DIGITAL BEHAVIORAL-FINGERPRINT FOR USER ATTRIBUTION IN DIGITAL FORENSICS: ARE WE THERE YET?

Adeyemi R. Ikuesan* and Hein S.Venter

Digital Forensic Research Group, Department of Computer Science, School of Information Technology, Faculty of Engineering Built Environment and Information Technology, University of Pretoria, South Africa
*Corresponding author. rikuesan@ieee.org

Abstract: the need for a reliable and complementary identifier mechanism in a digital forensic analysis is the focus of this study. Mouse dynamics have been applied in information security studies, particularly, continuous authentication and authorization. However, the method applied in security is void of specific behavioral signature of a user, which inhibits its applicability in digital forensic science. This study investigated the likelihood of the observation of a unique signature from mouse dynamics of a computer user. An initial mouse path model was developed using non-finite automata. Thereafter, a set-theory based adaptive two-stage hash function and a multi-stage rule-based semantic algorithm were developed to observe the feasibility of a unique signature for forensic usage. An experimental process which comprises three existing mouse dynamics datasets were used to evaluate the applicability of the developed mechanism. The result showed a low likelihood of extracting unique behavioral signature which can be used in a user attribution process. Whilst digital forensic readiness mechanism could be a potential approach that can be used to achieve a reliable behavioral biometrics modality, the lack of unique signature presents a limitation. In addition, the result supports the logic that the current state of behavioral biometric modality, particularly mouse dynamics, is not suitable for forensic usage. Hence, the study concluded that whilst mouse dynamics-based behavioral biometrics may be a complementary modality in security studies, more will be required to adopt it as a forensic modality in litigation. Furthermore, the result from this study finds relevance in other human attributional studies such as user identification in recommender systems, e-commerce, and online profiling systems, where the degree of accuracy is not relatively high.

Keywords: User-attribution, digital forensic readiness, behavioral fingerprint, mouse dynamics, a hash function.

Introduction

Biometric modalities, either physiological or behavioral, have been applied in the science of forensics for user identification [1,2] for decades. While physiological attributes involve the usage of human physiological composition, behavioral attributes, on the other hand, involve the usage of invariant behavior-based features associated with a human in interaction with the device or environment. A behavior-based biometric recognition system usually comprises identity identification and identity verification process. The verification process is usually a one-to-one matching process. However, the former, identity identification, involves the process of identifying a singular identity from a larger sample of the user (a 1: N matching process). Forensic science in a digital medium often involves identification, while security such as authorization and authentication involves verification. User attribution in the digital forensic analysis (DFAn) is the process of identifying the individual (and or device in some cases) 'who did what' on a given system under observation. The science behind this process is crucially important in a forensic investigation since such a process requires that the method employed should be reliable and repeatable. Reliability in forensic science is generally measured using error rate with specific acceptable threshold/bounds. This notion is held by various forensic scientists and has recently

been affirmed as a procedure that should be satisfied before the admissibility of such an evidence[3]. Mouse dynamics studies that focus on information security -authentication and authorization- do employ error rates such as equal error rate (EER), receiver operating characteristic curve (ROC), false acceptance rate (FAR), as well false rejection rate (FRR) for reliability. These error rates (in the context of authentication and authorization) are extracted based on the combination of behavioral features through a machine learning process, either based on rules or other metrics of boundary establishment. The training process takes in a series of preprocessed data from the user and tries to observe frequent pattern based on the semantic relationship among features. This process is then used to populate a training model, through which testing (validation) is carried out. Measures based on such a process requires high volume of data for model training and testing, low probability of fingerprinting a user (since hyperplane boundaries are extremely disorderly and nonlinear), high volume of data for re-identification process (which is usually not feasible, as such a process may not generate sufficient data required), as well as frequent training-retraining process. Sadly, these error rates do not present a behavioral marker which satisfies the science employed for digital forensics, hence, cannot be integrated for a user attribution process (in its current form). As highlighted in a digital forensic manifesto[2], the reliability of forensic science is limited by the accuracy with which the underlying science can predict. However, the underlying science and the intuitive composition of mouse dynamics comprises critical behavioral markers[4–6] which can satisfy the criteria as a potential forensic attribute that can be used in a DFAn process[7]. The Daubert and Frye Standards are the two fundamental standards (a legal truth which is decided by the court. Such a legal truth is often deemed as the final authority unless changed in such an explicit manner [2]) employed for admissibility and reliability of evidence. As highlighted in [3], the Fryer standard can be defined within this five basis: *"(1) Testing: Has the scientific procedure been independently tested? (2) Peer Review: Has the scientific procedure been published and subjected to peer review? (3). Error rate: Is there a known error rate, or potential to know the error rate, associated with the use of the scientific procedure? (4) Standards: Are there standards and protocols for the execution of the methodology of the scientific procedure (5) Acceptance: Is the scientific procedure generally accepted by the relevant scientific community?"* It basically requires that a method of investigation be scientific and be generally accepted by the scientific community with a known error rate. Further clarification is presented by the metrics for evidence admissibility, in [8], and a recent report on the requirement for ensuring the scientific validity of a forensic method[9]. These metrics necessitates that scientific validity is contingent on the repeatability, reproducibility, accuracy, and reliability of the method employed. One key component of these metrics, particularly repeatability, is that it requires a model that can establish unambiguous accuracy and a reliable behavioral marker. Therefore, repeatability in this sense would require much more than a generic semantic relationship model which attempts to generate behavioral marker based on a high-level relationship among observable vectors. It will require a granular microscopic magnification of each behavioral feature for the possibility of the relatively consistent behavioral marker. A process that can generate a reliable threshold for the extraction of the behavioral marker(s) from mouse dynamics within the local optimum threshold, is thus asserted to present a probable method for digital forensics. The underlying composition of mouse dynamics is presented in the next section.

Mouse-Dynamics

Mouse dynamics is the process of extracting meaningful behavioral information from human-mouse (or any other form of computer-based pointing device) navigation [4], which is then used to generate a behavioral pattern of the movement pattern of the user. Behavioral attributes considered in mouse dynamics include base (or primary) and secondary features. Furthermore, it is safe to assert that mouse dynamics is a widely accepted measure of behavioral biometrics in the information security community which is gradually attaining the 0.001% FAR, and 1.00% FRR, of the European Standard for Commercial Biometric Technology [5]. A descriptive composition of these attributes used in mouse dynamics study is shown in Table 1.

Table 1: Summary of mouse characteristic features

Mouse feature	Description	Unit of measurement
Button-down & up	The time stamp when a button is pressed and when the is released	Milliseconds (*ms*)
Single & double click	The time between the mouse click-up and mouse click-down. This can be represented by either dwell time for single click or flight time for double click	*ms*
Raw mouse movement	Aggregated over a predefined delimiter such as session, clicks, or path. It forms the base on which other features are extracted.	Click, path or session in pixels and *ms*
Direction of movement	Depending on the sector size chosen, the direction is a vector quantity which shows the path of the mouse movement, from the beginning to the end. Few studies used 8-sectors (each at 45^0) to define the probable direction of mouse movement	categorical (Sector number)
Movement distance	The total number of pixels covered by the mouse movement between consecutive clicks or between the fundamental units of mouse movement.	Pixel/*ms*
Velocity of movement	The ratio of the movement speed between consecutive intervals and the duration of the interval measured in milliseconds	Pixel/*ms*
Duration	The time between the start of a movement and the end. Also, it could be time between subsequent clicks after the initial click.	*ms*
Number of silent	This refers to the number of times a silent action is recorded in a given mouse data. It can be aggregated over a predefined session or any other unit of mouse data. Furthermore, other statistical properties can be generated.	number
Scroll duration	The time between the first observation of scroll to the time of the last observation	*ms*
Scroll direction	This depicts an upward direction or a downward direction of the scroll	categorical
Scroll speed	The rate of the scroll in either an upward or downward direction	*ms*

Each feature characterized in Table 1 represents the base feature through which further statistical features, aggregated features, and the transformed feature can be generated. Generally, studies either utilized ordered moments of statistical analysis or arithmetic aggregates of behavioral data over a given segment/session. Aggregation of behavioral data provides a level of abstraction which could be further applied for forensic purposes. However, the probability of extrapolating behavioral signature explored in existing studies is based on the interaction of multiple features is primarily reliant on pattern-based multiple feature aggregation. While such multi-feature-based approach provides a substantial feature-vector for data mining process, it does not present a measure of reliable forensic identification. One major assumption in existing studies, particularly on user identification, is that behavioral attributes collected during the verification phase are sufficient for authentication/authorization. For instance, findings by Gamboa and Colleague[10] was

based on 10-strokes (where each stroke is a movement between two consecutive clicks) within an approximation of 10-seconds. Similarly, the study by Bours [11] detected imposter in an average of 182-strokes, while findings by Bailey and colleagues[12] was based on 987-strokes per 10-minutes sliding window. In reality, the feasibility of user attribution from the dataset available for authentication and authorization of an imposter in these studies may not be realistic, as asserted in [1,4]. This is because most malicious actions are carried out within brief action and under short duration, thus, user attribution based on mouse dynamics would require a process that is based on lesser data input (lesser number of strokes). One area of probable logic is the description of mouse dynamics as hand-gesture-based behavioral biometrics. Intuitively, hand-gesture-based biometrics could be synonymous with gait trajectory as both would comprise angle of inclination with reference to a starting coordinate, distance traveled, direction and duration of travel which can be parametrically represented by the expression $X_a = \left[x_a(t), z_a(t), \theta_a(t)\right]^T$ where $x_a(t)$, $z_a(t)$ describes the ankle position and $\theta_a(t)$ depicts the angle between the foot and the horizontal plane. A human gait, k, is defined when the heels of any foot leave the ground and finished when the same heel touches the ground again. Following the intuition of gait trajectory, a mouse movement from one point in Cartesian space to another point in the Cartesian space can be parametrically described by the mathematical expression $M_a = \left[x_a(t), y_a(t), \theta_a(t)\right]^T$ where $x_a(t)$, $y_a(t)$ depicts the position of the cursor at coordinates x, and y at mouse location a and $\theta_a(t)$ represents the angle of inclination of the mouse. A generic description of a mouse action is further presented in (1) while (2) depicts the mathematical expression of a mouse movement path. Furthermore, a mouse path can be described by a boundary condition which satisfied either a double click or time-lapsed criteria. This expression mimics a typical human behavioral movement/trajectory on a computer-based pointing device.

