Towards a Better Understanding of Impersonation Risks Anonymous

Anne Bumiller, Olivier Barais, Nicolas Aillery, Gael Le Lan

Towards a Better Understanding of Impersonation Risks

Anne Bumiller
University of Rennes 1/INRIA/IRISA
Rennes, France
Email: anne.bumiller@inria.fr

Olivier Barais
University of Rennes 1/INRIA/IRISA
Rennes, France
Email: olivier.barais@inria.fr

Nicolas Aillery
Orange Labs
Rennes, France
Email: nicolas.aillery@orange.com

Gael Le Lan
Orange Labs
Rennes, France
Email: gael.lelan@orange.com

***Abstract*—In many situations, it is of interest for authentication systems to adapt to context (*e.g.,* when the user's behavior differs from the previous behavior). Hence, during authentication events, it is common to use contextually available features to calculate an impersonation risk score. This paper proposes an explainability model that can be used for authentication decisions and, in particular, to explain the impersonation risks that arise during suspicious authentication events (*e.g.,* at unusual times or locations). The model applies Shapley values to understand the context behind the risks. Through a case study on 30,000 real world authentication events, we show that risky and non-risky authentication events can be grouped according to similar contextual features, which can explain the risk of impersonation differently and specifically for each authentication event. Hence, explainability models can effectively improve our understanding of impersonation risks. The risky authentication events can be classified according to attack types. The contextual explanations of the impersonation risk can help authentication policymakers and regulators who attempt to provide the right authentication mechanisms, to understand the suspiciousness of an authentication event and the attack type, and hence to choose the suitable authentication mechanism.**

***Index Terms*—Explainable AI, Shapley Values, Authentication, Impersonation Risk**

I. Introduction

Authentication technique weaknesses, like password-based authentication, are known, and service operators often implement additional authentication mechanisms to limit the restraints of the individual techniques [12], [17]. During authentication events, contextually available features are used to calculate a **risk score**. Such risk scores are typically classified into three buckets: low, medium, and high [5], [7], [10]. Additional authentication mechanisms are required if a high risk is detected [20]. Therefore, the impersonation risk scores are often referred to as **black boxes**, even if they do not contain any information about the context. The question of **which authentication mechanisms are suitable** (*e.g.,* appropriate for security and usability) in the context can not be answered only based on the score. For example, a "verification code send per SMS" can not be bypassed by an attacker who has stolen the password, but it can be bypassed by one who owns the victim's device. To decide which authentication mechanism to require, contextual explanations giving insights about the risk type (*e.g.,* password theft, device theft) in addition to the risk score are necessary.

Explainable AI models provide details or reasons to make the functioning of *Artificial Intelligence* (AI) straightforward or easy to understand. Explanations can answer different kinds of questions about *what the AI model is learning*, *which parts of the inputs are the most important for the prediction* and *can we trust the model's decision* [11]. From a mathematical viewpoint, "simple" statistical learning models, such as linear and logistic regression models, provide high **interpretability** but, possibly, limited predictive **accuracy**. On the other hand, "complex" machine learning models, such as neural networks, provide high predictive accuracy at the expense of a limited interpretability [11]. The same holds for impersonation risk scores calculated based on available contextual features during the authentication event. "Simple" statistical estimations are easy to understand, but when more "complex" models are used, it becomes hard to understand the risk prediction of an authentication event. Hence, it becomes difficult for authentication policymakers and regulators to provide the suitable authentication mechanisms. Among other dimensions, explainability models can be distinguished according to the scope of explications they provide. There are **global** explainability models which aim to explain the model as a whole and **local** explainability models that seek to explain individual predictions and that we propose to use to **explain the risk of a specific authentication event**. Also, we can distinguish between **model-specific** and **model-agnostic** explainability models. The latter, in contrast to the former, can be used without any knowledge about the AI model [11]. **Shapely values** provide local and model-agnostic explanations of AI algorithms by assuming that each feature is a player and the prediction is the outcome of the game. The Shapley value of a feature is the average of all its **marginal contributions** to all possible coalitions of contextual features [11]. Instead of how fair the distribution of a game's payout is, we want to analyze **how each contextual feature contributes to the risk**

score of an authentication event that estimates the risk of impersonation. We aim to answer the question of how to explain the risk of a suspicious authentication attempt. Our "game" is the risk score estimation. The "players" are the contextual features. They contribute to the risk score. The "gain" of one specific contextual feature is its marginal contribution to the risk score. Within this work, we propose a contextual feature engineering approach based on Shapley values. With the help of a case study on real-world authentication events, we show that the risk of impersonation can be explained differently and specifically for each authentication event. Through this case study we show that explainable machine learning models can effectively improve our understanding of impersonation risks. Authentication policymakers and regulators can use this explanations in addition to the risk score to identify risk types and to choose the suitable authentication mechanisms. Predicting when an authentication event is risky can be of use but more importantly, understanding the risk score can help to identify sets of clusters (authentication events with similar characteristics) that are at high risk. This will lead to better and appropriate authentication mechanisms adapted to the risk type of an authentication event. Our contribution consists of a case study to show that our proposed explainable AI model can help to understand the risks of impersonation. We provide a framework for authentication regulators and practitioners to apply the methodology of Shapley values to risky authentication events. We also propose a clustering of the explanations and a novel reasoning about risk types (authentication attacks) with the help of contextual information. Our application case study shows for 30,000 real world authentication events that it is possible to explain the risk of impersonation differently and specifically for each authentication event and that those explications can help authentication practitioners to reason about different attack types.

