IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 15 May 2024, accepted 29 July 2024, date of publication 7 August 2024, date of current version 20 August 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3440187

RESEARCH ARTICLE

Enhancing User Authentication: An Approach Utilizing Context-Based Fingerprinting With Random Forest Algorithm

AKRAM AL-RUMAIM AND JYOTI D. PAWAR
Goa Business School, Goa University, Taleigao, Goa 403206, India
Corresponding author: Akram Al-Rumaim (akramalrumaim@gmail.com)

ABSTRACT In the evolving world of Cyber Attacks, this research presents an innovative approach aimed at fortifying user authentication within the Application Programming Interfaces (APIs) Ecosystem. We employ a groundbreaking synthesis of Context-Based fingerprinting attributes alongside the Random Forest algorithm, assessing their efficacy in enhancing security in how the user is authenticated. The study emphasizes the model's potential to advance the precision of identifying legitimate login attempts, positioning it as a superior alternative to conventional methods. A model evaluation investigates the capacity of the Random Forest algorithm, augmented with these attributes, to discern the authenticity of user login attempts. Results unveil an exceptionally accurate model with 99.5 % accuracy, showcasing elevated F1 scores, precision, recall, and a notable MCC. The paper's insights underscore the Random Forest algorithm's potential as a robust tool for user authentication using the user's historical profile, significantly contributing to the domain of Cyber Security. As the digital landscape continues to evolve, this research endeavours to provide a pioneering solution, ensuring robust API security and endorsing the broader adoption of this groundbreaking model.

INDEX TERMS API ecosystem security, API security, context-based fingerprinting, cyber security, random forest, user authentication.

I. INTRODUCTION

In this digital age of ubiquitous cyber attacks, user authentication has become the linchpin of security, acting as the initial guardian of the gates to digital fortresses. It is the paramount access gateway, ensuring that only authorized users traverse the digital landscape [1], [2]. As we journey through the digital era, where data reigns supreme, and information is the currency, the importance of user authentication cannot be overemphasized. The security process commences at the very first interaction point: user authentication. Be it a username and password, biometric scan, or multi-factor authentication, this initial validation point is crucial in safeguarding the digital space [3]. It is the first line of defence against potential security threats, ensuring a protected and secure user experience. The ubiquitous login screen symbolizes the essence of this process, where the user interface and security meet [4]. While the fundamental importance of user authentication is paramount, its significance extends well beyond the individual user. In the broader realm of APIs, this process underpins the foundational structure of digital interactions [5]. APIs serve as the conduit connecting disparate applications, services, and users. Within this intricate web of data exchanges and interconnectivity, the strength and accuracy of user authentication play a pivotal role. As we embark on this research journey, we will explore the profound implications of user authentication, especially in the context of API-level security. With a focus on its essential role, we will delve into the intricate nuances and far-reaching impacts of this initial security checkpoint, acknowledging its place as the digital world's first gateway. User authentication has become increasingly

The associate editor coordinating the review of this manuscript and approving it for publication was Amjad Ali.

challenging in the face of evolving cyber threats [6]. Among these challenges, detecting anomalous login attempts stands as a paramount concern. The digital landscape is replete with malicious actors attempting to infiltrate systems, impersonate users, or gain unauthorized access. Traditional user authentication methods often struggle to differentiate between legitimate and suspicious login attempts. These issues underscore the need for innovative approaches to accurately discern potential security threats from authentic user interactions.

Compared to existing studies, this research provides significant contributions by integrating context-based fingerprinting with the Random Forest algorithm and incorporating effective attributes that significantly enhance user authentication mechanisms. This approach addresses attributes incorporating environmental and behavioural factors to improve model accuracy and robustness.

This research paper addresses the critical need for robust user authentication in the context of API-level security. It delves into the intricate relationship between user authentication and API security, emphasizing their intertwined significance. The overarching objective is to explore and evaluate advanced authentication methods that leverage Context-Based fingerprinting alongside the Random Forest algorithm. Through analysis, this research seeks to highlight the potential of these methods in significantly improving the accuracy of detecting anomalous login attempts.

This paper unfolds systematically to provide a nuanced exploration of the research landscape. Starting with this introduction, the groundwork is laid for subsequent sections. Section II scrutinizes the topic background through user authentication, delving into conventional methods. It also elaborates on adaptive authentication and its types, paving the way for the proposed model in Section IV. Section III reviews related work, contextualizing our research. Our model's architecture is detailed in Section IV, while Section V explains the methodology. Sections VI and VII present and discuss results, offering real-world insights. The concluding sections, VIII and IX, provide a discussion and future directions, highlighting the lasting contributions to Cyber Security and API Security.

II. BACKGROUND

A. USER AUTHENTICATION

User authentication is the foundational gateway to determine whether a user's login attempt is legitimate or potentially malicious [3]. It encompasses a spectrum of methods and mechanisms that verify and validate the identity of individuals seeking access to a system or platform. In the API domain, data is shared, and communication between different software components occurs, ensuring secure user authentication is paramount.

The relevance of user authentication to API-level security cannot be overstated. APIs enable the exchange of sensitive data and commands, making them a common target for cyber threats. If user authentication is not robust, unauthorized access and data breaches become distinct possibilities, jeopardizing system integrity and user privacy. Therefore, an effective user authentication system within APIs is the first line of defence in the battle against cyber threats.

- Knowledge-based authentication: This type requires the user to know something, such as a password, PIN, or passphrase.
- Property-based authentication: This type requires the user to possess something, such as a security token or smart card.
- Biometric authentication: This type of authentication uses a unique physical characteristic of the user, such as a fingerprint, facial scan, or voice print.

B. IMPORTANCE OF USER AUTHENTICATION

User authentication operates on multiple layers designed to address specific security aspects. These layers include something you know (e.g., passwords or PINs) [1], something you have (e.g., smart cards or tokens) [3], and something you are (e.g. biometric characteristics like fingerprints or facial recognition) [4]. These layers combine to form a multifaceted approach that enhances the security of user authentication [5]. The importance of these layers lies in their ability to create multiple barriers to potential threats. When an attacker attempts to breach a system, they must circumvent not just one but several authentication layers. This multi-layered approach enhances security, making it significantly more challenging for malicious actors to gain unauthorized access. These layers work harmoniously to verify the user's identity and only provide access to those meeting the established criteria.

C. THE DYNAMIC NATURE OF USER AUTHENTICATION

A dynamic user authentication system can continuously evaluate and reevaluate the legitimacy of login attempts. It adapts to various factors, including the user's historical context-based fingerprinting of the login attempt and real-time threat intelligence. The authentication system can identify anomalies and respond accordingly by incorporating dynamic elements. This dynamic approach is fundamental to ensuring API-level security, as it empowers systems to detect and mitigate emerging threats. In API security, dynamic user authentication can bolster defences and protect against known and unknown vulnerabilities, ultimately preserving the integrity and confidentiality of data and transactions.