$$
Mouse - Point_{(i,i+1)} = \begin{cases}
\Delta v_i = \dfrac{\sqrt{\left(x_{(i+1)} - x_i\right)^2 + \left(y_{(i+1)} - y_i\right)^2}}{t_{(i+1)} - t_i} \\
D_i = \{0,1,2,3,4,5,6,7\} \cong \begin{cases} upward, & 1 \\ neutral, & 0 \\ downward, & -1 \end{cases} \\
\Delta\theta_i = \tan^{-1}\left(\dfrac{y_{(i+1)} - y_i}{x_{(i+1)} - x_i}\right) \\
\Delta w_i = \left(t_{i+1} - t_i\right)^{-1} \begin{cases} \sqrt{\left(x_{(i+1)} - x_i\right)^2 + \left(y_{(i+1)} - y_i\right)^2} \times \left(\sin\theta\right), & D_i = 1 \\ \sqrt{\left(x_{(i+1)} - x_i\right)^2 + \left(y_{(i+1)} - y_i\right)^2} \times \left(\cos\left(360 - \theta\right)\right), & D_i = -1 \\ \sqrt{\left(x_{(i+1)} - x_i\right)^2 + \left(y_{(i+1)} - y_i\right)^2}, & D_i = 0 \end{cases}
\end{cases} \tag{1}
$$

$$\overrightarrow{Mouse\_Path}_{(i,i\pm1)} = \begin{cases} Weight_{(i,i\pm1)} = \sum_{i=2}^{n} \Delta w_{(i,n)} \\ Path \triangleq \begin{cases} Delay \begin{cases} Min \geq 3secs \\ Max \leq 10secs \end{cases} \\ 2\,consecutive\,clicks \end{cases} \\ Speed_{(i,i\pm1)} = \dfrac{\left(\sqrt{\left(x_{(i+1)} - x_i \right)^2 + \left(y_{(i+1)} - y_i \right)^2} \right)}{t_{(i+1)} - t_i} \\ Duration_{(i,i\pm1)} = \sqrt{\left(x_{(i+1)} - x_i \right)^2 + \left(y_{(i+1)} - y_i \right)^2} \\ Event_{type(i,i\pm1)} \equiv \begin{cases} click \rightarrow click \\ click \rightarrow move \\ move \rightarrow click \\ move \rightarrow move \end{cases}, scroll, silent \end{cases} \quad (2)$$

The path expression shown in Equation 2 is a depiction of mouse dynamics as a hand-gesture process, which is dependent on human disposition, amongst other factors. This study is anchored on the supposition of an inherent behavioral composition of human hand-gesture. This logic has been asserted in existing mouse dynamics studies [4,5,13]. However, these studies failed to identify behavioral markers suitable for 'forensic print' which can be applied for a digital forensic analysis process. A forensic print, in this instance, can be described as a behavioral marker that is extracted from a mouse dynamics data, which can then be compared against other mouse dynamics data for similarity. However, such a marker should be extracted from data within a given short interval. In addition, a marker is required to be consistent for a given user, and relatively distinct among other users. This aligns with the logic of maximized local minimum, and minimized global maximum in classical machine learning problems.

Research Methodology

The architectural depiction of the overall process employed in this study is shown in Figure 1. A generic model of human dynamics based on mouse action is first derived based on a non-deterministic finite state machine. The state machine model provides a generic sequence model suitable for the representation of a typical human action [14]. Such representation provides a higher order of behavioral abstraction. To further extract a lower abstraction, the next phase of the methodology considered the study of the development of individual behavioral marker extraction

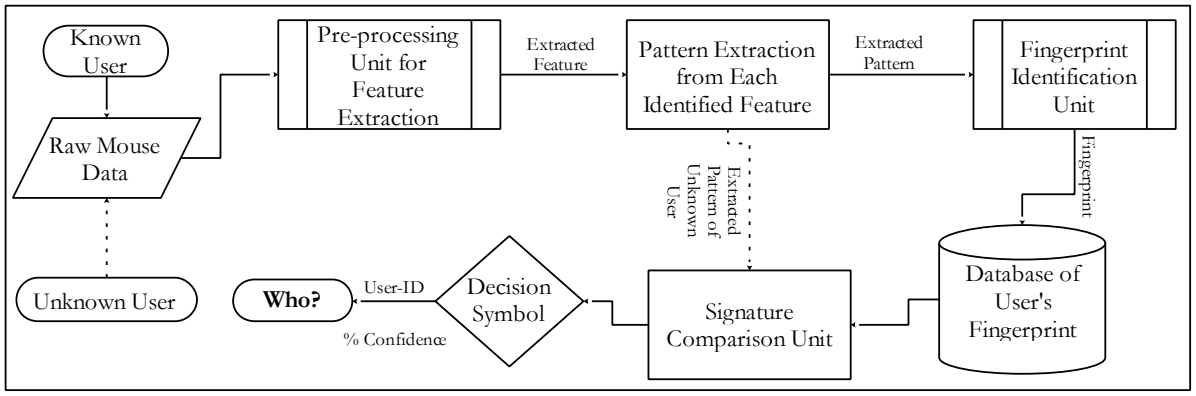

Fig. 1. Architectural layout of the Mouse Dynamics fingerprinting process

process. The behavioral-marker process is based on a two-stage hash function process. This is then followed by the validation of the reliability (on a hypothetical hypothesis) of the observed behavioral markers using Bayesian inference process.

Mouse Navigation Feature

As highlighted in Table 1, features of mouse dynamics encompass primary and secondary characteristics. The process of extracting behavioral signature based on these features require aggregation at different granular levels. For instance, in mouse pattern analysis, the use of the mouse path among a series of path curvature presents a feasible human pattern visualization which can be used in aggregation mechanisms. In this study, mouse-path is considered a fundamental component of mouse navigation. A mouse path is defined by the expression in (2) which include mouse movement, click, scroll and silent events which can be characterized by weight, speed and/or duration of the path.

Mouse Path Model based on Finite state machine

Mouse path formalization process is an abstraction process. The traversing from reality into digital representation is generally referred to as the abstraction approach. Models based on finite state machine have been widely applied to digital forensic language formalization. More specifically, models based on the finite state machine (FSM) have been used to create abstraction targeted at improving the efficiency of data representation and forensics process. A study conducted by James and colleagues [15], assert that a given system can be directly mapped to a model of FSM, which can be adapted for evidence reconstruction. Whilst it is asserted that a deterministic-based FSM (DFA) is capable of modeling a system, the current study posits that human behavior in mouse dynamics comprises sequences of states which satisfies the non-deterministic approach of FSM (NFA). FSM is deterministic if the probability that an input to a state will produce singular output, is 1(absolute). However, if the probability is less than 1, then such FSM can be defined as non-deterministic. Table 2 presents a more detailed comparison between DFA and NFA.

Table 2: DFA and NFA comparison

No.	Deterministic Finite Automata	Non-Deterministic Finite Automata
i	If the current state is known, and the input to the current state is known, the next state can be determined with absolute certainty	If the current state is known, and the input to the current is known, the next state cannot be determined with absolute certainty
ii	It cannot accept an empty state	It can accept an empty input to a state which leads to an empty state
iii	Based on i the above, the system can only have one unique next state	The system can accept multiple next states, and it can be complex.
iv	Based on iii the above, the system is not stochastic, non-random, non-chaotic and its degree of freedom is equivalent to 0	Random next state, which could have a variable degree of freedom, chaotic. It could also assume a parallel process next state.

Human dynamics usually convey a certain degree of freedom, which can be characterized by stochastic properties [16,17]. This property satisfies the interest-driven model of the form $\lambda(t) = \frac{a}{t}\int_a^t dt\lambda(t) \Leftrightarrow 25591x^{-0.372}$ [18] which also obeys the Power of the form $P(t) \approx T^{-\alpha}$. Properties

of mouse-dynamics, therefore, satisfy the non-deterministic characteristics, which can be modeled using an NFA. Formally, this study defines NFA for mouse dynamics, by extending the definition in [15] as follow:

Assume a given mouse path (P_m) has a path sequence (P_s) denoted by M. For a given mouse action from one state to another, it can be encoded as a triple (P_1, e, P_2) such that a sequence of mouse action can be computed. The model for path start and path end is denoted by the conventional FSM notation defined by $P_m \forall M = (Q, \Sigma, q_0, F, \delta)$ where;

$Q = Finite\ set\ of\ all\ possible\ states; event\ type\ \{click, movement, silent, scroll\}$

$\Sigma = Finite\ set\ of\ all\ possible\ input$ variables; $direction, speed,$

$q_0 = Initial\ state, \in Q;\ path\ beginning$

$F = Set\ of\ final\ states$

$\delta = Q \times \Sigma \to 2^Q$ $;transition\ function\ that\ returns\ the\ next\ state$

The number of possible states of a mouse action is defined in this study by four primary actions (click, move, silent, scroll). These four actions depict a summary of probable action that can be performed with a mouse. Other actions such as double-click, drag-and-drop, as well as highlight (drag only) are defined as integrative action. An integrative action is an action that combines two or more primary actions. The input variable of a mouse action comprises action attributes such as the direction of mouse movement and speed of the mouse. As defined in the existing studies on mouse dynamics [5,19], mouse direction can comprise 8-sectors$\{0,1,2,3,4,5,6,7\}$, $\theta = 45$. A mouse direction can also be defined as complex attributes which combine the angle of inclination, and movement location as a vector quantity. For the sake of simplicity, this study adopted the sectorial classes of direction during modeling and further categorized mouse speed into three classes $\{fast, \text{mod}erate, slow\}$. The silent state $\{\varphi\}$ is denoted as the initial state of the NFA, while the other states form the set of the final states $\{click(c), scroll(s), movement(m)\}$ of the NFA. A mouse-path can also be defined as a restriction (path expression over a sequence of transition triple) on the probable computation over the encoded NFA triple $\left(P_m = \{Q, \Sigma, \delta\}\right)$. Therefore, the transition function which generates the restriction for the next state of a mouse-action is defined as;

$$\delta_{next\ state} = \begin{cases} Q_i = Q \cup \{\varphi\},\ \varphi \in Q \equiv Q \\ \Sigma_i = (q, e, q_n);\ q \in Q,\ e \in \Sigma,\ q_n = \delta(q, e) \end{cases} \quad (3)$$

This proposed mouse path differs from the existing model in two forms. Firstly, the definition of the mode as an NFA, as opposed to DFA. Secondly, the initial state in the present study is defined as a subset of the finite set of all probable states (as shown in (3)) which contrast existing work [15] that defines the initial state as a disjoint set of the finite set of all probable states. These two distinctions differentiate the present study. The next subsection discussed the actual process of generating behavioral fingerprints from mouse-path model.

