайн-ресурсы от мошеннических, таких как электронная почта и веб-сайты. Поэтому крайне важно разработать эффективные решения для смягчения этих проблем. Многие элементы повседневной жизни, включая социальные сети, онлайн-банкинг, электронную коммерцию и другие виды деятельности, переместились в Интернет из-за быстрого развития глобальных сетевых и коммуникационных технологий. Однако открытый, частный и неконтролируемый характер Интернета также создает благоприятную среду для кибератак, создавая серьезную угрозу безопасности как для сетей, так и для обычных пользователей компьютеров, даже опытных. Трудно полностью предотвратить страдания людей от фишинговых мошенничеств, даже несмотря на то, что забота и навыки пользователей важны. Фишинговый веб-сайт – это вводящий в заблуждение и мошеннический

веб-сайт, целью которого является обман и манипулирование пользователями с целью разглашения конфиденциальной информации. Эти веб-сайты обычно замаскированы под законные веб-сайты или электронные письма и часто содержат поддельные страницы входа в систему или другие формы, предназначенные для кражи информации у ничего не подозревающих пользователей. Фишинговые веб-сайты обычно используют тактику социальной инженерии, чтобы заманить пользователей предоставить свою конфиденциальную информацию, например, выдавая себя за заслуживающее доверия учреждение, такое как банк, платформа социальных сетей или сайт электронной коммерции. Как только пользователь вводит свою информацию на поддельном веб-сайте, злоумышленники могут использовать эту информацию для кражи денег, личных данных или совершения других форм мошенничества [1]. Чтобы не стать жертвой фишинговых веб-сайтов, крайне важно проявлять осторожность при вводе личной информации в Интернете. Проверка URL-адреса веб-сайта, поиск индикаторов безопасности, таких как HTTPS и значок замка, а также воздержание от перехода по ссылкам в подозрительных электронных письмах – все это необходимые меры.

Методология исследования

Целью данного исследования является анализ различных методов машинного обучения для изучения потенциальных применений пяти различных моделей классификации для выявления фишинговых веб-сайтов. Основная цель состоит в том, чтобы разработать интеллектуальную модель, способную оценить подлинность веб-сайта и определить степень обмана. Предложенный метод проиллюстрирован на рисунке 1.

Рисунок 1. Предлагаемый метод

Сбор данных. "Набор данных о фишинговых сайтах" от Kaggle [2] представляет собой набор URL-адресов, классифицированных как легальные или фишинговые веб-сайты. Набор данных содержит 11 055 URL-адресов, из которых 5 643 помечены как законные, а 5 412 - как фишинговые. Он объединяет два источника: первый - законные URL-адреса с веб-сайтов, входящих в топ-1 миллиона сайтов Alexa, а второй - известные фишинговые URL-адреса из PhishTank, службы по борьбе с фишингом, созданной сообществом. Пользователи могут сообщать о подозрительных фишинговых веб-сайтах в PhishTank для проверки. Сайт добавляется в базу данных PhishTank после подтверждения.

Этот набор данных может быть использован для ряда приложений машинного обучения, включая бинарную классификацию, обнаружение аномалий и разработку функций для идентификации фишинговых веб-сайтов.

Предварительная обработка данных. Существует несколько методов предварительной обработки, которые можно применить к набору данных фишинговых веб-сайтов, чтобы

подготовить его к использованию в моделях машинного обучения. Например, очистка данных включает в себя удаление повторяющихся или нерелевантных записей. Например, удаление URL с недопустимым форматом. Извлечение из URL-адресов ценные характеристики, такие как длина, наличие определенных ключевых слов или количество специальных символов. Нормализация данных для обеспечения согласованности. Преобразование категориальных данных, такие как метка (фишинговая или законная), в числовой формат, используемый алгоритмами машинного обучения. Если набор данных несбалансирован (один класс содержит непропорционально больше выборок, чем другой), пересчитывание данных, чтобы сбалансировать классы. Эти методы предварительной обработки повышают качество набора данных для задач машинного обучения, таких как двоичная классификация. В рабочем процессе машинного обучения разделение данных на обучающие (80 %) и тестовые наборы (20 %) является критически важным этапом, поскольку оно позволяет оценить производительность модели и способность к обобщению.

