Journal of Infrastructure, Policy and Development 2024, 8(13), 9311.
https://doi.org/10.24294/jipd9311

Article

Advancing user classification models: A comparative analysis of machine learning approaches to enhance faculty password policies at the University of Buraimi

Boumedyen Shannaq[1,*], Oualid Ali[2], Said Al Maqbali[1], Afraa Al-Zeidi[1]

[1] University of Buraimi, Al Buraimi, Sultanate of Oman
[2] College of Arts & Science, Applied Science University, Manama, Kingdom of Bahrain
[*] **Corresponding author:** Boumedyen Shannaq, boumedyen@uob.edu.om

CITATION

Shannaq B, Ali O, Al Maqbali S, Al-Zeidi A. (2024). Advancing user classification models: A comparative analysis of machine learning approaches to enhance faculty password policies at the University of Buraimi. Journal of Infrastructure, Policy and Development. 8(13): 9311.
https://doi.org/10.24294/jipd9311

ARTICLE INFO

Received: 25 September 2024
Accepted: 10 October 2024
Available online: 11 November 2024

Abstract: In this paper, we assess the results of experiment with different machine learning algorithms for the data classification on the basis of accuracy, precision, recall and F1-Score metrics. We collected metrics like Accuracy, F1-Score, Precision, and Recall: From the Neural Network model, it produced the highest Accuracy of 0.129526 also highest F1-Score of 0.118785, showing that it has the correct balance of precision and recall ratio that can pick up important patterns from the dataset. Random Forest was not much behind with an accuracy of 0.128119 and highest precision score of 0.118553 knit a great ability for handling relations in large dataset but with slightly lower recall in comparison with Neural Network. This ranked the Decision Tree model at number three with a 0.111792, Accuracy Score while its Recall score showed it can predict true positives better than Support Vector Machine (SVM), although it predicts more of the positives than it actually is a majority of the times. SVM ranked fourth, with accuracy of 0.095465 and F1-Score of 0.067861, the figure showing difficulty in classification of associated classes. Finally, the K-Neighbors model took the 6th place, with the predetermined accuracy of 0.065531 and the unsatisfactory results with the precision and recall indicating the problems of this algorithm in classification. We found out that Neural Networks and Random Forests are the best algorithms for this classification task, while K-Neighbors is far much inferior than the other classifiers.

Keywords: password classification; machine learning; TF-IDF vectorization; random forest; K-Nearest Neighbors (KNN); decision tree; neural network; support vector machine

1. Introduction

In today's digital age, the security of information systems is paramount. With the increasing reliance on online services and the growing complexity of cyber threats, protecting user data has become a critical challenge (Amity University Uttar Pradesh, 2024; Boumedyen and Richmond, 2017; Shannaq, 2024a). Password-based authentication remains one of the most widely used methods for securing access to sensitive information (Ugwu et al., 2024). As we have seen, password security remains one of the common security challenges since users create weak passwords, pattern of passwords that are easily guessable, and user behavior defeat traditional authentication solutions (Blessing et al., 2024). The remaining challenges in password-based systems have fostered innovation in improving authentication techniques (George, 2024; Por et al., 2024; Sheng and Umejiaku, 2024). Of them, machine learning has risen to become a significant solution for providing new ways of identifying and preventing security (Ahammed and Labu, 2024; Atadoga et al., 2024). Thus, through analyzing

the passwords and the users' behavior, the machine learning models will define such threats and, thus, will enhance the general security nature of the authentications (Akinola et al., 2024; Alshamsi et al., 2024; Al-Shamsi et al., 2024; Farhan et al., 2024; Okoli et al., 2024; Rashid Al-Shamsi and Shannaq, 2024; Shannaq and Shakir, 2024; Shannaq, 2024c; Shannaq, 2024d).

1.1. Problem statement and research gap

Thus, using passwords is really fragile as it is so easily break by using advanced hacking strategies. When an attacker has finally compromised a legitimate user's password, they may decide to modify the password to the point where the owner cannot gain access. This is usually allowed by security systems since the right password is typed hence it often is hard to differentiate between the real user and the invader. However, by using MFA, one is somehow safe but users can still override and change passwords. The main problem is that current security systems can hardly differentiate the owner of the account from an imposter if the only thing entered is the right password. In the current academic literature concerned with cybersecurity, especially user authentication, new key directions are yet to be prioritized actively research, whereas conventional approaches are dominated by password policies, MFA, and encryption techniques (Alrawili et al., 2024; Baseer and Charumathi, 2024; Das and Singh, 2024; Hasan et al., 2024; Singla and Verma, 2024; Smith et al., 2024). However, much has not been done to enhance the password security by identifying user's authorization based on password history. However, machine learning is used in cybersecurity, comparative studies of the algorithms used for the classification of users on the basis of passwords are scarce (Almujahid et al., 2024; Shi and Wang, 2024). In addition, there is little emphasis on investigating the effectiveness of different machine learning algorithms in terms of accuracy with passwords sets, essential for building secure authentication systems (Aboukadri et al., 2024; Andelić et al., 2024; Shannaq et al., 2019). Just as passwords are text data, they can be enhanced by natural language processing or NLP. Nonetheless, specific message feature engineering research on password analysis, for example, employing TF-IDF is still insignificant (Chanthati, 2024; Dias et al., 2024; Mo et al., 2024). For this reason, there lies a myriad of possibilities for development and improvement of password safety through the use of superior NLP methods.

1.2. Proposed solution

This work proposes the development of a machine learning model to analyze and understand user behavior, particularly the preferences users exhibit when choosing passwords. For example, some users might incorporate their birth year, names of family members, or favorite cars into their passwords. The proposed machine learning classification model will be trained to recognize whether the person attempting to change the password is the legitimate user or a fake one based on their password patterns and behavior. Also, the research will compare and analyze different machine learning classification algorithm and eventually arrive at which is most appropriate for the identification of the authorized and unauthorized users. In **Figure 1** there shows how the developed work of the simulator.

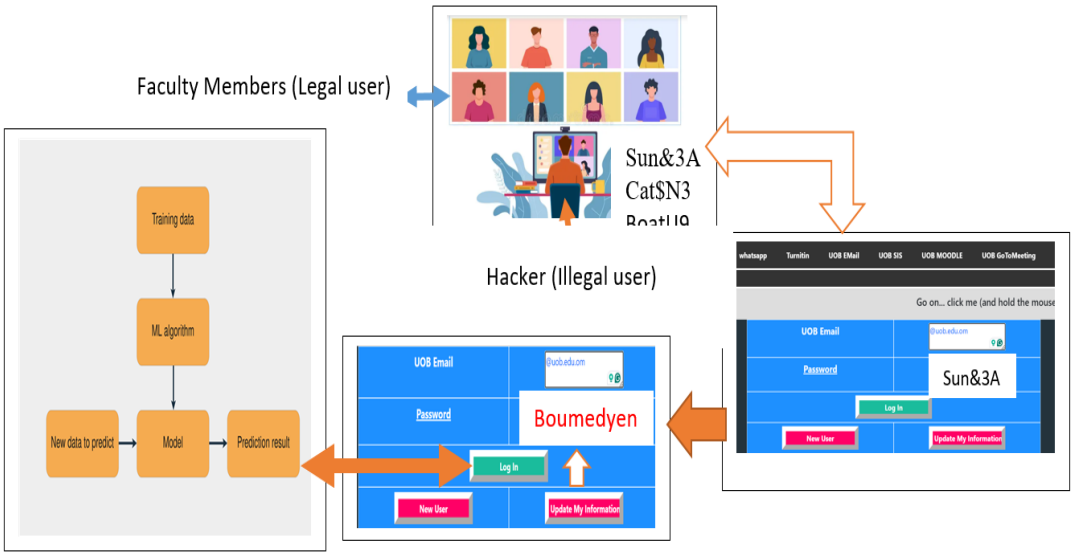

Figure 1. Simulator workflow.