Hash-based Behavioral Identity

The developed model in the previous section provides a higher abstraction which covers the general process of modeling human navigation. This section builds on the model to extrapolate behavioral identifiers which can be used in the user attribution process. To achieve this, a two-stage hash function process is defined as shown in Table 3. A hash function is a method of

producing a fixed-length representation of the variable-length message. The two-stage hash process adapted in this study comprises a locality-sensitive hashing algorithm (modular hash function), and a bijection injective hash function. The modular hashing algorithm intuition posits the logic that an integer/float from an arbitrary source have the probability (albeit, higher) of exhibiting self-similarity. For instance, the expression in (4) shows a congruence relationship between $200,\ 20,\ 10$ *and* 0. This congruence relationship provides a level of abstraction through which self-similarity can be observed among the variables of interest. Modular hashing function thus magnifies self-similarity in patterns, consequently, help to optimize probable behavioral signatures that would otherwise be ignored.

$$200 Mod 10 \cong\ 20 Mod 10 \cong 10 Mod 10 \cong 0 Mod 10 \tag{4}$$

This contrasts with the general string hashing process where self-similarity is not considered. Thus, modular hashing suggests a higher probability of capturing the rhythmic pattern in the mouse data of a user. A modular hash function is generally defined by $h(k)=kModm$ where k is an integer/float and m is the list size. If m is a power of 2; $m=2^p$, then $h(k)$ is the lowest-order bits of k. The second stage hash function (based on a fuzzy hash function) computes the hash of each n-bits (in this case, 8-bits) block of the output of the binary32 IEEE-754-standard sequence as illustrated in (5).

$$\begin{array}{l}\{10111000101010101100111110111000\} \\ \left.\begin{array}{l}\overbrace{10111000}^{block_a} \\ \qquad\overbrace{10101010}^{block_b} \\ \qquad\qquad\overbrace{11001111}^{block_c} \\ \qquad\qquad\qquad\overbrace{10111000}^{block_d}\end{array}\right\} 8\text{-}bit\ \ blocks\end{array} \tag{5}$$

The hash $\{h(a,b,c,d)\}$ of a block a,b,c,d is the expression:

$$\begin{cases} a_1\times\alpha^0 + a_2\times\alpha^1 + ... + a_n\alpha^{n-1}\ ,\forall a \in block_a \\ b_1\times\alpha^0 + b_2\times\alpha^1 + ... + b_n\alpha^{n-1}\ ,\forall b \in block_b \\ c_1\times\alpha^0 + c_2\times\alpha^1 + ... + c_n\alpha^{n-1}\ ,\forall c \in block_c \\ d_1\times\alpha^0 + d_2\times\alpha^1 + ... + d_n\alpha^{n-1}\ ,\forall d \in block_d \end{cases} \\ where\ \alpha = hash\ key \tag{6}$$

This approach mitigates the effect of *signed and unsigned* float input to the IEEE-754-standard output. However, hashing a 32-bits strings presents a potential towards the global optimum. To address this, block-wise (piecewise) hash computation is performed on the 32-bit stream as shown in (6). Such a process further allows for a block-wise comparison (fuzzy hashing) of behavioral features.

Table 3: TSHF Algorithm for Behavioral signature extraction

Algorithm-1: Two-Stage Hash Function (TSHF)
Input: multiple Mouse data for a single user, Mouse-path model, and preprocessed mouse-path features.
Output: sets of behavioral signatures and its respective weight
Step-1. f_i = extract individual feature from the mouse-path from each mouse data
Stage-1*: modular hash function*; to maximize local optimum within individual pattern
Step 2. $m(f_i)$ = compute the modular hash;

M= k *mod* m, where m = 2^p
P =8, signifies an 8-bit process
Stage-2*: injective hash function*; to eliminate the effect of float point data
Step 3. Compute M'= 32-bit IEEE 754 standard of M
M' = IEEE 754 equivalent of M
Split M′ into octets M_8'; to prevent having a global optimum problem, as shown in (5)
Step 4. Compute a new hash for each octet, hM_8'
hM_8'= binary hash with key size, α, α=3 in current study
Concatenate HM_8'; generate string of octet
Step 5. Compute identity function
modal frequency computation ≥3 octets
Step 6. Compute similarity(fingerprint) across different mouse data from the same user
Step 7. Create a matrix of a fingerprint, which also includes the minimum and maximum weight.
Return matches and weight of each match

The hash function ensures that for each block of 8-bit sequence, a computation which satisfies the expression defined in (7) is obtained. Using the block size illustrated in (5), the second stage hash will generate an identical hash digest for block-a and block-d, which satisfies the positive property for the piecewise hash digest. Conversely, the digest of block-b and block-c will satisfy the contrapositive property. Consequently, the hash digest of an individual block is then concatenated to generate a new hash digest as a representation of the original integer/float value.

$$\begin{cases} \forall p,q \in S,\ f(p)=f(q)\ \ iff\ \ p=q, & positive \\ \forall p,q \in S,\ f(p)\neq f(q)\ \ iff\ \ p\neq q, & contrapositive \end{cases} \tag{7}$$

This process provides a measure to prevent a problem of global optimum in behavioral signature observation. Global optimum occurs when all the respective hash digest is approximated to the nearest value such that every behavioral pattern appears to be the same. This is further illustrated in Table 4.

Table 4: Comparison of Forward Hash Function Process

Float/Integer Value	IEEE 754 Binary equivalent	Forward-count hash function (α=3)	Proposed approach (α=3)
892.6504954065829	01000100010111110010100110100010	208813466278209	246-3270-2277-739
807.8637642163036	01000100010010011111011101001000	238618024054770	246-2271-3199-84
698.8538674657473	01000100001011101011011010100110	277389245933859	246-1062-1009-982

$$\begin{aligned} value-P \in User_i &= 01000100/01011111/00101001/10100010 \equiv \underset{a_1}{246}/\underset{b_1}{3270}/\underset{c_1}{2277}/\underset{d}{739} \\ value-Q \in User_i &= 01000100/01001001/11110111/01001000 \equiv \underset{a_2}{246}/\underset{b_2}{2271}/\underset{c_2}{3199}/\underset{d_2}{84} \end{aligned} \tag{8}$$

The behavioral similarity is calculated based on the location alignment fuzzy hashing process. Location alignment of hash digest implies the pairwise comparison of the similar hash block. This process is illustrated in (8). As illustrated in(8), there is only one similarity match between the observed attributes $[a_1,a_2]$. This study considered the threshold of $\geq 3 identical\ blocks$ as the baseline similarity index between two attributes, while the $|blocks|=4$. Therefore, given a set of behavioral attributes from a single user, the expression described in (9) can be used to define the similarity of behavioral features.

$$\forall P_{(a,b,c,d)}, Q_{(a,b,c,d)} \in User_i : \exists \{a_i, b_i, c_i, d_i\} : \Pr(P \mid Q := 4),$$
$$where\ X_{k\{i\}} = Seqence\ of\ behavioral\ markers\ of\ the\ mouse\ dynamics\ of\ User_i \tag{9}$$

The output of (9) forms the basis for the process of user identification. This process is covered in the next subsection.

User Identification Process

For each user, a sequence of behavioral signature $(X_{k\{i\}})$ from features (f_j) is observed based on the probability of extracting two identical hash digest that satisfies the expression in (9), in a given mouse-path data. This identical hash digest is then observed for all session of each user. The process of user identification defined in this study is intuitively based on the Bayesian theorem, with a general rule as depicted in (10).

$$P(A|B) = \frac{P(B|A) \times P(A)}{P(B)}, where$$
$$P(A) \Rightarrow prior\ probability\ (what\ is\ known\ about\ A, before\ any\ observation)$$
$$P(B \mid A) \Rightarrow likelihood\ (the\ observed\ factor\ B,\ for\ all\ values\ of\ A) \tag{10}$$
$$P(A \mid B) \Rightarrow Posterior\ probability\ (what\ is\ known\ about\ A, after\ observing\ B)$$

Bayesian statements are probabilistic statements which attempt to quantify and analyze the subjective degree of belief. It follows directly from the rule of conditional probability; $\{P(A \cap B) = P(A|B) \times P(B) = P(B|A) \times P(A)\}$. Bayes rule can be centrally stated as a proportional expression of the form; $posterior\ probability \propto prior\ probability \times likelihood$. Given that Bayes theorem assumes uniform weight for all variables, the study further defined a weighting factor for each feature based on inverse document frequency (idf) technique, and weight factor for the given user based on term frequency technique. The *idf* of a feature (f) is parametrically defined by the expression in (11).

$$idf_f = 1 + \log_{10}\left(N / df_f\right), where$$
$$df_f \Rightarrow count\ of\ signature\ of\ feature,\ f \tag{11}$$
$$N \Rightarrow Total\ count\ of\ signature\ in\ the\ dataset$$

The $idf - technique$ essentially provide a scaling factor for feature ranking, where features with the highest number of the signature count are appended with a low scaling factor, in relation to the converse. This logic posits that signature collision among user is proportional to the signature count in any feature. In other words, the higher the observed signature in a given feature, the more its probability of belonging to multiple users. Given the probability of signature collision, the likelihood of the Bayesian theorem defined in (10) is further modified as follow;

Suppose that the likelihood of an unknown behavioral fingerprint with weight, belonging to a known user with weight, w_i (also referred to as term frequency), is given by the expression

$P_{(i)} = \frac{(w_i \times idf) + w_j}{(w_i \times idf) \times w_j}$, and if the probability of an instance weight of user, (u_i), be denoted by $p_{(u_i)}$, then the likelihood that an unknown user instance weight $(w_{\mu i})$ to belong exclusively to u_i among all user, u_{n-i}, can be denoted by (12).

$$P(u_i|u_{n-i}) = \frac{\left|\sum P_{(n-i)} - P_{(i)}\right|}{\sum P_{[n]}}, \text{ where } i, \text{ and } n \in \text{UserIdentifiers} \tag{12}$$

Substituting (12) into (10), a new expression for the computation of the Bayesian statement is generated as shown in (13).

$$\begin{aligned} P(n-i|i) &= \frac{\frac{\left|\sum P_{(n-i)} - P_{(i)}\right|}{\sum P_{[n]}} \times P_{(n-i)}}{P_{(i)}}, \text{ for simplicity}, \sum P_{[n]} = \eta \\ \Rightarrow \therefore P(n-i|i) &= \frac{\left[\left|\sum P_{(n-i)} - P_{(i)}\right| \times \eta P_{(n-i)}\right] P_{(i)}}{\eta} \end{aligned} \tag{13}$$

In summary, in order to achieve the user inference process, the behavioral marker will be generated using the expression in (9). Subsequently, each feature of the behavioral marker will be ranked using the expression in (11). Consequently, a ranked count vector of features in the $\mathbb{N}^{|v|}$ in the natural vector space will be generated. Inference will then be generated for every unknown input using the output of the expression in (13). Inference in this sense implies the summation of all the observed ranked factor weights for each user. In order to attribute a given signature to an individual among a pool of known user, the expression (14) takes the sum of the overall likelihood of a signature. The given pattern is attributed to the user with the highest overall likelihood as highlighted in Algorithm-1 in Table 3. A summary of the attribution process is further described in Algorithm-2 in Table 5.

