Electronic and Cyber Defense
Vol. 13, No.4, 2026 (Serial No. 52) .pp.57-72
ISSN: 2322-4347 , E-ISSN: 2980-8979

Enhancing Digital Password Security Using Machine Learning: A Comparative Analysis of Classification Algorithms

Mahnaz Doroudi[1], Seyed Hasan Mortazavi[2*], Fatemeh Zare Mehrjardi[3], Mohsen Sardari Zarchi[4]

[1] MSc. Student, Faculty of Technology and Engineering, University of Meybod, Meybod, Iran, Email: mahnazdorodi8@gmail.com
[2] Assistant Professor, Faculty of Technology and Engineering, University of Meybod, Meybod, Iran, (*Correspondence: hassanmortazavi@meybod.ac.ir)
[3] Assistant Professor, Faculty of Technology and Engineering, University of Meybod, Meybod, Iran, Email: fzare@meybod.ac.ir
[4] Associate Professor, Faculty of Technology and Engineering, University of Meybod, Meybod, Iran, Email: sardari@meybod.ac.ir

ARTICLE INFO

Article history:
Article Type: Research paper
Received: 16 April 2025
Revised: 23 June 2025
Accepted: 3 September 2025
Available online: 29 December 2025

Keywords:
Machine Learning, Cybersecurity, Strength Password, Random Forest.

ABSTRACT

In the digital age, passwords remain one of the primary methods of authentication in information systems. Despite their critical role in protecting personal and organizational data, the use of weak passwords such as those that are overly simple, short, or repetitive poses a significant threat to cybersecurity. This study proposes a machine learning-based approach to enhance the accuracy of password security assessment. The proposed method consists of three main steps: preprocessing, new feature extraction and password classification. In this study, a dataset with 669,880 passwords is analyzed. First, in the preprocessing step, operations such as removing missing data, encoding text to number, and fixing the problem of unbalanced classes using smote are performed. After data cleaning, 10 new features are extracted from each password. Finally, the processed dataset is partitioned for training and testing operations and strength classification of passwords are predicted using various machine learning classifiers such as Decision tree, Random Forest, XGBoost, Logistic regression and others. Experimental results shows that the Random Forest and XGBoost algorithms achieved the highest F1-score with values of 99.665% and 99.642% respectively rather than other models. The outcome of this research is a scalable and efficient framework for identifying password vulnerabilities and reinforcing security policies in digital systems. This framework can play a key role in designing robust passwords and mitigating the risk of unauthorized access.

his article: M. Doroudi, S. H. Mortazavi, F. Zare Mehrjardi, M. Sardari zarchi, "Enhancing Digital Password Security using Machine Learning: A Comparative Analysis of Classification Algorithms," Journal of Electronic and Cyber Defense, vol. 13(4), pp. 57-72, **DOI:** https://doi.org/10.47176/ECDJ.2025.16

Publisher: Imam Hossein University.

1. Introduction

Passwords are the most fundamental authentication tool in information systems, making their security a critical concern [1]. With increasing cyber threats, attackers employ sophisticated techniques to breach password protection [2]. The quality of user-chosen passwords significantly impacts system security, yet users struggle to create passwords that are both memorable and resistant to attacks [3]. Behavioral factors like using common terms and reusing passwords exacerbate this weakness [4]. Traditional password strength evaluation methods are insufficient, necessitating more advanced solutions. Machine learning (ML) offers a promising approach, enabling analysis of complex features and identification of hidden patterns in passwords [5]. ML algorithms can classify passwords by considering length, character composition, sequential patterns, and structural complexity [6]. This study aims to design an intelligent framework for classifying password strength using ML algorithms. The main contributions include:

- Utilizing a large dataset of 669,880 labeled passwords
- Extracting ten novel structural features,
- Evaluating seven ML classifiers,
- Applying SMOTE for class balancing,
- Presenting a deployable framework for real-world authentication systems.

2. Methodology

The proposed methodology for identifying password strength consists of three main stages: preprocessing, feature extraction, as explained in the following subsections.

2.1. Da taset Description

This research utilizes a publicly available dataset from Kaggle, comprising 669,880 records. Each record contains two primary features: the password as a text string and its strength as an integer label, categorized into three classes: weak (0), medium (1), and strong (2). Sample passwords include "123456" (weak), "kzde5577" (medium), and "StrongP@ss!1" (strong). This dataset was chosen for its high volume, sample diversity, and accurate labeling, providing a suitable foundation for password security research.

2.2. Data Preprocessing

The preprocessing phase began with data cleaning by removing any missing or null values to ensure data quality. Since machine learning algorithms cannot process raw text directly, each password was converted into a numerical format using ASCII encoding. In this method, every character in a password is mapped to its corresponding ASCII code, transforming each password into a list of numerical values that represent its constituent characters.

Initial analysis of the dataset revealed a significant class imbalance problem. The 'medium' strength class contained considerably more samples (496,801) compared to the 'weak' class (89,701) and the 'strong' class (83,137). This imbalance could bias machine learning models toward the majority class, leading to poor generalization for minority classes. To address this issue, the Synthetic Minority Over-sampling Technique (SMOTE) was applied. SMOTE generates

synthetic samples for the minority classes by interpolating between existing samples and their nearest neighbors, rather than simply duplicating existing data. This approach creates new, realistic password samples for the weak and strong classes, resulting in a perfectly balanced dataset with an equal number of samples (496,801) for all three strength categories. This balancing ensures that the trained models learn equally from all classes and generalize better to unseen data.

