Preprints.org

Article

Not peer-reviewed version

An Intelligent Browser History Forensics Method for Automated Analysis of Web Activity Logs, Credentials, and User Behavioral Profiles

Leila Rzayeva *, Aliya Zhetpisbayeva *, Alisher Batkuldin *, Nursultan Nyssanov *, Alissa Ryzhova *, Faisal Saeed

Posted Date: 11 December 2025

doi: 10.20944/preprints202512.0899.v1

Keywords: cybersecurity; digital forensics; machine learning; deep learning; browser history; data analysis; web activity logs; credentials; suspicious user activity; security analytics; data privacy; security analytics

Preprints.org is a free multidisciplinary platform providing preprint service that is dedicated to making early versions of research outputs permanently available and citable. Preprints posted at Preprints.org appear in Web of Science, Crossref, Google Scholar, Scilit, Europe PMC.

Copyright: This open access article is published under a Creative Commons CC BY 4.0 license, which permit the free download, distribution, and reuse, provided that the author and preprint are cited in any reuse.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 December 2025 doi:10.20944/preprints202512.0899.v1

Disclaimer/Publisher's Note: The statements, opinions, and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions, or products referred to in the content.

Article

An Intelligent Browser History Forensics Method for Automated Analysis of Web Activity Logs, Credentials, and User Behavioral Profiles

Leila Rzayeva [1,2,*], Aliya Zhetpisbayeva [1,*], Alisher Batkuldin [1,*], Nursultan Nyssanov [1,*], Alissa Ryzhov [1,*] and Faisal Saeed [2]

[1] Research and Innovation Center "CyberTech", Astana IT University, Kazakhstan
[2] Birmingham City University, Birmingham, United Kingdom
* Correspondence: l.rzayeva@astanait.edu.kz (L.R.); a.zhetpisbayeva@astanait.edu.kz (A.Z.); a.batkuldin@astanait.edu.kz (A.B.); nursultan.nyssanov@astanait.edu.kz (N.N.); 913443@stud.unive.it (A.R.)

Abstract

In digital forensics, one of the complicated task is analyzing web browser data due to different types of devices, browsers and no updated approaches. Browsers store a large amount of information about user activity because users most often access the internet through them. However, existing approaches to analyzing this browser data still have gaps. One of the main problem developed platforms based on the old methods can not show complete information about the user's activity and have issues with precision. The article discusses the internal architecture of the browser, which is stored in the memory drives inside devices, for instance, computers or mobile devices. The research paper offers solution with developed module based on new method which integrates machine learning algorithms, such as K-NN algorithm and Naive Bayes. The main purpose of the paper it is shows new method which can automatically analyzes browser's data, detects suspicious login activity, and generates user behavior profile. The results show that the proposed new method , on which the developed platform is based, demonstrates user's profile by interests, emotional state and financial state. Also it possible to see list of top visited domain and main user's favorite website categories. It has been found that our methods shows with high accuracy 99.9% . Also the result of new method , on which the developed platform is based shows the suspicious web-sites and user's logins. Compared to Oxygen Forensics and Nirsoft which less capabilities., the proposed method provides increased accuracy , automated user profiling and detection of suspicious user's activity.

Keywords: cybersecurity; digital forensics; machine learning; deep learning; browser history; data analysis; web activity logs; credentials; suspicious user activity; security analytics; data privacy; security analytics

1. Introduction

The constant growth of Internet usage among users has led to a significant increase in the number of potential crimes investigated in the digital forensic environment. Millions of users around the world rely on web browsers as their primary means of accessing the Internet every day. Despite the existence of other highly targeted applications such as messengers, online games, and financial services, web browsers remain the most common interface for Internet access. During operation, browsers continuously create and store digital data about user actions, known as browser artifacts: browsing history, cookies, cache, bookmarks, downloaded files, and saved passwords. These data have an important role in web forensics, as they allow a person's online activity to be reconstructed and verified. Malicious actors also use the Internet for illegal activities, leaving digital footprint that can be extracted and investigated. Such information is crucial for the security, especially in digital investigations, where such data can serve as evidence.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 December 2025 doi:10.20944/preprints202512.0899.v1

The relevance of the research is related to the development of a new method for analyzing browser artifacts, using machine learning algorithms. Unlike existing solutions based on classical data processing methods, the proposed approach uses machine learning to more accurately identify suspicious activity. Since attackers often masquerade as ordinary users, analyzing their digital footprints can help identify possible malicious intent in the future.

According to recent research in digital forensics, a person's emotional and financial condition can motivate actions that may violate rules or laws [1]. For example, people who experience emotional instability may be more vulnerable and more easily influenced by cybercriminals. Similarly, users who are facing financial difficulties may be inclined to commit illegal acts to obtain money. By analyzing browser artifacts, the proposed method can infer aspects of users' psycho-emotional and financial states, thereby expanding the scope of forensic analysis beyond purely technical traces.

The contributions of this research are threefold. First, this study presents a critical analysis of existing methods for web browser forensics, identifying key limitations in their ability to detect suspicious login activities and perform holistic user profiling. Second, a novel hybrid methodology is proposed that integrates machine learning with established forensic techniques. This methodology facilitates the construction of a comprehensive user profile, including inferred interests, sentiment patterns, and indicators of financial activity, for use in digital forensic investigations. Third, the proposed methodology is validated by its implementation as a prototype software tool. This tool, developed in Python, automates the extraction and processing of browser artifacts to perform the behavioral profiling analysis.

In many cases, the tools available are tailored to specific tasks and offer only minimal support for various data formats. In contrast, the proposed method differs from existing solutions in several key dimensions, including support for extended data formats, improved accuracy of artifact detection, more detailed behavioral analysis, and enhanced detection of suspicious URLs. The most valuable aspect of the proposed approach is the introduction of a new method to analyze user' socio-economic status, personal interests, and psychological behavior based on browser activity. Furthermore, the framework supports multiple browser types, handles heterogeneous data sources, and provides automated profiling capabilities. Although most existing methods focus solely on data extraction, this research combines the development of a novel analytical methodology with its practical implementation in a Python-based software module. As a result, the proposed solution is well suited for application in digital forensics, law enforcement agencies, and the broader field of cybersecurity.

The novelty of this scientific research lies in the implementation of machine learning methods for automated analysis of web browser data, which accurately detect suspicious user activity, recognize potentially malicious logins on websites, and automatically create user profiles based on psychological and emotional state. By presenting both a methodological framework and a practical software solution, this research significantly expands the potential applications of browser data analysis in web forensics.

The structure of the paper is as follows: Section 2 examines the existing literature and related work in this domain. In Section 3, the proposed methodology is detailed, including the processes for data extraction, artifact analysis, and the application of machine learning. In Section 4, the experimental results generated by the module are presented. Section 5 provides a discussion of these findings and includes a comparative analysis with existing tools. Finally, Section 6 offers the general conclusions derived from this research.

2. Literature Review

2.1. Overview of Browser

Web browsers are one of the most common applications for accessing the Internet. They allow users to interact with web resources, which are primarily created using HTML, CSS, JavaScript, diverse frameworks and different markup technologies. Nowadays, on desktop and mobile devices, browsers can process and display text, hyperlinks, interactive components, and multimedia material. The idea behind how they work is the same regardless of how they are implemented: when a user uses a URL

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 December 2025 doi:10.20944/preprints202512.0899.v1

to request a web page, the browser downloads the relevant HTML document, runs the required scripts, and then displays the content [1].

From an architectural point of view, a browser consists of several basic modules. The graphical user interface (GUI) provides visual interaction with web pages. The server part of the interface controls the internal mechanisms of the browser and interacts with the components of the operating system. The browser engine is responsible for coordinating input/output operations and working with the rendering engine. The rendering engine performs network operations, JavaScript processing, XML parsing, and interaction with plugins [3]. All these components together ensure the correct retrieval, processing, and display of web content, as well as data exchange between the client-side and the server.

Two elements of browser architecture are particularly important for digital forensics: data persistence and the network subsystem. Data persistence covers various browser artifacts: browsing history, cookies, cache, and saved accounts. The network subsystem is related to information that is transmitted and recorded during user interaction with web servers. Both aspects play an essential role in forensic investigations, as they preserve digital data that can serve as evidence when restoring user activity and detecting suspicious or malicious behavior.

In this research work take into account two important aspects (data persistence and network functions), shown in Figure 1. These aspects are crucial in web forensics.

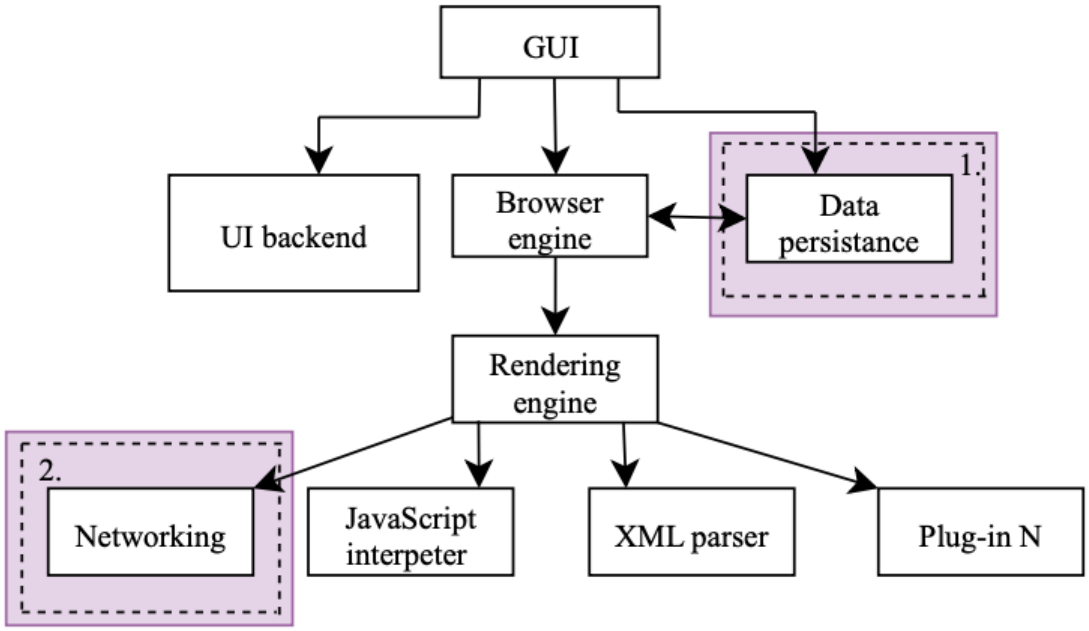

Figure 1. Browser architecture.