Table 5: syntactic pattern User Attribution Process

Algorithm-2: Inference based user Attribution mechanism

Input: Matrix of the behavioral signature weight of known users $\{I_k^1, I_k^2, I_k^3, ..., I_k^n \quad size\, n \times 4\}$ for min and max weight

and Matrix of the unknown user $\{I_k^m, size\, m \times 4\}$

Output: Attribute range inference scale (most probable)

$$A_r = \sum P((n-i)|i) \tag{14}$$

$$Output(A_r) = \begin{cases} 1; individual\ with\ the\ highest\ range \Rightarrow most\ probable \\ \prec 1; other\ ranked\ order\ of\ individual \Rightarrow likely\ probable \\ 0;\ individual\ with\ the\ lowest\ range \Rightarrow least\ probable \end{cases}$$

Step-1*: i.* Create a new matrix of known users based on the common behavioral signature of the unknown user

$I_k^{1'}, I_k^{2'}, I_k^{3'}, ..., I_k^{n'} \qquad size\, n \times 4, where\ I_k^{1'} = I_k^1 \cap I_k^m$

ii. generate a normalized fingerprint for each user data

a. Compute the count of signature of each feature
b. Compute the bias based on the highest count:

$$\kappa = highest\_count\_value \Big/ individual\_count\_value$$

c. Compute a localized threshold for κ using:

$\tau = m(x_i) + \sigma(x_i)$: the sum of the mean and the standard deviation

d. Compute a new fingerprint bias, κ', using:

$$\kappa' = \begin{cases} if\ \kappa \succ \tau, bias = \tau \\ else, bias = \kappa \end{cases}$$

e. Generate a new fingerprint by scaling each instance of a feature by κ'

$$\kappa' = instance\_weight + \kappa$$

Step-2*:* Compute the scaling factor for each feature using (11), and generate the inference weight using (13)

$$Inference\ weight = \begin{array}{c} \begin{matrix} u_1 & \cdots & \cdots & u_n \end{matrix} \\ \begin{pmatrix} w_{11} & \cdots & \cdots & w_{1n} \\ \vdots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \vdots \\ w_{m1} & \cdots & \cdots & w_{mn} \end{pmatrix} \end{array} \text{ for min and max weight.}$$

Step-3: Compute the attribute range from the inference weight:

a. Compute a range scale using 3-sigmal-rule on the inference weight of each user

b. Compute the overall inference weight using the expression:

$$A_r = \begin{cases} A_{r(\min)} = \Sigma\, inference\ weight\ percentile_{\min} \\ A_{r(\max)} = \Sigma\, inference\ weight\ percentile_{\max} \end{cases}$$

Step-4. Compute standardize ranking; $range: 0 \to 1$

Return inference scale of the most probable, likely probable and least probable match to the unknown user.

The structural depiction of the overall process for user attribution through mouse event action is further illustrated in Figure 1. Given that the defined attribution process defined in Table 2 lacks the semantic relationship among features of mouse action, a further attribution mechanism is defined in Table 3. Arguably, the introduction of the semantic properties of a component can provide a more reliable platform for forensic integration. Therefore, the expression in (1), and (2) are used to extract mouse dynamic semantic components for user attribution, as presented in Table 3. Summary of the features is shown in Table 4. The semantic attribution process is illustrated in Figure 2. This process comprises two core functional compositions: behavioral model development process and the behavioral pattern validation process. The composition of each process is further discussed in the discussion section of this manuscript. In order to achieve higher accuracy and a reliable attribution process, a baseline accuracy threshold was defined as rule-inclusion property. The inclusion property, based on the Biometric standard[5], is further expressed in (15). These criteria select rules within a given rule-based classification process which satisfies the false acceptance rate of 0.001, 0.01 and 0.1 respectively. However, a minimal rule weight is further defined. This approach addresses cases when the rule satisfies the criteria, but the weight of the rule is below a given threshold. In this study, a minimal threshold of 5-rule-weight is considered.

$$Inclusion_{criteria} = \begin{cases} c_1 \leftarrow \sigma Rule \in FAR \le 0.001,\ \forall Rules \\ c_2 \leftarrow \sigma Rule \in FAR \le 0.01,\ \forall Rules \\ c_3 \leftarrow \sigma Rule \in FAR \le 0.1,\ \forall Rules \end{cases} \quad (15)$$

The overall semantic-process, as shown in Table 6, is based on the integration of off-the-shelf supervised human-readable machine learning algorithm. To select the off-the-shelf machine learning algorithms, a 10-fold cross validation 10-iteration experimental process was carried out using the Waikato environment for knowledge analysis (WEKA) tool. The best performing human readable (rule or tree-based algorithms) were then considered. Mouse dynamics data is used as input, and the process generates a matrix of the degree of user-certainty of a given instance. The semantic algorithm comprises four phases, as shown in Table 6.

Table 6: Semantic Rule-based Attribution process

Algorithm-3: Attribution based on Semantic properties of Mouse dynamics

Input: Y : Mouse dynamics feature dataset

Output: Matrix of the degree of certainty of attribution $size\, n \times 3$

Step-I: Split dataset into development (P), and validation (Q) dataset

i. $P \leftarrow Sampling(90\%\text{ of } Y \subseteq Y)$: ordered sampling without replacement

ii. $Q \leftarrow Sampling(10\%\text{ of } Y \subseteq Y)$

Step-II: Explore applicable rule-based classifier

i. Dataset $\leftarrow P$

ii. Acceptance criteria $\overset{def}{\equiv} \{AUC \geq 0.8; Accuracy \geq 80\%\}$

Step-III: Extract Rule-set to build a forensic readiness database

i. Forensic inclusion criteria (Ψ) $\overset{def}{\equiv} accuracy \vdash FAR \leq 0.01 \& FRR \leq 0.1$

$t\Re \leftarrow$ total Rule

ii. Sort classifier into the hierarchical order of accuracy

iii. Extracted Rule $\leftarrow \forall Rule \subseteq t\Re, \exists Rule \in \Psi$

forensic Rule $\left(f\Re\right) \therefore \leftarrow \forall_c \left(\forall Rule \subseteq \Re, \exists Rule \in \Psi\right)$ c $\Rightarrow$ classifier

Step-IV: attribution process:

i. Dataset $\leftarrow Q$

ii. Extract instances that match $f\Re$

$k \leftarrow$ count of the classifier,

$x \leftarrow$ instance of Q

while $k \geq 1$ do:

a. Check for matches such that;

$$output = \begin{cases} Correct \leftarrow \sum\limits_{x} \forall x \subseteq Q, \exists ! x \in f\Re \\ Incorrect \leftarrow \sum\limits_{x} \forall x \subseteq Q, \exists x \in f\Re, \notin class_{label} \\ Unknown \leftarrow \sum\limits_{x} \forall x \subseteq Q, \exists x \notin f\Re \end{cases}$$

b. Define the conflict resolution process for different classifier

for $\forall \{x_{ik} \neq x_{jk}\}$ of $Output_{Correct}$, **do**;

prior probability $\left(c_i\right) \leftarrow$ accuracy weight of rule match for x_i from

prior probability $\left(c_j\right) \leftarrow$ accuracy weight for rule-match for x_j

if $\mathrm{P}(c_i | c_j) > \mathrm{P}(c_j | c_i)$

$c_i \leftarrow conflict$, else, $c_j \leftarrow conflict$

end **for**

$k = k - 1;$

end while

c. Perform simple majority vote:

$$\begin{bmatrix} q_i & f\Re_i & l_i \\ . & . & . \\ . & . & . \\ q_n & f\Re_n & l_n \end{bmatrix} \begin{matrix} q_i \Rightarrow validation\,instance \\ f\Re_i \Rightarrow forensic\,Rule \\ l_i \Rightarrow matching\,label \end{matrix}$$

return

The first phase, the data splitting phase, splits the input data into model development (training and testing) dataset and the validation dataset in the proportion of 90%:10% respectively. 10-fold cross validation is considered for the development process. Rule-based classifier exploration process and forensic-rule extraction process are carried out in the second and third phases respectively. The classifier exploration process employs standard machine learning metrics for classifier performance. These metrics include the area under the receiver operating characteristic curve (AUC), accuracy, kappa statistics, and sensitivity. Based on the output of the classifier exploration and forensic criteria phases, the fourth phase enumerate the process of attributing a given unknown instance to a known user. The attribution comprises the pattern matching, conflict resolution, and voting processes. The conflict resolution process is based on Bayesian inference process which considers the posterior probability of a given instance based on the corresponding weight of the matching rule (as the prior probability). The sequence representation of these phases is further shown in Figure 2. Instances from the validation process that does not match any are categorized as unknown. These unknowns can be defined as noise after the last stage of the rule conflict handling process. The unknown instance, therefore, depicts a statistically random behavioral (noise) that is usually associated with most forms of the behavior-based identification process. In other words, an unknown instance could be a new pattern, or a completely random pattern of a User; albeit unknown.