This study explores the application of machine learning techniques to classify users based on their password patterns. Our research aims to fill the gap in the existing literature by providing a systematic approach to password classification, utilizing a variety of machine learning algorithms to assess their effectiveness. We focus on transforming raw password data into meaningful features using TF-IDF vectorization, a method commonly used in text analysis but not widely applied in cybersecurity contexts (Alketbi et al., 2024; Othman et al., 2024; Shannaq, 2024b).

1.3. Contribution

Lies in its potential to enhance the security of password-based systems through the integration of advanced machine learning methods. By evaluating the performance of different classifiers, we aim to identify the most effective model for this task, thereby contributing to the development of more resilient authentication mechanisms. Our findings not only address a critical need in the field of information security but also offer a foundation for future research aimed at further improving the security of user authentication processes.

1.3.1. Novel application of TF-IDF for password classification

Prior to this, we applied a Term Frequency-Inverse Document Frequency (TF-IDF) vectorizer to convert passwords to feature vectors, which is a novel contribution as the presented technique originates from text processing. This approach is useful in the analysis of password data as it keeps the essence of password information by also considering the positions of the characters and forms of the passwords.

1.3.2. Comparative evaluation of multiple ML models

We conducted a comprehensive evaluation of three different machine-learning models: is used the following algorithms Supports Vector Machine (SVM), Neural Networks (NN), Random Forest (RF), K Nearest Neighbors (KNN), Decision Tree (DT) for the classification of users from passwords. The current research makes a contribution to the comparative analysis of available models, which will help to

determine which models are most suitable for solving this particular problem and versus which they can be used as a benchmark in further studies and in implementing new practical developments in the field of cybersecurity.

1.3.3. Enhanced understanding of password patterns and user behavior

The current work therefore brings into view insights as to which patterns and features are most informative in password data by comparing how various models gravitate towards classifying users using passwords. This serves to augment the databases of user choice in password and hopefully help in improving the security of the authentication.

1.3.4. Improving cybersecurity practices

The implications of our results are evident in the way that cybersecurity professionals can adopt our recommendations. Thus, by determining which of the machine learning models can be used for user classification based on passwords, we have a further way to designing better and more secure authentication systems capable of detecting and eliminating fraudsters, or preventing unauthorized penetration.

1.3.5. Foundation for future research

Therefore, the results of this research may be used as a basis for further research of the interaction between machine learning and cybersecurity. It creates a possibility to expand knowledge about more complex issues such as higher level of feature engineering, model tuning, and application of these models in operational security environments. The comparison of models also provides a great source of reference for a researcher who would like to investigate similar applications in other domains for instance fraud detection or behavioral analysis.

1.3.6. Contributing to the body of knowledge in ML and cybersecurity

Our research enhances the literature on the use of machine learning in cyber security prevention and detection. Through filling the highlighted research gaps, the present work contributes to the further developments of the understanding of the possibilities of machine learning in regards to password protection. The addition of employing TF-IDF in password analysis and the comparison of multiple ML models make this study meaningful for both academic and practical future developments in the field.

1.3.7. Interdisciplinary impact

Since this study focuses on text data analysis and the approach developed and tested in the research can be applied to cybersecurity and possibly many other fields apart from cybersecurity such as NLP, digital forensic and behavior analytics. This interdisciplinary impact underlines the body of research as having wider applications and importance, and as such, the research will be useful to various individuals across various disciplines.

The following is the outline of the paper: A brief of password security and the use of machine learning in cybersecurity is as follows Section 2. Section 3 provide detailed method that involves the processes of data collection, feature extraction and applying the model. Section 4 contains the results of gained experiments which are based on the comparison of different classifiers. In Section 6, we revisit the findings

and discuss potential future work directions, while in Section 5, we discuss the implications of these results for the cybersecurity domain.

2. Literature review

Credentials based authentication hence have now become more at risk largely because of poor passwords, repeated passwords, and improved hacking techniques such as the phishing attacks an deficiency based attack among others. There are several ways to increase password protection, and the use of the ML approach is one of the research trends in this direction (Babar and Chen, 2024; Okoli et al., 2024).

2.1. Password security challenges

This research reveals that the use of easy to guess passwords and underrate password signals present considerable security threats. Research such as that authored by (Bonneau et al., 2012; Rooney et al., 2024; Shakir et al., 2024) establishes the fact that the password tends to be weak and vulnerable to attacks. In the same way, (Lykousas and Patsakis, 2024; Wasfi et al., 2024) established that the users choose easily guessable passwords thus adding a lot of danger to the whole authentication phase [2]. Additionally, Das and Singh (2024). looked at the risks surrounding password reuse, whereby through a single password, the attacker has access to several other accounts (Gautam et al., 2024; Lykousas and Patsakis, 2024).

2.2. Machine learning for password security

Password security has been the biggest beneficiary of machine learning as a powerful technique. According to (Okoli et al., 2024; Vanila et al., 2024), it was also hypothesized that various password datasets produced by users might be predicted by ML techniques. After this, (Atzori et al., 2024) used statistical models of relevant text and their context to estimate likely passwords, though introducing a clock scheme, illustrating that ML also might be used for password cracking as well as for password protection.

The more recent studies have therefore shifted to how the application of the ML approach can be used to identify anomalies in terms of users. For example, the work (Altulaihan et al., 2024) proposes an anomaly detection framework that utilises various ML algorithms to identify unusual patterns of login attempts that may portend a security breach and additional related work explored in (Komadina et al., 2024). Furthermore, Martín et al. (2022) proposed a system that integrates user level with password pattern to enhance the authentication techniques and as similar work done in (Papaspirou et al., 2023) have highlighted the prevalence of weak passwords and their susceptibility to attacks. Similarly, (Lykousas and Patsakis, 2024; Wasfi et al., 2024) demonstrated that users often select passwords that are easy to guess, thereby compromising the security of the authentication process [2]. Furthermore, Das and Singh (2024) explored the vulnerabilities associated with password reuse, which allows attackers to gain access to multiple accounts once a single password is compromised (Gautam et al., 2024; Lykousas and Patsakis, 2024).

2.3. Machine learning for password security

Machine learning has emerged as a potent tool for enhancing password security. (Okoli et al., 2024; Vanila et al., 2024) were among the first to suggest that ML models could predict password strength by analyzing large datasets of user-generated passwords. Following this, (Atzori et al., 2024) applied statistical models to infer likely passwords, highlighting the potential for ML to crack passwords as well as defend against such attacks.

More recent studies have focused on applying ML to detect anomalies in user behavior. For example, (Altulaihan et al., 2024) work on anomaly detection uses ML algorithms to flag unusual login attempts, which may indicate a security breach and more similar work investigated in (Komadina et al., 2024). Moreover, (Martín et al., 2022) developed a system that combines user behavior with password patterns to strengthen authentication mechanisms and more similar work investigated in (Papaspirou et al., 2023). Feature Extraction Techniques in Password Security.