2.3. Feature Extraction

To enrich the dataset and improve model learning, ten novel structural features were extracted from each password. These features were specifically designed to capture the complexity, patterns, and distinguishing characteristics that differentiate weak, medium, and strong passwords. The extracted features include:

- Password length: Total number of characters in the password, as longer passwords are generally more secure.
- Uppercase count: Number of uppercase letters, indicating character diversity.
- Lowercase count: Number of lowercase letters.
- Digit count: Number of numeric characters (0-9).
- Special character count: Number of non-alphanumeric characters such as !, @, #, $, etc.
- Palindrome property: A boolean feature indicating whether the password reads the same forwards and backwards (ignoring case), which can indicate weakness.
- Sequential property: A boolean feature detecting whether the password contains sequential patterns like "123", "abc", or "qwerty" that are easily guessable.
- Repetition property: A boolean feature identifying whether any character repeats within the password, as excessive repetition weakens security.
- Unique character count: Number of distinct characters in the password, measuring character diversity.
- ASCII encoding: The numerical representation of the password as a list of ASCII values, preserving the original character sequence information.

For example, for the password "AaBbCc123!!!121", analysis yields: length of 15 characters, 3 uppercase letters (A, B, C), 3 lowercase letters (a, b, c), 6 digits (1,2,3,1,2,1), 3 special characters (!, !, !), palindrome property is false (the string does not read the same backwards), sequential property is false (although "123" appears, the entire password lacks consistent sequential order), repetition property is true (characters 1, 2, and ! repeat), and unique character count is 10 (A, a, B, b, C, c, 1, 2, 3, !). These features collectively provide a comprehensive representation of password structure and complexity.

2.4. Classification

After feature extraction, the dataset was split into training and testing sets using an 80/20 ratio, with 80% of the data used for model training and 20% reserved for final evaluation. This split ensures that models are trained on sufficient data while maintaining a separate unseen dataset

for unbiased performance assessment. Then seven even different machine learning classifiers such as Logistic Regression (LR), K-Nearest Neighbors (KNN), AdaBoost, Decision Tree (DT), Gradient Boosting (GB), XGBoost, and Random Forest (RF) were employed to predict password strength, representing a diverse range of algorithmic approaches. Finally model performance was evaluated using standard classification metrics: Accuracy (overall correct predictions), Precision (positive predictive value), Recall (sensitivity), and F1-score (harmonic mean of precision and recall). These metrics provide a comprehensive assessment of each model's classification capability across all three password strength categories.

3. Results and Discussion

Table 1 presents the performance comparison of all models. Ensemble methods significantly outperformed simpler models. Random Forest achieved the highest F1-score of 99.665%, followed closely by XGBoost with 99.642%. Logistic Regression showed the weakest performance (92.67%) due to its inability to model non-linear relationships.

Table 1. Performance comparison of classifiers

Algorithm	Accuracy (%)	F1-Score (%)
Random Forest	99.672	99.665
XGBoost	99.653	99.642
Gradient Boosting	99.650	99.640
Decision Tree	99.560	99.553
KNN	98.380	98.340
AdaBoost	97.760	97.750
Logistic Regression	92.670	93.50

Table 2 compares our results with previous studies. Our Random Forest model achieved 99.672% accuracy, substantially outperforming prior approaches (83.4% to 94.1%). This improvement validates the effectiveness of our feature engineering and SMOTE balancing.

Table 2. Comparison with previous studies

Study	Algorithm	Accuracy (%)

Proposed method	Random Forest	99.672
Proposed method	XGBoost	99.653
Jiao et al. [7]	RF & Naive Bayes	83.4
Nosenko et al. [8]	RNN	83.0
Demenongo et al. [9]	RF & Naive Bayes	94.1
Mallet et al. [10]	SVM, RF, KNN	82.0
Pryor et al. [11]	RF, SVM, Logistic regression	86.0
Anwer et al. [12]	SVM, RF, DT	85.34

The superior performance of ensemble methods stems from their ability to combine multiple weak learners, resist noise, prevent overfitting, and capture complex non-linear patterns in password data.

4. Conclusion

This research successfully developed a machine learning framework for password strength classification. Using ten structural features and SMOTE balancing, ensemble methods particularly Random Forest (99.665% F1-score) and XGBoost (99.642% F1-score) achieved near-perfect accuracy, significantly outperforming prior studies. The framework offers high accuracy, scalability, and deployability for real-world authentication systems. Limitations include reliance on quantitative features and lack of semantic analysis. Future work should explore hybrid approaches incorporating semantic analysis and deep learning architectures.

«پدافند الکترونیک و سایبری»
سال سیزدهم، شماره ۴، زمستان ۱۴۰۴؛ ص ۷۲-۵۷
شاپا الکترونیکی: ۸۹۷۹-۲۹۸۰ شاپا چاپی: ۴۳۴۷-۲۳۲۲

افزایش امنیت گذرواژه دیجیتالی با استفاده از یادگیری ماشین: تحلیلی مقایسه‌ای بر الگوریتم‌های طبقه‌بندی

مهناز درودی[۱]، سید حسن مرتضوی[۲*]، فاطمه زارع مهرجردی[۳]، محسن سرداری زارچی[۴]

[۱] دانشجوی کارشناسی ارشد، دانشکده فنی و مهندسی، دانشگاه میبد، میبد، ایران mahnazdorodi8@gmail.com
[۲] استادیار، دانشکده فنی و مهندسی، دانشگاه میبد، میبد، ایران (نویسنده مسئول) hassanmortazavi@ meybod.ac.ir
[۳] استادیار، دانشکده فنی و مهندسی، دانشگاه میبد، میبد، ایران fzare@meybod.ac.ir
[۴] دانشیار، دانشکده فنی و مهندسی، دانشگاه میبد، میبد، ایران sardari@meybod.ac.ir