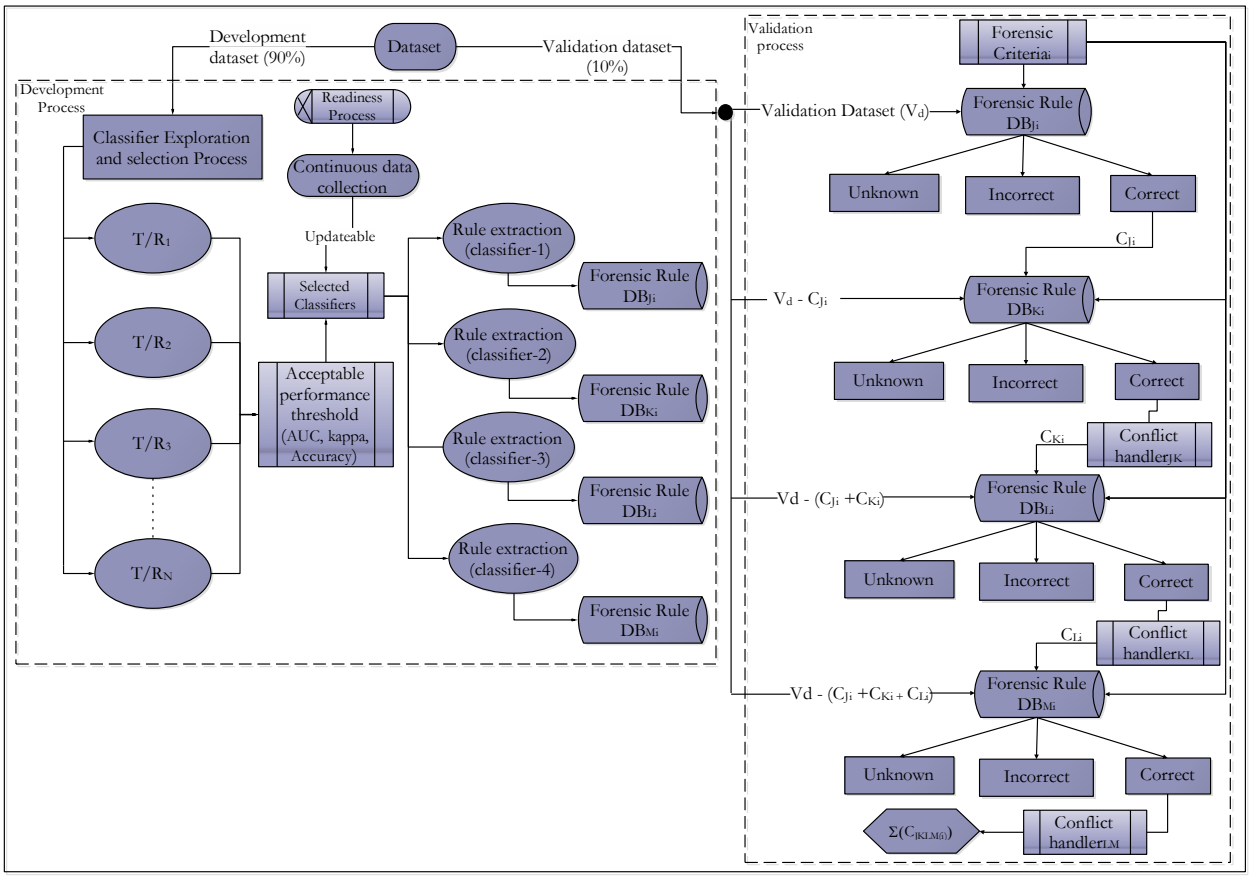

Figure 2: Process of semantic behavioral pattern attribution

Evaluation Metrics:

The measure of evaluating human forensic-attribution mechanism is like the traditional security measure of identification and authentication. The general measures used in the mouse-dynamics study is also considered in this study. These include the false acceptance rate (FAR), and false rejection rate (FRR) [4,5,20]. The FAR for each user is defined as the ratio of the falsely accepted inference by the attribution to the total number instances of all another user excluding the user under investigation, as parametrically depicted in the expression in (16).

$$FAR_i = \frac{Falsely\,accepted\,instance}{Falsely\,accepted\,Instance + Truely\,rejected\,instance} \tag{16}$$

If an instance $X \in session_x$, is known to belong to $User - X$, and instance $Y \in session_y$, is known to belong to $User - Y$, then the FAR, therefore, depicts the probability of attributing an instance of a known user to a wrong user or unknown user (example, instance Y to $session_x$). Conversely, the FRR is defined as the ratio of the falsely rejected inference by the attribution, to the total number instances of the genuine user. A parametric expression of the FRR is presented in (17).

$$FRR_i = \frac{Falsely\,Rejected\,instance}{Truely\,accepted\,Instance + Falsely\,rejected\,instance} \tag{17}$$

These FAR and FRR expressions provide a clearer granularity for evaluating performance efficiency. Whilst classical forensic statistics, In order to evaluate the attribution processes defined in Table 2, Table 3 and Figure 2, using the expression in (16) and (17), one Baseline and two off-the-shelf datasets are considered. Detail of these datasets is presented in the next section.

Dataset

A. Baseline Data

The Mouse dynamics data used in Ernsberger et al.[7] was adopted as a baseline for this study. This is based on the ease of access to both the tool and the data. The study developed is a Java-based mouse navigation and visualization tool that captures each mouse path of a user-action while surfing the Internet. Specifically, the tool was designed to capture mouse navigation in an HTTP-based website. The data was collected over a period of two weeks from 10-volunteers. One major limitation in mouse data collection process is the impact of hardware. To overcome this limitation, the tool was developed such that "*the coordinates, resolution and scrolled pixels are captured with clientX, clientY, clientWidth, clientHeight, and pageYOffset methods. These methods return the value in Cascading Style Sheets (CSS) pixels. A CSS pixel is a software pixel which forms the unit of measurement, whereas a hardware pixel is an individual dot of light on the screen. A CSS pixel can contain a few hardware pixels and is designed to be the same size across different devices. The CSS pixels defines uniform size irrespective of the hardware pixel resolution. A CSS pixel is a software pixel which forms the unit of measurement, whereas a hardware pixel is an individual dot of light on the screen*"[7].

B. Benchmark Mouse Data

Two existing benchmark mouse datasets are considered in this section. The benchmark data include the Ahmed and Traore[13,19] (hereinafter referred to as Benchmark-1), and the Balabit mouse dynamics challenge dataset[21] (hereinafter referred to as Benchmark-2). The Benchmark-1 dataset comprises class-imbalance data collected from 48-respondents in two tranches (2004 and 2008). The highest, lowest and average class instance of the Benchmark-1 are 75895, 895 and 15931-instances respectively. Given that the Benchmark-

1 comprises the class-wide margin of class-imbalance structure, further selection and instance-resampling were carried out on the data. For the class selection process, the expression $\forall_{class}, \sigma_{class} \geq 10\%(highest\ class\ instance)$ is utilized. It states that for all classes in the dataset, select the class whose instance is greater than or equal to 10% of the highest-class instance in the dataset. Percentage split of 90:10 was implemented on the selected class of Benchmark-1 for training and testing dataset respectively. The Benchmark-2 comprises mouse action data collection from 10-respondents during a mouse challenge competition. It contains multiple mouse action session, and each session contains a short duration of user action. Instances from two users in the experimental data were significantly lower than other users in the dataset. Summary of the datasets used in this section is presented in Table 7.

Table 7: Description of the dataset used in the study

Dataset Source	Statistics			
	Number of users	Highest class instance	Lowest class instance	Total instance
Benchmark-1	25	68339	6438	625257
Benchmark-2	10	14047	4864	80700
Baseline dataset	9	948	148	3156

Experimental Results and Analysis

To evaluate the effectiveness of the developed mouse path modeling and the method of user attribution, logical relation and experimental process was carried out. This section discusses the outcome of the result. First, the logical sequence of the mouse path model is presented. This is then followed by the experimental process of validating the attribution model. Furthermore, the attribute model is applied to datasets from two existing studies. The result of the probability of extracting unique individual behavioral signature (as observed on both the existing and experimental dataset) is presented and discussed.

Mouse path model: Based on the notation of the NFA, a representative navigation model of mouse-path is presented in Figure 3. The composition of the NFA comprises all likely primary mouse actions. Primary mouse action is defined in this sense to refer to a mouse event, which occurs as an output to direct human action during mouse navigation.

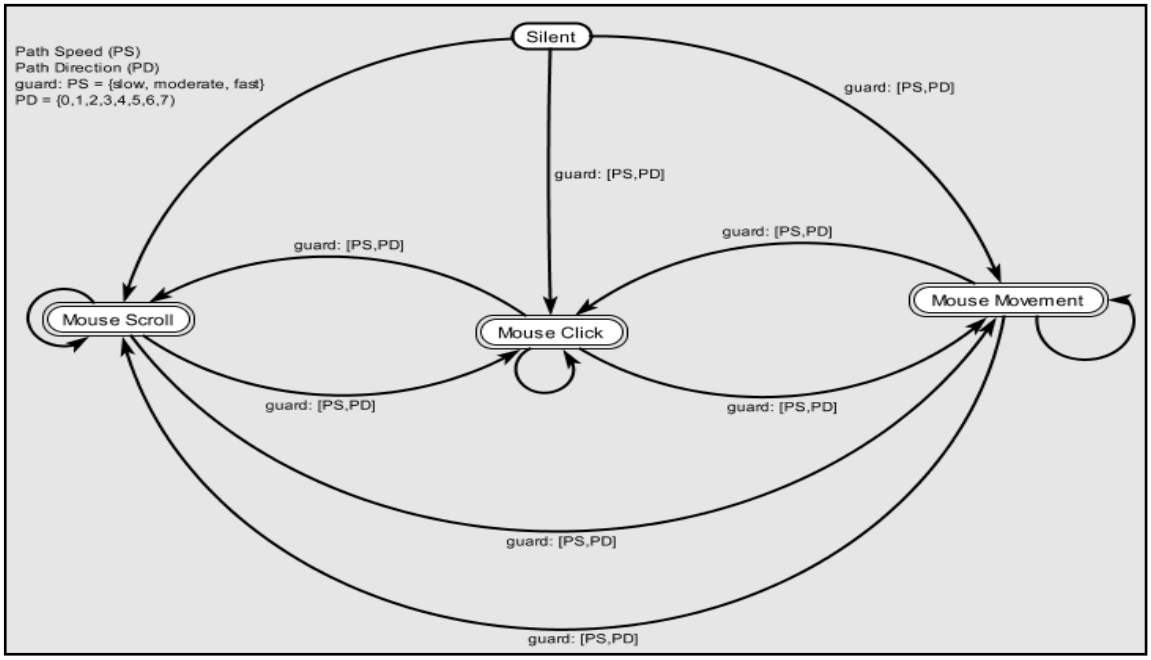

Fig.3: Non-Finite Automata model for Mouse-path

These actions include mouse movement-event, click-event, and scroll-event. Each event is considered as both a likely start state and an end-state of mouse action, as depicted with a double circle in Figure 3. A mouse-path sequence can start and end with each primary event. A mouse-path trigger from silent to any mouse action is observed when there is an input of mouse path direction and/or mouse-path speed.

Table 8: Example of Mouse-path from the initial state through a transition model.