Feature extraction is an essential step in the development of strong ML models (Escobar-Linero et al., 2022; Fraser et al., 2024; Ng et al., 2023; Veras et al., 2021) an attempt was made to introduce sequential characteristics by extracting n-grams as the main features which were proved to enhance the accuracy of password strength prediction models. Similarly, the study done by (Liao and Yu, 2016) on frequency analysis employed the result to derive password pattern which was utilize in feature space (Zhou et al., 2024).

TF-IDF vectorization, which is associated with text mining, is suggested for improving the process of feature extraction in cybersecurity fields (Othman et al., 2024; Pendela et al., 2024). For instance, (Harshita and Leema, 2024) implemented TF-IDF algorithm in distinguishing phishing emails—showing that this technique can be applied to various security-related tasks (Aouedi et al., 2024).

Using AI and cyber security in the evaluation of ML models the following are the outcomes of the experiments conducted with a cross-section of ML algorithms to determine the utility of the afforded procedures for improving password security. Nowadays random forests, support vector machines (SVM) and neural networks are gaining much attention from researcher. In (Shi and Wang, 2024), the authors identified that SVMs are well suited for classifying password patterns, their accuracy is high and their computational cost is relatively low. On the other hand, random forests were proved to be less sensitive to overfitting when dealing with large password databases, as mentioned in (Etzler et al., 2024; Maçãs et al., 2024).

2.4. Deep learning models

Using password datasets (Han, 2024) proposed the use of CNNs, proving that they are the best in password strength prediction. Likewise, (Kaur and Kaur, 2022) incorporated recurrent neural networks (RNNs) for better identification of sequential pattern in passwords for façade better accuracy results in password strength prediction. Similarly, (Kaur and Kaur, 2022) used recurrent neural networks (RNNs) to capture sequential dependencies in passwords, improving classification accuracy.

2.5. Challenges and future directions

However, these advancements have raised different issues as to the applicability of ML in password safety. One of the major flaws is to extend models to multiple datasets. Unfortunately, many models that tend to perform better with certain datasets may not have the same impact on other users and this is highlighted by (He and Liu, 2024; Zhou et al., 2024) that there is a want for models that can work across different datasets that are not limited in capacity. The present issue, another one, relates to the complication that entails the training of large scale ML models, including the use of deep learning models. As accurately mentioned by (Bakhtiarnia et al., 2024; Bello et al., 2024; Wang et al., 2024) deep learning models come up with high accuracy, for this, they need enough computational power that can be limiting.

2.6. Novelty and contribution of this work

Despite the abundance of work on the role of ML in password security, the current study makes several theoretical contributions.

First, we employ TF-IDF vectorization to extract feature from raw password data, which involves converting passwords from textual data into numerical vectors, a procedure usually applied to text mining but seldom investigated in the field of cybersecurity. This approach enables the capturing of details of password usage that may not be easily extracted by employing the standard approach to feature extraction.

Second, we also present a comparative study of multiple commonly used Multiple ML classifiers such as SVM, NN, DT, RF, and KNN to identify the suitable model for Password classification.

Last, we consider the generalization problem in our work by evaluating models on various password datasets. This makes it possible for us to come up with more generalized conclusions since our results will be tested under different condition. The usage of TF-IDF vectorization alongside improved model assessment opens up a new approach of subnet.

3. Methodology

The methods employed in the study investigation entitled “Comparative Analysis of Machine Learning Models for Password-Based User Classification Using TF-IDF Vectorization.” The proposed methodology provides a systematic presentation of the several phases, beginning with issue identification and culminating with the evaluation of machine learning models according to their performance. The initial stage of the process is known as Problem Identification and Objective Setting, which entails the identification of vulnerabilities in password-based systems. Specifically, unauthorized users can compromise credentials, resulting in security breaches. The main goal is to categorize computer users according to their password habits using machine learning algorithms, therefore improving security protocols. The second phase involves doing an investigation of the literature review. The current step involves a comprehensive study of pertinent research and established solutions pertaining to password categorization, user behavior analysis, and text vectorization techniques such as TF-IDF. Assessing the strengths and drawbacks of current methodologies is crucial for

establishing the basis for model development. The next step is Data Collection, which involves gathering a dataset including usernames and their corresponding passwords.

3.1. Dataset description

Figure 2 show a part of the dataset, the dataset consisted of 733 records and included two columns: usernames and their corresponding passwords (the previous passwords that had been changed and updated by each user). The total number of users was 147.

This dataset comprises of password change records from workers who have quit from an organization. For privacy and security considerations, usernames have been encoded, and no personal or identifying information is included. This anonymized dataset was utilized by a research data laboratory center, focused on research without exposing the true identity of the consumers.

Relation: testpassword Attributes: 2
Instances: 733 Sum of weights: 733

No.	Label	Count	Weight
1	USER70	4	4
2	USER71	4	4
3	USER72	7	7
4	USER73	5	5
5	USER74	3	3
6	USER75	21	21
7	USER76	4	4
8	USER77	3	3
9	USER78	6	6

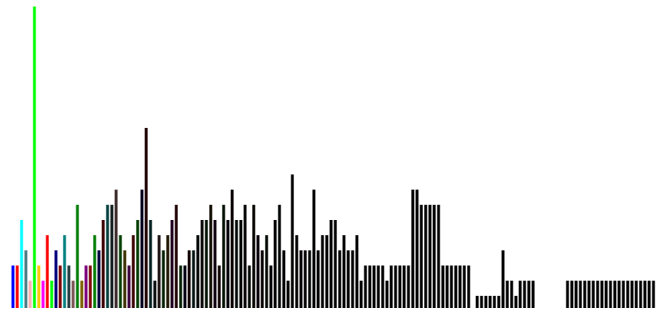

Figure 2. Sample of dataset for class users distribution.

3.1.1. The dataset's structure has two primary columns

User: Represents the encoded username of each employee (e.g., USER70, USER71, etc.), providing anonymity. Old-Password: Shows a succession of old passwords previously used by the various users. Each user contains many entries, signifying different times when they changed their passwords.

3.1.2. Analyzing password change frequency

Password change tracking by user

Each person in the sample has undergone many password changes. The changes are presented in chronological order, with the most recent password showing at the bottom of each user's list. Each user has a unique set of passwords, following various protocols. While many passwords comprise combinations of letters, numbers, and

special characters (following strong password regulations), the format and pattern varies amongst users.

Illustration of password modification frequency

USER70: Changed their password 4 times, moving from: Sun&3A → Boat#M8 → Moon*U5 → Cat$N4.

USER75: Changed their password 12 times, moving from: Cat$M3 → Dog#N6 → BBoatO4 → Moon$T8 → Sun22&N5 → Ttar#U3Y → Dog#M2 → BoatU9 → MoAAn$A6 → Sun&O4 → HHar@B8 → Cat$N3.

USER91: Changed their password 6 times, moving from: Wi-FiRouter#X4 → PowerBank$Y5 → Earbuds@Z6 → USBFlashDrive#M7 → E-reader$N8 → FitnessTracker@O9.

General patterns observation

The majority of users have changed their passwords between 4 and 8 times. The information indicates adherence to company password standards requiring unusual characters, digits, and mixed case letters. There is a wide spectrum of password difficulty, with users regularly cycling between similar patterns, typically modifying special characters or shifting between words/numbers.