1. Data persistance means that data can be saved locally even after restarting or closing tabs. This data includes browsing history logs, cached files, cookies, and local storage/session storage. Thus, this data helps to display the user's activity in the web browser. Timestamps such as the last visit, the previous update cookies, and the cache are also used to reconstruct the timeline of user events.

2. Network interactions include all network interactions that form connections between the client and the server. Without network connections, users will not be able to connect to the Internet. Therefore, without network connections, the web browser will stop working. This plays a crucial role in detecting interactions with suspicious servers in web forensics [4].

An analysis tool is needed in web forensics. Therefore, this study considers data security and network interactions to analyze the timeline, user activity, search history, and URLs. This provides information about whether the user has performed a search or performed suspicious actions on the Internet. The combination of data retention and network functions is the basis for web forensics.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 December 2025 doi:10.20944/preprints202512.0899.v1

These two key aspects are essential for web forensics. Because there is key user data in these aspects. Consequently, a special tool is needed to automatically analyze the data in these two aspects, obtaining a complete profile of the user's actions.

2.2. Analyzing the Browser History Format

Researches already existing in the field of web forensics focuses on analyzing data extracted from various web browsers [5,6]. This enables the collection of sufficient evidence characterizing user behavior on the Internet. This study focuses on data extracted from the most popular browsers. According to global statistics, the most popular web browsers in Central Asia are Google Chrome, Safari, Internet Explorer, and Mozilla, as shown in Figure 1. The prevalence of browser usage in each country may vary depending on the purpose of use and convenience [7].

Each web browser is running and can be installed depending on the operating system version, as shown in Table 1. Getting information from the browser is affected if a later version of the operating system is installed on the computer or mobile phone. This worsens the ability to extract information because the operating system is outdated and lacks modern extraction utilities. The browser constantly interacts with the operating system, both on the computer and on mobile devices [10].

A fundamental difference exists between desktop and mobile operating systems, which directly impacts forensic data acquisition. On mobile devices, extracting web browser artifacts is contingent upon obtaining root-level access. Without such privileges, data extraction is severely hindered or rendered impossible, with the specific limitations varying by the mobile OS [11]. The default storage format is the location in the browser where the user's browsing history is stored. Most modern browsers (as shown in Table 1) store the history in an SQLite database file. It can be accessed and analyzed using database tools. Some older or system-limited browsers (such as Internet Explorer or mobile versions of Chrome/Safari) use secure storage formats (such as ESE or system-locked files). However, access to them is difficult or impossible without root rights or special tools. This means that browser extensions or add-ons can provide access to the browsing history and export it to standard formats such as CSV or JSON. These materials are helpful and convenient for analysis in a forensic browser.

Table 1. Supported OS and browser history storage format.

Browser	Operating System	Default storage format	Export to CSV/JSON
Google Chrome	Windows / macOS	SQLite	Yes
	Android	No access (root required)	Access only with root
	iOS	No access (root required)	No
Mozilla Firefox	Windows / macOS	SQLite	Yes
Internet Explorer	Windows / macOS	ESE	Need parser
Safari	macOS	SQLite	Yes
	iOS	No access (root required)	No

In Table 1:

- "Yes" means that installation of an extension to export browser history is possible.
- "No" means that the browser either restricts access or does not support this feature (especially on iOS or Android devices without root access).
- "Need parser" means that additional tools or scripts are required.
- "Access only with root" means that data can only be accessed if the device has superuser privileges.

In this research work, considering these formats because various operating systems, programming languages, databases, and data analysis tools widely support CSV and JSON formats. Therefore, CVS and JSON formats are convenient for exporting, distributing, and importing browser history to other systems. These formats are also helpful for analysis in browser-based forensics to identify potentially suspicious user activity. These formats work with programming languages such as Python and JavaScript, and allow researchers to analyze browser history, visualize patterns, or create automated

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 December 2025 doi:10.20944/preprints202512.0899.v1

systems. In particular, JSON is a standard format for APIs and modern web applications. Integration with cloud services and data processing systems is made possible by exporting history to JSON [12].

2.3. Web Forensics Methods

Finding evidence related to Internet searches is often crucial in forensic investigations around the world. Almost any online activity can leave traces on digital devices. Therefore, examining this data can provide investigators with valuable information. Thus, extracting data such as download lists and browser history can make it easier to analyze a suspect's actions [13]. In web forensics, there are various methods for analyzing browser artifacts. When examining the browser history, one can find data such as visited websites, online activity, timestamps, frequency of visits, and registration for suspicious websites. Bookmark analysis makes it possible to determine the main actions of users on the network by identifying frequently visited or saved websites. Password analysis will be helpful in restoring saved credentials and identifying potential vulnerabilities. Malware analysis, meanwhile, focuses on detecting malicious websites, scripts, or uploaded files from intruders. The Table 2 shows these types of analysis and the main browser artifacts [14].

The purpose of these methods (Table 2) is to examine various aspects of browser activity to determine the user's behavioral potential. Digital forensics specialists use various methods to analyze and extract relevant data. These methods focus on multiple aspects of browser activity to identify potential clues, monitor user behavior, and identify security threats [15].

Table 2. Types of analysis methods and recoverable contents.

Type of analysis method	Recoverable browser artifacts
Browser History Analysis	Visited websites, Timestamps, Page titles, Count of visits, Search queries
Bookmarks Analysis	Saved website URLs, Bookmark titles, Creation timestamps
Password Analysis	Stored login credentials, Websites, Login, Timestamps of password creation/update, Encryption info (if recoverable)
Malware Analysis	Malicious websites, Known malware signatures

Examining existing web forensics methods showed that they focus on extracting browser artifacts and performing fundamental analysis [16]. However, they rarely utilize machine learning methods for automated user data mining or user profiling. Thus, a gap remains in current research related to the integration of algorithms into existing web forensics methods.

In reality, the existing methods described are implemented in various software programs. For example, Nirsoft provides a set of utilities for extracting saved logins and passwords from browsers, as well as basic browsing history information. NetAnalysis supports the extraction of artifacts from various browsers but does not perform automated analysis. Modern programs such as Oxygen Forensic Detective offer tools for analyzing browser activity, allowing user to recover browsing history, cookies, search queries, and saved passwords on both desktop systems and mobile devices. However, these programs focus on collecting artifacts and cannot analyze user behavior based on browser history [17].

Therefore, existing solutions(Nirsoft, NetAnalysis and Oxygen Forensic) typically implement only a limited set of web forensic methods: some focus on browsing history, others on passwords, and still others on identifying malicious websites. Although existing tools have the potential and value for practical use, they have a primary problem. This poses a serious problem, as virtually none of the existing tools use machine learning methods that could improve data detection accuracy, automate user profiling, and provide more comprehensive analysis. The mentioned limitation highlights the need for a new approach that combines traditional web forensics methods with machine learning algorithms. Solving this research problem requires the development of a methodology that integrates classical forensic approaches with modern ML technologies [18].

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 December 2025 doi:10.20944/preprints202512.0899.v1

3. Method

3.1. Data Exstraction and Recovery

Research on storage devices such as hard disk drives (HDDs) and solid-state drives (SSDs) shows that they store various information, such as web browser data. If data is deleted or the operating system is reinstalled, file fragments, cache, cookies, and user credentials are retained. On HDDs, these records are often recovered from unallocated or unused space due to the magnetic nature of data storage, which allows partially overwritten sectors to be preserved. In SSDs, the situation is complicated by the TRIM command and wear-leveling mechanisms that actively erase unused memory blocks; however, forensic methods can still extract browser artifacts by analyzing controller behavior, NAND remnants, or system snapshots. These recovered artifacts can include visited URLs, timestamps, downloaded files, search queries, and session IDs, which can be crucial for reconstructing user activity during cybercrime investigations or behavioral research. Thus, both hard drives and solid state drives serve as important sources of evidence in digital forensics, although data recovery from solid state drives requires special techniques due to their dynamic data management [8].

It's important to note that after extracting and reconstructing browser artifacts from hard drives or SSDs, the next step is analyzing this large volume of data in various formats. Therefore, in order to analyze this raw, recovered data into forensic information about user online behavior and detect suspicious activity, a new method is necessary. This requires developing an automated method based on this method that automatically creates a user profile based on their browsing history and queries, and also automatically analyzes suspicious online activity.

Different types of computers and mobile phones contain their own storage drives. Therefore, it is important to consider recovering browser data not only from computers, but also from mobile devices. Mobile devices typically use storage drives such as eMMC and UFS. Those storage drives operates with the same principles as SSD in computers [3].