Initial State	**Transition function**	**Probable Next state**	**Mouse-path**
$\varphi \to c$ $c \to c$	$\{D[0,1,2,3,4,5,6,7],$ $S[s, m, f]\}$	$\forall\{c, s, m, \varphi\} \in \mathrm{Q},$ $\exists![c,s,m,\varphi]: P(\mathrm{c, s, m}, \varphi)$	$\exists!\{\varphi c, \{D,S\}, c[\mathrm{D,S}]\},$ $\{\varphi c, \{D,S\}, s[\mathrm{D,S}]\},$ $\{\varphi c, \{D,S\}, m[\mathrm{D,S}]\},$ $\{\varphi c, [\varphi]\}$
$\varphi \to m$ $m \to m$			$\exists!\{\varphi m, \{D,S\}, m[\mathrm{D,S}]\},$ $\{\varphi m, \{D,S\}, s[\mathrm{D,S}]\},$ $\{\varphi m, \{D,S\}, c[\mathrm{D,S}]\},$ $\{\varphi m, [\varphi]\}$
$\varphi \to s$ $s \to s$			$\exists!\{\varphi s, \{D,S\}, s[\mathrm{D,S}]\},$ $\{\varphi s, \{D,S\}, c[\mathrm{D,S}]\},$ $\{\varphi s, \{D,S\}, m[\mathrm{D,S}]\},$ $\{\varphi s, [\varphi]\}$
$\varphi \to \varphi$	ϕ	$\forall\{c, s, m, \varphi\} \in \mathrm{Q},$ $\exists![c,s,m,\varphi]: P(\varphi) \equiv \phi$	$\exists \varphi, [\varphi]$
$m \leftrightarrow c$	$\{D[0,1,2,3,4,5,6,7],$ $S[s, m, f]\}$	$\forall\{c, s, m, \varphi\} \in \mathrm{Q},$ $\exists![c,s,m,\varphi]: P(\mathrm{c, s, m}, \varphi)$	$\exists!\{mc, \{D,S\}, m[\mathrm{D,S}]\}$ $\{mc, \{D,S\}, s[\mathrm{D,S}]\}$ $\{mc, \{D,S\}, c[\mathrm{D,S}]\}$ $\{mc, [\varphi]\}$
$s \leftrightarrow c$	$\{D[0,1,2,3,4,5,6,7],$ $S[s, m, f]\}$	$\forall\{c, s, m, \varphi\} \in \mathrm{Q},$ $\exists![c,s,m,\varphi]: P(\mathrm{c, s, m}, \varphi)$	$\exists!\{sc, \{D,S\}, s[D,S]\},$ $\{sc, \{D,S\}, c[D,S]\},$ $\{sc, \{D,S\}, m[D,S]\},$ $\{sc, [\varphi]\}$
$m \leftrightarrow s$	$\{D[0,1,2,3,4,5,6,7],$ $S[s, m, f]\}$	$\forall\{c, s, m, \varphi\} \in \mathrm{Q},$ $\exists![c,s,m,\varphi]: P(\mathrm{c, s, m}, \varphi)$	$\exists!\{ms, \{D,S\}, m[\mathrm{D,S}]\},$ $\{ms, \{D,S\}, s[\mathrm{D,S}]\},$ $\{ms, \{D,S\}, c[\mathrm{D,S}]\},$ $\{ms, [\varphi]\}$

It was observed that the first path of a mouse action begins with silent, and the last path of mouse action ends with a silent. This is logical, as human mouse action typically starts after a period of long inaction. The transition function $\{D[0,1,2,3,4,5,6,7], S[s,m,f]\}$ simply implies for a given mouse action, a user must navigate through any of the 8-sectors of the direction, and through any of the speed classes. The speed classes cover both scroll speed and mouse navigation speed. This could be in the slow, moderate or fast class. The input of this tuple in the sequence of the input generates a probable next state as parameterized in (3). This outcome is presented in the third column of Table 5. It can be explained thus; for any state which belongs to the set of all probable states $\forall\{c,s,m,\varphi\}\in Q$, there exists a unique next state which belongs to any of the probable states $\exists![c,s,m,\varphi]: P(c,s,m,\varphi)$. Consequentially, the effect of the transition function applied to the initial state which resulted in a next state resulted in the mouse-path, defined in the fourth column of Table 8. The mouse-path can be explained as follows; consider an initial state from $\varphi \to c$, for example, there exists a unique mouse-path whose initial state, transition function, and next state is a sequence of $silent \to click$, which resulted in either a $click, scroll, movement, or\ silent$.

A. **Syntactic Attribution**:

Algorithm 1 is applied to the Benchmark-2 to explore the probability of unique fingerprint among the classes. The benchmark-2 dataset was considered based on the smaller sample size and the relatively larger instance size of the dataset. The result, as shown in Table 9, suggests that the proposed hash-based signature is not capable of extracting discriminative properties for each user. Suggestively, the observed FRR for User-35, which are lesser than 0.05 in all the evaluation metrics, could be attributed to chance. This supposition is further supported by the FAR of all observed users.

Table 9: Result of the evaluation of Algorithm 1 on Benchmark-2

User	Evaluation									
	Duration		Movement		Flight		Path length		Path weight	
	FAR	FRR	FAR	FRR	FAR	FRR	FAR	FRR	FAR	FRR
User-7	0.031	0.354	0.500	0.500	0.725	0.000	0.363	0.031	0.359	0.031
User-9	0.479	1.000	0.500	0.333	0.718	0.059	0.460	0.659	0.449	0.659
User-12	0.435	0.684	0.458	0.000	0.723	0.051	0.441	0.684	0.438	0.684
User-15	0.458	0.810	-	-	0.717	0.008	0.475	0.810	0.466	0.810
User16	0.369	0.115	0.458	0.667	0.726	0.000	0.395	0.115	0.382	0.115
User-20	0.494	0.991	0.458	0.000	0.833	0.880	0.503	0.991	0.494	0.991
User-23	0.487	0.928	0.564	1.000	0.754	0.206	0.499	0.928	0.487	0.928
User-29	0.457	0.667	0.563	1.000	0.721	0.014	0.467	0.667	0.457	0.667
User-35	0.379	0.044	0.440	0.000	0.722	0.044	0.389	0.029	0.379	0.029

The result shown in Table 9 reveals that the syntactic composition of mouse dynamics has a very low probability of discriminating individuals. In essence, the logic of defining hash-based properties for a given user should integrate semantic characteristics of the mouse action. A semantic characteristic attempt to extract behavior based on the interactive properties of the mouse action. The result of the semantic process is presented in the next section.

B. **Semantic Attribution**:

To evaluate the semantic-based mouse dynamics algorithm defined in Table 6, the Benchmark-1, Benchmark-2, and the experimental dataset are utilized. Sequel to the exploration of applicable classifiers, three rule-based classifiers: C4.5 decision tree, Random Forest, Random Tree, and partial decision tree (PART) algorithms were adopted in this section. The three algorithms were applied to each of the datasets using a seed-size of 10, and 10-fold-cross-validation. The seed-size is the random number used to select the attribute for each run. For the C4.5 algorithm, a confidence interval of 0.95 was used for decision tree pruning to prevent model-overfitting. Following Steps, I and II of the semantic rule-based attribution

processes presented in Table 6, the characteristics of the AUC for each data is shown in Figure 4, 5, and 6. From the exploratory result of the AUC, only result from the Benchmark-2 data partially met the inclusion criteria for the application of the proposed attribution process. Statistically, the Benchmark-2 comprises average instance size, class size, and minimum class size of 4866, 4839 and 4867 respectively. This process also corresponds to the developmental phase of Figure 2. The Stage-III of the developed algorithm was then carried out on the rules generated by the Random Forest classifier. This corresponds to the validation process of Figure 2. The overall outcome of the result is presented in Table 10. A major challenge encountered during the application of the semantic algorithm was the extraction of a reliable signal to noise ratio accuracy. An iterative process was employed to measure the appropriate threshold for noise filtration, such that common rules among multiple users are minimized, while unique rules in individual output, are maximized. It was observed that the combination minimum of 10 rule weight, which has >99% uniqueness and the minimum number of attributes >4, achieved an optimal signal to noise ratio.

Table 10: Result of the proposed Attribution Process on Benchmark-2

Users	Dataset Instance		Output of development data on Random Forest algorithm		Validation data Output			
					Random Forest		Proposed Attribution	
	Development	Validation	AUC	Accuracy	FAR	FRR	FRR	FAR
User-7	4867	261	0.968	0.782	1	1	0.998	0.981
User-29	4865	261	0.869	0.524	0.948	0.916	0.989	0.941
User-35	4866	263	0.903	0.652	0.912	0.723	0.952	0.919
User-23	4866	261	0.780	0.293	0.800	0.854	0.980	0.834
User-21	4839	264	0.829	0.396	0.895	0.864	0.992	0.940
User-15	4866	263	0.842	0.435	0.852	0.825	0.967	0.893
User-16	4866	267	0.873	0.496	0.968	0.996	0.999	0.878
User-20	4864	264	0.885	0.481	0.944	0.981	0.994	0.926
User-9	4866	261	0.966	0.748	0.893	0.942	0.998	0.700
User-12	4867	262	0.817	0.381	0.925	0.943	0.997	0.874