چکیده

در عصر دیجیتال، رمزهای عبور یکی از رایج‌ترین و حیاتی‌ترین روش‌های احراز هویت در سامانه‌های اطلاعاتی هستند. بااین‌حال، انتخاب و استفاده از رمزهای عبور ضعیف و غیر ایمن، همچون رمز عبورهای ساده، کوتاه یا تکراری، تهدیدی جدی برای امنیت سایبری محسوب می‌شود. این پژوهش باهدف بهبود دقت در ارزیابی سطح امنیت رمزهای عبور، چارچوبی مبتنی بر یادگیری ماشین ارائه می‌دهد. روش پیشنهادی شامل سه مرحله‌ی اصلی پیش‌پردازش، استخراج ویژگی‌های جدید و طبقه‌بندی رمزهای عبور است. در این پژوهش، مجموعه‌ای شامل ۶۶۹۸۸۰ رمز عبور مورد تحلیل قرارگرفته است. ابتدا در مرحله پیش‌پردازش مراحلی مانند حذف داده‌های گم‌شده، کدگذاری متن به عدد و متعادل‌سازی داده‌های کلاس‌ها با استفاده از الگوریتم SMOTE انجام‌شده است. بعدازاینکه داده‌ها تمیز شدند از هر رمز عبور ده ویژگی ساختاری جدید استخراج‌شده است. درنهایت داده‌ها به دودسته آموزش و آزمایش تقسیم‌شده و دسته‌بندی سطح امنیتی رمزهای عبور در سه کلاس ضعیف، متوسط و قوی با استفاده از الگوریتم‌های یادگیری ماشین مانند درخت تصمیم، جنگل تصادفی، XGBoost و رگرسیون لجستیک و غیره صورت گرفته است. نتایج تجربی نشان داد که الگوریتم جنگل تصادفی و روش XGBoost با معیار F1 به ترتیب با مقدار ۹۹,۶۶۵٪ و ۹۹,۶۴۲٪ درصد عملکرد برتری نسبت به سایر مدل‌ها داشته‌اند. دستاورد این تحقیق، ارائه مدلی کارآمد و مقیاس‌پذیر برای شناسایی نقاط ضعف رمزهای عبور و ارتقاء سیاست‌های امنیتی در سامانه‌های دیجیتال است که می‌تواند به شکل مؤثری در طراحی رمزهای عبور مقاوم، کاهش ریسک نفوذ، و ارتقاء سطح امنیت سایبری نقش‌آفرینی کند.

مشخصات مقاله

تاریخچه مقاله:
نوع مقاله: علمی پژوهشی
دریافت: ۱۴۰۴/۰۷/۲۷
بازنگری: ۱۴۰۴/۰۸/۳۰
پذیرش: ۱۴۰۴/۰۹/۲۰
ارائه آنلاین: ۱۴۰۴/۱۰/۰۸

کلیدواژه‌ها:
یادگیری ماشین، امنیت سایبری، قدرت رمز عبور، جنگل تصادفی.
...

استناد: درودی، مهناز، مرتضوی، سید حسن، زارع مهرجردی، فاطمه، سرداری زارچی، محسن. افزایش امنیت گذرواژه با استفاده از یادگیری ماشین: تحلیل مقایسه‌ای بر الگوریتم‌های طبقه‌بندی. پدافند الکترونیک وسایبری. ۱۳(۴). ۷۲-۵۷

DOI https://doi.org/10.47176/ECDJ.2025.1652

ناشر: دانشگاه جامع امام حسین(ع).

۱- مقدمه

در دنیای دیجیتال امروزی، گذرواژه یا رمزهای عبور همچنان به‌عنوان اساسی‌ترین و پرکاربردترین ابزار احراز هویت در سامانه‌های اطلاعاتی شناخته می‌شوند و امنیت آن‌ها دغدغه‌ای جدی در حوزه فناوری اطلاعات و ارتباطات به شمار می‌رود [۱]. با توسعه روزافزون خدمات اینترنتی و هم‌زمان با افزایش تهدیدات سایبری، حملات هدفمند علیه رمزهای عبور نیز رشد یافته و مهاجمان از فن‌های پیچیده‌تری برای نفوذ بهره می‌برند [۲-۴]. در این میان، انتخاب و استفاده از گذرواژه‌های قوی و غیرقابل حدس نقش تعیین‌کننده‌ای در جلوگیری از دسترسی‌های غیرمجاز دارد، چراکه افشای اطلاعات احراز هویت می‌تواند پیامدهایی جبران‌ناپذیر برای کاربران و زیرساخت‌های حیاتی به همراه داشته باشد [۵].

بخش بزرگی از امنیت کلی یک سامانه، وابسته به کیفیت گذرواژه انتخاب‌شده توسط کاربر است. بااین‌حال، مطالعات نشان می‌دهند که کاربران در ایجاد گذرواژه‌هایی که در عین سهولت در یادآوری، از مقاومت کافی در برابر حملات برخوردار باشند، با چالش مواجه‌اند [۶]. برای مثال، بیش از ۶۹ درصد کاربران از گذرواژه‌های خود به‌صورت مکرر در چندین حساب کاربری استفاده می‌کنند که این رفتار، ریسک آسیب‌پذیری را به‌شدت افزایش می‌دهد. این در حالی است که گذرواژه‌های متنی هنوز هم پرکاربردترین روش احراز هویت محسوب می‌شوند و ضعف‌های ذاتی آن‌ها نظیر قابلیت حدس زنی با استفاده از اطلاعات عمومی (مانند تاریخ تولد، الگوهای صفحه‌کلید و واژه‌های رایج فرهنگ لغت) آن‌ها را به هدفی سهل‌الوصول برای مهاجمان تبدیل می‌کند [۷].