II. Motivational Scenarios

In order to determine the suitable authentication mechanism for a particular context, which is the role of *Risk Based Authentication* (RBA) approaches, it is crucial to understand the context behind a risk of impersonation beyond the score. The relevance of authentication mechanisms cannot simply be determined by a one-dimensional risk score, as different types of risks need to be differentiated.

a) Scenario 1: Let us assume a legitimate user who authenticates regularly with username, password, and an *One Time Password* (OTP). An attacker was able to get the user credentials through a phishing attack. Using social engineering, the attacker calls the user and convinces him to give away an OTP. Then, the attacker enters the credentials and types in the OTP, getting access to the protected resource.

b) Scenario 2: Another possible scenario is a user who authenticates regularly with username, password, and push-authentication[1]. The attacker hacked the phone, and malware ended up being installed by an attacker, giving him complete control of the user's phone. Push is not protected by a PIN or biometric. The attacker would use stolen credentials to authenticate, while monitoring the user's phone. When the push arrives, the attacker will use the control of the phone to approve the push and get access to the resource.

The two scenarios illustrate that for high-risk authentication attempts, there are different types of attacks. These differences are not considered when the context information is exclusively used to calculate a one-dimensional risk score. Therefore, there is a need for a modelling framework that enables a complex and fine-grained mapping between context information and authentication mechanisms.

The following example further illustrates the importance of taking into account context information for authenticating legitimate users in different contexts and not only denying access in the case of high-risk.

c) Scenario 3: Let us consider Bob, a German traveler in Spain. He checks his e-mails at 2:00 am in a poorly lit room. He enters the username and password correctly. His e-mail provider can acquire contextual information: geolocation, luminosity, time, and typing speed. Bob's e-mail provider determines some threats: Bob is not located in Germany as usual, he is checking his e-mails at an unusual time, it is dark around him, and he is typing slower than usual. All these threats make the e-mail provider assume that there is a risk that an intruder who has Bob's password might try to access Bob's e-mails. Bob has registered facial recognition and fingerprint as authentication mechanisms. Password-based authentication can be bypassed by the intruder who has stolen Bob's password. Face recognition is not efficient to use in the dark. Bob needs to be authenticated with his fingerprint.

The three presented scenarios would all have led to a high score in a risk-score-based approach. However, we see that to properly fend off attackers and allow legitimate users access, more information about the conetxt is necessary.

III. Methodology

In this section, we first present a **statistical approach to measure impersonation risks**, which is proposed in [5]. Since RBA is not a standardized procedure, multiple solutions exist in practice. We focus on Freeman et al.'s [5] implementation, since other works showed good performance [20]. Also, this RBA model is known to be widely used, *e.g.,* by popular online services like Amazon, Google, and LinkedIn [19], [20]. Afterwards, we explain how to exploit the explanatory context information contained in the impersonation risk score with the help of **Shapley values**. The Shapley method is **agnostic** (model neutral) applied to the predictive output, regardless of which model generated it. Hence, the method can be applied to any impersonation risk score estimation model. Also, with the Shapley method, **local explanations** can be obtained, and we can hence **explain every authentication event separately**. There are other local, model-agnostic explanation methods, *e.g., Individual Conditional Expectation* (ICE), Local Surrogate (LIME), Counterfactual Explanations and Scoped Rules

[1] A mobile-centric authentication mechanism whereby the service provider sends the user a notification and the user responds to the challenge by performing an action (*e.g.,*"OK" button)

(Anchors) [11]. According to [11], Shapley values might be the only method to deliver a **full explanation**, which is based on a **solid theory**. The problem of **allocating responsibility for risks** plays an important role in other domains as well (*e.g.,* in finance to evaluate the risk of an individual asset in a portfolio). We identify a set of works proposing the use of Shapley values for allocating responsibility for risks [2], [3], [13].

A. Statistical Approach to Measure Impersonation Risks

Impersonation risk models are usually employed to estimate the expected risk of impersonation of an authentication event a of a given user $u \in U$. The most important component of an impersonation risk model is a **risk score**, which is usually estimated statistically employing context scoring models [20]. $s_a = p(X_a, u, Y_a)$ is the risk score of an authentication event a of a user u, where $X_a = (x_a^1, ..., x_a^d) \in X$ indicates a d-dimensional vector of explanatory context information characterizing an authentication event (e.g. *IP address*, *user agent*). $Y_a \in G, I$ is the class label of a genuine authentication event (G) or an imposter authentication event (I) and f is a classification function $f : s \rightarrow Y$.