D. ADAPTIVE AUTHENTICATION AND ITS TYPES

Adaptive authentication represents a progressive approach to user authentication that tailors security measures based on the unique characteristics and context of each login attempt [7]. In the context of API-level security, adaptive authentication is paramount to safeguarding sensitive data and system resources [8]. It recognizes that not all login attempts are equal; some may carry a higher risk of being malicious.

The relevance of adaptive authentication to API Ecosystem security stems from the need to identify and respond to anomalous login attempts effectively. Traditional static authentication methods are ill-suited to this task, often leading to false positives or overlooking subtle signs of malicious activity [9]. Adaptive authentication, in contrast, enhances security by continuously analyzing a variety of factors, such as user behaviour, device information, and real-time threat data, to make informed access decisions [10]. This dynamic and context-aware approach aligns perfectly with the nuanced challenges of securing APIs. Adaptive authentication encompasses several types, each designed to address specific security aspects. These types are characterized by their unique attributes and the factors they consider when evaluating login attempts:

- Behavioral Based Authentication: This type of adaptive authentication focuses on the user's behaviour and habits during the login process. It assesses patterns like typing speed, mouse movements, and navigation choices. Any deviations from established behavioural norms may trigger additional authentication steps.
- Context-Based Authentication: Context-based authentication leverages contextual information such as the user's Geo-location, IP address, and Device Type to assess the legitimacy of a login attempt. It evaluates whether the context aligns with the expected or typical environment of the user. The dynamic nature of adaptive authentication allows for a tailored approach that responds to evolving threats and provides robust security in an API environment.

III. RELATED WORK

Paper [11] addresses the challenges in authenticating a myriad of heterogeneous IoT devices in their respective trust domains. It critiques traditional methods like passwords and pre-defined keys and proposes context-based authentication, utilizing ambient physical properties of co-located devices. The study assesses such solutions' security and context quality requirements using real-world data from typical IoT environments. The key findings highlight the potential of contextual information as a shared secret and emphasize the need for robust security measures in IoT authentication. Methodologically, empirical data from IoT environments is analyzed to quantify achievable security. The results underscore the viability of context-based authentication, paving the way for more secure IoT ecosystems. However, limitations include potential vulnerabilities in context quality and the dynamic nature of IoT environments that could impact the effectiveness of context-based approaches.

The research [12] focuses on enhancing risk-based static authentication in web applications by combining behavioural biometrics and session context analytics. It critiques the fragility of password-based authentication and proposes a fusion of fingerprinting and behavioural dynamics using machine learning. The study evaluates the approach on a dataset containing mouse, keyboard, and session context information, showcasing a significant increase in accuracy compared to individual analyses. Key findings emphasize the efficacy of combining fingerprinting and behavioural dynamics for heightened security. Methodologically, machine learning techniques are employed for accuracy, and the results demonstrate a robust authentication mechanism. Limitations include potential false alarms that necessitate manual inspection and the need for continual evaluation against evolving cyber threats.

This paper [13] delves into the challenges of device identification and authentication in the Internet of Things (IoT). It introduces a methodology for IoT device behavioural fingerprinting, utilizing features extracted from network traffic to train a machine-learning model. The findings showcase high identification rates and accuracy, positioning behavioural fingerprinting as an effective solution. Methodologically, the study employs five-fold cross-validation to validate the approach, demonstrating robust results. Limitations include the need for ongoing refinement as IoT landscapes evolve and the potential impact of diverse network conditions on behavioural fingerprinting accuracy.

The research [14] addresses the escalating need for secure data access on the web by proposing a risk-based authentication system. The study integrates a risk engine with machine learning algorithms to examine users' past login records, generating a risk level that adapts authentication methods. Key findings emphasize the adaptability of risk-based authentication according to user risk profiles. Methodologically, machine learning algorithms are utilized for risk assessment and enhancing security. Limitations involve the challenge of accurately determining the risk level and potential biases in the machine learning models, requiring ongoing refinement for optimal performance in diverse scenarios.

IV. PROPOSED SOLUTION

This section introduces our innovative approach to user authentication within APIs. We leverage a Random Forest machine learning model incorporating context-based fingerprinting attributes. This combination of attributes sets the stage for a more accurate user authentication mechanism.

Context-based authentication considers the user's environment, including their geographical location, IP address, and device. Our model seamlessly integrates both attributes, creating a profile of the user's contextual attributes such as User ID, IP Address, Country, Region, City, ASN, Browser Name and Version, OS Name and Version, and Device Type.

The Random Forest model, known for its robustness and adaptability, becomes the linchpin of our user authentication system. It excels at handling complex, high-dimensional datasets and is well-suited to the nuances of API-level security. By leveraging this machine learning model, we can develop a precise and resilient adaptive authentication system. Our proposed model resonates with the principles of adaptive authentication. By blending context-based

authentication attributes, we build a system that responds dynamically to the characteristics of each login attempt. The model is trained to evaluate patterns and context, adapting its response based on evolving threats and user-specific fingerprints.

Adaptive authentication recognizes the limitations of traditional static authentication methods and the importance of accounting for contextual factors. It embodies a context-aware, risk-based approach to user authentication. Our model aligns with these principles by evaluating various factors, enabling it to make informed and dynamic access decisions. It embodies the essence of adaptive security measures by staying agile and evolving with the user's contextual attributes.

A. HYPERPARAMETER TUNING

To optimize the performance of the Random Forest model, we fine-tuned several key hyperparameters. The primary hyperparameters adjusted were the number of trees in the forest (n_estimators), the maximum depth of the trees (max_depth), the minimum number of samples required to split an internal node (min_samples_split), and the minimum number of samples needed to be at a leaf node (min_samples_leaf).

Number of Trees (n_estimators): The current setting is ten trees. Increasing the number of trees generally enhances performance, as more trees can capture a broader range of data patterns.

Maximal Number of Considered Features (max_features): Currently set at five features. Adjusting this value helps balance the model's complexity and performance.

Maximal Tree Depth (max_depth): The current setting is ten. While unlimited depth can lead to overfitting, setting a maximum depth, such as ten or twenty, helps control the model's complexity and prevents overfitting.

Stop Splitting Nodes with Maximum Instances (min_samples_split): Currently set at five instances. This value should be adjusted based on the dataset size. Smaller values may cause overfitting, while larger values may result in underfitting.

By carefully tuning these hyperparameters, we aim to balance model accuracy and generalizability, enhancing the overall performance of the Random Forest model.

B. THE MODEL SUITABILITY FOR API SECURITY AND ANOMALOUS LOGIN DETECTION

Our proposed model, rooted in the Random Forest machine learning algorithm, is particularly well-suited for enhancing API security. The versatility of the Random Forest model allows it to adapt to the complexities and evolving threats commonly encountered in API environments.