عوامل رفتاری کاربران نیز به این ضعف‌ها دامن زده‌اند؛ استفاده از اصطلاحات تکراری [۸]، بهره‌گیری مکرر از گذرواژه‌های پیشین [۹] و تمایل به انتخاب گذرواژه‌هایی باقابلیت حفظ آسان [۱۰و۱۱] ازجمله عادات نادرست کاربران هستند که موجب کاهش امنیت کلی گذرواژه‌ها می‌شوند.

تحقیقات اولیه موریس و تامپسون [۱۲] که طی یک مطالعه گسترده ۳۲۸۹ گذرواژه را بررسی کردند، نشان داد که بالغبر ۸۹ درصد آن‌ها ضعیف بوده‌اند. آدام و ساس نیز در مطالعات خود [۱۳] اثبات کردند که اعمال محدودیت‌های امنیتی بر انتخاب گذرواژه ممکن است به‌جای ارتقاء امنیت، منجر به کاهش قابلیت یادآوری و افزایش افشای آن شود. در همین راستا، برخی پژوهش‌ها به بررسی راهکارهای جایگزین مانند استفاده از رمزهای گرافیکی [۱۴و۱۵] یا تحلیل الگوهای رفتاری کاربران [۱۶و۱۷] پرداخته‌اند، اما نتایج نشان داده است که بدون ابزارهای کمکی هوشمند، همچنان سطح امنیت پایین باقی می‌ماند. لذا رویکردهای سنتی در ارزیابی قدرت گذرواژه دیگر کافی نبوده و نیاز به راهکارهای پیشرفته‌تر و تطبیق‌پذیرتر به‌شدت احساس می‌شود.

یکی از راه‌کارهای نوین و مؤثر در این حوزه، بهره‌گیری از یادگیری ماشین است. الگوریتم‌های یادگیری ماشین باقابلیت تحلیل ویژگی‌های پیچیده و شناسایی الگوهای پنهان در داده‌ها، می‌توانند در فرآیندهایی همچون ارزیابی قدرت گذرواژه، تولید رمزهای ایمن، تشخیص رمزهای تکراری و حتی بازیابی گذرواژه‌های فراموش‌شده به کار گرفته شوند [۱۸و ۱۹]. این الگوریتم‌ها قادرند با در نظر گرفتن شاخص‌هایی مانند طول گذرواژه، ترکیب کاراکترها، توالی منطقی، تنوع و پیچیدگی ساختاری، گذرواژه‌ها را به‌طور دقیق طبقه‌بندی کنند و از خطاهای انسانی جلوگیری نمایند [۲۰و۲۱].

هدف این مطالعه، طراحی و پیاده‌سازی چارچوبی دقیق و هوشمند برای شناسایی و طبقه‌بندی قدرت امنیتی گذرواژه‌ها با استفاده از الگوریتم‌های یادگیری ماشین است. این چارچوب با بهره‌گیری از ویژگی‌های ساختاری رمز عبور و تکنیک‌های نوین پیش‌پردازش داده، قادر است قدرت گذرواژه‌ها را با دقت بالا پیش‌بینی کند و به‌عنوان ابزاری پشتیبان برای سیستم‌های احراز هویت مدرن مورداستفاده قرار گیرد.

مدل‌های یادگیری ماشین قادرند با بهره‌گیری از ویژگی‌های ساختاری رمزهای عبور، آن‌ها را به‌طور دقیق در سه سطح ضعیف، متوسط و قوی تفکیک کنند. به‌کارگیری فرآیندهای پیش‌پردازش همچون نرمال‌سازی داده‌ها و متعادل‌سازی کلاس‌ها نقش مؤثری در ارتقای عملکرد این مدل‌ها دارد، به‌طوری‌که استخراج دقیق ویژگی‌هایی نظیر طول، تنوع کاراکترها و الگوهای تکراری می‌تواند به‌عنوان عوامل کلیدی در بهبود دقت طبقه‌بندی مورداستفاده قرار گیرد.

در این پژوهش، چندین مشارکت علمی و نوآوری ارائه‌شده است که به ارتقاء دانش موجود در حوزه امنیت گذرواژه‌ها و یادگیری ماشین منجر می‌گردد:

استفاده از یک مجموعه داده گسترده و واقعی شامل ۶۶۹۸۸۰ نمونه گذرواژه دارای برچسب قدرت (ضعیف، متوسط، قوی)، که تنوع ساختاری قابل‌توجهی را در رمزهای عبور پوشش می‌دهد و اعتبار نتایج حاصل از مدل‌های یادگیری را تضمین می‌نماید.

طراحی و استخراج ده ویژگی ساختاری و معنادار از رمزهای عبور ازجمله طول، تنوع کاراکترها، میزان تکرار، الگوهای ترتیبی و پالین‌دروم یا متقارن بودن، که باهدف تقویت توانایی تفکیک

مدل‌های طبقه‌بندی طراحی‌شده‌اند و نقش کلیدی در افزایش دقت پیش‌بینی ایفا می‌کنند.

به‌کارگیری و ارزیابی چندین الگوریتم پیشرفته یادگیری ماشین شامل جنگل تصادفی، XGBoost، درخت تصمیم و رگرسیون لجستیک، به‌منظور مقایسه عملکرد آن‌ها در طبقه‌بندی قدرت گذرواژه‌ها و شناسایی مدل‌های بهینه انجام‌شده است.