B. Exploiting the Explanatory Context Information

We now explain **how to exploit the explanatory context information** contained in the risk score with the help of **Shapley values**. For an authentication event a, we propose to calculate the Shapley value for each contextual feature $x_a^i \in X_a = \{x_a^1, ..., x_a^d\}$ characterising a. For each feature x_a^i, the Shapley value is defined as

$$\theta_{x_a^i}(a) = \sum_{z_a \subseteq X_a \setminus x_a^i} \frac{|z_a|!\,(d - |z_a| - 1)!}{d!} \left[p(z_a \cup x_a^i, u, Y_a) - p(z_a, u, Y_a)\right] \quad (1)$$

where p is the risk score estimation model and X_a the input vector of all d contextual features characterising the authentication event a. $z_a \subseteq X_a \setminus x_a^i$ is a subset of X_a that does not contain the contextual feature x_a^i for that the Shapley value is calculated. The quantity

$$\left[p(z_a \cup x_a^i, u, Y_a) - p(z_a, u, Y_a)\right] = MC_{x_a^i, z_a} \quad (2)$$

is the contribution of the contextual feature x_a^i to the impersonation risk estimation in the coalition $z_a \cup x_a^i$. This contribution is calculated as the difference between the risk score p estimated from $z_a \subseteq X_a \setminus x_a^i$ and p estimated from $z_a \cup x_a^i$. In Equation 2, we summarize the marginal contributions of x_a^i to all possible subsets $z_a \subseteq X_a \setminus x_a^i$. The fraction

$$\frac{|z_a|!\,(d - |z_a| - 1)!}{d!} \quad (3)$$

is a weighting function of $MC_{x_a^i, z_a}$. Depending on the number of contextual features in the subset z_a and the total number of contextual features d, $MC_{x_a^i, z_a}$ is weighted differently. If a contextual feature is added to an already large number of contextual features in z_a and yet the risk score is strongly influenced, then this must be weighted more than if the contextual information is added to an empty set. In the latter case, it is normal that the risk score is then strongly influenced.

a) Example: Let us take the example of $X_a = \{device, IP, location\}$ with $d = 3$ illustrated in Table I.

We want to calculate the Shapley value of $x_a^{location}$. There are four subsets $z_a \subseteq X_a \setminus x_a^{location}$ (column 1). The risk score can be estimated for these four sets (column 2)and then for the union of these four sets and $x_a^{location}$ (column 4). $MC_{x_a^{location}, z_a}$ (column 5) is the difference between the estimated risk scores. Depending on the number of contextual features in z_a we calculate the weighting function (column 6). To obtain $\theta_{x_a^{location}}(a)$ we multiply $MC_{x_a^{location}, z_a}$ with the weight (column 7) and sum up all the values: $\theta_{x_a^{location}}(a) = 0.198 + 0.051 + 0.034 + 0.099 = 0.382$. The Shapley value of $x_a^{location}$ feature is equal to 0.382. We can calculate the Shapley values for all the contextual features and compare their values to understand which contextual features contribute the most to the impersonation risk.

The Shapley values can be used to indicate **which contextual features contribute more to the prediction of the impersonation risk** of an authentication event. Not only in general, as it is typically done by statistical models, but differently and specifically **for each authentication event**.

b) Appropriateness of Shapley Values for Risk Attribution: Before explaining how we applied our methodology to a real-world dataset, we now explore, why the Shapley value properties (efficiency, symmetry, linearity, null-player) are **appealing in the context of risk attribution** [13]. *Efficiency* means that the sum of the Shapley values of all features equals the value of the coalition of all features, so that all the gain (risk) is distributed among the features [11]. Hence, the Shapley values reflect the risk diversification at the system level (at the authentication event level in our case). *Symmetry* means that for two equal features x_a^i and x_a^j $MC_{x_a^i, z_a} = MC_{x_a^j, z_a} \forall z_a \in X_a$ [11]. The symmetry property means that the labeling of individual components does not affect their measured contribution to system-wide risk. *Linearity* means that when two risk estimation models described by p_1 and p_2 estimate the risk of impersonation, then $\theta_{(x_a^i, p_1)+(x_a^i, p_2)} = \theta_{(x_a^i, p_1)} + \theta_{(x_a^i, p_2)}$ and $\theta_{(x_a^i, p_1)*\lambda} = \lambda * \theta_{(x_a^i, p_1)}$ [11]. The linearity property is useful in contexts, where model and parameter uncertainty calls for robust estimates. Such estimates are often obtained by combining the outcomes of competing risk estimation models. The linearity property of the Shapley Value implies that a robust estimate of a contextual feature's contribution to the impersonation risk of an authentication event would be the (weighted) average of the Shapley values for this feature across different risk estimation models. A contextual feature is a *null-player* if $p(z_a \cup x_a^i, u, Y_a) = p(z_a, u, Y_a) \forall z_a$. The Shapley value of a null-player is zero [11]. Given a player set X_a, the Shapley value is the only map from the set of all risk score estimations to risk score vectors that satisfies all four properties: **efficiency**, **symmetry**, **linearity** and **null-player**. Given the one-to-one