This adaptability is vital for detecting anomalous login attempts. It leverages context-based attributes to profile users uniquely. Any deviations from established patterns or out-of-context requests can trigger additional authentication steps, thus providing a robust defence against potentially malicious activities. By enhancing the precision and recall of detecting anomalous login attempts, our model raises the security bar for API systems.

Furthermore, our model not only excels at identification but also prediction. It can predict potential risks by evaluating historical login data and training on the full range of attributes. This predictive capability ensures that security measures are proactively adjusted to the user's evolving behaviour and the dynamic threat landscape of the API ecosystem. The fusion of context-based attributes within the Random Forest model creates a powerful and forward-thinking user authentication mechanism, exemplifying the future of API-level security.

Fig 1 illustrates the adaptive authentication system, encompassing the User, User Interface (UI), and a core Machine Learning Model. As the User initiates a login attempt, the UI forwards a request to the Adaptive Authentication System. Our machine learning model calculates the risk score by considering context-based parameters. Threshold evaluation categorizes risk, influencing Decision Logic to grant access, prompt verification, or deny entry. The system's Response to the UI communicates these decisions, ensuring a dynamic and secure authentication process tailored to perceived risk levels.

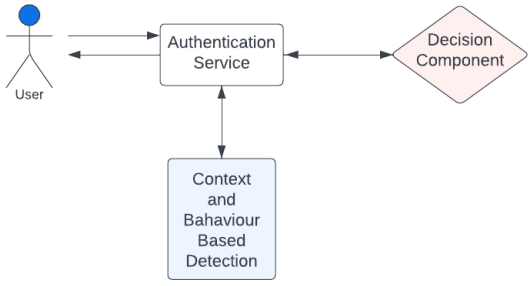

FIGURE 1. Model architecture.

Fig 2 shows the system's diagram unfolds the process of API security through adaptive authentication, orchestrated by the API Gateway. The journey initiates with a request for user data, wherein the API Gateway seeks information from the Authentication Service. Our model, nestled within the Authentication Service, commences its pivotal role at this juncture. Subsequently, the Authentication Service delves into stored historical data, and our model analyzes this information to comprehend the user's past login behaviour. Guided by a set of rules, the Authentication Service, driven by our model, calculates a risk score for the ongoing login attempt. Our model considers various context-based parameters to generate a nuanced score indicative of potential risk. Categorisation into risk levels—Low, Medium, or High—follows the adaptive authentication model. The final step involves the authentication service communicating the risk assessment to the API gateway, and our model crucially contributes to this response. Our model's versatility, integration of context-based parameters, anomaly detection, and prediction capability align seamlessly with the intricacies

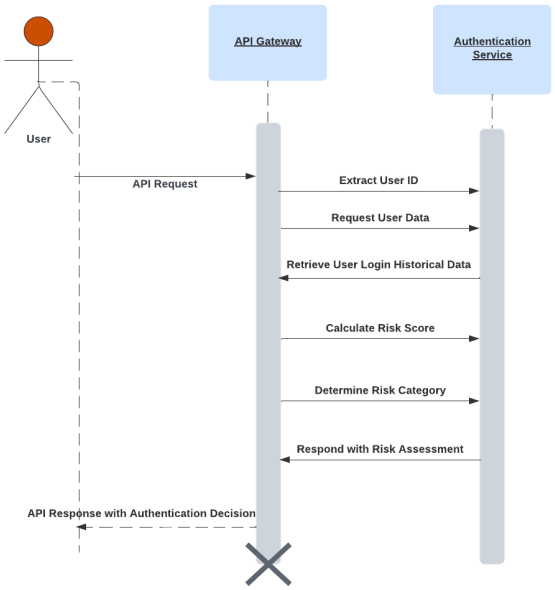

FIGURE 2. System's diagram.

of API-level security, making it a robust and adaptive safeguard against potential threats.

V. METHODOLOGY

This research employs a robust and methodical approach, amalgamating machine learning, data pre-processing of publicly available RBA Context-Based dataset [15], and rigorous model evaluation.The context-based attributes data was synthesized from the real-world login behaviour of over 3.3M users at a large-scale single sign-on (SSO) online service in Norway for a year. The users used this SSO to access sensitive data provided by the online service, e.g., cloud storage and billing information. This data set and the contents of this repository are licensed under the Creative Commons Attribution 4.0 International (CC BY 4.0)Â license. The collected data is preprocessed and cleaned, settling for about 19200 records by removing noise, eliminating missing and incomplete data, converting the data type and normalizing values before being integrated into the Random Forest model, enhancing its decision-making process by providing more accurate information. This also makes it computationally feasible. The methodology unfolds in well-defined steps, ensuring the research objectives are met effectively. The choice of the ML Random Forest algorithm is the cornerstone of this research methodology. This decision is fortified by the algorithm's outstanding characteristics and adaptability. Random Forest excels at managing intricate, high-dimensional datasets, making it a natural fit for the nuanced realm of API Ecosystem security. Furthermore, its intrinsic capability to handle categorical, numerical, and mixed data types positions it as an ideal choice. The model's ensemble learning approach, which leverages multiple decision trees, adds a layer of robustness to detect anomalous login attempts. This is especially vital in security contexts, where precision and recall are paramount.

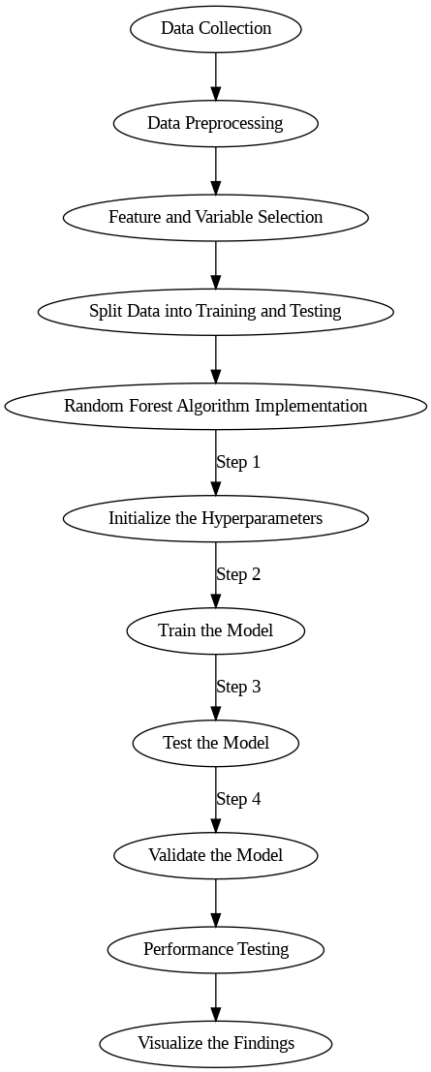

FIGURE 3. Random forest implementation flowchart.