استفاده از تکنیک[1] SMOTE به‌منظور رفع مشکل عدم توازن کلاس‌ها در مجموعه داده و افزایش دقت و تعمیم‌پذیری مدل‌ها از طریق تولید نمونه‌های مصنوعی در کلاس‌های اقلیت صورت گرفته است.

ارائه یک چارچوب جامع، مقیاس‌پذیر و عملیاتی برای تحلیل و ارزیابی خودکار قدرت گذرواژه‌ها، که می‌تواند به‌عنوان زیرساختی کاربردی در سامانه‌های احراز هویت و سیاست‌گذاری امنیتی جهت جلوگیری از انتخاب رمزهای ضعیف مورد بهره‌برداری قرار گیرد.

در ادامه در بخش دوم این تحقیق، مطالعات پیشین بررسی‌شده است. سپس در بخش سوم، روش پیشنهادی و در بخش چهارمقاله نتایج روش پیشنهادی آورده شده است. درنهایت در بخش پنجم، نتیجه‌گیری از تحقیق ارائه‌شده است.

۲- مطالعات پیشین

مطالعات پیشین تلاش‌های متعددی را برای بهبود امنیت رمزهای عبور از طریق روش‌های مختلف ارائه کرده‌اند. به‌عنوان نمونه، در پژوهش [۲۲] از منطق فازی به‌عنوان ابزاری جهت تقویت امنیت گذرواژه‌ها استفاده‌شده است که نتایج مطلوبی در کاهش آسیب‌پذیری‌ها به همراه داشته است.

جیائو و همکاران [۲۳] با بهره‌گیری از الگوریتم جنگل تصادفی، عملکرد این مدل را در زمینه تشخیص نفوذ مورد ارزیابی قراردادند. آن‌ها به‌دقت ۸۳/۴ درصد دست یافتند که نشان‌دهنده ارتباط معنادار میان ویژگی‌های امنیتی و قابلیت تشخیص نفوذ است. هم‌چنین آن‌ها بر نقش آگاهی امنیتی کاربران تأکید کردند و سهم آن را حدود ۶ درصد از عوامل تأثیرگذار بر امنیت کلی سیستم برآورد کردند.

در مطالعه‌ای دیگر، آبدرابو و همکاران [۲۴] با تحلیل رفتار ۴۹ کاربر در هنگام ایجاد حساب کاربری در دو بستر متفاوت (ایمیل و وب‌سایت خبری)، مدل‌های یادگیری ماشین را برای تشخیص استفاده مجدد از گذرواژه‌ها بررسی کردند. دقت مدل‌ها در این زمینه به ترتیب ۸۷/۷ درصد با استفاده از تحلیل نگاه، ۸۸/۷۵ درصد با تحلیل رفتار نوشتاری، و با ترکیب هر دو، به بیش از ۸۹ درصد رسید.

پارکینسون و همکاران [۲۵] به بررسی روش‌های احراز هویت بر پایه بیومتریک ضربه‌های کلید پرداختند. آن‌ها با جمع‌آوری داده‌هایی از ۴۲ کاربر و ۴۰ رمز عبور مختلف، نشان دادند که سیستم طراحی‌شده با هفت ویژگی زمانی و فشاری می‌تواند دقتی بین ۸۹ درصد تا ۹۴ درصد داشته باشد، درحالی‌که نرخ خطای مساوی آن بین ۶ درصد تا ۱۱ درصد متغیر بوده است. هم‌چنین مشخص شد که رمزهای کوتاه‌تر در این سامانه عملکرد بهتری دارند.

در راستای ترکیب روش‌های سنتی و نوین، سوروتی و همکاران [۲۶] سامانه‌ای مبتنی بر بیومتریک دست‌نویس را معرفی کردند که در آن کاربران به‌جای واردکردن عدد با صفحه‌کلید، اعداد را روی صفحه لمسی می‌نویسند. آزمایش این سیستم با استفاده از دیتاست e-BioDigit نشان داد که نرخ خطابه حدود ۰/۴ درصد کاهش‌یافته و این روش در مقایسه با سامانه‌های سنتی دقت بالاتری دارد.

اسحاق و همکاران [۲۷] نیز الگوریتم جنگل تصادفی را برای پیش‌بینی وب‌سایت‌های فیشینگ به کاربردند. آن‌ها دقت پیش‌بینی مدل را حدود ۶۶/۶۶ درصد گزارش کردند و علت دقت پایین‌تر را هم‌بستگی بالای برخی ویژگی‌ها دانستند که در آینده نیازمند بهینه‌سازی خواهد بود.

نوسینکو و همکاران [۲۸] با استفاده از شبکه عصبی بازگشتی روشی جدید برای پیش‌بینی رمزهای عبور مبتنی بر متن ارائه دادند. آن‌ها با بررسی مجموعه داده‌ای شامل ۲۸/۸ میلیون کاربر و ۶۱/۵ میلیون گذرواژه، به دقتی حدود ۸۳ درصد در کمتر از پنج تلاش دست یافتند. این پژوهش به‌عنوان نخستین مطالعه در استفاده از شبکه عصبی بازگشتی برای پیش‌بینی لحظه‌ای گذرواژه‌ها شناخته می‌شود.

درنهایت، الورشاسی و همکاران [۲۹] مدلی ترکیبی شامل الگوریتم‌های بیز ساده و جنگل تصادفی را برای شناسایی گذرواژه‌های ضعیف و فریب‌های اینترنتی توسعه دادند. آن‌ها با استفاده از رأی‌گیری میان مدل‌ها به دقتی بیش از ۹۰ درصد دست یافتند که نشان از کارایی بالای روش ترکیبی در مقابله با تهدیدات امنیتی دارد. به‌طورکلی، شواهد حاکی از آن است که الگوریتم‌هایی نظیر جنگل تصادفی، شبکه‌های عصبی بازگشتی، و مدل‌های بیومتریک، توان بالایی در تحلیل و پیش‌بینی رمزهای عبور و هم‌چنین شناسایی تهدیدات مرتبط با آن‌ها دارند.