z_a	$p(z_a, u, Y_a)$	$z_a \bigcup x_a$	$p(z_a \bigcup x_a^i, u, Y_a)$	$MC_{x_a^i, z_a}$	$\frac{\lvert z_a \rvert!(d-\lvert z_a \rvert-1)!}{d!}$	
{device, IP}	0.3	{device, IP, location}	0.9	0.6	0.33	0.198
{device}	0.4	{device, location}	0.7	0.3	0.33	0.051
{IP}	0.2	{IP, location}	0.6	0.4	0.33	0.034
{}	0.1	{location}	0.4	0.3	0.17	0.099

Table I
EXAMPLE: CALCULATION OF THE SHAPLEY VALUE OF $x_a^{location}$ FOR $X_a = \{device, IP, location\}$

mapping between two risk estimation models, we henceforth focus exclusively on the risk attribution problem and thus on the risk measure, s.

IV. APPLICATION CASE STUDY

We now present the application case study of our methodology on real-world authentication events. Within this case study we demonstrate that our proposed framework can be applied to a dataset of real-world authentication events of a telecommunication company and that the obtained explanations are useful for the authentication policymakers and regulators to make adapted authentication decisions according to the attack types. Authentication policymakers from other companies can apply our framework in the same way and thus get information about attack types relevant to them.

We first describe the dataset which we used to test our model and explain the proposed method in detail. Afterwards, we present the obtained results. By describing our approach in detail, this chapter provides a framework that can be used by authentication regulators to apply the method in the same way on their data.

A. Data

We test our model to data supplied by a telecommunication company. In summary, the analysis relies on a dataset composed of contextual information on **30,000 authentication events** mostly based in France for the year 2022. The context information contains twelve categorical contextual features (see Table II).

B. Method

Based on [5], we have constructed a statistical estimation of the impersonation risk. For every authentication event, we calculate the logarithmic probability that it is a legitimate event and an imposter event. The actual **risk score** describes the difference between these two log probabilities.

$$s_a = log(p(X_a, u, I)) - log(p(X_a, u, G)) \tag{4}$$

We choose a threshold λ for the risk score to label the events as **genuine and imposter events**.

First, we calculate some basic **descriptive statistics** to summarise the central tendencies, and to analyze how the values of the contextual features are spread off.

Then, we split the authentication event data between a **training set (80%) and a test set (20%)**, using random sampling without replacement. We then get 24,000 training samples and 6,000 test samples.

On these samples, we run a **Logic Regression**, a **Random Forest Classifier**, a **Decision Tree Classifier**, and a ***Support Vector Machines* (SVC) Classifier**. To obtain Y_a, the estimated impersonation risk probability is classified into "genuine" (G) or "imposter" (I), depending on whether the threshold is passed or not. For a given threshold *T*, one can then count the frequency of the **possible outputs**: *False Positives* (FP): authentication events predicted to imposter, that are genuine; *False Negatives* (FN): authentication events predicted to genuine, which are imposter; *True Positives* (TP): authentication events predicted as imposter, which are imposter; *True Negatives* (TN): authentication events predicted as genuine, which are genuine. The **misclassification rate** of a model can be calculated as

$$\frac{FP + FN}{TP + TN + FP + FN} \tag{5}$$

and it characterizes the proportion of wrong predictions among the total number of predictions. *False Positive Rate* (FPR) and *True Positive Rate* (TPR) are then calculated as follows:

$$\frac{FP}{FP + TN} \tag{6}$$

$$\frac{TP}{TP + FN} \tag{7}$$

Further, we analyze the ***Receiver Operating Characteristic* (ROC) curves** of the four classifiers. They plot the FPR on the *Y* axis against the TPR on the *X* axis for a range of threshold values. The ideal ROC curve coincides with the Y axis, a situation which cannot be realistically achieved. The best model will be the one closest to it. The ROC curve is usually summarised with the *Area Under The Curve* (AUC), a number between 0 and 1. The higher the AUC, the better the model.

Next, we calculate the **Shapley value explanations** of the authentication event logs in the test set, using the values of their explanatory contextual features. In particular, we use the **Tree*SHapley Additive exPlanations* (SHAP) method** in combination with Random Forest. Tree SHAP is a fast and exact method to estimate Shapley values for tree models and ensembles of trees [11]. We calculate the local Shapley values for the 6,000 authentication events of our test sample. We get 6,000 arrays consisting of two sub-arrays. In the first sub-array, we get the Shapley values for our first class (**imposter authentication events**). In the second sub-array, we get the Shapley values for the second class (**genuine authentication events**).