3.2. New Forensic Method for Automated User Profiling

The method proposed in this research paper, shown in Figure 2. This research paper proposes a new forensics method that has not been used in other developed platforms. This new method involves the use of machine learning algorithms. To identify suspicious activity and create a forensic profile of the user. This focus highlights the scientific and practical novelty of the research, as existing approaches rarely integrate multi-format browser data into a unified analytical framework. Browser artifacts (such as browsing history, cookies, saved credentials, and bookmarks) are valuable data because they provide evidence in investigations. They store information about visited websites, authentication information, and timestamps of user activity (Figure 2). This logic-diagram of presented method combines three main approaches, which are:

First, the rules-based detection module identifies suspicious online activity, such as URLs containing IP addresses instead of domain names, excessively long or confusing strings, misuse of subdomains, or the presence of keywords commonly associated with fraudulent activity (e.g., gambling, loans, or explosive substances) [20].

Secondly, machine learning is used using algorithms like K-Nearest Neighbors (k-NN) and Naive Bayes together to classify all visited domains and websites, as well as to identify malicious websites based on characteristics such as URL structure, temporal activity, and frequency of visits. Together, these modules support the creation of a user profile for digital forensics that reflects behavioral activity, personal interests, financial activity, and psycho-emotional indicators based on machine learning algorithms [21].

To achieve this research goal, a new multi-level methodology was implemented. It included empirical data collection from real browsers, preprocessing of multi-format artifacts, and deployment of a modular forensic analysis module [22].

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 December 2025 doi:10.20944/preprints202512.0899.v1

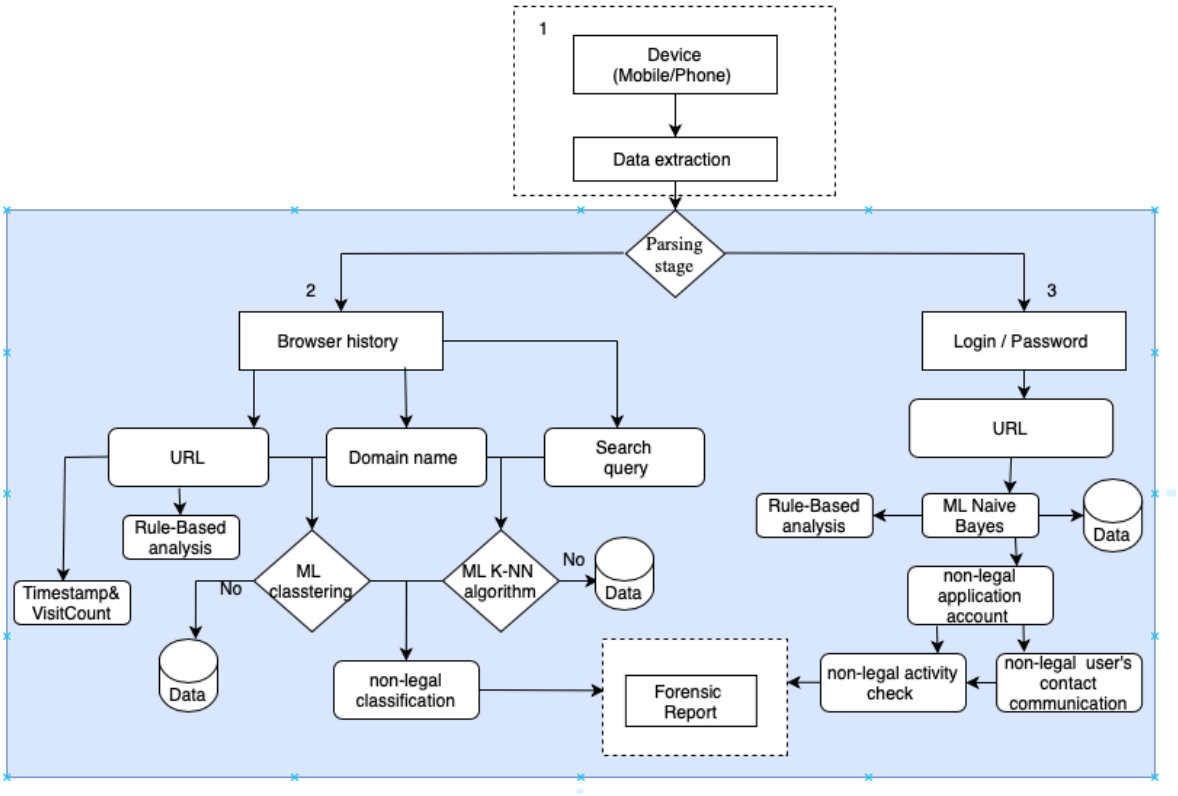

Figure 2. Logic-diagram of proposed method.

As mentioned earlier, that in new method integrated machine algorithms such as K-Nearest Neighbors (k-NN) and Naive Bayes. Those two machine learning algorithms was used because, firstly, the K-Nearest Neighbors algorithm used not only for regression task but it also used for classification. The principle of this algorithm is to find "K" closest points to a given input and makes a predictions on the majority classification. One of the main advantage of this algorithm that it uses for classification which is important aspect to create as user's profile. Through K-Nearest Neighbors (k-NN) algorithm becomes possible to observe patterns of user activity in browser shows on the Figure 3. Another justification for using this algorithm in this research work is that K-Nearest Neighbors (k-NN) is reliable machine algorithm. Because it classifies new data by comparing it to the existing training dataset therefore identifying the most similar patterns shows on the Figure 3. Another machine learning algorithm that used is called Naive Bayes which is predicts the category of a data point using probability. The main principle that this algorithm uses Bayes' Theorem to classify data which are based on the probabilities of different classes. This classes given from the features of the data picture.

To process big data, which includes arrays and logs, the Naive Bayes algorithm was chosen. This algorithm requires less energy resources and works quickly in time, therefore this Naive Bayes machine learning algorithm was also integrated to develop the new method.

The backend was implemented in Python using the Flask framework, with dedicated modules for history parsing, input handling, and user profiling.

The interface, designed using HTML, JavaScript, and CSS, provides interactive visualization (Chart.js). It also supports exporting results to PNG format. This integration provides a visually understandable format of evidence for judicial investigations [30].

In contrast, the method developed as part of this study automatically analyzes datasets that are artifacts of a web browser taken from devices such as a phone or computer. This is done in order to create a forensic user profile. This profile provides information about the user's interests, financial activity, and mental and emotional state. Therefore, the new method in this scientific work expands

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 December 2025 doi:10.20944/preprints202512.0899.v1

the possibilities of traditional web-based forensic analysis. This one demonstrates both scientific and practical novelty in web forensics [29].

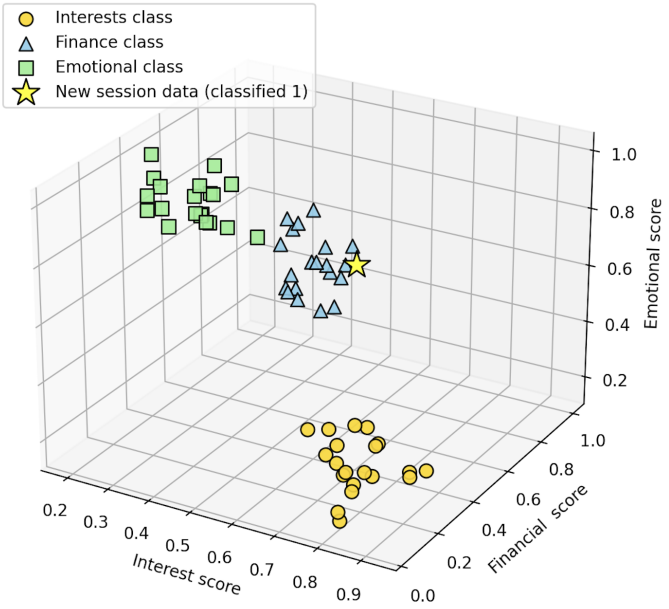

Figure 3. Illustration of the K-NN algorithm.

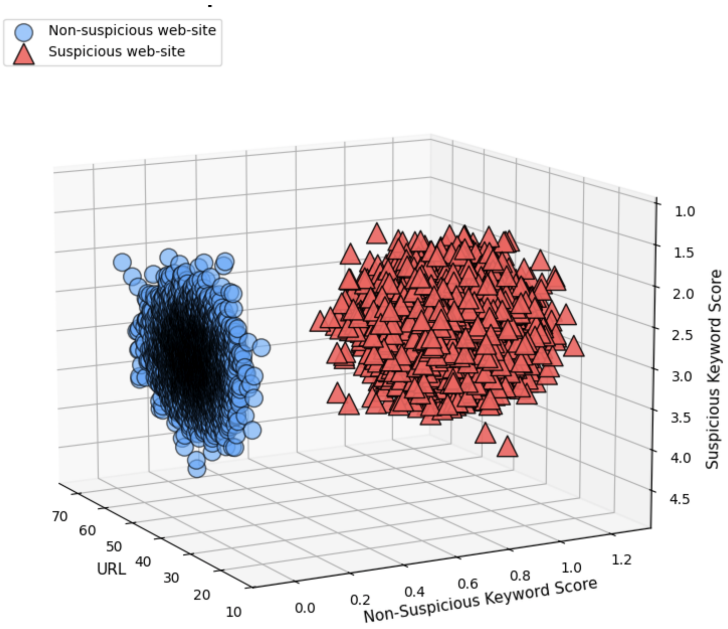

Figure 4. Illustration of Naive Bayes algorithm.

While existing tools in web forensics, such as Nirsoft, Web Historian, and NetAnalysis, are mainly focused on extracting web browser datasets without automated analysis, without using machine learning, and without creating a user profile. The proposed method pays special attention to the integrated processing of file formats (JSON, CSV) in combination with analysis using machine learning algorithms. Existing tools are often limited to specific browsers or file formats and are not capable of creating an automated user profile [23]. The research paper also uses a newly developed method for automatically analyzing user logins and passwords. Consequently, the platform automatically analyzes the user's login and determines whether the registration is suspicious shows on the Figure 5. The figure shows how the Naive Bayes algorithm is applied, which makes a determination based on

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 December 2025 doi:10.20944/preprints202512.0899.v1

probability theory where specified features. Key indicators include the length of the domain, the URL, and the presence of suspicious word matches within the domain itself.