3.1.3. Security implications

The information implies that password changes occurred routinely, presumably in conformity with organizational regulations or best security practices. Password modifications are vital in maintaining security, and users typically adjust only sections of the password, indicating that memorability could have been a role in their password design practices.

3.1.4. Dataset use case

This dataset gives significant insights on how employees handle password changes, frequency of updates, and probable patterns in choosing new passwords. By studying such patterns it will be possible to update security policies and enhance further user education and support as to create stronger passwords as well as reduce the discernable pat-terns in users' behavior. The information collected as part of this study was of subjects which could be utilised for research purposes without infringing on the privacy of the subject.

Subsequent to collection is the data that undergoes Preprocessing as a part of the Analysis stage. This occurs through the labeling of usernames using numerical labels as utilized in label encoding while transforming passwords into numerical vectors through the use of TF-IDF. These vectors offer an indication on the importance of password components based on the frequency of use.

Furthermore, the Model Selection and Evaluation stage involves the implementation of three distinct classification algorithms: NN, SVM, KNN, FR and DT. Therefore, an evaluation of each model is done using accuracy and classification criterion such as precision and recall.

Finally, in the Ranking phase, based on the algorithms, the models classify the people correctly by considering their password profiles. Such type of analysis helps in understanding which of the machine learning model could be suitable for given problem. **Figure 3** illustrates the stages of the developed algorithm.

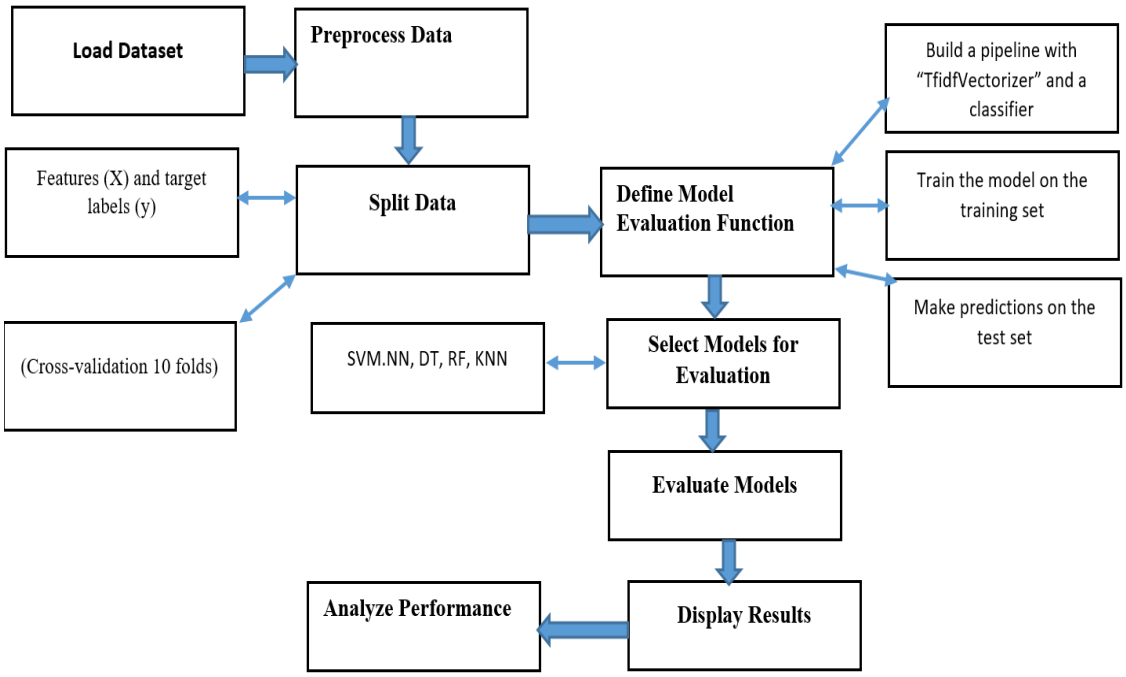

Figure 3. The proposed algorithm.

4. Experimental procedure and code

The purpose of this work was to determine the efficiency of many machine learning classification algorithms on our dataset through experiments done using Python code. The Python code was developed in a way to include not just a part of the machine learning pipeline, but several levels of its functionality, namely data pre-processing, model training and evaluation followed by sorting of the resulting models according to set performance metrics.

In the preprocessing phase, the original data set was transformed into a format that was can be fed into a machine learning algorithm. This involved converting categorical data including usernames to numerical to the form of labels through label encoding. Additionally, based on the fact that passwords could be treated as text, we proceeded by applying a TF-IDF vectorization. This allowed us to transform the passwords into numerically encoded strings, in terms of frequency and importance level of recurrence.

Following data preprocessing, many classification models were employed, including Support Vector Machines (SVM), Neural Networks (NN), Random Forest (RF), K-Nearest Neighbors (KNN) and Decision Tree (DT) classifiers. All the models were trained using cross-validation (10 folds) for all the models. This train-test split in their implementation allowed us to evaluate the models' capacity to perform on unseen data.

The evaluation for each model was made using conventional metrics, including accuracy, precision, recall, and F1-score indices. These measurements give a broad view of the effectiveness and inefficiency of the models in grouping users based on password behavior. After the models were built, they were further ranked based on their degree of dependency accuracy in order to determine the best algorithm.

Due to the limited word count feasible in an article, only a part of the code is shown as a screen shot. For the full implementation, readers can write to the authors through email and ask for the full source code. This methodology not alone saves physical space but also ensures that whoever would wish to replicate or scale up the experiments has equal access to the basic resources.

Table 1 Illustrates the steps involved in importing all necessary Python libraries for conducting the experiments, along with a description of the process for loading the dataset.

Table 1. Importing python libraries.

```
import pandas as pd
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.metrics import accuracy_score, precision_score, recall_score, f1_score
from sklearn.svm import SVC #SVM
from sklearn.ensemble import RandomForestClassifier #RF
from sklearn.neighbors import KNeighborsClassifier #KNN
from sklearn.tree import DecisionTreeClassifier #DT
from sklearn.neural_network import MLPClassifier #NN
```

Table 2. Presents the continuation of the code, which includes loading the dataset, encoding the user column, splitting the data into training and testing sets, and defining a function to create a pipeline, train the model, and evaluate its performance.

Table 2. Loading the dataset.

```
data = pd.read_csv('/content/testpassword.csv', header = 0, sep = ",")
label_encoder = LabelEncoder()
data['user'] = label_encoder.fit_transform(data['user'])
X = data['oldpassword']
y = data['user']
def evaluate_model_cv(model, X, y, cv = 10):
pipeline = Pipeline([('tfidf', TfidfVectorizer()), ('classifier', model)])
```

Table 3 Illustrates the screenshot of the Python code, showcasing the steps to train the model, predict the results, evaluate the model, define the models to be tested, and evaluate each model while storing the results.

Table 3. Code for predict the results, evaluate the model.

```
for model_name, model in models.items():
accuracy, precision, recall, f1 = evaluate_model_cv(model, X, y)
results.append({'Model': model_name, 'Accuracy': accuracy, 'Precision': precision, 'Recall': recall,
'F1-Score': f1})
```

The developed simulator in this work is presented in the **Figures 4**–**6**.