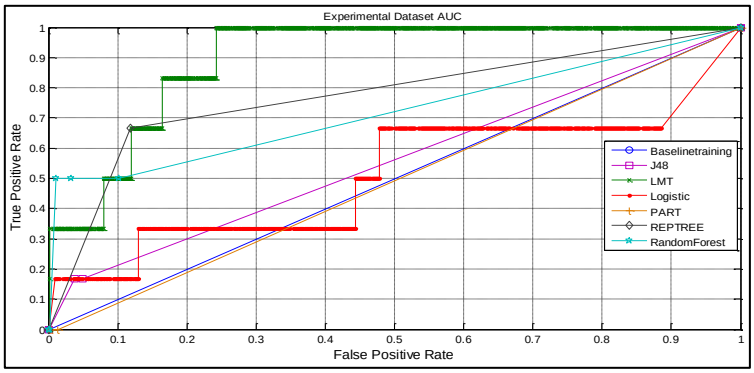

Fig.4: Area under the ROC Curve for Baseline Data

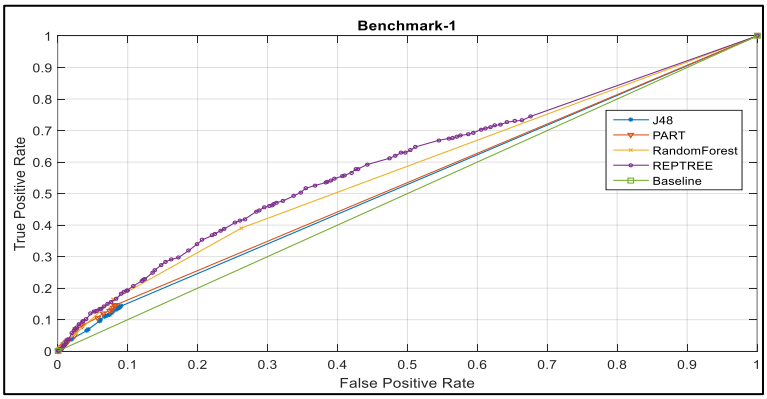

Fig.5: Area under the ROC Curve for Benchmark-1

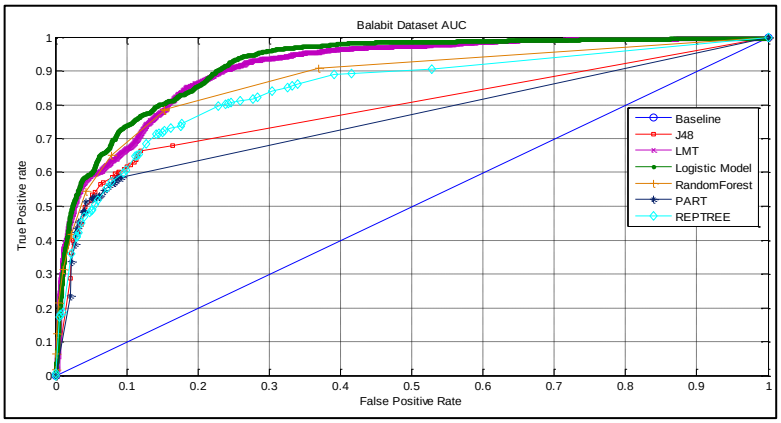

Fig.6: Area under the ROC Curve for Benchmark-2

The result from Table 10 clearly shows the partial fulfillment of the Benchmark-2 to the forensic inclusion rule defined in Table 3. All classes satisfied the $\geq 0.8\,AUC$ criterion. However, no class satisfied the $\geq 80\%Accuracy$ criterion. An exhaustive exploration of the classification of other classifiers also shows similar accuracy result as none satisfied the accuracy threshold of the forensic inclusion criteria. Hence, the obtained result from the proposed approach, as shown in Table 10, exhibit poor performance on the validation instance in similitude with the performance of existing approach (Random Forest algorithm in this instance). To ascertain the effectiveness of the proposed attribution process, behavioral data from an existing study was considered[18]. The data comprises web request characteristics of 11-users. Accordingly, a total of fifteen features were defined and extracted in the study. Summary of the distribution of the training/development and testing/validation data is shown in Table 11. Whilst the Random Forest algorithm produced higher accuracy among the explored classifiers in the result presented in Table 10, the C4,5 decision tree algorithm was observed to generate higher accuracy. This result is also in tandem with the obtained result in Adeyemi et. al.[18]. Only the C4.5 algorithm is leveraged in the proposed approach, as illustrated in Figure 2. The result exhibits a relatively better FAR and FRR result than the base algorithm. However, upon further computation, the proposed approach demonstrated a promising mechanism for classification accuracy. Relative to the result obtained in Table 10, the result shown in Table 11 shows that the proposed mechanism is capable of improving the efficiency of the classification process.

Table 11: Comparative analysis of the proposed attribution mechanism

Users	**Dataset Instance**		**Output of development data on C4.5 algorithm**				**Validation data Output**			
							C4.5 Algorithm		**Proposed Attribution**	
	Development	**Validation**	**AUC**	**Accuracy**	**FAR**	**FRR**	**FAR**	**FRR**	**FAR**	**FAR**
User-1	466	211	0.992	0.979	0.044	0.021	0.0377	0.033	0.043	0.043
User-2	365	129	0.883	0.556	0.450	0.444	0.589	0.480	0.581	0.760
User-3	614	293	0.989	0.974	0.015	0.026	0.007	0.003	0.007	0.031
User-4	263	111	0.973	0.897	0.092	0.103	0.117	0.117	0.111	0.207
User-5	363	163	0.895	0.625	0.407	0.375	0.465	0.528	0.363	0.641
User-6	313	154	0.932	0.764	0.201	0.236	0.204	0.240	0.162	0.396
User-7	437	193	0.953	0.863	0.180	0.137	0.222	0.166	0.188	0.285
User-8	652	220	0.988	0.965	0.026	0.035	0.031	0.009	0.014	0.032
User-9	312	124	0.954	0.869	0.117	0.131	0.150	0.129	0.111	0.220

User-10	403	177	0.966	0.928	0.048	0.072	0.025	0.119	0.019	0.136
User-11	224	112	0.882	0.683	0.275	0.317	0.227	0.330	0.075	0.262

Discussion

Attribution concept (and context) is a major source of concern among security and forensic stakeholders[22]. Conceptually, attribution relates to the technical expertise and mechanism to append causation (event or action) to a known effect (actor, recipient, or source). Two fundamental assertions on attribution context are found in the literature. Whilst one logic posit that attribution can be limited to the system (system attribution)[23], others assert that attribution should be delimited to the actual individual behind a given system (human attribution)[24]. Both assertions, however, conveys the logic that attribution is a technical process that should eventually lead to the apprehension of a target. Human identification through biometric modality has been defined as one primary source that can be used to identify the actual human behind the system[22,25]. Such biometrics include physiological and behavioral. A physiological biometric modality, such as facial, iris, vein, fingerprint, palm print, and DNA, requires additional physical constraints. This limits the applicability of physiological biometric modalities in user attribution in digital forensics. Behavioral biometrics-based attribution technique presents a potential mechanism for identifying the actual individual behind a given digital system. Example of such behavioral modality includes keystroke[26], web-click patter[27], web-browsing pattern, and mouse dynamics. Research on mouse dynamics owes its inception to the research conducted by Everitt and colleague[28]. This has generated studies, majorly, on active authentication[5,12,29] and continuous authentication[30,31]. More recently, it has been extended to digital forensic readiness processes[7]. A synopsis of studies on mouse dynamics is presented in Table 12.

Table 12: Summary of existing studies in Behavioral Biometrics

Study	**Dataset composition**	**Classification technique**	**Metric of evaluation**				
			Accuracy	**EER**	**FRR**	**FAR**	**AUC**
Shen et al.[5]	37-users, 5550-instances	One-class SVM, Kernel PCA	-	HTER: 8.35	7.96%	8.74%	-
Ahmed & Traore[13]	22-Users, 998-sessions	Neural Network	-	-	2.461%	2.465%	-
Traore et al.[19]	12-Users, 447625-instances	Bayesian Network and fusion	-	22.41%	-	-	-
Jorgensen & Yu [4]	17-Users, 30-minutes session	Neural network, and Logistic regression	96.7% & 97.8%	-	-	-	-
Gamboa et al[32]	50-Users, 10-minutes per user	Sequential Forward Selection algorithm	-	-	6.2%	-	-
Feher et al.[33]	25-users, 100-actions, and 30-actions	A hierarchical feature of mouse action verification method		7.5%, 8.53%			
Mondal & Bours[34]	25-users	SVM, DT, ANN, Counter-propagation ANN, pairwise user coupling	61.3%				
Ernsberger et al.[7]	3-Users	LibSVM, ANN, C4.5 DT, Random Forest	78.1%			0.21	
Current Study	10-Users	C4.5 DT, Random Forest, and Heuristic algorithm	<80%		>0.01	>0.001	>0.8

As asserted[34], behavioral data is characterized by high intra-class variation, a stochastic property which defines most behavioral biometric modalities. Mondal and Bours[34] observed a very poor performance of statistical analysis, support vector machine (SVM), artificial neural network (ANN), and counter-propagation ANN on such type of data.

Human attribution based on behavioral biometrics is relatively a new scope of research in digital forensics. Behavioral biometrics such as keystroke dynamics, mouse dynamics, gait biometrics, stylometrics, and surfing navigation style, are common biometrics used in existing studies in information security. Research outcome in [35] attempted to define a forensic readiness modality for the implementation of behavioral biometrics. However, the outcome of the study is limited to the framework and recommendation without a quantitative outcome to support assertions. The proposition from the current research posits that there is a high likelihood of attributing behavioral biometrics data to an individual. One major limitation observed in this study is the collection and identification of informative dataset capable of inter-person discrimination. Furthermore, the noise constraint in the data remains a major source of research challenges. One consideration for this challenge is the development of significant instance identification mechanism, as an alternative to the generic significant feature identification. Approaches such as the component analysis, SVM attribute evaluation, information gain, gain ratio, and significant attribute evaluation are limited to a higher abstraction of instance identification. The exploratory process carried out in this study support the assertion that these attribute selection approaches are vulnerable to noise. A noise in this instance refers to mouse behavioral actions which are random and does not constitute any consistent pattern. Such actions are carried out by users during normal mouse action and do not require any specific operational innuendo. This is typical of behavioral data, as some human actions are a result of random reflects and context specific. Therefore, considering the high signal to noise ratio in a typical mouse behavioral biometric dataset, approaches which leverage signal to noise ratio filtration could provide a better technique for significant instance identification.

Another probable consideration for this limitation is the implementation of a digital forensic readiness process in an organization over a given duration. Such a longitudinal study can be used to generate informative data with minimized noise. Noise reduction was not considered in the developed heuristics algorithm. A mechanism to address inaccuracy due to noise in data remains an open research challenge in most behavioral biometric data. The underlying question attempted by this study is contingent on the readiness of the usage of digital behavioral biometrics. Whilst the obtained results can be used to complement security apparatus and mechanisms, it, however, falls short of the forensic requirement for litigation. This, therefore, implies that the integration of behavioral biometrics in digital forensics is currently limited to the strengthening of security. Furthermore, a visualization of the mouse path using the non-finite automata model expressed in Table 8 presents a logical examination of path characteristics. Such characteristic behavior can be used to aid behavioral evaluation; a security mechanism that can be leveraged during incident analysis. This can be achieved through the use of a digital forensic readiness approach. A digital forensic readiness approach is a proactive mechanism implemented to collect, preserve, preprocess and store potential information which could provide corroborative insight during incident analysis. In addition, a digital forensic readiness mechanism is also suitable for 1:N identity matching process. This contrast the typical 1:1 identity matching approach such as the use of psychological profiling approach. However, given that the psychological approach of human behavioral components is baselined to a specific locus, a digital forensic readiness approach could therefore, provides a comprehensive and robust locus of mouse dynamics. In addition, the psychological approach is limited to deviation from a context-based locus of observation which therefore limits its application to areas such as lie detection where expected and unexpected behavioral sequence can be analyzed. Such an approach does not present the loci for the analysis of human behavioral navigation, particularly, in a free-browsing setting that is synonymous to real human navigation behavior.