Context Information	Type	Description
changingIP	Boolean (0,1)	the IP address is known or unknown from the user's history
gatewayOwner	Boolean (0,1)	the user is or is not behind his or her own line
changingDevice	Boolean (0,1)	the device is known or unknown from the user's history
internISP	Boolean (0,1)	the *Internet Service Provider* (ISP) is or is not the telecommunication company
app	Category (#38)	the accessed resource (*e.g.,* Mail)
fastLocationChange	Boolean (0,1)	successive connections from two countries in a short time or not
authenticationMethod	Category (#3)	the used authentication method (*e.g.,* password)
country	Category (#136)	the country that the authentication attempt originates from
robot	Boolean (0,1)	regularity of successive connections is or is not detected
changingSim	Boolean (0,1)	the SIM card has been changed or not
knownUser	Boolean (0,1)	the user is known (has been previously seen) or is connecting for the first time
changingLocation	Boolean (0,1)	the location is known or unknown from the user's history

Table II
DESCRIPTION OF THE USED CONTEXTUAL FEATURES

The last part of the analysis involves using the Shapley value vectors that correspond to each authentication event, and look for the presence of **clustering structures** that group together **similar risky authentication events**. To this aim, we employ a **K-means** clustering algorithm[2]. We cluster the Shapley values calculated for the authentication events of our test sample to find patterns that can lead to appropriate authentication mechanisms. We are using the elbow method[3] to decide how many clusters are a good fit for our data. Next, we fit the K-Means model to our Shapley values for the test sample with four as the number of clusters. We then map for each authentication event (data point) which cluster was assigned to it based on its training and look for the presence of clustering structures that group together similar authentication events.

As you can see from Figure 1, our methodology consists of first, to fit a complex model like Random Forest to the data consisting of a number of features to predict whether an authentication event is risky or not, and then use the TreeSHAP method to get Shapley values for each authentication event in the test sample, and then cluster them to find patterns that can lead to better authentication decisions.

C. Results

Figure 2 displays exemplary the distribution of the "changingIP" feature regarding the risk score. We observe that the medium risk score is higher if the "changingIP" feature takes the value 1 (unknown IP) than if the value is 0 (known IP).

Figure 3 shows that all classifiers outperform the SVC classifier. Indeed the comparison of the **Area Under the ROC curve (AUC)** for the four classifiers indicates an increase from 0.90 (SVC) to 0.94 (Random Forest, Decision Tree). For further analysis we choose the **Random Forest** Classifier because it outperforms the Decision Tree Classifier in terms of accuracy (92.1 versus 90.8).

For single authentication events, we can visualize the explications as illustrated in Figure 4 for an imposter authentication event. Features that **push the risk score higher** (to the right) are shown in red, and those **pushing the prediction lower** are in blue. The **output value** is the prediction for that authentication event (0.92). The **base value** is the value that would be predicted if we did not know any features for the current authentication event. In other words, it is the mean risk prediction. Our base value is 0.1843. This is because the mean of the risk scores in our test sample is 0.1843. In the exemplary authentication event at risk (see Figure 4), the contextual features that drive the score up the most are *changingDevice*, *changingIP* and *gatewayOwner*.

Rather than referring to the risk score as a black box to choose the suitable additional authentication mechanisms for a high-risk authentication event, the **explanations** can be used. These explain the contextual background of the risk, which is necessary to reason about the suitability of additional authentication mechanisms.

According to the elbow method, we choose 4 as number of clusters for the k-means clustering (Figure 5).

In Figure 6, we plot the scatterplot of the first two principal components of the Shapley values, attributing each authentication event to one of the four clusters. The obtained clusters are

[2]A method of vector quantization that aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean

[3]Heuristic used in determining the number of clusters in a data set consisting of plotting the explained variation as a function of the number of clusters, and picking the elbow of the curve as the number of clusters to use

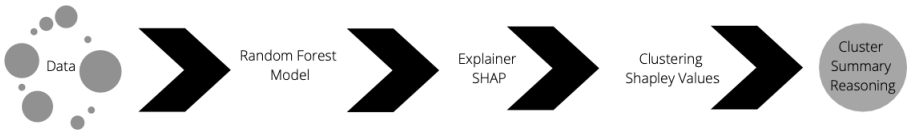

Figure 1. Methodology to Identify Similar High-Risk Authentication Events through Clustering