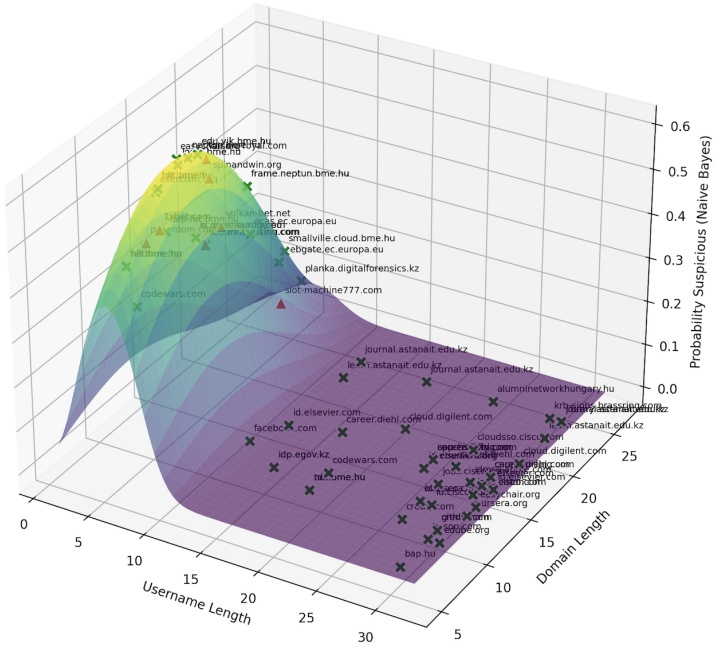

Figure 5. Illustration of Naive Bayes algorithm for logins/passwords .

3.3. Device and Data Extraction

The process begins with identifying the digital device from which browser data can be extracted. In most cases, this is either a personal computer or a mobile phone. These two type of digital devices were chosen because they are the most common means of internet access for most users worldwide. Both of those types consists storage drives that have user's history browser [27]. Web browsers such as Google Chrome, Mozilla Firefox, Safari, and others are widely used on mobile and computer devices. A digital device can be viewed as a repository of data generated through the user's daily activities. These browser artifacts form a structured set, defined as:

$$A = \{H, C, L, B\} \tag{1}$$

where H = browsing history, C = cookies, L = login/password credentials, and B = bookmarks. Artifacts are extracted in formats $F = \{\text{SQLite}, \text{JSON}, \text{CSV}\}$:

$$Extract(D) \rightarrow A \tag{2}$$

This operation illustrates that the extraction function applied to digital device D defers a complete set of artifacts A. This representation guarantees the process's reproducibility and its technology independence, providing a formal abstraction that is independent of the specific browser implementation. The availability of various formats ($F = \text{SQLite}, \text{JSON}, \text{CSV}$) enables detailed analysis for digital forensics.

3.4. Parsing Stage

After the extraction stage, the collected browser artifacts undergo a parsing procedure. Parsing is an intermediate stage in which raw data is converted into semantically meaningful information that can then be passed on to forensic analysis. The initial set of artifacts usually includes browser history, cookies, saved credentials (logins and passwords), and bookmarks. Each of these elements plays an important role in conducting a detailed investigation.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 December 2025 doi:10.20944/preprints202512.0899.v1

A thorough analysis of both the browsing history and user authentication data is required for a complete examination. During processing, the extracted artifacts are structured and classified, after which they are divided into two main categories:

$$Parse(A) \rightarrow \{H, L\} \tag{3}$$

This enables specialized processing: H is analyzed for user's browser history, while L is analyzed for user's login/passwords.

3.5. Browser History Analysis

Browser history (H) represents one of the most informative artifacts for digital forensics, as it reflects the sequence of user interactions with online resources. In the proposed approach, the history dataset is decomposed into three main components: URL, search queries, and domain names. Such decomposition ensures that each element can be subjected to a specialized form of analysis.

URL. Complete web addresses are examined using rule-based analysis, where structural anomalies (e.g., excessive length, use of special symbols, or embedded IP addresses) are identified. In addition, metadata such as visit times and frequency are extracted from the artifacts. This information helps to reconstruct the user's behavioral habits and perform a detailed temporal analysis of their Internet activity.

Search queries. The queries that users enter into search engines reflect their interests, cognitive state, emotional background, and possible intentions. Analyzing these queries allows us to identify sensitive topics (such as finance and health issues) and classify actions that may pose a threat or indicate illegal activity.

Domain names. The domain part of the URL is analyzed separately, as it is the key identifier of the web resource. Domains are grouped using machine learning methods to divide them into categories of legitimate and potentially suspicious services. To improve the accuracy of illegal domain classification, the K-Nearest Neighbor (K-NN) algorithm is used, which compares the domains under investigation with known patterns of malicious activity.

URL Rule-Based Detection and Time detection. Each record in the browser history can be represented as a triplet consisting of the visited URL, the time of access (timestamp), and the number of visits (visit count). This structure allows data to be normalized and automatically analyzed.

To identify suspicious characteristics of URLs, a rule-based detection method is applied. A set of heuristic rules is being developed to identify potentially malicious or disguised addresses:

Rule 1 (Hyphen rule). Legitimate domains rarely contain hyphens. Their presence often indicates impersonation, where attackers create visually similar domains (e.g., sbi-online.com instead of sbi.com).

Rule 2 (IP-based rule). URLs that directly use IP addresses (e.g., http://205.53.73.105/fake.html) are flagged as suspicious, since this technique is frequently used to bypass filtering systems.

Rule 3 (Length rule). Abnormally long URLs may indicate obfuscation, inclusion of hidden parameters, or malicious scripts.

Rule 4 (Subdomain rule). An excessive number of nested subdomains suggests redirection chains or hosting of phishing pages. Each history record $h \in H$ is represented as:

$$h = (u, t, v), \quad u \in U,\ t = \text{timestamp},\ v = \text{visit count}. \tag{4}$$

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 December 2025 doi:10.20944/preprints202512.0899.v1

Suspicious features are flagged with rules:

$$R_1(u) = \begin{cases} 1, & \text{if "-" in Domain}(u) \\ 0, & \text{otherwise} \end{cases} \tag{5}$$

$$R_2(u) = \begin{cases} 1, & \text{if Domain}(u) \in IPv4/IPv6 \\ 0, & \text{otherwise} \end{cases} \tag{6}$$

$$R_3(u) = \begin{cases} 1, & \text{if } |u| > \theta \\ 0, & \text{otherwise} \end{cases} \tag{7}$$

$$R_4(u) = \begin{cases} 1, & \text{if SubdomainCount}(u) > \delta \\ 0, & \text{otherwise} \end{cases} \tag{8}$$

The overall suspiciousness score is:

$$S_{RB}(u) = \sum_{i=1}^{4} R_i(u) \tag{9}$$

A URL is flagged if $S_{RB}(u) \geq 1$. Each rule produces a binary result: if the condition is met, the URL is considered suspicious under that criterion. The general "suspiciousness score" of a URL is calculated as the sum of triggered rules. If at least one rule is activated, the URL is flagged as potentially unsafe.

In addition to structural analysis, the proposed solution takes into account temporal characteristics such as timestamps and number of visits. Temporal irregularities, such as frequent access to the same resource in a short period or activity at unusual hours, may indicate compromised accounts or automated scripts. Likewise, a high level of visits to resources associated with flagged domains increases the likelihood of malicious activity. Domain Name Analysis. The domain name component of a URL is a key feature for forensic browser analysis, as it uniquely identifies the resource accessed by the user. For systematic processing, the set of all domains extracted from history is denoted as:

$$D = \{d_1, d_2, \ldots, d_n\}, \quad d_i \in \text{Domain}(H). \tag{10}$$

To detect suspicious domains, a machine learning clustering technique is applied. Each domain is represented by a feature vector $\phi(d)$, which may include attributes such as:

- length of the domain name,
- number of subdomains,
- presence of digits or special characters,
- occurrence of suspicious keywords.

Formally, the feature space is defined as: Each domain is represented as a feature vector:

$$\phi(d) = (f_1, f_2, \ldots, f_m) \in \mathbb{R}^m, \tag{11}$$

where features include length, number of subdomains, presence of digits, and occurrence of suspicious keywords.

Domains are clustered into k groups using unsupervised machine learning:

$$Cluster(D) \to \{C_1, C_2, \ldots, C_k\}. \tag{12}$$

where each cluster corresponds to a group of domains with similar structural and semantic properties. Clusters containing domains with high suspiciousness scores are flagged for further investigation.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 December 2025 doi:10.20944/preprints202512.0899.v1

For refinement, the K-Nearest Neighbor (K-NN) algorithm is applied to classify new domains against previously labeled data. Given a new domain d_{new} with features $\phi(d_{new})$, the classification is defined as: For classification of new domains, the K-Nearest Neighbor (K-NN) algorithm is applied:

$$Class(d_{new}) = \text{majority}\big(y(d_i) \mid d_i \in N_k(d_{new})\big), \tag{13}$$

where $N_k(d_{new})$ denotes the set of k nearest neighbors in the feature space, and $y(d_i)$ is the label (legitimate or suspicious). URLs are converted into feature vectors $f(u)$. Classification is based on k nearest neighbors:

$$Label(u) = \arg\max_{c \in C} \sum_{i=1}^{k} w_i \cdot \mathbb{1}(c_i = c) \tag{14}$$

with distance-based weights:

$$w_i = \frac{1}{1 + dist(f(u), f(u_i))} \tag{15}$$

This two-step process, clustering followed by K-NN classification, ensures both unsupervised discovery of domain groups and supervised precision in identifying malicious resources. Search Query Analysis. For domain $d = Domain(u)$ and query q:

Fuzzy matching:

$$FM(d) = \min_{d^* \in D^*} \text{Levenshtein}(d, d^*) \tag{16}$$

If $FM(d) < \epsilon$, then d is suspicious.