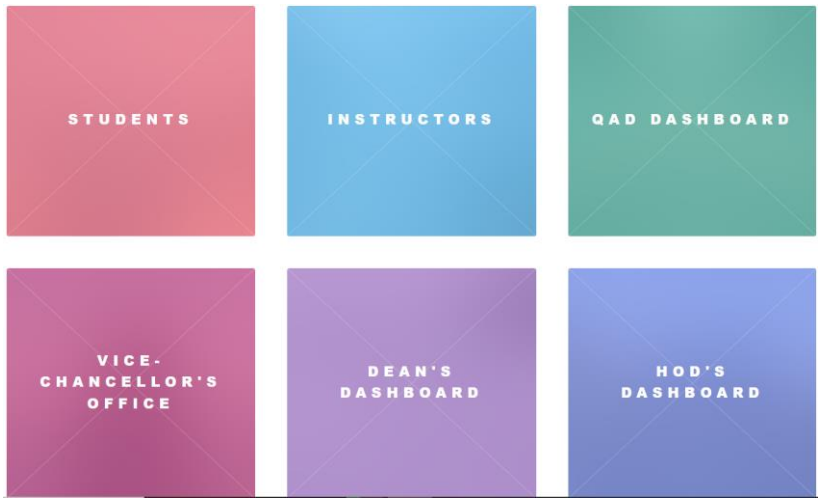

Figure 4. Main Dashboard: http://aerstore-001-site2.itempurl.com/main/index.html.

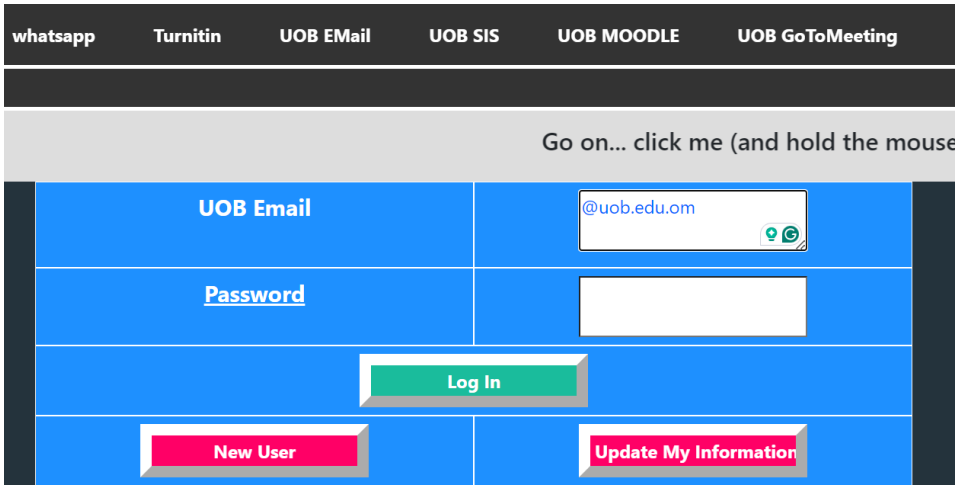

Figure 5. Login screen.

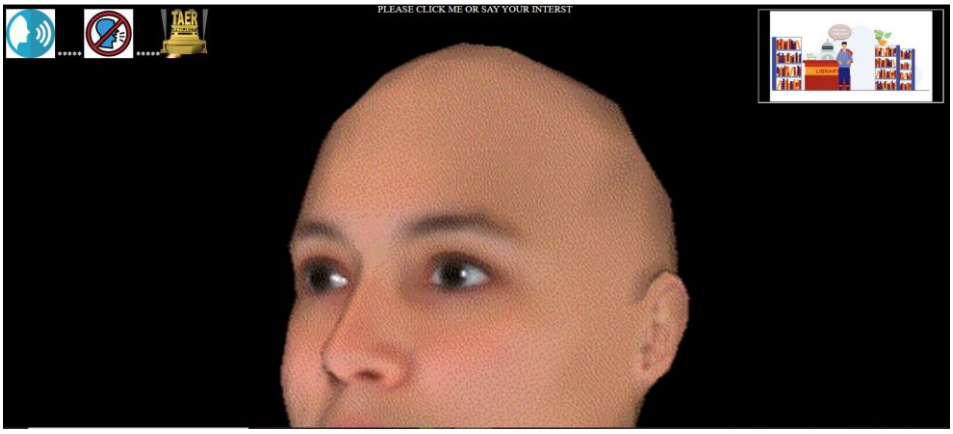

Figure 6. ML password testing: https://taerproject.com//squ/robot.aspx (developed by the author).

4.1. Rationale (python code)

- Data Loading: The dataset must be imported into a Pandas Data Frame. The data must be inserted into the user and old password columns.

- Label Encoding: The user column is encoded to numerical values as machine learning models demand numerical input.

The data is partitioned into training and testing sets by using cross-validation (10 folds).

- Evaluation Function of the Model: One function is written to establish a pipeline consisting of TfidfVectorizer for text input and the classifier. This pipeline is then used to train the model, make predictions on the test set, and assess the results using accuracy and classification reports.
- The model training process involves evaluating 5 distinct classifiers: SVM, NN, RF, KNN and DT.
- Results Printing: Individual accuracy and classification reports are generated for each model.
- Model Ranking: Ultimately, the models are evaluated according to their respective levels of accuracy.

4.2. Evaluation measures

Table 4 shows the results obtained for all 5 classification models, ranked according to their performance based on the weighted average precision measure.

Table 4. Classification model results.

Rank	Model	Accuracy	Precision	Recall	F1-Score
1	Neural Network	0.129526	0.112146	0.129545	0.118785
2	Random Forest	0.128119	0.118553	0.126805	0.120582
3	Decision Tree	0.111792	0.100776	0.118567	0.100887
4	SVM	0.095465	0.066457	0.095465	0.067861
6	K-Neighbors	0.065531	0.053219	0.065531	0.056132

Precision: Precision is the ratio of genuine positive predictions to the total number of positive predictions generated by the model. This metric measures the fraction in the overall projected positive cases that were actually positive. That means when the model has any positive value it will be quite accurate. Recall is the ratio of true positive predictions to the overall observed positives. The metric measures how effectively all the positive cases are identified by the model. It is high recall, meaning we are quite confident that the model can correctly classify most of the actual positive cases.

F1-Score: F1 score was defined as the average of Accuracy and Recall, it is the F measure of the test. This way we create a measure that combines both accuracy and recall to give a comprehensive end product. There is F1_score which will help to strike a balance between the precision and the recall and it is more useful where the cases are not split 50/50.5.

5. Results and discussion

While advancing in cybersecurity, one of the most important objectives to improve is the user authentication systems. In this study we propose the use of deep learning models especially long short-term memory (LSTM) for classification of users from their passwords while Term Frequency-Inverse Document Frequency (TF-IDF) are used to capture the textual features of passwords. The main goal is to determine

the states of the machine learning algorithms for this purpose and advance the present mechanisms of authentication.

Neural Network: It has emerged on the top owing to its highest accuracy of 0.129526 and F1-Score of 0.118785. They show that this model has the highest f1-score among all the models and that means we get the highest measure of recall by keeping as much of the precision as required. Random Forest: Close down to the second place with the accuracy of 0.128119, higher precision score is 0.118553. Although it yields lower recall than Neural Networks, the model's overall capability proves it as a contender, which can work with high-order components in the set.

Decision Tree: They are positioned third with an accuracy of 0.111792. It has produced higher recall value (0.118567) than SVM which ensure true positive formation but it's precised value reveal that it may produce more noise than top two models i.e., it could be forming more false positives than them.

SVM: Ranked fourth it has the least accuracy (0.095465) and F1-Score (0.067861) compared with the previous models. Its accuracy is relatively low, which has led to the assumption that it is inefficient at predicting the right classes.