Future works

As part of ongoing research work, the digital forensic readiness approach will be considered as metrics for data collection and data preprocessing for data de-noising. Furthermore, an extensive study will be carried out on other behavioral biometrics. This will include the development of attribution process for keystroke dynamics, hand gesture, and the integration of these behavioral modalities to a more adaptive multimodal attribution process for digital forensic processes. The integration of physiological demographics such as handedness and gender will constitute other aspects of future works. A further area of improvement would involve the exploration of other reliable fingerprinting techniques, such as the short tandem repeat (STR) sequencing technique, motifs and common subsequence analysis from behavioral features, and other similar fingerprinting process for user attribution. Such techniques, if successful, could present a potent scientific and statistical premise for a reliable result that can provide a consistent behavioral fingerprint measure of a known user. Furthermore, studies on the probability of integrating behavioral fingerprint into digital forensic readiness models could be considered. Models on how to effectively and proactively harness the potential of the behavioral fingerprint could be developed and implemented. Such model would conform to the standard forensic procedure for evidence acquisition, collection, preservation, and analysis, with the hope of generating a result that has statistically acceptable, and technically sufficient evidential value in a legal proceeding. One area that was not considered in this study is the potential impact of the sample size of the data, and the effect of the number of instances required for attribution. This could be part of future study. A study that will attempt to understand the impact, if any, of the sample size of the data on the attribution process.

Conclusion

User attribution process through behavioral fingerprinting is a promising methodology for digital forensics, which requires intense scrutiny of technical protocol, result analysis, and factual interpretation. However, human behavior is a stochastic characteristic behavior with a low probability of a unique pattern. Though a classical machine learning approach of model development and validation have shown significant improvement over the years, the approach used in this study has not provided a substantially reliable accuracy applicable for forensic purpose. This is in part, due to the inherent noise of a stochastic process. Therefore, to address such stochastic behavior with classical machine learning process without recourse to the inherent noise could generate higher false results. This study developed an attribution model based on heuristics built on classical machine learning algorithms, with emphasis on the probability of extracting behavioral consistency from the stochastic data of human behavior. Digital investigation, specifically in computer and network forensic stand to gain much support from such behavioral attribution process. The result presented in this study leverages human kinematics in the form of mouse behavior, to evaluate the preponderance of mouse dynamics for user attribution. Behavioral fingerprinting in this regard shows a promising mechanism for the identification of a human, which is beyond the classical username and password-based timeline profile. Therefore, this study has successfully demonstrated, in tandem with existing studies, the viability of the integration of mouse dynamics for a reliable user identification process. Consequently, mouse-based behavioral fingerprint can provide a strong scientific value to mouse dynamics evidence which can be used to further strengthens the evidential weight in a digital investigation

References

1. Ahmed, A. A. E., Traoré, I. & Almulhem, A. Digital Fingerprinting Based on Keystroke Dynamics. in *Second International Symposium on Human Aspects of Information Security & Assurance (HAISA 2008) Digital* (2008).
2. Olivier, M. S. Digital Forensic Science: A Manifesto. *South African Comput. J.* **28**, 46–59 (2016).

3. Insa, F. The Admissibility of Electronic Evidence in Court (A.E.E.C.): Fighting against High-Tech Crime—Results of a European Study. *J. Digit. Forensic Pract.* **1**, 285–289 (2007).

4. Jorgensen, Z. & Yu, T. On mouse dynamics as a behavioral biometric for authentication. in *Proceedings of the 6th ACM Symposium on Information, Computer and Communications Security - ASIACCS '11* 476–482 (2011). doi:10.1145/1966913.1966983

5. Shen, C., Cai, Z., Guan, X., Du, Y. & Maxion, R. A. User authentication through mouse dynamics. *IEEE Trans. Inf. Forensics Secur.* **8**, 16–30 (2013).

6. Weiss, A., Ramapanicker, A., Shah, P., Noble, S. & Immohr, L. Mouse movements biometric identification: A feasibility study mouse movement biometric system. in *Proceedings of StudentFaculty Research Day CSIS Pace University* 1–8 (2007).

7. Ernsberger, D., Ikuesan, A. R., Venter, H. S. & Zugenmaier, A. A Web-Based Mouse Dynamics Visualization Tool for User Attribution in Digital Forensic Readiness. in *9th EAI International Conference on Digital Forensics & Cyber Crime* 1–13 (Springer Berlin Heidelberg, 2017).

8. Ayers, D. A second generation computer forensic analysis system. *Digit. Investig.* **6**, S34–S42 (2009).

9. Technology, P. C. of A. onScience &. Forensic Science in Criminal Courts: Ensuring Scientific Validity of Feature-Comparison Methods, Report to the President, Executive Office of the President. 1–134 (2016).

10. Gamboa, H. & Fred, A. L. N. An Identity Authentication System Based On Human Computer Interaction Behaviour. *Proc. Int. Work. Pattern Recognit. Inf. Syst.* 46–55 (2003).

11. Bours, P. Continuous keystroke dynamics: A different perspective towards biometric evaluation. *Inf. Secur. Tech. Rep.* **17**, 36–43 (2012).

12. Bailey, K. O., Okolica, J. S. & Peterson, G. L. User identification and authentication using multi-modal behavioral biometrics. *Comput. Secur.* **43**, 77–89 (2014).

13. Ahmed, A. A. E. & Traore, I. A New Biometric Technology Based on Mouse Dynamics. *IEEE Trans. Dependable Secur. Comput.* **4**, 165–179 (2007).

14. Dargham, J. & Al-Nasrawi, S. FSM behavioral modeling approach for hypermedia web applications: FBM-HWA approach. *Proc. Adv. Int. Conf. Telecommun. Int. Conf. Internet Web Appl. Serv. AICT/ICIW'06* **2006**, 199 (2006).

15. James, J., Gladyshev, P., Abdullah, M. T. & Zhu, Y. Analysis of Evidence Using Formal Event Reconstruction. in *1st International Conference on Digital Forensics and Cyber Crime* (ed. S. Goel) 85–98 (Springer International Publishing, 2010). doi:10.1007/978-3-642-11534-9_9

16. Wang, Q. & Guo, J. L. Human dynamics scaling characteristics for aerial inbound logistics operation. *Phys. A Stat. Mech. its Appl.* **389**, 2127–2133 (2010).

17. Fan, C., Guo, J.-L. & Zha, Y.-L. Fractal Analysis on Human Behaviors Dynamics. *arXiv Prepr. arXiv1012.4088* 22 (2010).

18. Adeyemi, I. R., Razak, S. A., Salleh, M. & Venter, H. S. Observing consistency in online communication patterns for user re-identification. *PLoS One* **11**, 1–27 (2016).

19. Traore, I., Woungang, I., Obaidat, M. S., Nakkabi, Y. & Lai, I. Combining mouse and keystroke dynamics biometrics for risk-based authentication in web environments. *Proc. - 4th Int. Conf. Digit. Home, ICDH 2012* 138–145 (2012). doi:10.1109/ICDH.2012.59

20. Anjomshoa, F., Aloqaily, M., Kantarci, B., Erol-Kantarci, M. & Schuckers, S. Social Behaviometrics for Personalized Devices in the Internet of Things Era. *IEEE Access* **3536**, 1–1 (2017).

21. Fülöp, Á., Kovács, L., Kurics, T., Windhager-Pokol, E. Balabit Mouse Dynamics Challenge data set. 1 (2016).

22. Cohen, F. Attribution. in *IFIP Advances in Information and Communication Technology* **337 AICT**, 321–442 (2013).

23. Hauger, W. & Olivier, M. Determining Trigger Involvement during Forensic Attribution in Databases. in *11th IFIP International Conference on Digital Forensics (DF). ADVANCES IN DIGITAL FORENSICS XI* (ed. Shenoi., G. P. S.) 163–177 (IFIP Advances in Infor- mation and Communication Technology, AICT-462, 2015).

24. Boebert, W. A Survey of Challenges in Attribution. in *Proceedings of a workshop on Deterring CyberAttacks* 41–52 (2010).

25. Cohen, F. B. Attribution of messages to sources in digital forensics cases. *Proc. Annu. Hawaii Int. Conf. Syst. Sci.* 1–10 (2010). doi:10.1109/HICSS.2010.75

26. Shukla, P. & Solanki, R. Web Based Keystroke Dynamics Application for Identifying Emotional State. *Ijarcce.Com* **2**, 4489–4493 (2013).

27. Padmanabhan, B. & Yang, Y. Clickprints on the web: Are there signatures in web browsing data? *Available SSRN http//ssrn.com/abstract=931057 or http//dx.doi.org/10.2139/ssrn.931057* 21–22 (2007). doi:10.2139/ssrn.931057

28. Everitt, R. A. J. & McOwan, P. W. Java-based internet biometric authentication system. *IEEE Trans. Pattern Anal. Mach. Intell.* **25**, 1166–1172 (2003).

29. Mondal, S. & Bours, P. Combining keystroke and mouse dynamics for continuous user authentication and identification. in *ISBA 2016 - IEEE International Conference on Identity, Security and Behavior Analysis* **6**, 55–58 (2016).

30. Mondal, S. & Bours, P. Neurocomputing A study on continuous authentication using a combination of keystroke and mouse biometrics. **230**, 1–22 (2017).

31. Stanic, M. Continuous User Verification Based on Behavioral Biometrics Using Mouse Dynamics. in *Proceedings of the ITI 2013 35th International Conference on Information Technology Interfaces* 251–256 (SRCE University Computing Centre, University of Zagreb, 2013). doi:10.2498/iti.2013.0505

32. Gamboa, H., Fred, A. L. N. & Jain, A. K. Webbiometrics: User verification via web interaction. in *2007 Biometrics Symposium, BSYM* (2007). doi:10.1109/BCC.2007.4430552

33. Feher, C., Elovici, Y., Moskovitch, R., Rokach, L. & Schclar, A. User identity verification

via mouse dynamics. *Inf. Sci. (Ny).* **201**, 19–36 (2012).

34. Mondal, S. & Bours, P. Combining keystroke and mouse dynamics for continuous user authentication and identification. in *ISBA 2016 - IEEE International Conference on Identity, Security and Behavior Analysis* 1–8 (2016). doi:10.1109/ISBA.2016.7477228

35. Adeyemi, I. R. & Venter, H. S. Digital Forensic Readiness Framework Based on Behavioral-Biometrics for User Attribution. in *IEEE Conference on Applications, Information and Network Security (AINS)* 54–59 (IEEE Comput. Soc, 2017).