The Random Forest algorithm can effectively manage a combination of context-based attributes for user authentication, underlining its suitability for the research goals. The model adapts dynamically to the evolving threat and user patterns, embodying the principles of adaptive security. Before implementing the ML Random Forest model, the selected RBA Context-Based dataset is preprocessed. Data preprocessing is fundamental to ensuring data quality and readiness for model training. This phase involves handling missing values, encoding categorical variables, scaling numerical features, and partitioning the dataset into training and testing subsets. Fig 3 illustrates the steps toward implementing the random forest algorithm.

The risk score (RS) is calculated as the sum of the products of each attribute value and its corresponding weight:

$$RS = \sum_{i=1}^{n} (X_i \times W_i) \tag{1}$$

breaking it down,

$$RS = (X_1 \times W_1) + (X_2 \times W_2) + \ldots + (X_n \times W_n)$$

The weights (W) and attributes (X) are represented as:

$$W = \{W_{\text{User ID}}, W_{\text{IP Address}}, W_{\text{Country}}, \ldots, W_{\text{n}}\}$$
$$X = \{X_{\text{User ID}}, X_{\text{IP Address}}, X_{\text{Country}}, \ldots, X_{\text{n}}\}$$

Finally, the risk score is categorized into risk levels using predefined thresholds based on the sensitivity of the system wanted threshold to be adjusted:

Categorize the risk scores df['Risk Category'] = pd.cut(df['Risk Score'], bins=[thresholds['Low Risk'], thresholds['Medium Risk'], thresholds['High Risk'], float ('inf')], labels=['Low Risk', 'Medium Risk', 'High Risk'])

Categorize risk scores based on thresholds

The following algorithm outlines the key steps involved in calculating the risk score and categorizing it based on predefined thresholds:

Algorithm 1 Calculate Risk Score and Risk Category

Require: User login attempt data with attributes (X)

Ensure: Risk Score (RS), Risk Category (Low/Medium/High)

1: Initialize Weights (W) for each attribute.
2: Calculate Risk Score (RS) using the formula: $RS = \sum_{i=1}^{n}(X_i \times W_i)$ for all i.
3: Categorize the risk scores based on predefined thresholds:
4: **if** $RS <$ Low Risk Threshold **then**
5: Assign "Low Risk" category.
6: **else if** Low Risk Threshold $\leq RS <$ High Risk Threshold **then**
7: Assign "Medium Risk" category.
8: **else**
9: Assign "High Risk" category.
10: **end if**
11: **return** RS, Risk Category.

To evaluate the research hypothesis and train the Random Forest model, we employed a split of 70 % for the training data and 30 % for testing. This division balances model training and testing, enabling robust evaluation. This methodology aligns seamlessly with the research goals due to the following aspects:

- Data Suitability: The research harnesses a diverse dataset encompassing various context-based parameters. This diversity caters to the nuances of API Ecosystem security.
- Precision and Recall: The emphasis on precision and recall in model evaluation is imperative in security contexts—the Random Forest algorithm's ability to provide high precision and recall positions it as an ideal choice.
- Adaptive Authentication: In the rapidly evolving era of Cyber Security, the methodology's focus on strengthening the adaptive authentication ensures a resilient defence against evolving threats and emerging user fingerprinting.

This crafted methodology constitutes a robust approach to achieving the research objectives. Combining machine learning, data preprocessing, and dynamic model evaluation yields insights that significantly contribute to Cyber Security and user authentication within the API layer.

VI. RESULTS AND FINDINGS

In this section, we present the results of hypothesis testing and the performance of the proposed Random Forest model for user authentication within the API layer. The model evaluation encompasses a range of crucial metrics that provide insights into the effectiveness of our approach.

The hypothesis under investigation sought to determine whether the integration of context-based fingerprinting features into the API Ecosystem authentication mechanism significantly impacts detecting anomalous login attempts. Specifically, the null hypothesis (H0) posited no significant difference in detection performance between the integrated approach and traditional methods. Conversely, the alternative hypothesis (H1) suggested that the integrated approach significantly enhances the detection of anomalous login attempts compared to traditional methods. Our methodology employed a logistic regression model, with a significance level set at $\pm = 0.05$. The logistic regression model was trained using the preprocessed dataset and evaluated using various metrics.

A. HYPOTHESIS TESTING METRICS

Statistical Significance: All independent variables, encompassing context-based historical user profiles, were statistically significant in explaining the target variable "Is Attack" This outcome decisively rejected the null hypothesis.

B. MODEL PERFORMANCE METRICS

Subsequently, we implemented the Random Forest ML Algorithm to predict the legitimacy of user login attempts, and it showed an impressive spike in detection accuracy that outperformed other models. The model was rigorously evaluated using a suite of performance metrics:

Accuracy: This metric measured the overall effectiveness of the model in correctly identifying both legitimate and malicious login attempts. Our model achieved an accuracy of 99.5%, reflecting its exceptional performance. Precision: Precision quantified the proportion of true positive predictions among all positive predictions, highlighting the model's ability to minimize false positives. With a precision of 99.5%, our model exhibited a remarkable ability to minimize false alarms. Recall: Recall, often referred to as sensitivity, measures the model's capacity to identify true positives among all actual positives correctly. Our model demonstrated a recall of 99.5%, indicating its excellence in capturing true positives. F1-Score: The F1-score provided a balanced assessment of a model's precision and recall.

Algorithm 2 Adaptive Authentication Algorithm

Require: User login attempt parameters (X), Historical data (H)
Ensure: Decision on access (Grant/Deny), Adaptive Actions
Step 1: User Initiates Login
Action: X represents the user's login attempt data
Step 2: Data Collection
Action: Collect relevant data for X
Step 3: Historical Data Retrieval
Action: H is retrieved based on the User ID
Step 4: Freeman Algorithm Calculation
Action: Calculate Risk(X, H) using the Freeman algorithm
Step 5: Assigning Risk Category
Action: Categorize risk into Low, Medium, or High using Category(X, H)
Step 6: Decision Making
Action: Compare Risk(X, H) with historical data
if Risk is Low **then**
    Grant access
else
    Proceed to Step 7 for adaptive actions
end if
Step 7: Adaptive Actions
Action: Based on the risk level, take adaptive actions
if Risk is Medium **then**
    Request additional authentication
else if Risk is High **then**
    Deny access and initiate further verification steps
end if
Step 8: Logging and Monitoring
Action: Log relevant information for monitoring
Step 9: Continuous Learning
Action: Incorporate feedback for continuous learning

It served as a reliable indicator of overall model performance, with our model achieving an F1-score of 99.4%. AUC (Area Under the Curve): The AUC metric quantified the model's ability to distinguish between legitimate and malicious login attempts. Our model exhibited an AUC of 93.3%, signifying its robustness in distinguishing between the two. MCC (Matthews Correlation Coefficient): Our model achieved an MCC of 80.5%, reflecting a substantial correlation between observed and predicted classifications, as in Fig 3.