K-Neighbors: In sixth place, the formula gave an accuracy of 065531. The built K-Neighbors model appears poorly performing in the aspects of precision and recall illustrating most actual instances to be misclassified. The ranking also signifies the considerations of precision and recall shape it, these measures both dictate the general performance of the models in classification of data. Neural Networks and Random Forests proved themselves to be the two best models in this regard, but K-Neighbors suffers from poor Prediction capabilities. The results showed in the study show that the proposed model of Neural Network performs better than the others to classify the users according to password pattern with the highest accuracy level. This means that is capable of properly addressing different classes in the password field, and more importantly, it has the ability in identifying the unique characteristics of passwords.

6. Conclusion

The obtained results determine that the Neural Network model outperforms the others in classifying users based on password patterns, achieving the maximum precision. This indicates that Neural Network is particularly effective at handling different classes of passwords and capturing their distinctive patterns. The comparison reveals that, for example, Random-Fort and Decision Tree give worse performance but the proposed approach is still helpful in the task of password-based user classification. In totality, this research demonstrates how value can be derived in the future from avant-garde machine learning methodology to improve on user authentication systems based on the ability of the system to distinguish between authentic and impostor users from their password behaviors. This research work would serve as a base to other future research works that is intended to enhance password security beyond any attack with machine learning.

Author contributions: Conceptualization, BS; methodology, BS and OA; software, BS; validation, BS, SAM and AAZ; formal analysis, BS and OA; investigation, SAM and AAZ; resources, SAM and AAZ; data curation, BS and OA; writing—original

draft preparation, BS; writing—review and editing, OA, SAM and AAZ; visualization, BS and AAZ; supervision, BS; project administration, BS; funding acquisition, BS, OA, SAM and AAZ. All authors have read and agreed to the published version of the manuscript.

Acknowledgments: The author would like to express sincere gratitude to the University of Buraimi for their invaluable support and funding of this research. This work has been made possible through the university's internal project titled "Advancing Faculty Competencies: A Theoretical Framework and Model Development for Collaborative Learning and Multicultural Faculty Twinning Utilizing an Interactive Social Media Information System." The continued support from the University of Buraimi has been instrumental in advancing research and fostering academic excellence.

Conflict of interest: The authors declare no conflict of interest.

References

Aboukadri, S., Ouaddah, A., and Mezrioui, A. (2024). Machine learning in identity and access management systems: Survey and deep dive. Computers & Security, 139, 103729. https://doi.org/10.1016/j.cose.2024.103729

Akinola, O., Akinola, A., Ifeanyi, I. V., Oyerinde, O., Adewole, O. J., Sulaimon, B., and Oyekan, B. O. (2024). Artificial Intelligence and Machine Learning Techniques for Anomaly Detection and Threat Mitigation in Cloud-Connected Medical Devices. International Journal of Scientific Research and Modern Technology (IJSRMT), 1–13. https://doi.org/10.38124/ijsrmt.v3i3.26

Alketbi, S., BinAmro, M., Alhammadi, A., and Kaddoura, S. (2024). A Comparative Study of Machine Learning Models for Classification and Detection of Cybersecurity Threat in Hacking Forum. 2024 15th Annual Undergraduate Research Conference on Applied Computing (URC), 1–6. https://doi.org/10.1109/URC62276.2024.10604519

Almujahid, N. F., Haq, M. A., and Alshehri, M. (2024). Comparative evaluation of machine learning algorithms for phishing site detection. PeerJ Computer Science, 10, e2131. https://doi.org/10.7717/peerj-cs.2131

Alrawili, R., AlQahtani, A. A. S., and Khan, M. K. (2024). Comprehensive survey: Biometric user authentication application, evaluation, and discussion. Computers and Electrical Engineering, 119, 109485. https://doi.org/10.1016/j.compeleceng.2024.109485

Al-Shamsi, I. R., Shannaq, B., Adebiaye, R., & Owusu, T. (2024). Exploring biometric attendance technology in the Arab academic environment: Insights into faculty loyalty and educational performance in policy initiatives. Journal of Infrastructure, Policy and Development, 8(9), 6991. https://doi.org/10.24294/jipd.v8i9.6991

Alshamsi, I., Sadriwala, K. F., Ibrahim Alazzawi, F. J., & Shannaq, B. (2024). Exploring the impact of generative AI technologies on education: Academic expert perspectives, trends, and implications for sustainable development goals. Journal of Infrastructure, Policy and Development, 8(11), 8532. https://doi.org/10.24294/jipd.v8i11.8532

Altulaihan, E., Almaiah, M. A., and Aljughaiman, A. (2024). Anomaly Detection IDS for Detecting DoS Attacks in IoT Networks Based on Machine Learning Algorithms. Sensors, 24(2), 713. https://doi.org/10.3390/s24020713

Amity University Uttar Pradesh. (2024). Cyber Security Threats and Countermeasures in Digital Age. Journal of Applied Science and Education (JASE), 4(1), 1–20. https://doi.org/10.54060/a2zjournals.jase.42

Andelić, N., Baressi Šegota, S., and Car, Z. (2024). Robust password security: A genetic programming approach with imbalanced dataset handling. International Journal of Information Security, 23(3), 1761–1786. https://doi.org/10.1007/s10207-024-00814-2

Aouedi, O., Vu, T.-H., Sacco, A., Nguyen, D. C., Piamrat, K., Marchetto, G., and Pham, Q.-V. (2024). A Survey on Intelligent Internet of Things: Applications, Security, Privacy, and Future Directions. IEEE Communications Surveys & Tutorials, 1–1. https://doi.org/10.1109/COMST.2024.3430368

Atadoga, A., Sodiya, E. O., Umoga, U. J., and Amoo, O. O. (2024). A comprehensive review of machine learning's role in enhancing network security and threat detection. World Journal of Advanced Research and Reviews, 21(2), 877–886. https://doi.org/10.30574/wjarr.2024.21.2.0501

Atzori, M., Calò, E., Caruccio, L., Cirillo, S., Polese, G., and Solimando, G. (2024). Evaluating password strength based on information spread on social networks: A combined approach relying on data reconstruction and generative models. Online Social Networks and Media, 42, 100278. https://doi.org/10.1016/j.osnem.2024.100278

Bakhtiarnia, A., Zhang, Q., and Iosifidis, A. (2024). Efficient High-Resolution Deep Learning: A Survey. ACM Computing Surveys, 56(7), 1–35. https://doi.org/10.1145/3645107

Baseer, S., and Charumathi, K. S. (2024). Multi-Factor Authentication: A User Experience Study. SSRN Electronic Journal. https://doi.org/10.2139/ssrn.4840295

Bello, H. O., Ige, A. B. and Ameyaw, M. N. (2024). Deep learning in high-frequency trading: Conceptual challenges and solutions for real-time fraud detection. World Journal of Advanced Engineering Technology and Sciences, 12(2), 035–046. https://doi.org/10.30574/wjaets.2024.12.2.0265

Blessing, J., Hugenroth, D., Anderson, R. J., and Beresford, A. R. (2024). SoK: Web Authentication in the Age of End-to-End Encryption (Version 1). arXiv. https://doi.org/10.48550/ARXIV.2406.18226