Классификация. Классификация данных включает в себя систематическую организацию и категоризацию данных по отдельным группам или классам, используя сходства или различия в их особенностях или характеристиках. Целью классификации данных является обеспечение эффективного и действенного управления, анализа и принятия решений [3]. В машинном обучении классификация данных — это общая задача, которая включает в себя обучение модели изучению отношений между признаками данных и метками классов или категориями. Обученная модель может быть впоследствии применена для составления прогнозов по классу метки новых и неопубликованных данных. Различные алгоритмы, такие как логистическая регрессия, дерево решений, XGBoost, среди прочих, могут быть использованы для классификации данных, при этом выбор соответствующего алгоритма зависит от характеристик данных и конкретной цели классификации. После успешного обучения и валидации модели она может быть применена для классификации новых данных в режиме реального времени.

В целом, классификация данных является важнейшим аспектом анализа данных и машинного обучения, обеспечивающим эффективное управление и использование больших и сложных наборов данных. Цель этой работы — использовать различные алгоритмы машинного обучения, такие как логистическая регрессия, адаптивный бустинг, дерево решений, случайный лес и XGBoost, для классификации набора данных фишинговых веб-сайтов. Каждый из этих алгоритмов обладает своими уникальными преимуществами и ограничениями, что потенциально делает их более подходящими для конкретных типов данных или задач классификации.

Логистическая регрессия. Логистическая регрессия – это популярный метод, используемый для задач двоичной классификации, где она моделирует вероятность исхода, принадлежащего к одному из двух классов. Он часто используется в задачах машинного обучения, связанных с двоичной классификацией, где цель состоит в том, чтобы разделить данные на одну из двух групп на основе набора характеристик. Основываясь на предоставленных независимых переменных, модель логистической регрессии использует сигмоидальную или логистическую функцию для прогнозирования вероятности того, что зависимая переменная будет равна 1. Сигмоидальная функция сопоставляет вещественные входные данные в диапазоне от 0 до 1, представляя вероятность. Алгоритм логистической регрессии использует оценку максимального правдоподобия для оценки коэффициентов независимых переменных. Затем эти коэффициенты используются для вычисления вероятности того, что зависимая переменная равна 1 на основе входных данных. На практике логистическая регрессия может быть применена к различным приложениям, таким как кредитный рейтинг, диагностика заболеваний и обнаружение мошенничества. Благодаря своей простоте, интерпретируемости и устойчивости, это делает его популярным методом.

Дерево решений. Древовидная модель обычно используется в машинном обучении для задач, связанных с регрессией и классификацией. В нем наглядно изображены варианты

принятия решений на основе обстоятельств и их результатов. Структура состоит из узлов, ветвей и листьев. Узлы представляют тесты на входных признаках, ветви отображают возможные результаты, а листья представляют окончательные решения или классификации. Алгоритм строит дерево путем рекурсивного разбиения данных на подмножества на основе значений входных признаков [4]. Как многоклассовая, так и двоичная категоризация могут выиграть от универсальности деревьев решений. Цель состоит в том, чтобы разработать модель, которая прогнозирует целевую переменную путем последовательного принятия решений на основе входных признаков. Алгоритм изучает оптимальные правила принятия решений на основе обучающих данных, чтобы свести к минимуму ошибку классификации и повысить точность прогнозирования. Примечательно, что деревья решений интерпретируемы, поскольку они предлагают ясные и интуитивно понятные средства для визуализации процесса принятия решений. Это облегчает понимание факторов, влияющих на окончательное решение или результат. Кроме того, деревья принятия решений могут обрабатывать как непрерывные, так и категориальные входные функции, а также демонстрируют устойчивость к отсутствующим значениям и выбросам. Тем не менее, деревья решений могут быть подвержены переобучению, что приводит к потенциальным проблемам с обобщением на новые, невидимые данные.