Anomaly detection in queries (TF–IDF + k-NN):

$$Label(q) = \arg\max_{c \in C} \sum_{i=1}^{k} \frac{1}{dist(tfidf(q), tfidf(q_i))} \tag{17}$$

Temporal Analysis. User activity in hour h is modeled as:

$$Activity(h) = \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}(Hour(t_i) = h) \tag{18}$$

Visit frequency per domain:

$$Freq(d) = \frac{1}{T} \sum_{i=1}^{N} \mathbb{1}(Domain(u_i) = d) \tag{19}$$

Forensic User Profile for the final stage of analysis in the proposed solution is the construction of a forensic user profile, which integrates information derived from multiple categories of browser artifacts. The user profile serves as a consolidated representation of behavioral, financial, and psycho-emotional characteristics inferred from browsing history and associated data. Formally, the profile can be expressed as:

$$P = f(H) = (I, F, E) \tag{20}$$

where I = interests, F = financial indicators, and E = psycho-emotional indicators. Interests (I). This component reflects the user's thematic preferences, identified based on frequently visited websites, saved bookmarks, and predominant search query clusters. Examples include interests in technology, social media, online shopping, or gambling. Such data provides valuable contextual information for both security monitoring and digital investigations.

Financial indicators (F). This component reflects information about the user's financial behavior and possible vulnerability. Indicators are extracted from visits to online banking, credit or loan services, and online stores. Repeated queries such as "quick loan," "pawnshop," or "urgent money"

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 December 2025 doi:10.20944/preprints202512.0899.v1

may indicate financial difficulties, which in turn may be associated with risky online behavior or an increased propensity for fraud.

Psycho-emotional indicators (E). This component depicts the user's emotional and cognitive state based on their search queries and online behavior patterns. For example, frequent searches related to health, emotional stress, or high-risk activities (such as gambling or illegal activities) may indicate psychological stress or behavioral abnormalities. Combined with a temporal analysis of browser activity, this allows for the creation of a psycho-emotional profile of the user.

By integrating these three dimensions, the forensic profile P provides a comprehensive perspective on the user's digital footprint. This structured model enables investigators to move beyond raw artifact analysis toward a higher level of forensic interpretation, which can be applied in law enforcement, cybersecurity risk assessment, and user awareness studies.

3.6. Analysis Logins for Identify Suspicious Domens

The second branch of the proposed workflow focuses on analyzing login credentials and passwords (L) extracted from the browser. Unlike browsing history, which reflects general user activity patterns, authentication artifacts are directly linked to login events and specific online accounts. This data is always considered in the context of the corresponding URL of the resource that was accessed, which allows the legitimacy of the account to be assessed in the context of its use. Formally, each login entry can be represented as:

$$l = (u, c), \quad u \in U,\ c = (\text{login}, \text{password}), \tag{21}$$

where u denotes the associated URL and c corresponds to the credential pair consisting of a login and a password. Rule-Based Analysis.

Structural irregularities in the login record are detected using heuristic rules. For instance, accounts linked to URLs previously flagged as suspicious, the reuse of identical passwords across multiple domains, or empty/weak password fields are treated as risk indicators. The rule-based function can be expressed as:

$$R(u, c) = \begin{cases} 1, & \text{if } u \in U_{suspicious} \text{ or } Weak(c) = 1 \\ 0, & \text{otherwise} \end{cases} \tag{22}$$

where $U_{suspicious}$ is the set of domains flagged by the URL detection stage, and Weak(c) checks whether the password violates security policies (e.g., too short, dictionary-based).

Naïve Bayes Classification.

Beyond rule-based checks, login records are further processed using a probabilistic machine learning classifier. The Naïve Bayes algorithm evaluates the likelihood that a credential corresponds to an illegitimate account or non-legal activity. For each login $l = (u, c)$, the posterior probability is calculated as:

$$P(y \mid x) = \frac{P(y) \prod_{i=1}^{n} P(x_i \mid y)}{P(x)}, \tag{23}$$

where $x = (x_1, x_2, \ldots, x_n)$ is the feature vector derived from u and c (e.g., password length, entropy, URL category), and $y \in \{\text{legitimate}, \text{illegitimate}\}$ is the classification label.

A login record is classified as non-legal if:

$$\hat{y} = \arg\max_{y} P(y \mid x) = \text{illegitimate}. \tag{24}$$

This comprehensive approach allows the system not only to identify suspicious use of accounts (weak or reused passwords), but also to detect cases of illegal activity, such as attempts to authenticate on fraudulent domains, hidden services, or blacklisted modules.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 December 2025 doi:10.20944/preprints202512.0899.v1

By integrating rule-based heuristic method with Naïve Bayes classification, the login/password branch achieves both interpretability (clear forensic rules) and adaptability (probabilistic learning from new evidence), thus strengthening the overall accuracy of the approach. This methodology ensures that the system detects unsafe account usage (weak or reused credentials) and also identifies cases of non-legal active behavior, such as attempts to authenticate into fraudulent domains, shadow services, or blacklisted module.

4. Results

The module demonstrates the ability to systematically extract, parse, and analyze: browser artifacts, including history, cookies, login credentials, and bookmarks, across multiple formats such as SQLite, JSON, and CSV. The results generated by the module, (Figure 6, Figure 7, Figure 8, Figure 9, Figure 10, Figure 11), illustrate the structured visualization of browser data, the identification of suspicious URLs, and the application of both rule-based and machine learning techniques for deeper analysis. These outcomes confirm that the method extends beyond simple data extraction by providing comprehensive forensic interpretation, thereby enabling the construction of a detailed user profile for investigative purposes.

Figure 6 shows the forensic analysis results of the method proposed in this research paper. A user profile is automatically generated based on extracted browser data (browser artifacts) for subsequent analysis. Unlike existing tools such as Nirsoft and Oxygen, the developed method uses a combination of rule-based analysis and machine learning algorithms to more accurately determine the user's behavior and psychological state.

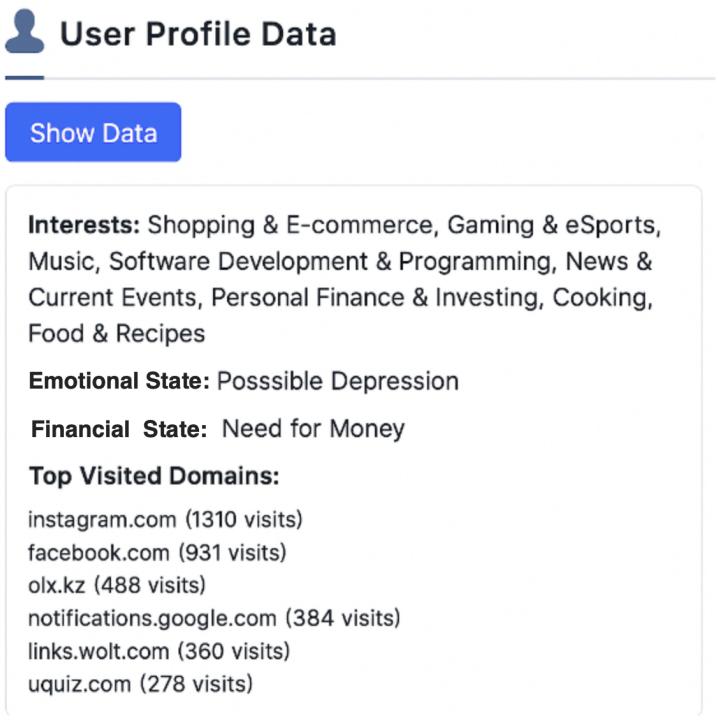

Figure 6. User profile data.

As shown in the Figure 6, the method developed in this research paper accurately identified the user's dominant interests, including shopping and e-commerce, gaming and e-sports, music, programming, financial activities, and cooking. In addition to interests, the solution was able to determine the user's financial status (possible depression) and identify markers of emotional and psychological state (need for money). These conclusions are based on browsing frequency patterns, domain analysis, and the identification of keywords associated with the user's financial or emotional characteristics.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 December 2025 doi:10.20944/preprints202512.0899.v1

The profile also displays the most frequently visited domains and their corresponding visit counts, such as instagram.com (1,310 visits), facebook.com (931 visits), and olx.kz (488 visits). Based on domain frequency, a user's interests and behavioral patterns can be determined. The structured presentation of both categorical data (e.g., interests, status indicators) and numerical indicators (frequency of visits) provides a comprehensive understanding of the user's online activity.

The results show that the solution proposed in the study goes beyond the usual analysis of browser history and also covers forensic examination of visited pages. The profiles created, which include the user's psycho-emotional characteristics, can serve as valuable evidence in digital forensics, allowing investigators to assess the socio-economic conditions and potential illegal activity relevant to the investigation.

Figure 7 illustrates the output of the developed forensic analysis module responsible for identifying the user's potential interests and hobbies based on the extracted artifacts from web browsers. The system automatically classifies the detected categories using a rule-based analysis and machine learning algorithms (K-Nearest Neighbors and Naïve Bayes).