Bonneau, J., Herley, C., Oorschot, P. C. V., and Stajano, F. (2012). The Quest to Replace Passwords: A Framework for Comparative Evaluation of Web Authentication Schemes. 2012 IEEE Symposium on Security and Privacy, 553–567. https://doi.org/10.1109/SP.2012.44

Boumedyen, S. and Richmond, A. (2017). A Security Analysis To Be Technology Architecture for Ministry of Regional Municipalities and Water Resources (MRMWR) Sultanate of Oman. International Journal of Research in Social Sciences, 7(4), 247-258. https://www.ijmra.us/project%20doc/2017/IJRSS_APRIL2017/IJMRA-11393.pdf

Chanthati, S. R. (2024). How the power of machine – machine learning, data science and NLP can be used to prevent spoofing and reduce financial risks. Global Journal of Engineering and Technology Advances, 20(2), 100–119. https://doi.org/10.30574/gjeta.2024.20.2.0149

Chen, H., and Babar, M. A. (2024). Security for Machine Learning-based Software Systems: A Survey of Threats, Practices, and Challenges. ACM Computing Surveys, 56(6), 1–38. https://doi.org/10.1145/3638531

Escobar-Linero, E., Luna-Perejón, F., Muñoz-Saavedra, L., Sevillano, J. L., and Domínguez-Morales, M. (2022). On the feature extraction process in machine learning. An experimental study about guided versus non-guided process in falling detection systems. Engineering Applications of Artificial Intelligence, 114, 105170. https://doi.org/10.1016/j.engappai.2022.105170

Etzler, S., Schönbrodt, F. D., Pargent, F., Eher, R., and Rettenberger, M. (2024). Machine Learning and Risk Assessment: Random Forest Does Not Outperform Logistic Regression in the Prediction of Sexual Recidivism. Assessment, 31(2), 460–481. https://doi.org/10.1177/10731911231164624

Farhan, Y. H., Shakir, M., Tareq, M. A., & Shannaq, B. (2024). Incorporating Deep Median Networks for Arabic Document Retrieval Using Word Embeddings-Based Query Expansion. Journal of Information Science Theory and Practice, 12(3), 36–48. https://doi.org/10.1633/JISTAP.2024.12.3.3

Fraser, W., Broadbent, M., Pitropakis, N., and Chrysoulas, C. (2024). Examining the Strength of Three Word Passwords. In N. Pitropakis, S. Katsikas, S. Furnell, & K. Markantonakis (Eds.), ICT Systems Security and Privacy Protection (Vol. 710, pp. 119–133). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-65175-5_9

Gautam, A., Yadav, T. K., Seamons, K., and Ruoti, S. (2024). Passwords Are Meant to Be Secret: A Practical Secure Password Entry Channel for Web Browsers (Version 1). arXiv. https://doi.org/10.48550/ARXIV.2402.06159

George A. S. (2024). The Dawn of Passkeys: Evaluating a Passwordless Future. https://doi.org/10.5281/ZENODO.10697886

Hagui, I., Msolli, A., Ben Henda, N., Helali, A., Gassoumi, A., Nguyen, T. P., and Hassen, F. (2023). A blockchain-based security system with light cryptography for user authentication security. Multimedia Tools and Applications, 83(17), 52451–52480. https://doi.org/10.1007/s11042-023-17643-5

Han, J. (2024). CNN-Based Multi-Factor Authentication System for Mobile Devices Using Faces and Passwords. Applied Sciences, 14(12), 5019. https://doi.org/10.3390/app14125019

Harshita, B., and Leema, N. (2024). ESD: E-mail Spam Detection using Cybersecurity-Driven Header Analysis and Machine Learning based Content Analysis. International Journal of Performability Engineering, 20(4), 205. https://doi.org/10.23940/ijpe.24.04.p2.205213

Hasan, M. K., Weichen, Z., Safie, N., Ahmed, F. R. A., and Ghazal, T. M. (2024). A Survey on Key Agreement and Authentication Protocol for Internet of Things Application. IEEE Access, 12, 61642–61666. https://doi.org/10.1109/ACCESS.2024.3393567

Kaur, K., and Kaur, P. (2022). SABDM: A self-attention based bidirectional-RNN deep model for requirements classification. Journal of Software: Evolution and Process, e2430. https://doi.org/10.1002/smr.2430

Komadina, A., Kovačević, I., Štengl, B., and Groš, S. (2024). Comparative Analysis of Anomaly Detection Approaches in Firewall Logs: Integrating Light-Weight Synthesis of Security Logs and Artificially Generated Attack Detection. Sensors, 24(8), 2636. https://doi.org/10.3390/s24082636

Labu, R. and Ahammed, F. (2024). Next-Generation Cyber Threat Detection and Mitigation Strategies: A Focus on Artificial Intelligence and Machine Learning. Journal of Computer Science and Technology Studies, 6(1), 179–188. https://doi.org/10.32996/jcsts.2024.6.1.19

Liu, Z., and He, K. (2024). A Decade's Battle on Dataset Bias: Are We There Yet? (Version 1). arXiv. https://doi.org/10.48550/ARXIV.2403.08632

Lykousas, N., and Patsakis, C. (2024). Decoding developer password patterns: A comparative analysis of password extraction and selection practices. Computers & Security, 145, 103974. https://doi.org/10.1016/j.cose.2024.103974

Maçãs, C., Campos, J. R., Lourenço, N., and Machado, P. (2024). Visualisation of Random Forest classification. Information Visualization, 14738716241260745. https://doi.org/10.1177/14738716241260745

Manthiramoorthy, C., Khan, K. M. S., and A, N. A. (2023). Comparing Several Encrypted Cloud Storage Platforms. International Journal of Mathematics, Statistics, and Computer Science, 2, 44–62. https://doi.org/10.59543/ijmscs.v2i.7971

Martín, A. G., de Diego, I. M., Fernández-Isabel, A., Beltrán, M., and Fernández, R. R. (2022). Combining user behavioural information at the feature level to enhance continuous authentication systems. Knowledge-Based Systems, 244, 108544. https://doi.org/10.1016/j.knosys.2022.108544

Mo, Y., Li, S., Dong, Y., Zhu, Z. and Li, Z. (2024). Password Complexity Prediction Based on RoBERTa Algorithm. https://doi.org/10.5281/ZENODO.11180356

Ng, C. K., Al-Quraishi, T., and Souza-Daw, T. D. (2023). Application of Sequential Analysis on Runtime Behavior for Ransomware Classification. Applied Data Science and Analysis, 2023, 126–142. https://doi.org/10.58496/ADSA/2023/012

Norman, D., Mouleeswaran, S. K, and Reeja, S. R. (2024). Natural language processing and stable diffusion model based graphical authentication using passphrase. Intelligent Decision Technologies, 18(2), 935–951. https://doi.org/10.3233/IDT-230279

Okoli, U. I., Obi, O. C., Adewusi, A. O., and Abrahams, T. O. (2024). Machine learning in cybersecurity: A review of threat detection and defense mechanisms. World Journal of Advanced Research and Reviews, 21(1), 2286–2295. https://doi.org/10.30574/wjarr.2024.21.1.0315

Othman, R., Rossi, B., and Barbara, R. (2024). A Comparison of Vulnerability Feature Extraction Methods from Textual Attack Patterns (Version 2). arXiv. https://doi.org/10.48550/ARXIV.2407.06753

Papaspirou, V., Papathanasaki, M., Maglaras, L., Kantzavelou, I., Douligeris, C., Ferrag, M. A., and Janicke, H. (2023). A Novel Authentication Method That Combines Honeytokens and Google Authenticator. Information, 14(7), 386. https://doi.org/10.3390/info14070386