باوجود تلاش‌های گسترده در تحقیقات پیشین برای ارتقاء امنیت گذرواژه‌ها از طریق منطق فازی، تحلیل‌های بیومتریک، شبکه‌های

[1] Synthetic Minority Over-sampling Technique(SMOTE)

عصبی و مدل‌های ترکیبی، بسیاری از این رویکردها با محدودیت‌هایی مواجه بوده‌اند. برخی روش‌ها دقت کافی در تفکیک گذرواژه‌های ضعیف و قوی را ارائه نمی‌دهند [۲۷]، برخی دیگر نیازمند تجهیزات خاص یا داده‌های رفتاری پیچیده هستند که قابلیت پیاده‌سازی عمومی در محیط‌های واقعی را کاهش می‌دهد [۲۵و۲۶]، و گروهی از آن‌ها فاقد مقیاس‌پذیری لازم برای پردازش داده‌های حجیم و متنوع‌اند [۲۳و۲۸]. هم‌چنین، اکثر مطالعات تمرکز خود را صرفاً بر جنبه‌هایی خاص از امنیت گذرواژه مانند شناسایی یا تولید رمزها گذاشته‌اند، بدون آن‌که چارچوبی یکپارچه برای تحلیل جامع قدرت رمز عبور ارائه دهند.
در مقابل، روش پیشنهادی این پژوهش با اتکا به یک مجموعه داده بزرگ و واقعی، استخراج ویژگی‌های ساختاری هدفمند، بهره‌گیری از الگوریتم‌های قدرتمند و قابل تفسیر همچون جنگل تصادفی و XGBoost، و اعمال تکنیک‌های پیش‌پردازش پیشرفته مانند SMOTE توانسته است مدلی دقیق، عملیاتی و قابل‌تعمیم برای طبقه‌بندی قدرت گذرواژه‌ها ارائه دهد. این رویکرد نه‌تنها از منظر دقت عملکردی از بسیاری از روش‌های پیشین برتر است، بلکه قابلیت پیاده‌سازی آسان در سامانه‌های واقعی و کمک به تدوین سیاست‌های امنیتی را نیز داراست.

۳- روش تحقیق

در پژوهش جاری برای شناسایی قدرت رمز عبورها از روش‌های مبتنی بر هوش مصنوعی و الگوریتم‌های موجود در یادگیری ماشین استفاده‌شده است. ویژگی‌های اصلی رمز عبورها تولید و تحلیل‌شده‌اند تا مدلی برای طبقه‌بندی قدرت آن‌ها ایجاد شود.
پس از ارزیابی چندین الگوریتم، مدل برتر به‌عنوان راهکار نهایی برای پیش‌بینی قدرت رمز عبورها پیشنهادشده است. این رویکرد با ترکیب ویژگی‌های متعدد، امکان ارزیابی جامع‌تری از امنیت رمز عبورها را فراهم کرده و درنهایت، به بهبود کلی امنیت کاربران کمک می‌کند. شکل (۱) خلاصه‌ای از روش پیشنهادی را نمایش می‌دهد. در ادامه مراحل روش پیشنهادی با توضیحات کامل آورده شده است.

شکل (۱). مراحل روش پیشنهادی

۳-۱. مجموعه داده‌های موردبررسی

در این پژوهش، از مجموعه داده‌ای معتبر و پرکاربرد که از وب‌سایت Kaggle گردآوری‌شده است، استفاده‌شده است. این مجموعه داده شامل ۶۶۹۸۸۰ رکورد است که هر رکورد متشکل از دو ویژگی اصلی است. ستون Password که شامل مقادیر متنی از نوع رشته‌ای بوده و بیانگر گذرواژه‌های کاربران است، و ستون Strength که نمایانگر سطح قدرت هر گذرواژه از نوع عدد صحیح است و در سه کلاس ضعیف (۰)، متوسط (۱) و قوی (۲) دسته‌بندی‌شده است. قدرت گذرواژه معیاری است برای سنجش میزان امنیت یک رمز عبور و نشان می‌دهد که چقدر یک رمز عبور می‌تواند در برابر حملات سایبری و تلاش‌های غیرمجاز برای دسترسی به‌حساب کاربری مقاومت کند. در جدول (۱) تعدادی از رمزهای عبور و قدرت آن‌ها آورده شده است.
هدف اصلی استفاده از این مجموعه داده، ارزیابی و طبقه‌بندی قدرت گذرواژه‌ها با استفاده از الگوریتم‌های یادگیری ماشین است. در گام ابتدایی، داده‌ها بارگذاری شده و مقادیر خالی یا ناقص و گم‌شده حذف شدند تا کیفیت داده‌ها برای تحلیل افزایش یابد. سپس، بر اساس برچسب قدرت گذرواژه، داده‌ها به سه گروه مجزا تقسیم گردیدند. این دسته‌بندی، مبنای اصلی برای آموزش و ارزیابی مدل‌های طبقه‌بندی قدرت گذرواژه به شمار می‌رود. مجموعه داده مذکور به دلیل حجم بالا، تنوع نمونه‌ها، و برچسب‌گذاری دقیق، بستری مناسب برای پژوهش‌های مرتبط با امنیت رمز عبور و تحلیل رفتار کاربران در انتخاب گذرواژه فراهم می‌سازد.