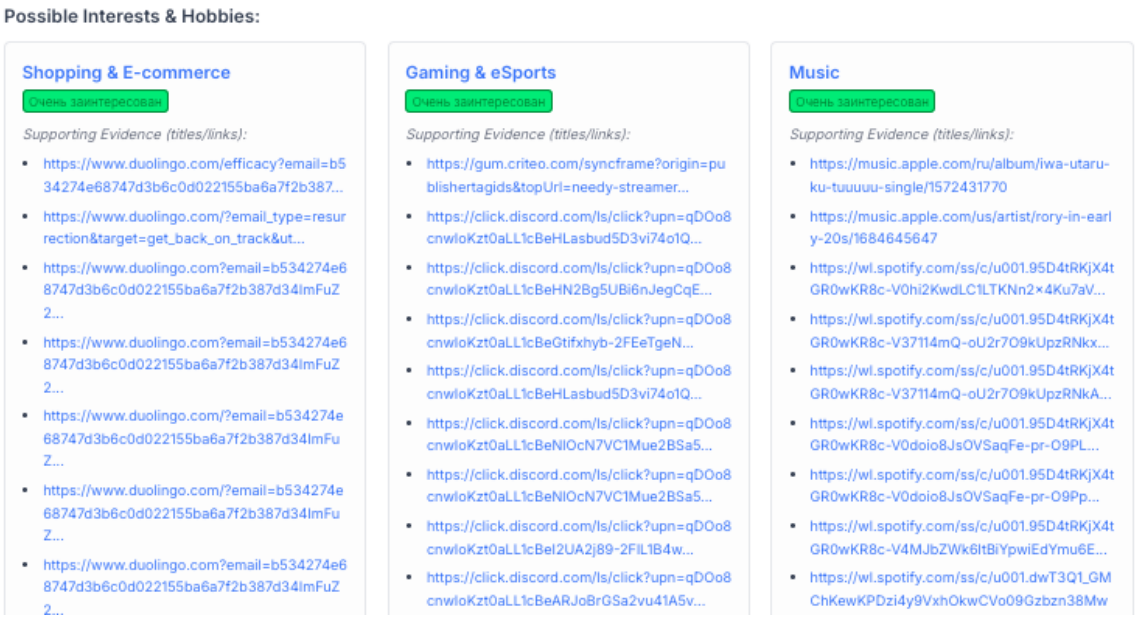

Figure 7. Possible user's interests and Hobbies.

In this case, the module generated three primary categories of user interests: Shopping, E-commerce, Gaming, eSports, and Music. Each category is accompanied by a list of supporting evidence in the form of URL links extracted from the browser's history, cookies, and related metadata. The classification confidence level ("Highly Interested") is automatically assigned based on the frequency of visits, dwell time, and URL contextual similarity across multiple sources.

The presented visualization demonstrates the system's ability to correlate heterogeneous web artifacts and form a structured profile of the user's behavioral tendencies. Such an approach enhances the forensic interpretation of digital traces and allows investigators to infer psychological and lifestyle indicators relevant to digital identity profiling.

Figure 8 presents the output of the Suspicious Website Detection Module integrated into the developed forensic module. The interface demonstrates a dynamically generated list of URLs classified as potentially malicious or high-risk domains. The extracted addresses from the user's browsing history are automatically flagged based on rule-based heuristics and machine learning indicators.

The system detects suspicious activity by analyzing URL structures, domain names, embedded parameters, and behavioral patterns such as frequent redirects, gambling-related keywords, and domain obfuscation (e.g., random alphanumeric strings or subdomain chains). Each detected entry is

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 December 2025 doi:10.20944/preprints202512.0899.v1

displayed in a structured format with hyperlinks, allowing investigators to examine the sources in detail or export the findings as a PDF report.

This result confirms that the proposed system is capable of automating the detection of potentially dangerous or fraudulent web resources among browser artifacts, which helps forensic experts detect traces of risky online behavior, phishing attacks, and interaction with illegal content.

Figure 8 shows the interface of the developed browser forensic module, illustrating the module responsible for displaying the full browsing and search history of the analyzed user.

Figure 8. Suspicious visited websites.

The presented table automatically lists all extracted URLs and associated metadata, including the domain name, full link, visit timestamp, and visit count. The data were parsed from browser artifacts such as the History SQLite database and subsequently formatted into a structured log view for visualization.

This component enables investigators to perform temporal analysis of the user's online activity: identifying specific access times, repetitive domain visits, and behavior patterns across different sessions. In this example, the Figure 9 shows multiple visits to educational websites (e.g., znanija.com), with detailed timestamps recorded on 6–7 October 2025. Such chronological mapping supports the reconstruction of user activity timelines and enhances the interpretability of digital evidence within forensic investigations.

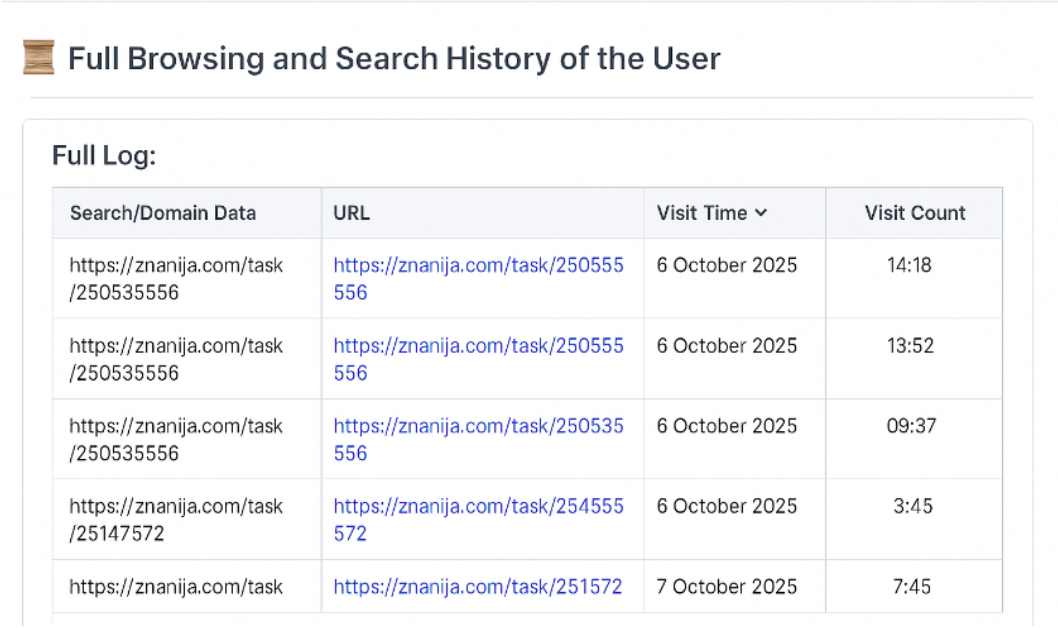

Figure 9. Browsing history.

Figure 10 presents the interface of the Login and Password Analysis Module implemented in the developed browser forensic module. This element automatically extracts and displays authentication-

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 December 2025 doi:10.20944/preprints202512.0899.v1

related artifacts (usernames, email addresses, and encrypted passwords) recovered from browser databases such as Login Data (for Chrome) and logins.json (for Firefox).

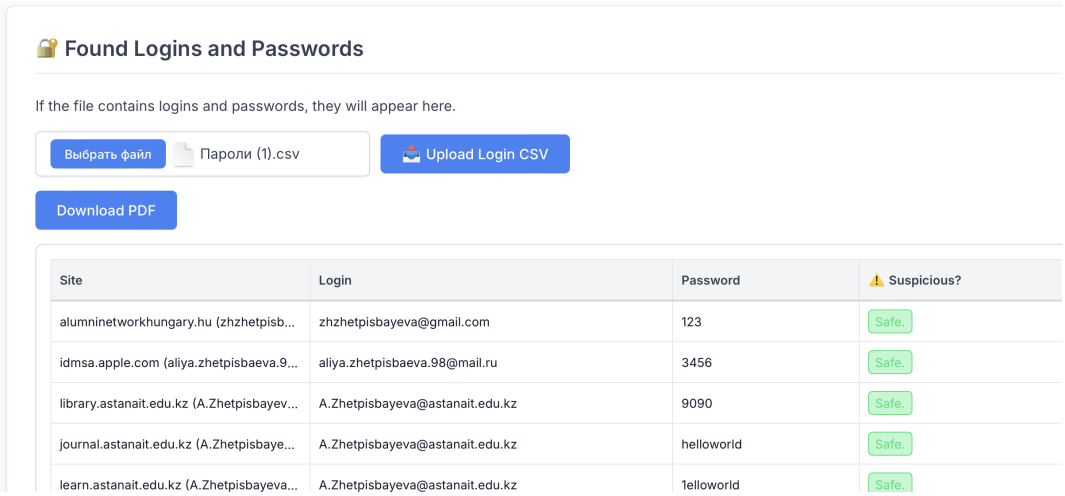

Figure 10. Sensitive search in browser history.

The module enables investigators to review and verify stored credentials, assess their potential exposure, and identify whether the same login credentials are reused across multiple websites. The table includes columns for the website domain, login, password (masked for privacy), and a column labeled "Suspicious?", which marks potentially compromised or unsafe credentials based on rule-based heuristics and similarity detection algorithms.

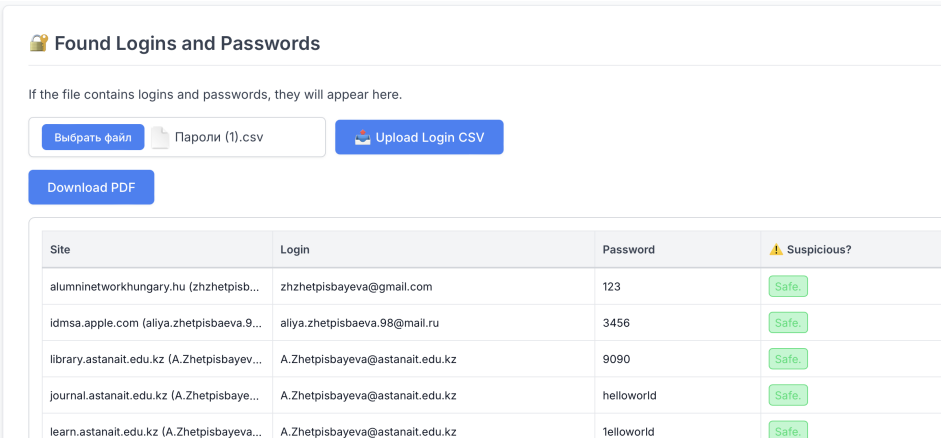

Figure 11. Suspicious user's logins/password.