C. MODEL EVALUATION

The confusion matrix offers a more detailed perspective on the model's performance, allowing us to visualize the distribution of True Positives(TP), True Negatives(TN), False Positives(FP), and False Negatives(FN). The confusion matrix is a vital tool for evaluating the performance of a classification model, such as the Random Forest model we used for user authentication. It allows us to visualize the model's classifications in a tabular format and helps us understand how well it distinguishes between different classes. Fig 4 shows the prediction performance of the model.

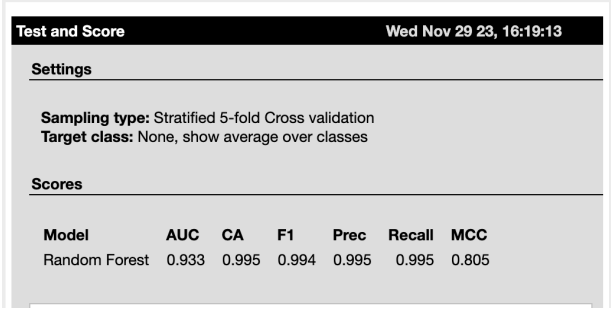

FIGURE 4. Model's performance.

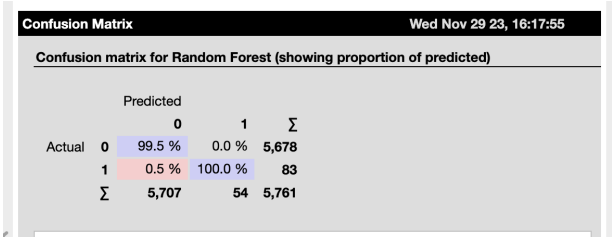

FIGURE 5. Model's prediction performance.

D. OBSERVATIONS

Our model showcases exceptional performance metrics. True Negatives (TN), representing 99.5 % in the top-left cell, affirm the accurate identification of legitimate users, significantly minimising false alarms. False Positives (FP) stand at 0.0%, emphasizing the model's remarkable precision in avoiding misclassifications of non-malicious attempts. There are no instances where the model wrongly categorized normal logins as malicious. False Negatives (FN) at 0.5% in the bottom-left cell highlight the model's ability to effectively identify actual malicious attempts, demonstrating its robustness in reducing the risk of overlooking security threats. True Positives (TP) reach 100.0 % in the bottom-right cell, showcasing the model's unwavering capability to identify and respond to malicious login attempts precisely. These values underscore the model's heightened accuracy, specificity, and sensitivity, contributing significantly to the overall security robustness.

Fig 5 illustrates the Receiver Operating Characteristic (ROC) curve, offering valuable insights into the efficacy of our model. Within this ROC curve, several crucial observations come to light. The curve effectively portrays the trade-off between the True Positive Rate (TPR) and the False Positive Rate (FPR). A point positioned closer to the top-left corner signifies a high TPR, albeit accompanied by a high FPR, whereas a point closer to the top-right corner indicates a low FPR at the cost of a diminished TPR. Notably, our ROC curve demonstrates an area under the curve (AUC) of 0.93, signifying the model's proficiency in discerning between attacks and non-attack instances. This metric attests

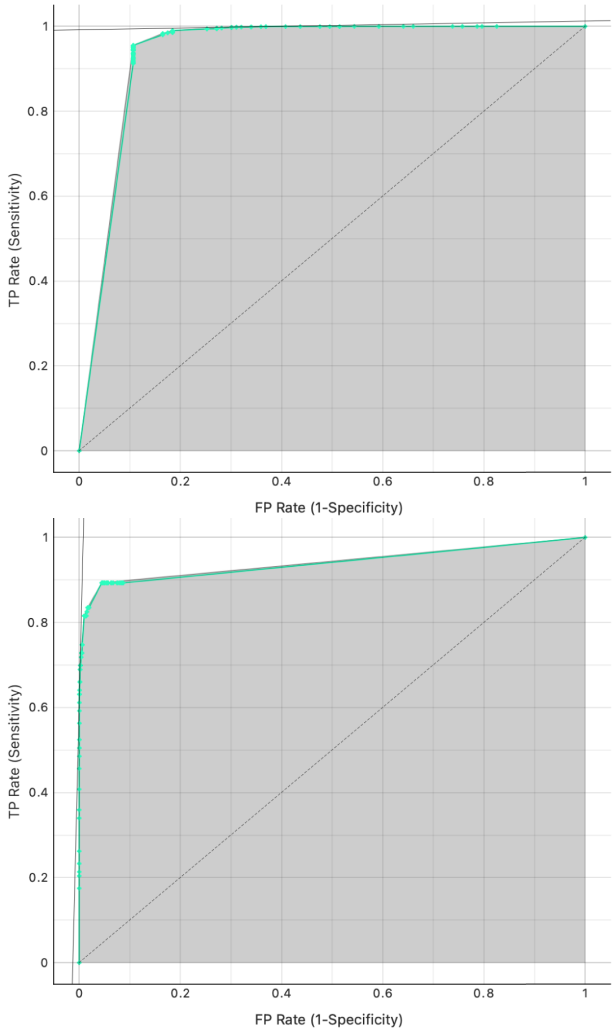

FIGURE 6. The receiver operating characteristic (ROC) curve.

to the robust performance of our model. Furthermore, the ROC curve empowers us to tailor the model's threshold, balancing sensitivity and specificity to our security and usability requisites. The versatility of the ROC curve extends beyond mere evaluation; it serves as a strategic instrument for meticulous model assessment and refinement. We can fine-tune the model by adjusting thresholds and aligning it with evolving security demands. This adaptability is pivotal for addressing the dynamic landscape of security challenges. The ROC curve's utility is underscored by its capacity to evaluate and enhance the model through parameter modifications or additional data assimilation.

Fig 7 suggests the Precision-Recall curve reveals essential aspects of our Random Forest model's performance. Despite a high overall accuracy of 95.5%, the Precision-Recall AUC of 0.76 highlights that our model effectively identifies positive instances. The slight drop in precision at a recall threshold of 0.892 indicates a critical balance point, suggesting that our model maintains a good trade-off between precision and recall. This curve demonstrates that our model performs well in distinguishing positive cases and, with further

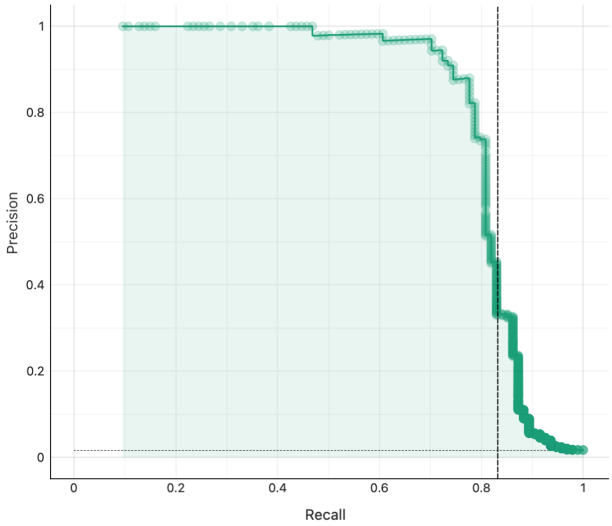

FIGURE 7. Precision recall curve.

tuning, has the potential to enhance its precision without compromising recall, thereby solidifying its robustness in security applications.