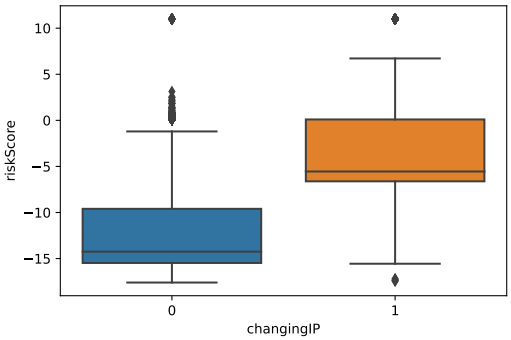

Figure 2. Boxplot Displaying the Distribution of the "changingIP" Feature Regarding the Risk Score

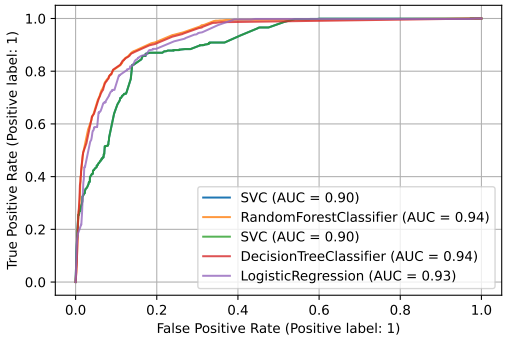

Figure 3. Receiver Operating Characteristic (ROC) curves for the Logic Regression, the Random Forest Classifier, the Decision Tree Classifier, and the SVC Classifier

clearly differentiated and balanced, confirming the advantage of using our proposed method.

Furthermore, we take a closer look at the authentication events of our four clusters. For most of the authentication events which have been assigned to **cluster 0** the user is not behind his or her own line and the ISP is not the telecommunication company. For most of the authentication events which have been assigned to **cluster 1**, the IP address is unknown and the device is unknown. For most of the authentication events which have been assigned to **cluster 2**, the IP address and the geolocation are unknown. For most of the authentication events which have been assigned to **cluster 3**, the user is connecting for the first time. We can see, that the different cluster represent different contextual situations.

V. Using Explanations to Differentiate Between Different RBA Attacks Types

The contextual explanations of the impersonation risk can help authentication policymakers to understand the suspiciousness of a high-risk authentication event in terms of the attack type, which is behind the risk. Hence, they can choose authentication mechanisms that are suitable for the attack type.

We take three **attack types** based on known ones in the RBA context presented in [18]. All attackers possess the victim's login credentials, none of the attackers possesses the complete context of the legitimate user (see Figure 7).

Wiefling et al. [18] describe the attack types. We here analyze them further regarding six exemplary contextual features. We can see in Table III that the values that the contextual features take depend on the attack type. This illustrates that contextual explanations of the impersonation risk score can help to choose an authentication mechanism that is suitable for the attack type. Common risk factors of the attack types (*e.g., fastLocationChange*) are evident from the explanations but not from a risk score itself.

VI. Related Work

The works related to our work fall into two categories: **proposals for using Shapley values for allocating responsibility for risks** (1) and **works that consider different types of attacks in the RBA context** and evaluate authentication mechanisms in terms of resilience against the different attack types (2).

A. Shapley Values for Allocating Responsibility for Risks

The main part of proposals for using Shapley values for **allocating responsibility for risks** relate to the **finance domain**. Bussman et al. [3] propose an explainable AI model based on Shapley values that can be used in **credit risk management**. They group risky and not risky borrowers according to financial characteristics. Nikola Tarashev et al. [13] use Shapley values to derive measures of **banks' systematic importance**. In [2], the authors propose a method to **allocate risk capital** to divisions or lines of business within a firm. Wang et al. [16] suggest an **income allocation scheme** for farmers, insurance institutions, and futures institutions. Also in other domains than finance, we identify works proposing the use of Shapley values for allocating responsibility for risks.

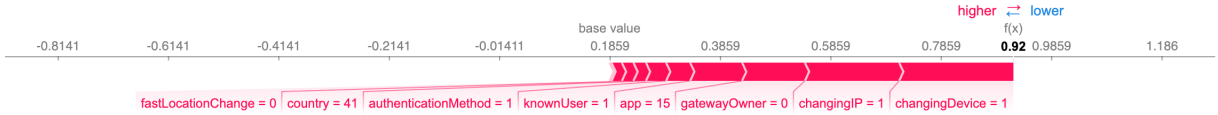

Figure 4. Local Explication for an Authentication Event at Risk

Contextual Feature	Naive Attacker	VPN Attacker	Targeted Attacker
IP address	randomly located	located in the victim's country	located in the victim's city
browser	random popular browser	random popular browser	the victim's browser
device	random popular device	random popular device	the victim's device
keystrokes	unknown	unknown	unknown
changingLocation	1	1	0
fastLocationChange	1	0	0