In addition, the system is equipped with built-in functions for exporting data to CSV format and generating reports in PDF format, allowing forensic experts to archive analysis results or include them as evidence in digital investigation reports. This visualization demonstrates the system's ability to process confidential authentication artifacts in a secure and structured manner, preserving privacy and ensuring the accuracy of forensic analysis.

The proposed solution was evaluated in terms of its ability to systematically extract and analyze browser artifacts in various formats (SQLite, JSON, CSV) . The purpose of the evaluation was to demonstrate the advantages of the developed method, which combines rule-based approaches with machine learning algorithms, compared to existing digital forensics tools.

In particular, the evaluation was conducted by comparing the developed module with several widely used digital forensics tools, including NirSoft BrowserView, Oxygen Forensic Detective, and NetAnalysis. Although these tools provide useful functions for extracting artifacts and visualizing them, their capabilities in terms of analytics are limited. In contrast, the proposed solution implements

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 December 2025 doi:10.20944/preprints202512.0899.v1

automatic classification of suspicious URLs, behavioral profiling based on search queries, and probabilistic assessment of the security of credentials (logins and passwords), which allows for the creation of a complete forensic profile of the user.

5. Discussion

The comparative analysis presented in Table 3 and Figure 12 discusses the performance and analytical depth of the proposed browser forensic module in relation to two widely used tools, NirSoft Browser Tools and Oxygen Forensic Suite. Previous studies in the field of browser artifact analysis [1,3] have demonstrated that most available forensic utilities are limited to static data extraction and manual inspection, often lacking automation, context interpretation, and cross-artifact correlation. In contrast, the developed module introduces an intelligent hybrid model that integrates rule-based reasoning with machine learning algorithms (K-Nearest Neighbors and Naïve Bayes) for a deeper understanding of behavioral patterns.

From the comparison, it is evident that the proposed module surpasses traditional tools in several key dimensions:

1. it supports multiple artifact formats (SQLite, JSON, CSV), enabling broader compatibility across browsers and operating systems;
2. it automates the detection of suspicious URLs using pattern recognition and clustering techniques;
3. it integrates a user profiling module that infers psychological, financial, and interest-related indicators from browsing behavior.
4. it automatically checks user's logins/passwords to identify suspicious domains.

These features allow for a more holistic interpretation of user activity, aligning with recent research emphasizing behavioral forensics and data-driven digital investigation [4,5].

Table 3. Comparison between the proposed module, NirSoft, and Oxygen Forensic.

Feature	Proposed module	NirSoft	Oxygen Forensic
Supported formats	SQLite, JSON, CSV	SQLite only (limited)	SQLite, proprietary formats
Artifact coverage	History, Cookies, Logins, Bookmarks	History, Cookies	History, Cookies, Bookmarks
Detailed analysis	Rule-based + ML (K-NN, Naïve Bayes)	None (depens on tool)	Limited keyword search, filters
Suspicious URL detection	Yes	No	Partial
Security analysis	Yes	No	No
User profiling	Yes (interests, financial, psycho-emotional)	No	No
Languages	ENG/RUS/KAZ	EN	Proprietary formats
Report generation	Automated PDF/HTML report	Text/CSV export	Proprietary formats

The findings suggest that combining heuristic rules with machine learning yields higher detection accuracy and improves the interoperability of forensic evidence. This supports the hypothesis that intelligent hybrid methods can substantially enhance forensic efficiency and reduce manual workload. However, limitations remain — particularly in real-time analysis of encrypted artifacts and integration with cloud-based browser data, which will form the focus of future work.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 December 2025 doi:10.20944/preprints202512.0899.v1

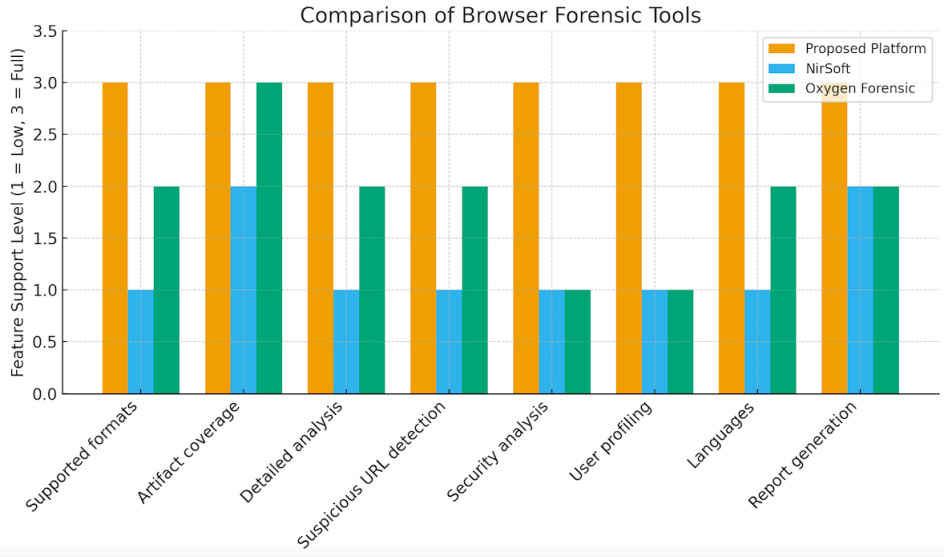

Figure 12. Example of inserted picture.

Future research directions will include:

1. expanding the module to support mobile browser artifacts and cross-device synchronization logs;
2. integrating deep learning models (e.g., CNN, LSTM) for adaptive URL risk classification;
3. developing a standardized forensic reporting API for law enforcement interoperability.

Such improvements aim to move digital forensics from data extraction toward fully automated behavioral intelligence systems.

6. Conclusions

In this paper a new method for the forensic analysis of web browser data was developed and presented. This method's primary contribution is its hybrid architecture, which effectively integrates rule-based analysis with machine learning algorithms, specifically k-nearest neighbors (k-NN) and Naïve bayes. This combination is designed to leverage the precision of heuristic rules while harnessing the adaptive pattern-recognition capabilities of machine learning. The proposed method has been implemented as a fully automated forensic module.A key practical feature of this method is to ability to natively process and interpret digital artifacts from diverse and common formats, including SQlite databases, JSON files, and CSV exports, ensuring broad applicability across different browsers and operating systems.

The developed method successfully combines various established web forensic approaches into a single, cohesive workflow. this integration enables a detailed and multi-dimensional analysis of user activity, allowing for the correlation and formation of relationships between disparate web digital traces—such as browsing history, cookie data, and saved credentials. consequently, the system transcends static data processing. it facilitates a deep behavioral interpretation of user actions, which includes the automated identification of suspicious URLs, the detection of recurring behavioral patterns, and the mapping of temporal regularities in online activity. this allows an investigator to move from answering "what" a user did to understanding "how" and "when" they did it.

Experimental validation confirmed the method's effectiveness and robustness in analyzing a comprehensive set of browser artifacts, including browsing history, cookies, logins, and bookmarks. a significant outcome of this analysis is the integrated profiling module. based on the aggregated results, this module constructs a forensic behavioral profile of the user. this profile encapsulates not only technical indicators but also inferred personal interests, potential financial characteristics, and psycho-emotional indicators. thus, the proposed method substantially expands the capabilities of traditional digital forensics, moving beyond simple artifact collection to provide actionable intelligence about user behavior and potential motivation.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 December 2025 doi:10.20944/preprints202512.0899.v1

A comparative analysis against existing solutions, such as the Nirsoft toolset and Oxygen Forensic detective, highlighted the distinct advantages in developed system. The solution demonstrated broader artifact coverage, greater analytical flexibility, and a significantly higher degree of automation. the proposed new method in this research paper effectively eliminates the critical bottleneck of manual browser data analysis. this automation not only accelerates the investigative process but also improves the accuracy and objectivity of the obtained data by reducing the potential for human error or bias.

Overall, the proposed method and the implemented solution represent a significant step forward in the development of intelligent and explainable digital forensics. This work provides a powerful tool for investigators by synergizing automation with the accuracy of user behavior analysis. Future work will focus on expanding the machine learning models to include deep learning techniques for more nuanced anomaly detection and on integrating support for artifacts from cloud-synced browser profiles, which represent a growing challenge in the field.

Author Contributions: Conceptualization:Leila Rzayeva,Aliya Zhetpisbayeva, Nursultan Nyssanov; Methodology: Leila Rzayeva, Aliya Zhetpisbayeva; Validation:Leila Rzayeva,Aliya Zhetpisbayeva,Nursultan Nyssanov,Alisher Batkuldin,Alissa Ryzhova;Formal analysis:Leila Rzayeva,Aliya Zhetpisbayeva,Nursultan Nyssanov,Alisher Batkuldin,Alissa Ryzhova;Resources:Leila Rzayeva,Aliya Zhetpisbayeva,Nursultan Nyssanov, Alisher Batkuldin,Alissa Ryzhova;Data curation:Aliya Zhetpisbayeva;Writing and editing:Leila Rzayeva,Aliya Zhetpisbayeva,Nursultan Nyssanov,Alisher Batkuldin,Alissa Ryzhova;Visualization:Leila Rzayeva,Aliya Zhetpisbayeva;Funding acqusition:Leila Rzayeva; Supervision: Faisal Saeed; Project Administration: Faisal Saeed. All authors have read and agreed to the published version of the manuscript.