Table 1. provides insights into the efficacy of various models for user authentication. Among the Behavior-Based models, the Naive Bayes model demonstrates a high accuracy of 93.56%, relying predominantly on mouse movement for user legitimacy detection. However, the model combining keystrokes and mouse movement using an Artificial Neural Network achieves a lower accuracy of 58.9%, suggesting that this combination may be less effective for user authentication.

In contrast, the Context-Based models showcase notable improvements. The model incorporating application usage, keystrokes, and mouse movement using a Random Forest algorithm achieves an accuracy of 84.08%, indicating that the inclusion of application context enhances the authentication mechanism. Furthermore, the model integrating session context information, keystrokes, and mouse dynamics attains a higher accuracy of 97.50%, emphasizing the positive impact of session context on authentication accuracy.

The most compelling performance comes from our proposed Context-Based model, which incorporates a set of role-playing parameters, including Login Time, User ID, Round Trip Time, IP Address, Country, Region, City, ASN, User Agent String, Browser Name and Version, OS Name and Version, and Device Type. This model achieves an exceptional accuracy of 99.5%, signifying its robustness in elevating user authentication legitimacy. The extensive set of context parameters enhances security within the API ecosystem, contributing to an overall elevation in API Ecosystem Security.

E. COMPARISON AND BENCHMARKING

Table 2 compares the performance of different models using the RBA Dataset [15] fed with the same data dimension and size; the Random Forest model consistently outperforms other models regarding classification accuracy (CA). The

TABLE 1. Performance elevation.

Fingerprinting Type	Parameters	Algorithm	Accuracy
Behavior-Based [17]	Mouse movement.	Naive Bayes	93.56%
Behaviour-Based [18]	Keystroke dynamics, mouse movement.	Artificial Neural Network	58.9%
Context-Based [19]	Application usage, keystroke dynamics, and mouse movement.	Random Forest	84.08%
Context-Based [20]	Session Context information and Keystroke dynamics, mouse dynamics.	Random Forest	97.50%
Our proposed (Context-Based model)	Login Time, user ID, Round Trip Time, IP Address, Country, Region, City, ASN, User Agent String, Browser Name and Version, OS Name and Version, and Device Type. dynamics.	Random Forest	99.5%

differences in CA between Random Forest and the other models range from 0.012 to 0.014, translating to percentage differences of approximately 1.22% to 1.43%. Specifically, the Random Forest model shows a 1.43% higher CA compared to Naive Bayes, a 1.22% higher CA compared to SVM and Decision Tree, a 1.32% higher CA compared to kNN and Neural Network, and a 1.22% higher CA compared to Logistic Regression. These results highlight our proposed Random Forest model's superior accuracy in classifying illegitimate and legitimate login attempts.

TABLE 2. Methods' performance comparison.

Model	AUC (%)	CA (%)	F1 (%)	Precision (%)	Recall (%)	MCC (%)
Our proposed Random Forest	94.9	99.5	99.4	99.5	99.5	82.5
Naive Bayes	80.6	98.1	97.4	96.6	98.1	-0.5
SVM	74.7	98.3	97.5	96.6	98.3	0.0
Decision Tree	49.4	98.3	97.5	96.6	98.3	0.0
kNN	75.3	98.2	97.7	97.5	98.2	21.1
Logistic Regression	79.7	98.3	97.5	96.6	98.3	0.0
Neural Network	86.8	98.2	97.5	97.1	98.2	7.2

VII. MODEL'S IMPACT ON API SECURITY AND CYBER SECURITY

The outstanding performance metrics and the confusion matrix demonstrate that our Random Forest model effectively predicts and classifies user legitimacy within the API layer. The high accuracy, precision, recall, and F1-score values indicate the model's reliability and effectiveness in detecting and predicting legitimate users while minimizing false alarms and ensuring malicious attempts are correctly flagged.

The implications of this model on API security are profound. It significantly enhances the security of API access points by providing a robust and dynamic authentication mechanism. Accurately detecting anomalous login attempts and reducing false alarms ensures that legitimate users experience a smooth login process while efficiently protecting the system against potential security threats. This model's contribution becomes even more apparent when integrated into adaptive authentication systems. The dynamic nature of adaptive authentication aligns perfectly with our model's capabilities. It can adapt its response based on the evolving security landscape, effectively differentiating between genuine users and malicious actors.

Our research findings highlight the potential of the Random Forest model integrated context-based attributes in significantly enhancing API Ecosystem Security. The model's accurate prediction of user legitimacy, coupled with a low false positive rate, showcases its suitability for the dynamic API security landscape. This research contributes to the robustness of user authentication and advocates for the broader adoption of such innovative approaches in Cyber Security.

Ultimately, the effective user authentication model presented in this research is poised to elevate the security of API authentication and advocate for its broader adoption across various application architectures, guaranteeing a higher level of protection and user experience.

The impact of the proposed solution on the field of cybersecurity extends beyond API security, bringing significant positive advancements in various aspects of user authentication and access control. Context-based authentication enhances traditional methods by incorporating login time, location, and device factors. This approach reduces reliance on static credentials, making it more difficult for attackers to impersonate legitimate users. Integrating behavioural biometrics with our solution through analysing mouse movements, keystrokes, and session context allows for continuous authentication and effective anomaly detection.

Context-based authentication dynamically adjusts security requirements, offering seamless access in low-risk scenarios and additional verification in high-risk situations. This model can be implemented within adaptive security systems, enabling organizations to protect their data and system assets dynamically. Such a model may be part of the adaptive systems' policy to tailor their security measures based on contextual information, such as enforcing stricter protocols during suspicious network activities. Although real-world deployment presents challenges like data quality and scalability, regular updates and continuous monitoring ensure the model's effectiveness and reliability, contributing to robust cyber security defences across various applications.

VIII. DISCUSSION

- - Interpretation of Results and Findings: The results and findings of this research indicate a promising direction for enhancing API security, specifically concerning user

authentication. We initially tested the hypothesis to determine the significance of integrating context-based fingerprinting attributes into the authentication process. The results decisively rejected the null hypothesis, affirming that these integrated features significantly impact the detection of anomalous login attempts.

The subsequent implementation of the Random Forest model reinforced the potential of this integration. The model demonstrated exceptional accuracy, precision, recall, and F1-score. The model accurately distinguishes between legitimate and malicious login attempts while minimizing false alarms. The substantial MCC further underscores its robustness in user authentication.