Случайный лес (Random Forest). Случайный лес – это метод ансамблевого обучения, который включает в себя деревья решений, обученные независимо на рандомизированных подмножествах обучающих данных и входных признаков. Результаты получаются путем агрегирования выходных данных деревьев, обычно путем усреднения или голосования большинства. Такой подход смягчает переобучение и может привести к повышению точности прогнозирования и надежности производительности модели [5]. Фундаментальная концепция, лежащая в основе случайных лесов, заключается в решении проблемы переобучения и повышении точности модели путем объединения нескольких деревьев решений. Уникальное подмножество входных характеристик и обучающих данных используется для обучения каждого дерева в лесу, тем самым уменьшая дисперсию модели и улучшая ее производительность обобщения. Объединяя выходные данные этих отдельных деревьев, как правило, путем усреднения или голосования большинства, подход случайного ансамбля леса смягчает переобучение и дает более надежную и точную модель прогнозирования или классификации. Алгоритм случайного леса работает путем выбора случайного подмножества обучающих данных и входных признаков в каждом узле каждого дерева. Затем он строит дерево решений на основе выбранных данных и признаков [6]. Эта процедура повторяется несколько раз, в результате чего получается коллекция или «лес» деревьев решений. На этапе прогнозирования случайный лес консолидирует результаты всех деревьев путем усреднения или получения большинства голосов, чтобы получить окончательный прогноз или классификацию. Точность и устойчивость модели повышаются благодаря этому методу ансамбля, использующему коллективную силу принятия решений несколькими деревьями.

Адаптивный бустинг (AdaBoost). Адаптивный бустинг, также известный как AdaBoost, является популярным методом ансамблевого метода, который сочетает слабые классификаторы для создания более надежного и точного общего классификатора. AdaBoost итеративно комбинирует прогнозы нескольких слабых классификаторов для создания более сильного классификатора, адаптивно корректируя веса обучающих выборок, чтобы придать большее значение неправильно классифицированным выборкам. Такой подход повышает производительность классификатора, позволяя ему обрабатывать сложные шаблоны данных и достигать более высокой точности по сравнению с отдельными слабыми классификаторами. AdaBoost особенно эффективен при работе со сложными наборами данных, содержащими множество входных признаков и классов. Обучающие данные делятся на разные подмножества для каждого слабого классификатора, и веса обучающих выборок динамически корректируются во время каждой итерации, чтобы придать более высокую важность выборкам, которые были неправильно классифицированы предыдущими слабыми

классификаторами. Этот процесс создает новую обучающую выборку, смещенную в сторону образцов, которые ранее были неправильно классифицированы, заставляя слабые классификаторы сосредоточиться на этих выборках и улучшить их производительность. Конечным результатом работы алгоритма является взвешенная сумма предсказаний всех слабых классификаторов, при этом веса определяются точностью каждого слабого классификатора. AdaBoost имеет ряд преимуществ по сравнению с другими методами ансамбля, такими как случайный лес и бэггинг. Он менее подвержен переобучению, хорошо работает с многомерными данными и может обрабатывать зашумленные и неполные данные. Одним из основных ограничений AdaBoost является его чувствительность к выбросам и шуму в обучающих данных. Если данные содержат много выбросов или зашумленных выборок, алгоритм может переподгонять эти выборки и плохо работать с новыми данными. Кроме того, AdaBoost может быть ресурсоемким, требуя обучения множества слабых классификаторов на нескольких подмножествах обучающих данных.