Table III
CONTEXTUAL CHARACTERISATION OF RBA ATTACK MODELS

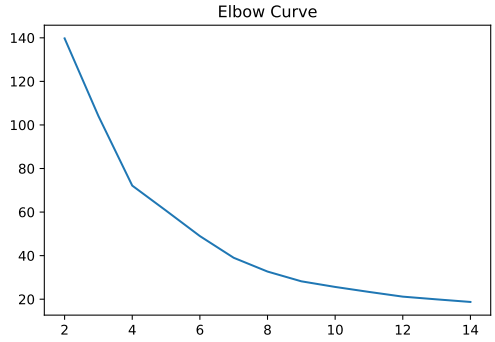

Figure 5. Elbow Curve to Choose the Right Number of Clusters (4)

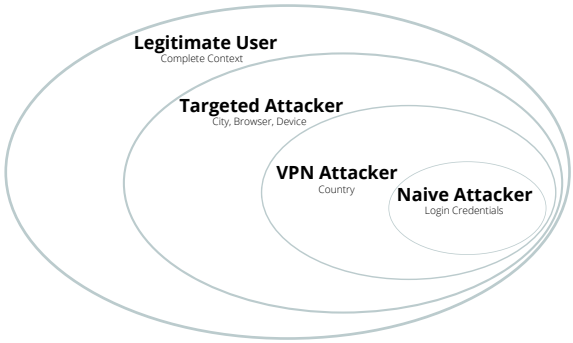

Figure 7. Overview of the Attack Types that Can be Distinguished based on Contextual Explanations of the Impersonation Risk Score [18]

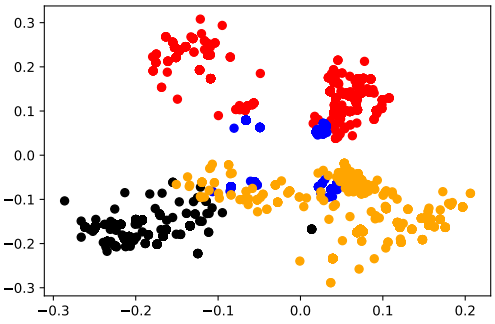

Figure 6. Scatterplot of the First Two Principal Components of the Shapley Values

Ginger Y. Ke et al. [9] aim to mitigate the risk of **possible incidents caused by the storage and transportation of hazardous materials** with the help of Shapley values. In [8], the authors investigate the **anthropometric characteristics of patients** with chronic diseases (diabetes, hypertension, cardiovascular disease, heart attacks, and strokes) and find the factors affecting these diseases with the help of Shapley values.

B. RBA Attack Types

With the help of the contextual explanations of the impersonation risk score, we aim to help authentication policymakers to get insights about the **attack type** behind the risk of impersonation and hence help them to choose the suitable authentication mechanism. We identify a set of works differentiating between multiple attack types in the context of RBA [1], [4], [6], [14], [15]. The **authentication mechanisms are evaluated in terms of resilience against the different attack types**.

Table IV shows an overview of attack types against which the resilience of authentication mechanisms is evaluated in different works. The set of works is not exhaustive because we do not aim to identify all the literature on the evaluation

	Physical Observation	Targeted Impersonation	Throttled Guessing	Unthrottled Guessing	Internal Observation	Leaks from Other Verifiers	Phishing	Physical Theft	VPN Attack	DoS Attack	Parallel Session Attack	Replay Attack	Reflection Attack	Observation from Third Parties
Doerfler et al. (2019) [4]	x		x	x		x	x	x						x
Velasquez et al. (2018) [14]							x					x		x
Wang et al. (2016) [15]		x	x		x	x		x		x	x	x	x	
Bonneau et al. (2012) [1]	x	x	x	x	x	x	x	x						x
Halunen et al. (2016) [6]	x	x	x	x	x	x	x	x	x	x	x	x	x	x

Table IV
OVERVIEW OF ATTACK TYPES AGAINST WHICH THE RESILIENCE OF AUTHENTICATION MECHANISMS IS EVALUATED IN DIFFERENT WORKS

of authentication mechanisms. But the existence of works evaluating authentication mechanism according to their resilience against different attack types shows that contextual explanations can help choosing the right authentication mechanism.

That authentication mechanisms can be evaluated in terms of their resilience against different attack types, underlines the importance of contextual explanations of the impersonation risk score. Only based on them, not on the risk score itself, can it be deduced what type of attack it is, and hence the correct resilient authentication mechanism can be chosen.

VII. CONCLUSION

In order to improve the understanding of impersonation risks, we have proposed a novel methodology that can be embedded within an authentication service. The methodology, which is based on a model agnostic interpretability tool (Shapley Values), leads to a powerful segmentation of authentication events. we show through a case study that our approach brings several advantages and, in particular, the ability to perform segmentation that is based on the risk similarity existing between authentication events. We showed with the help of a case study on 30,000 real world authentication events that risky and non-risky events can be grouped according to similar contextual features, which can explain the risk of impersonation differently and specifically for each authentication event. The research suggests that explainable machine learning models can effectively improve our understanding of impersonation risks. We have used TreeSHAP for the implementation of our model, because of its accuracy and its availability in open-source packages. With our proposal, we aim to **help authentication policymakers and regulators** in attempt to propose the suitable additional authentication mechanisms in case of high impersonation risks. While **risk score** estimation models only provide information about the probability of impersonation, our explanations can effectively **advance the understanding of the determinants of the impersonation risk** and therefore to differentiate between **attack types**. We show within this work that a reasoning about risk types (authentication attacks) with the help of contextual information is possible and can help to choose the right authentication mechanism in case of a high risk. Our proposed methodology should be **extended** to other datasets and other risk score estimation models. We were only able to carry out our analysis on one dataset, as we do not have another available. It would be interesting for future research to work on a public dataset. We also plan to further investigate the observed clusters to detail their mapping to attack types.