- Practical Implications of Integrating Our Proposed Model: The practical implications of integrating our proposed model into API authentication are substantial. API security is the first line of defence for any system, where the login process acts as the gatekeeper. Our model's effectiveness in correctly identifying legitimate users while efficiently detecting and responding to malicious attempts bolsters API security. Reducing the number of false alarms and ensuring that potential threats are promptly flagged significantly enhances the overall security of the application architecture.

  Moreover, when our model is integrated as part of an Adaptive Authentication system, its dynamic nature aligns seamlessly with the evolving security landscape. It can adapt its response and authentication requirements based on the observed user fingerprinting, device characteristics, and context. This makes it well-suited for real-world scenarios where security threats continuously evolve. Users benefit from a seamless and secure authentication process, while administrators can be confident that potential risks are being actively mitigated.
- Addressing Limitations and Future Research Directions: While the results are promising, it's essential to acknowledge the limitations of this research. One limitation is that the model's effectiveness depends on the quality and quantity of historical user data. The model's performance may be affected in scenarios with limited data or rapidly changing user patterns. Furthermore, our model assumes that anomalies in user fingerprinting are primarily indicative of malicious intent. While this assumption is sometimes valid, some anomalies may result from legitimate user actions or system changes. Future research could explore more sophisticated anomaly detection techniques to further refine the model's accuracy.

  Additionally, research can extend the model to include more authentication layers. While context-based authentication provides robust security, adding additional layers such as biometric authentication or adaptive multi-factor authentication could further be integrated to enhance security in high-risk scenarios.

Finally, our research provides a compelling case for integrating context-based attributes and Random Forest Algorithm API Ecosystem's User authentication. It significantly enhances security, minimizes false alarms, and adapts to the evolving threat landscape. However, ongoing research is needed to address specific limitations and explore additional security layers, ensuring that API security remains at the forefront of Cyber Security in an ever-changing digital landscape.

IX. CONCLUSION

In summary, our research has introduced a groundbreaking model that harnesses the power of context-based fingerprinting alongside the Random Forest algorithm to fortify User Authentication within the API Ecosystem. The findings underscore the potential of this model to markedly enhance the precision of identifying anomalous login attempts, positioning it as a formidable alternative to conventional methods. It reaffirms the pivotal role of User Authentication as the first gateway to any system, making it imperative for robust security. The proposed model, adept at distinguishing legitimate from malicious login attempts, holds vast promise for enhancing API security, particularly within the framework of Adaptive Authentication. The contributions of this research extend not only to the domain of Cyber Security but also advocate for the broader adoption of innovative approaches to security.

X. LIMITATIONS AND FUTURE SCOPE

The current investigation provides a foundation for prospective inquiries within the domain of User Authentication and Adaptive Authentication:

Subsequent research endeavours may concentrate on refining User Authentication models, with a particular emphasis on augmenting detection and prediction capabilities, while concurrently exploring inventive behavioural and context-based attributes. Additionally, there exists an opportunity to expand the scope of this study by investigating the integration of supplementary authentication layers, such as biometrics or geolocation, to fortify the model's security framework and mitigate potential risks.

Furthermore, the research can extend its purview to encompass diverse security domains, transcending beyond APIs to IoT devices, mobile applications, and cloud-based systems. This broader scope offers a more holistic approach to Cyber Security, addressing security concerns thoroughly. Evaluating the applicability of the proposed model through real-world implementations across various industries will be imperative, substantiating its potential for widespread adoption and substantial impact.

This research sets the stage for the evolution of adaptive authentication frameworks capable of dynamically responding to evolving security landscapes. As we navigate the ever-changing digital terrain, our research aspires to catalyze innovation in User Authentication, providing a robust

TABLE 3. Potential challenges and limitations and potential solution.

Challenge	Description	Limitation	Potential Solution
Data Quality and Availability	Obtaining high-quality and diverse datasets that include various Context-Based attributes and Behaviour-Based Attributes.	The model's performance heavily relies on the quality and diversity of the training and validation data.	Collaborating with data providers and developing techniques for data augmentation to enhance the dataset quality and diversity.
Context Attribute Variability	Variability in context-based attributes such as user behaviour and environmental factors can be significant across different users and over time.	Variability can introduce noise and potentially reduce model accuracy if not properly managed.	Implementing adaptive algorithms that can dynamically adjust to changes in context attributes and incorporate robust preprocessing techniques.
Real-time Data Processing	Processing and integrating context-based attributes in real-time for user authentication requires substantial computational resources.	Potential latency issues in environments demanding instantaneous authentication.	Optimizing algorithms for real-time processing and leveraging edge computing to reduce latency.
Scalability	Handling many users and context data points while maintaining model performance.	Ensuring scalability without compromising accuracy and speed can be challenging.	Employing distributed computing and scalable architecture designs to manage large-scale data processing and model training efficiently.
Privacy Concerns	Collecting and utilizing context-based attributes for user authentication raises privacy and data protection issues.	Ensuring user consent and maintaining data privacy is not an easy task.	Implementing robust data anonymization techniques and ensuring compliance with privacy regulations to protect user data.

groundwork for advancing API security and fostering future breakthroughs in the broader field of Cyber Security.

Future research could explore enhancements to the model by incorporating additional context-based attributes, such as user location and device type, plus behaviour-based attributes, such as keystrokes, mouse and mouse clicks and movement. Moreover, real-world implementation challenges, such as scalability and data privacy, should be addressed. Further studies could also investigate the impact of different machine learning algorithms on the performance of the context-based fingerprinting approach.

A. ENHANCEMENTS TO THE MODEL

Temporal Context: Consider incorporating temporal context, such as the time of day or day of the week. Users' behaviour may vary based on these factors, and including them could improve accuracy.

Dynamic Feature Selection: Explore adaptive feature selection. Some context attributes may be more relevant in certain scenarios. An intelligent model that dynamically selects features based on context could enhance performance.

Ensemble Methods: Combine multiple context-based models (e.g., your proposed model with other algorithms) to create an ensemble. Ensemble methods often yield better results by leveraging diverse approaches.

B. ADDITIONAL CONTEXT-BASED ATTRIBUTES

Network Context: Analyze network-related features, such as latency, packet loss, or connection type (e.g., Wi-Fi, cellular). Unusual network behaviour could signal security threats.

Device Context: Include device-specific attributes (e.g., device fingerprint, hardware details). Device changes or anomalies might indicate compromised accounts.

User Behavior Patterns: Investigate long-term behaviour patterns. For instance, does a user consistently log in from specific locations or exhibit consistent typing speed? These patterns can enhance accuracy.

C. REAL-WORLD IMPLEMENTATION CHALLENGES

Data Privacy: Collecting extensive context data raises privacy concerns. Striking a balance between security and user privacy is crucial. Implement robust data anonymization and consent mechanisms.

Scalability: Ensure your model scales well as the user base grows. Efficient algorithms and distributed computing can address scalability challenges.