جدول (۱). نمایش تعدادی رمز عبور و قدرت آن‌ها

ردیف	گذرواژه	میزان قدرت گذرواژه
۱	kzde5577	۱
۲	abcd1234	۰
۳	StrongP@ss!1	۲
۴	123456	۰
۵	lamborghin1	۱

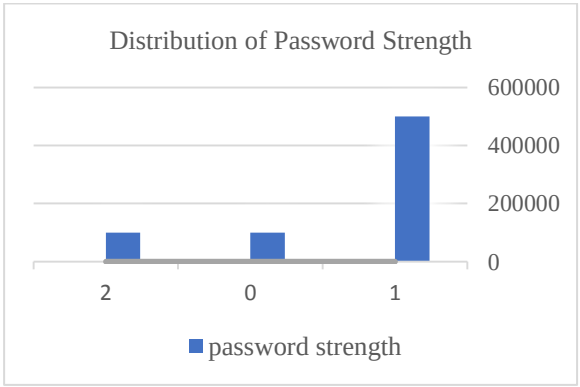
با توجه به جدول (۱) می‌توان دریافت که رمزهای عبوری که شامل اعداد یا حروف به‌صورت ترتیبی هستند، دارای قدرت ضعیف هستند. برای رمزهای عبور دارای ترکیبی از حروف و اعداد به‌صورت غیر ترتیبی قدرت متوسط و برای رمزهای عبوری که شامل هر سه کاراکتر حروف و اعداد غیر ترتیبی و علائم باشند قدرت قوی در نظر گرفته‌شده است.

۳-۲. تبدیل متن به عدد

تبدیل داده‌های متنی به فرم عددی یکی از مراحل اساسی در پیش‌پردازش داده‌ها برای الگوریتم‌های یادگیری ماشین به شمار می‌رود. زیرا این الگوریتم‌ها قادر به پردازش مستقیم متن نیستند و تنها با داده‌های عددی و ساختارمند سازگارند. هرچند متن برای انسان قابل‌درک است، اما برای تحلیل و یادگیری توسط ماشین‌ها باید به شکل عددی بازنمایی شود. در این پژوهش، از روش کدگذاری ASCII استفاده‌شده است. در کدگذاری ASCII، هر کاراکتر به یک عدد خاص نگاشت می‌شود، درنتیجه هر رمز عبور به‌صورت فهرستی از کدهای ASCII کاراکترهای تشکیل‌دهنده‌ی آن تبدیل می‌شود. این نوع کدگذاری به ساده‌سازی فرایند یادگیری کمک کرده و بستر لازم را برای آموزش مدل‌های طبقه‌بندی فراهم می‌سازد.

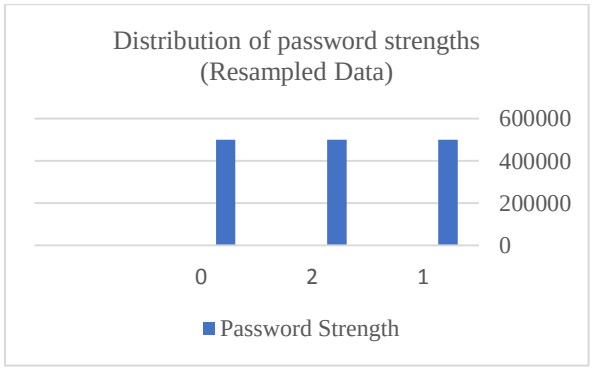
۳-۳. متعادل‌سازی داده‌ها

بررسی اولیه مجموعه داده، مطابق با آنچه در شکل (۲) مشاهده می‌شود، نشان می‌دهد که توزیع نمونه‌ها در میان سه کلاس قدرت گذرواژه (ضعیف، متوسط و قوی) نامتوازن بوده و کلاس متوسط دارای بیشترین تعداد نمونه است. این عدم تعادل میان کلاس‌ها می‌تواند منجر به سوگیری مدل یادگیری ماشین به‌سوی کلاس غالب شود. ازاین‌رو، لازم است پیش از آموزش مدل، این مسئله با به‌کارگیری روش‌های متناسب متعادل‌سازی برطرف گردد.



شکل (۲). نمایی از نامتعادل بودن کلاس‌های مجموعه داده

در این بخش، به‌منظور رفع مشکل عدم توازن میان کلاس‌ها، از تکنیک افزایش نمونه‌برداری استفاده‌شده است. در این روش، با افزودن نمونه‌های بیشتر به کلاس‌های دارای تعداد کمتر (کلاس‌های اقلیت)، تلاش می‌شود تعادل آماری بین کلاس‌های مختلف برقرار گردد و از تمایل مدل به کلاس غالب جلوگیری شود. افزایش نمونه‌برداری می‌تواند از طریق تکرار ساده داده‌ها یا تولید مصنوعی داده‌های جدید صورت گیرد. ازجمله روش‌های رایج در این زمینه می‌توان به افزایش نمونه‌برداری ساده، روش آداسین و روش اسموت اشاره کرد [۳۰و۳۱]. در این پژوهش، از روش اسموت که مخفف عبارت افزایش نمونه‌برداری مصنوعی کلاس اقلیت است، بهره گرفته شده است. در این روش، داده‌های جدید با استفاده از ویژگی‌های نمونه‌های موجود در کلاس اقلیت و بر پایه شباهت با نزدیک‌ترین همسایگان آن‌ها تولید می‌شود. شکل (۳) نمایی از متعادل بودن داده‌های کلاس‌ها پس از استفاده از روش Smote را نشان می‌دهد.