REFERENCES

[1] Bonneau, J., Herley, C., Van Oorschot, P.C., Stajano, F.: The quest to replace passwords: A framework for comparative evaluation of web authentication schemes. In: 2012 IEEE Symposium on Security and Privacy. pp. 553–567. IEEE (2012)

[2] Boonen, T.J., De Waegenaere, A., Norde, H.: A generalization of the aumann–shapley value for risk capital allocation problems. European Journal of Operational Research **282**(1), 277–287 (2020)

[3] Bussmann, N., Giudici, P., Marinelli, D., Papenbrock, J.: Explainable machine learning in credit risk management. Computational Economics **57**(1), 203–216 (2021)

[4] Doerfler, P., Thomas, K., Marincenko, M., Ranieri, J., Jiang, Y., Moscicki, A., McCoy, D.: Evaluating login challenges as adefense against account takeover. In: The World Wide Web Conference. pp. 372–382 (2019)

[5] Freeman, D., Jain, S., Dürmuth, M., Biggio, B., Giacinto, G.: Who are you? a statistical approach to measuring user authenticity. In: NDSS. vol. 16, pp. 21–24 (2016)

[6] Halunen, K., Häikiö, J., Vallivaara, V.: Evaluation of user authentication methods in the gadget-free world. Pervasive and Mobile Computing **40**, 220–241 (2017)

[7] Hurkała, A., Hurkała, J.: Architecture of context-risk-aware authentication system for web environments. The Third International Conference on Informatics Engineering and Information Science (2014)

[8] Jafari, H., Shohaimi, S., Salari, N., Kiaei, A.A., Najafi, F., Khazaei, S., Niaparast, M., Abdollahi, A., Mohammadi, M.: A full pipeline of diagnosis and prognosis the risk of chronic diseases using deep learning and shapley values: The ravansar county anthropometric cohort study. Plos one **17**(1), e0262701 (2022)

[9] Ke, G.Y., Hu, X.F., Xue, X.L.: Using the shapley value to mitigate the emergency rescue risk for hazardous materials. Group Decision and Negotiation pp. 1–16 (2021)

[10] Molloy, I., Dickens, L., Morisset, C., Cheng, P.C., Lobo, J., Russo, A.: Risk-based security decisions under uncertainty. In: Proceedings of the second ACM conference on Data and Application Security and Privacy. pp. 157–168 (2012)

[11] Molnar, C.: Interpretable machine learning. Lulu. com (2020)

[12] Morris, R., Thompson, K.: Password security: A case history. Communications of the ACM **22**(11), 594–597 (1979)

[13] Tarashev, N., Tsatsaronis, K., Borio, C.: Risk attribution using the shapley value: Methodology and policy applications. Review of Finance **20**(3), 1189–1213 (2016)

[14] Velásquez, I., Caro, A., Rodríguez, A.: Kontun: A framework for recommendation of authentication schemes and methods. Information and Software Technology **96**, 27–37 (2018)

[15] Wang, D., Gu, Q., Cheng, H., Wang, P.: The request for better measurement: A comparative evaluation of two-factor authentication schemes. In: Proceedings of the 11th ACM on Asia conference on computer and communications security. pp. 475–486 (2016)

[16] Wang, H., Xu, S.: Income allocation of 'insurance+ futures' with risk modified shapley value method. Applied Economics Letters pp. 1–11 (2022)

[17] Weber, J.E., Guster, D., Safonov, P., Schmidt, M.B.: Weak password security: An empirical study. Information Security Journal: A Global Perspective **17**(1), 45–54 (2008)
[18] Wiefling, S., Dürmuth, M., Iacono, L.L.: What's in score for website users: A data-driven long-term study on risk-based authentication characteristics. arXiv preprint arXiv:2101.10681 (2021)
[19] Wiefling, S., Dürmuth, M., Lo Iacono, L.: More than just good passwords? a study on usability and security perceptions of risk-based authentication. In: Annual Computer Security Applications Conference. pp. 203–218 (2020)
[20] Wiefling, S., Iacono, L.L., Dürmuth, M.: Is this really you? an empirical study on risk-based authentication applied in the wild. In: IFIP International Conference on ICT Systems Security and Privacy Protection. pp. 134–148. Springer (2019)