Экстремальный градиентный бустинг (XGBoost). Экстремальный градиентный бустинг — это высокоэффективный алгоритм машинного обучения, используемый для задач классификации. В качестве метода ансамблевого обучения он объединяет прогнозы из нескольких слабых моделей, часто в форме деревьев решений, для создания более точного и устойчивого окончательного прогноза. Известный своей превосходной эффективностью, скоростью и способностью обрабатывать большие наборы данных, XGBoost завоевал популярность в области машинного обучения. Этот процесс продолжается до тех пор, пока не будет достигнута требуемая степень точности в течение определенного количества итераций. XGBoost также включает в себя несколько методов регуляризации, таких как регуляризация L1 (Lasso) и L2 (Ridge), для предотвращения переобучения и улучшения производительности генерализации модели. Эти методы регуляризации наказывают большие веса или сложные модели, тем самым продвигая более простые и стабильные модели. После того, как ансамбль деревьев построен, прогнозы делаются путем агрегирования прогнозов всех деревьев. Как правило, XGBoost использует комбинацию взвешенного голосования или усреднения для получения окончательных прогнозируемых вероятностей классов. XGBoost стал предпочтительным вариантом для многочисленных задач классификации благодаря своему мастерству в управлении несбалансированными наборами данных, эффективной обработке отсутствующих значений и проведении анализа важности признаков. Этот анализ помогает определить наиболее важные характеристики, которые способствуют точным прогнозам. В результате XGBoost завоевал популярность в области машинного обучения благодаря своим уникальным возможностям в решении этих общих проблем. Чтобы классифицировать набор данных фишинговых веб-сайтов, каждый из этих алгоритмов может быть обучен и оценен с использованием созданных ранее наборов для обучения и тестирования. Такие метрики, как полнота, точность и прецизионность, могут использоваться для измерения производительности каждого алгоритма. На основе полученных результатов может быть выбран наиболее точный и эффективный алгоритм для развертывания в реальных сценариях.

Результаты исследования

Оценка эффективности классификации – это процесс измерения точности и эффективности модели классификации в прогнозировании правильной метки класса для заданного набора входных данных. Это важный шаг в процессе разработки модели, позволяющий получить представление о сильных и слабых сторонах модели и помочь определить области для улучшения. Наиболее распространенными показателями оценки эффективности классификации являются [7]:

1. *Accuracy*: Основываясь на соотношении правильно идентифицированных выборок ко всем выборкам набора данных, он обеспечивает целостную оценку производительности модели путем количественной оценки ее способности делать точные прогнозы.
2. *Precision*: способность модели классификации изолировать только релевантные

элементы данных. В математике точность вычисляется путем деления общего числа истинно положительных результатов на сумму истинно положительных и ложных срабатываний.

3. *Recall*: способность модели находить все соответствующие экземпляры в источнике данных. Полнота вычисляется математически как произведение числа истинно положительных результатов на сумму истинно положительных и ложноотрицательных результатов. Он указывает на долю положительных прогнозов, сделанных моделью, которые на самом деле являются истинно положительными.

4. *F1-Score*: Оценка F1 иллюстрирует компромисс между полнотой и точностью, вычисляя среднее гармоническое между каждой парой. Следовательно, он рассматривает наблюдения, которые являются как ложноположительными, так и ложноотрицательными.

Используя эти показатели, мы можем оценить производительность модели классификации и выбрать наиболее эффективную модель для нашей конкретной проблемы. На Рис. 2 показана точность моделей.

Рисунок 2. Точность моделей

Метод Случайного леса с точностью 0,97 демонстрирует самую высокую точность, за ней следуют методы дерево решений и XGBoost с оценкой точности 0,94. Модель логистической регрессии имеет точность 0,92, что ниже, чем у трех других моделей. Модель Adaboost имеет самую низкую точность – 0,91. Превосходная производительность модели Случайного леса обусловлена несколькими ключевыми факторами. Во-первых, в качестве ансамблевого метода он агрегирует несколько деревьев решений, уменьшая переобучение и повышая стабильность за счет комбинации отдельных прогнозов. Случайность признаков метода – использование подмножеств признаков для каждого дерева – обеспечивает различные точки зрения на данные, предотвращая доминирование одного влиятельного признака. Этот метод умело фиксирует сложные, нелинейные отношения в наборе данных, что делает его легко адаптируемым к различным сценариям. Random Forest также предоставляет метрики важности признаков, помогая в понимании данных и выборе признаков. Его устойчивость к переобучению, устойчивость к шуму и простота настройки гиперпараметров способствуют его надежности и точности в задачах классификации. Важно отметить, что точность сама по себе не всегда является лучшим показателем для оценки работы классификатора. Другие метрики, такие как точность и полнота памяти, также следует учитывать, чтобы убедиться, что модель хорошо работает во всех аспектах данных. Кроме того, при выборе метода нужно учитывать контекст и конкретные требования проблемы