Adversarial Attacks: Consider adversarial scenarios where attackers intentionally manipulate context features. Robustness against such attacks is essential.

Deployment Complexity: Integrating context-based authentication into existing systems requires careful planning. Compatibility, user education, and system updates are critical.

ACKNOWLEDGMENT

The authors sincerely thank Dr. Anand Upadhyay and their fellow researchers for their invaluable contributions. Their willingness to engage in discussions and share their expertise played a crucial role in completing this work.

REFERENCES

[1] R. Mayrhofer and S. Sigg, "Adversary models for mobile device authentication," *ACM Comput. Surveys*, vol. 54, no. 9, pp. 1–35, Dec. 2022.

[2] S. A. Almaqashi, S. S. Lomte, S. Almansob, A. Al-Rumaim, and A. A. A. Jalil, "The impact of icts in the development of smart city: Opportunities and challenges," *Int. J. Recent Technol. Eng.*, vol. 8, no. 3, Sep. 2019. [Online]. Available: https://www.ijrte.org/wp-content/uploads/papers/v8i3/B3154078219.pdf

[3] S. Kavianpour, B. Shanmugam, S. Azam, M. Zamani, G. N. Samy, and F. De Boer, "A systematic literature review of authentication in Internet of Things for heterogeneous devices," *J. Comput. Netw. Commun.*, vol. 2019, pp. 1–14, Aug. 2019.

[4] I. Anastasaki, G. Drosatos, G. Pavlidis, and K. Rantos, "User authentication mechanisms based on immersive technologies: A systematic review," *Information*, vol. 14, no. 10, p. 538, Oct. 2023.
[5] N. Siddiqui, L. Pryor, and R. Dave, "User authentication schemes using machine learning methods—A review," in *Proc. Int. Conf. Commun. Comput. Technol. (ICCCT)*. Singapore: Springer, 2021, pp. 703–723.
[6] M. Papathanasaki, L. Maglaras, and N. Ayres, "Modern authentication methods: A comprehensive survey," *AI, Comput. Sci. Robot. Technol.*, vol. 2022, pp. 1–24, Jun. 2022.
[7] K. A. A. Bakar and G. R. Haron, "Adaptive authentication: Issues and challenges," in *Proc. World Congr. Comput. Inf. Technol. (WCCIT)*, Jun. 2013, pp. 1–6.
[8] R. Pramila, M. Misbahuddin, and S. Shukla, "A survey on adaptive authentication using machine learning techniques," in *Data Science and Security*. Singapore: Springer, 2022, pp. 317–335.
[9] A. Hassan, B. Nuseibeh, and L. Pasquale, "Engineering adaptive authentication," in *Proc. IEEE Int. Conf. Autonomic Comput. Self-Organizing Syst. Companion (ACSOS-C)*, Sep. 2021, pp. 275–280.
[10] B. Vibert, C. Rosenberger, and A. Ninassi, "Security and performance evaluation platform of biometric match on card," in *Proc. World Congr. Comput. Inf. Technol. (WCCIT)*, Jun. 2013, pp. 1–6.
[11] M. Miettinen, T. D. Nguyen, A.-R. Sadeghi, and N. Asokan, "Revisiting context-based authentication in IoT," in *Proc. 55th Annu. Des. Automat. Conf.*, Jun. 2018, pp. 1–6.
[12] J. Solano, L. Camacho, A. Correa, C. Deiro, J. Vargas, and M. Ochoa, "Combining behavioral biometrics and session context analytics to enhance risk-based static authentication in web applications," *Int. J. Inf. Secur.*, vol. 20, no. 2, pp. 181–197, Apr. 2021.
[13] B. Bezawada, M. Bachani, J. Peterson, H. Shirazi, I. Ray, and I. Ray, "Behavioral fingerprinting of IoT devices," in *Proc. Workshop Attacks Solutions Hardw. Secur.*, Jan. 2018, pp. 41–50.
[14] M. Misbahuddin, B. S. Bindhumadhava, and B. Dheeptha, "Design of a risk based authentication system using machine learning techniques," in *Proc. IEEE SmartWorld, Ubiquitous Intell. Comput., Adv. Trusted Comput., Scalable Comput. Commun., Cloud Big Data Comput., Internet People Smart City Innov. (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI)*, Aug. 2017, pp. 1–6.
[15] S. Wiefling, P. R. Jørgensen, S. Thunem, and L. L. Iacono, "Pump up password security! Evaluating and enhancing risk-based authentication on a real-world large-scale online service," *ACM Trans. Privacy Secur.*, vol. 26, no. 1, pp. 1–36, Feb. 2023.
[16] O. A. Salman and S. M. Hameed, "Using mouse dynamics for continuous user authentication," in *Proc. Future Technol. Conf. (FTC)*, vol. 1. Cham, Switzerland: Springer, 2018, pp. 776–787.
[17] S. Mondal and P. Bours, "Combining keystroke and mouse dynamics for continuous user authentication and identification," in *Proc. IEEE Int. Conf. Identity, Secur. Behav. Anal. (ISBA)*, Feb. 2016, pp. 1–8.
[18] H. Zhang, D. Singh, and X. Li, "Augmenting authentication with context-specific behavioral biometrics," in *Proc. 52nd Hawaii Int. Conf. Syst. Sci.*, 2019, pp. 1–10.
[19] J. Solano, L. Camacho, A. Correa, C. Deiro, J. Vargas, and M. Ochoa, "Risk-based static authentication in web applications with behavioral biometrics and session context analytics," in *Proc. Appl. Cryptogr. Netw. Secur. Workshops (ACNS)*. Bogota, Colombia: Springer, 2019, pp. 3–23.

AKRAM AL-RUMAIM studied the H.S.C. from Sana'a, Yemen. He received the B.Sc. and M.Sc. degrees in information technology from the University of Mumbai, Mumbai, India, in 2018. He is currently pursuing the Ph.D. degree in computer science and technology (in process) with Goa University, Goa, India.

He has participated in many international conferences and published multiple research papers in international conferences and journals. His research interests include cyber security, cloud security, computer science, the IoT security, data science, and API security. He is a member of ACM.

JYOTI D. PAWAR received the M.C.A. and Ph.D. degrees from Goa University, in July 1990 and November 2005, respectively. Her Ph.D. thesis was titled, "Design and Analysis of Subspace Clustering Algorithms and Their Applicability."

She is currently a Professor and the Dean of the Goa Business School, Goa University. She is also a member of the Board of Studies in computer science at Goa University. She mentors the M.C.A. and Ph.D. students in computer science and technology at Goa University. She has over 71 publications distributed over a variety of reputed journals and conferences in her area of research. Her research interests include problem solving, data structures, and data mining. She is a Life Member of the Computer Society of India. She is a member of various committees set up occasionally to assist in various activities/events at Goa University, and the Art of Living Foundation, Goa Chapter, and the Yoga Vedanta Seva Samiti.