Acknowledgments: This study was carried out with the financial support of the Committee of Science of the Ministry of Science and Higher Education of the Republic of Kazakhstan under Contract 388/PTF24-26 dated 01.10.2024 under the scientific project IRN BR24993232 “Development of innovative technologies for conducting digital forensic investigations using intelligent software-hardware complexes”.

Abbreviations

The following abbreviations are used in this manuscript:

GUI	Graphical User Interface
k-NN	k-Nearest Neighbors
ML	Machine Learning
SQLite	Structured Query Language
CSV	Comma-Separated Values
URL	Uniform Resource Locator
JSON	JavaScript Object Notation
HTML	Hypertext Markup Language
CSS	Cascading Style Sheets
JS	JavaScript
XML	Extensible Markup Language
HDD	Hard Disk Drive
SSD	Solid-State Drive
PNG	Portable Network Graphics

References

1. Chand, R.; Sharma, N.A.; Kabir, M.A. Advancing Web Browser Forensics: Critical Evaluation of Emerging Tools and Techniques. *SN Computer Science* **2025**, *6*(355), 1–28. https://doi.org/10.1007/s42979-025-03921-6.
2. Yermekov, Y.; Rzayeva, L.; Imanberdi, A.; Alibek, A.; Kayisli, K.; Myrzatay, A.; Feldman, G. Secure Chip-Off Method with Acoustic-based Fault Diagnostics for IoT and Smart Grid Data Recovery. *International Journal of Smart Grid (ijSmartGrid)* **2025**, *9*(3), 116–126. https://doi.org/10.20508/ijsmartgrid.v9i3.502.g392
3. Fitra, M.R.A.; Pratama, E.; Al-Kautsar, M.Z.; Harahap, F.A.; Ramadhani, F. Implementasi Struktur Data Stack untuk Pengelolaan Riwayat Penelusuran dalam Bentuk Ekstensi Web Browser Chrome. *Indonesian Journal of Education and Development Research* **2025**, *3*(1), 239–241.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 December 2025 doi:10.20944/preprints202512.0899.v1

4. Kovacs, G. Reconstructing Detailed Browsing Activities from Browser History. *arXiv preprint* **arXiv:2102.03742**, **2021**. Available online: https://arxiv.org/abs/2102.03742 (accessed on 10 October 2025).
5. Mu'Minina, M.; Anwara, N. Live Data Forensic Artefak Internet Browser (Studi Kasus Google Chrome, Mozilla Firefox, Opera Mode Incognito). *Buletin Sistem Informasi dan Teknologi Islam* **2020**, *1*(3), 130–138. https://doi.org/10.5281/zenodo.1234567.
6. Rahman, R.U.; Yadav, L.; Tomar, D.S. Phish-Shelter: A Novel Anti-Phishing Browser Using Fused Machine Learning. *Journal of Information Technology Research* **2022**, *15*(1), 1–20. https://doi.org/10.4018/JITR.2022010104.
7. Mohamed, A.A.E.M.; Ismail, I. A Performance Comparative on Most Popular Internet Web Browsers. In *Proceedings of the 4th International Conference on Innovative Data Communication Technology and Application (ICIDCA 2022)*; Elsevier Procedia Computer Science: Al Ain, UAE, 2023; pp. 1–7. https://doi.org/10.1016/j.procs.2023.08.123.
8. Zhetpisbayeva, A.; Shayea, I.; Baibussinov, A. The Effect of TRIM Function on Data Recovery from SSD Solid-State Drives. *International Journal of Networked and Distributed Computing* **2025**. DOI: 10.21203/rs.3.rs-7419937/v1 (preprint).
9. Rzayeva, L.; Ryzhova, A.; Zhaparkhanova, M.; Myrzatay, A.; Konakbayev, O.; Imanberdi, A.; Kozhakhmet, Z. Development of a Method for Determining Password Formation Rules Using Neural Networks. *Information* **2025**, *16*(8), 655. https://doi.org/10.3390/info16080655.
10. Zimmeck, S.; Goldelman, D.; Kaplan, O.; Brown, L.; Casler, J.; Jean-Charles, J.; Champeau, J.; Harkous, H. Website Data Transparency in the Browser. *Proceedings on Privacy Enhancing Technologies* **2024**, *2*, 211–234.
11. Raza, A.; Khan, M.; Ullah, F.; et al. Forensic Analysis of Web Browsers Lifecycle: A Case Study. *Journal of Information Security and Applications* **2024**. https://doi.org/10.1016/j.jisa.2024.103839.
12. Kim, H.; Lee, J.; Park, S. AIBFT: Artificial Intelligence Browser Forensic Toolkit. *Journal of Digital Forensics and Cybersecurity* **2021**. Available online: https://www.sciencedirect.com/science/article/abs/pii/S2666281720303930 (accessed on 10 October 2025).
13. Khan, M.A.; Ali, N.; Javed, S. Heuristic Machine Learning Approaches for Identifying Phishing. *Frontiers in Artificial Intelligence* **2024**. https://doi.org/10.3389/frai.2024.1414122.
14. Almeida, R.; Costa, M.; Pires, J. AntiPhishX: An AI-Driven Service-Oriented Ensemble Framework for Phishing Detection. *Information and Software Technology* **2025**. https://doi.org/10.1016/j.infsof.2025.107891.
15. Zhou, W.; Chen, Y.; Wang, L. AntiPhishStack: LSTM-Based Stacked Generalization Model for Optimized Phishing URL Detection. *arXiv preprint* **arXiv:2401.08947**, **2024**. Available online: https://arxiv.org/abs/2401.08947 (accessed on 10 October 2025).
16. Das, S.; Gupta, V.; Rao, K. Analysis of Web Browser for Digital Forensics Investigation. *International Journal of Computer Applications and Technology* **2021**. https://doi.org/10.1504/IJCAT.2021.114987.
17. Hussain, S.; Farooq, R.; Malik, M. Machine Learning-Based Criminal Behavior Analysis for Online Activity. *PLOS ONE* **2025**. https://doi.org/10.1371/journal.pone.0332802.
18. Nguyen, H.; Tran, Q.; Le, D. A Novel Framework for Effective Phishing URL Detection Using Siamese Networks. *Knowledge-Based Systems* **2025**. https://doi.org/10.1016/j.knosys.2025.112781.
19. Majeti, K.V.P.S.; Sai, Y.V.L.; Ulichi, S.S.; Mohanty, S.N.; Sudha, S.V. Digital Forensic Advanced Evidence Collection and Analysis of Web Browser Activity. *EAI Endorsed Transactions on Scalable Information Systems* **2023**, *10*(5), 1–10. https://doi.org/10.4108/eetsis.3357.
20. Lim, B.; Huerta, R.; Sotelo, A.; Quintela, A.; Kumar, P. EXPLICATE: Enhancing Phishing Detection through Explainable AI and LLM-Powered Interpretability. *arXiv preprint* **arXiv:2503.20796**, **2025**. Available online: https://arxiv.org/abs/2503.20796 (accessed on 10 October 2025).
21. Uddin, M.A.; Sarker, I.H. An Explainable Transformer-Based Model for Phishing Email Detection: A Large Language Model Approach. *arXiv preprint* **arXiv:2402.13871**, **2024**. Available online: https://arxiv.org/abs/2402.13871 (accessed on 10 October 2025).
22. Adelusi, J.B. Explainable AI in Phishing Detection and Prevention. *ResearchGate Preprint* **2024**. Available online: https://www.researchgate.net/publication/387207029_Explainable_AI_in_Phishing_Detection_and_Prevention (accessed on 10 October 2025).
23. Mia, M.; Derakhshan, D.; Pritom, M.M.A. Can Features for Phishing URL Detection Be Trusted Across Diverse Datasets? A Case Study with Explainable AI. In *Proceedings of the 11th International Conference on Networking, Systems, and Security (NSysS '24)*; ACM: New York, NY, USA, **2024**; pp. 137–145. https://doi.org/10.1145/3704522.3704532.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 December 2025 doi:10.20944/preprints202512.0899.v1

24. Pau, K.N.; Lee, V.W.Q.; Ooi, S.Y.; Pang, Y.H. The Development of a Data Collection and Browser Fingerprinting System. *Sensors* **2023**, *23*, 3087. https://doi.org/10.3390/s23063087.
25. García, B.; Ricca, F.; Torchiano, M.; Leotta, M. Enhancing Web Applications Observability through Instrumented Automated Browsers. *Journal of Systems and Software* **2023**, *111723*. https://doi.org/10.1016/j.jss.2023.111723.
26. Laperdrix, P.; Bielova, N.; Baudry, B.; Avoine, G. Browser Fingerprinting: A Survey. *ACM Transactions on the Web* **2020**.
27. Khan, I.; Hollebeek, L.D.; Xiao, S.H.; Sigurdsson, V. Mobile App vs. Desktop Browser Platforms: The Relationships among Customer Engagement, Experience, Relationship Quality and Loyalty Intention. *Journal of Marketing Management* **2023**.
28. Tommasi, F.; Catalano, C.; Taurino, I. Browser-in-the-Middle (BitM) Attack. *International Journal of Information Security* **2022**.
29. Brinson, R.; Wimmer, H.; Chen, L. Dark Web Forensics: An Investigation of Tracking Dark Web Activity with Digital Forensics. In *Proceedings of the 2022 International Conference on Interdisciplinary Research in Technology and Management (IRTM 2022)*, 2022. https://doi.org/10.1109/IRTM54583.2022.9791646.
30. Verma, G.; Sarkar, M.; Seth, D. Visualization of Online Social Dynamics for Forensic Investigation of User's Behavior. *Lecture Notes in Networks and Systems*, 2022.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.