Дискуссия

На основе предоставленной информации метод Random Forest оказался наиболее эффективным алгоритмом классификации среди оцениваемых. Его точность (97 %) превзошла производительность как метода дерева решений (94%), так и технологии XGBoost. Такая

производительность позволяет предположить, что метод Random Forest хорошо подходит для различных задач классификации и должна рассматриваться как основной выбор для таких приложений. К новым тенденциям в системах обнаружения фишинговых сайтов относится поведенческий анализ. Это включает в себя анализ поведения пользователя на веб-сайте для выявления подозрительных действий, таких как множественные попытки входа в систему, быстрые клики, необычные шаблоны навигации или аутентификация на основе домена. Поведенческий анализ собирает данные о взаимодействиях с пользователем, такие как движения мыши, нажатия клавиш, время, проведенное на страницах, шаблоны кликов и поведение навигации. Эти признаки можно извлечь и обработать для создания набора данных для обучения моделей машинного обучения. Интеграция поведенческого анализа с моделями машинного обучения обеспечивает упреждающий и ориентированный на пользователя подход к обнаружению фишинга, используя сильные стороны обеих областей. Это позволяет создать более комплексную и динамичную систему, которая адаптируется к меняющейся тактике злоумышленников, тем самым повышая общую защищенность от попыток фишинга.

Заключение

Системы обнаружения фишинговых веб-сайтов значительно развились в последние годы из-за растущей изощренности фишинговых атак. Фишинг представляет собой значительную угрозу в корпоративной среде, что приводит к значительным финансовым потерям. Несмотря на различные решения, предлагаемые и внедряемые авторитетными компаниями в области кибербезопасности, количество успешных фишинговых атак стремительно растет, что свидетельствует о том, что современные методы недостаточны для борьбы с этой проблемой. В этом документе разрабатываются различные методы обнаружения фишинга и смягчения его последствий, которые улучшают предыдущие подходы и обеспечивают более высокую точность и лучшие результаты.

References

*[1] Hannousse, S. Yahiouche (2021) Towards benchmark datasets for machine learning based website phishing detection: An experimental study, Engineering Applications of Artificial Intelligence// Engineering Applications of Artificial Intelligence.*Volume 104*, Page 104347. https://doi.org/10.1016/j.engappai.2021.104347*

[2] Gfk Healthcare, (2023) Phishing Websites Dataset. Kaggle. https://www.kaggle.com/akashkr/phishing-website-dataset.

[3] Alnemari, S.; Alshammari (2023) M. Detecting Phishing Domains Using Machine Learning// Appl. Sci. 13, 4649. https://doi.org/ 10.3390/app13084649

[4] A.A. Otahonov (2024) Obnaruzhenie i ocenka fishingovyh url-adersov s ispol'zovaniem algoritmov mashinnogo obuchenija [Detecting and assessing phishing urls using machine learning algorithms]. "Desendants of Al-Farghani" electronic scientific journal of Fergana of TATU named after Muhammad al-Khorazmi.ISSN 2181-4252.vol:1.lss:4. (In Russian)

[5] A. Mandadi, S. Boppana, V. Ravella, R. Kavitha (2022) Phishing Website Detection Using Machine Learning. 2022 IEEE 7th International conference for Convergence in Technology (I2CT). https://doi.org/10.1109/I2CT54291.2022.9824801

[6] E. Y. Boateng, D. A. Abaye (2019) A Review of the Logistic Regression Model with Emphasis on Medical Research// Journal of Data Analysis and Information Processing. Vol. 07, no. 4, pp. 190–207. https://doi.org/10.4236/jdaip.2019.74012

[7] G. Chugh, S. Kumar, N. Singh (2021) Survey on Machine Learning and Deep Learning Applications in Breast Cancer Diagnosis. Cognitive Computation, vol. 13, no. 6. pp. 1451–1470.