Pendela, N. P. S., Janet, K. A., Yadav, A. M. R., Subramanyam, C. B., Hariharan, S., and Kekreja, V. (2024). Enhancing Cyberbullying Detection: A Multi-Algorithmic Approach. 2024 International Conference on Advances in Data Engineering and Intelligent Computing Systems (ADICS), 1–5. https://doi.org/10.1109/ADICS58448.2024.10533585

Por, L. Y., Ng, I. O., Chen, Y.-L., Yang, J., and Ku, C. S. (2024). A Systematic Literature Review on the Security Attacks and Countermeasures Used in Graphical Passwords. IEEE Access, 12, 53408–53423. https://doi.org/10.1109/ACCESS.2024.3373662

Rashid Al-Shamsi, I., & Shannaq, B. (2024). Leveraging clustering techniques to drive sustainable economic innovation in the India–Gulf interchange. Cogent Social Sciences, 10(1), 2341483. https://doi.org/10.1080/23311886.2024.2341483

Rooney, M. J., Levy, Y., Li, W., and Kumar, A. (2024). Comparing experts' and users' perspectives on the use of password workarounds and the risk of data breaches. Information & Computer Security. https://doi.org/10.1108/ICS-05-2024-0116

Shakir, M., Farsi, M. J. A., Al-Shamsi, I. R., Shannaq, B., and Taufiq-Hail, G. A.-M., (2024). The Influence of Mobile Information Systems Implementation on Enhancing Human Resource Performance Skills: An Applied Study in a Small

Organization. International Journal of Interactive Mobile Technologies (iJIM), 18(13), 37–68. https://doi.org/10.3991/ijim.v18i13.47027

Shannaq, B. (2024a). Digital Formative Assessment as a Transformative Educational Technology. In K. Arai (Ed.), Advances in Information and Communication (Vol. 921, pp. 471–481). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-54053-0_32

Shannaq, B. (2024b). Unveiling the Nexus: Exploring TAM Components Influencing Professors' Satisfaction With Smartphone Integration in Lectures: A Case Study From Oman. TEM Journal, 2365–2375. https://doi.org/10.18421/TEM133-63

Shannaq, B. (2024c). Enhancing Human-Computer Interaction: An Interactive and Automotive Web Application - Digital Associative Tool for Improving Formulating Search Queries. In K. Arai (Ed.), Advances in Information and Communication (Vol. 921, pp. 511–523). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-54053-0_35

Shannaq, B. (2024d). Novel Algorithm for Differentiating Authorized Users from Fraudsters by Analyzing Mobile Keypad Input Patterns during Password Updates. TEM Journal, 13(4)

Shannaq, B., Shamsi, I. A., and Majeed, S. N. A. (2019). Management Information System for Predicting Quantity Martial's. TEM Journal, 8, 1143–1149. https://doi.org/10.18421/TEM84-06

Shannaq, B., Talab, M. A., Shakir, M., Sheker, M. T., and Farhan, A. M. (2023). Machine learning model for managing the insider attacks in big data. 020013. https://doi.org/10.1063/5.0188358

Shannaq, B., Adebiaye, R., Owusu, T., & Al-Zeidi, A. (2024). An intelligent online human-computer interaction tool for adapting educational content to diverse learning capabilities across Arab cultures: Challenges and strategies. Journal of Infrastructure, Policy and Development, 8(9), 7172. https://doi.org/10.24294/jipd.v8i9.7172

Shannaq, B., & Shakir, M. (2024). Enhancing Security with Multi-Factor User Behavior Identification Via Longest Common Subsequence Analysis. Informatica 48 (2024) 73–82 73, 48(16), 73–82. https://doi.org/10.31449/inf.v48i19.6529

Shi, Y., and Wang, Y. (2024). A Comparative Work to Highlight the Superiority of Mouth Brooding Fish (MBF) over the Various ML Techniques in Password Security Classification. International Journal of Advanced Computer Science and Applications, 15(5). https://doi.org/10.14569/IJACSA.2024.0150520

Singh, N., and Das, A. K. (2024). TFAS: Two factor authentication scheme for blockchain enabled IoMT using PUF and fuzzy extractor. The Journal of Supercomputing, 80(1), 865–914. https://doi.org/10.1007/s11227-023-05507-6

Singla, D., and Verma, N. (2024). Performance Analysis of Authentication System: A Systematic LiteratureReview. Recent Advances in Computer Science and Communications, 17(7), e121223224363. https://doi.org/10.2174/0126662558246531231121115514

Smith, L., Prior, S., and Ophoff, J. (2024). Investigating the Accessibility and Usability of Multi-factor Authentication for Young People. In C. Stephanidis, M. Antona, S. Ntoa, & G. Salvendy (Eds.), HCI International 2024 Posters (Vol. 2119, pp. 129–135). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-61966-3_15

Ugwu, C., Ukwandu, E., Ofusori, L., Ezugwu, A., Ome, U., Ezema, M., and Ndunagu, J. (2024). Factors Influencing The Experiences of End-users in Password-Based Authentication System. https://doi.org/10.21203/rs.3.rs-4438584/v1

Umejiaku, A. P., and Sheng, V. S. (2024). RoseCliff Algorithm: Making Passwords Dynamic. Applied Sciences, 14(2), 723. https://doi.org/10.3390/app14020723

Vanila, S., Jeyavathana, B., Rathinam, A., and Elango, K. (2024). Enhancing Password Security With Machine Learning-Based Strength Assessment Techniques: In J. A. Ruth, V. G. V. Mahesh, P. Visalakshi, R. Uma, & A. Meenakshi (Eds.), Advances in Information Security, Privacy, and Ethics (pp. 296–314). IGI Global. https://doi.org/10.4018/979-8-3693-4159-9.ch018

Veras, R., Collins, C., and Thorpe, J. (2021). A Large-Scale Analysis of the Semantic Password Model and Linguistic Patterns in Passwords. ACM Transactions on Privacy and Security, 24(3), 1–21. https://doi.org/10.1145/3448608

Wang, Y., Han, Y., Wang, C., Song, S., Tian, Q., and Huang, G. (2024). Computation-efficient deep learning for computer vision: A survey. Cybernetics and Intelligence, 1–24. Cybernetics and Intelligence. https://doi.org/10.26599/CAI.2024.9390002

Wasfi, H., Stone, R., and Genschel, U. (2024). Word-Pattern: Enhancement of Usability and Security of User-Chosen Recognition Textual Password. International Journal of Advanced Computer Science and Applications, 15(6). https://doi.org/10.14569/IJACSA.2024.0150605

Yu, X., and Liao, Q. (2016). User password repetitive patterns analysis and visualization. Information & Computer Security, 24(1), 93–115. https://doi.org/10.1108/ICS-06-2015-0026

Zhou, D.-W., Cai, Z.-W., Ye, H.-J., Zhan, D.-C., and Liu, Z. (2024). Revisiting Class-Incremental Learning with Pre-Trained Models: Generalizability and Adaptivity are All You Need. International Journal of Computer Vision. https://doi.org/10.1007/s11263-024-02218-0

Zhou, E., Peng, Y., Shao, G., Deng, F., Miao, Y., and Fan, W. (2024). Password cracking using chunk similarity. Future Generation Computer Systems, 150, 380–394. https://doi.org/10.1016/j.future.2023.09.013