شکل (۳). نمایی ازمتعادل بودن کلاس‌ها با روش smote

این فرایند موجب تنوع بیشتر در داده‌های اقلیت شده و به مدل اجازه می‌دهد تا الگوهای متنوع‌تری را از این کلاس‌ها بیاموزد. فرایند تعادل‌سازی با روش اسموت بر روی داده‌ها اعمال شده و توزیعی متعادل‌تر میان سه کلاس قدرت گذرواژه ایجاد کرده است. اکنون قدرت یا امنیت رمز در هر ۳ دسته‌ی ضعیف، متوسط و قوی متعادل و یکسان شده است. جدول (۲) و (۳) تعداد نمونه‌های هر کلاس را قبل و بعد از فرآیند متعادل‌سازی با روش smote را نمایش داده است.

جدول (۲). تعداد نمونه‌ها قبل از متعادل‌سازی

تعداد	قدرت گذرواژه
۴۹۶۸۰۱	۱
۸۹۷۰۱	۰
۸۳۱۳۷	۲

جدول (۳). تعداد نمونه‌ها قبل از متعادل‌سازی

تعداد	قدرت گذرواژه
۴۹۶۸۰۱	۱

۴۹۶۸۰۱	۰
۴۹۶۸۰۱	۲

۳-۴. ایجاد ویژگی‌های جدید

پس از مرحله پیش‌پردازش، ۱۰ ویژگی ساختاری مهم نظیر طول گذرواژه، تعداد حروف بزرگ و کوچک، تعداد ارقام، کاراکترهای ویژه و سایر شاخص‌های مرتبط، از ستون اصلی گذرواژه استخراج و به مجموعه داده افزوده شده‌اند. استخراج این ویژگی‌ها با هدف غنی‌سازی داده و ارتقای کیفیت آموزش مدل‌های یادگیری ماشین صورت گرفته است. افزودن چنین ویژگی‌هایی به‌طور مؤثر به الگوریتم‌های یادگیری ماشین کمک می‌کند تا با بهره‌گیری از الگوهای پیچیده‌تر و روابط پنهان موجود در ساختار گذرواژه‌ها، پیش‌بینی‌ها و طبقه‌بندی‌های دقیق‌تری انجام دهند. این امر در نهایت می‌تواند موجب ارتقای سطح امنیت سامانه‌های احراز هویت و کاهش ریسک‌های مرتبط با انتخاب رمز عبور در سطح کاربران و سازمان‌ها گردد [۳۶-۳۲].

برای درک بهتر، ابتدا هر ویژگی جدید تعریف شده و سپس بر روی رمز عبور " AaBbCc123!!!121 " ویژگی‌های مربوطه استخراج شده است.

Password_encoded: این ویژگی گذرواژه را به فرم عددی یا کدگذاری شده اسکی تبدیل می‌کند تا برای مدل‌های یادگیری ماشین قابل استفاده باشد.

Length: این ویژگی تعداد کل کاراکترها (حروف، اعداد، فاصله و نشانه‌ها) در یک گذرواژه را نشان می‌دهد.

Uppercase: تعداد حروف بزرگ در یک گذرواژه را نشان می‌دهد.

Lowercase: تعداد حروف کوچک در یک گذرواژه را نشان می‌دهد.

Digits: این ویژگی نشان‌دهنده مجموعه‌ای از کاراکترهای عددی (۰-۹) در یک گذرواژه است.

Special: این ویژگی به معنای خاص یا منحصر به فرد بودن یک شی، ویژگی، یا عملکرد اشاره دارد و آن را از بقیه متمایز می‌کند. نشانه های موجود در مثال ذکرشده برابر با ۳ (! ,! ,!) است.

Is-palindrome: این ویژگی بررسی می‌کند که آیا یک رشته بدون در نظر گرفتن فاصله‌ها، علائم نگارشی و حروف بزرگ، به صورت یکسان از ابتدا و انتها خوانده می‌شود یا نمی‌شود. مثال ذکرشده بدون علائم بزرگ و کوچک "aabbcc123121" بدین صورت است، در نتیجه این مثال پالین‌دروم نیست و نتیجه False است.

Is-sequential: این ویژگی به معنای این است که داده‌ها به صورت ترتیب‌دار و دنباله‌وار (زمانی یا منطقی) هستند، به طوری که ترتیب و توالی داده‌ها اهمیت دارند و تغییر در ترتیب می‌تواند معنای آن‌ها را تغییر دهد. بررسی ترتیب منطقی یا عددی در مثال ذکرشده نشان می‌دهد که اعداد "123" به صورت ترتیبی هستند، اما کل رشته ترتیب منطقی یا یکنواخت ندارد پس نتیجه False است.

Has-repeat: این ویژگی نشان می‌دهد که آیا کلمات یا عبارات خاصی در گذرواژه بیشتر از یک بار تکرار شده‌اند یا خیر.

در مثال ذکرشده کاراکترهای ! ,2 ,1 چند بار تکرار شده‌اند پس نتیجه True است.

Num-unique-chars: این ویژگی به تعداد کاراکترهای منحصر به فرد (یعنی کاراکترهای غیر تکراری) در یک گذرواژه اشاره دارد. این ویژگی به مدل کمک می‌کند تا تنوع کاراکترها در متن را اندازه‌گیری کرده و برای تحلیل‌های بیشتر مانند تشخیص الگو یا شبیه‌سازی گذرواژه استفاده شود.

کاراکترهای یکتای موجود در مثال A, a, B, b, C, c, 1, 2, 3, ! است درنتیجه تعداد یکتاها ۱۰ است.

جدول (۴) ویژگی‌های استخراج شده از یک مثال را نمایش می‌دهد.

جدول (۴). مثالی از نمایش ویژگی های استخراج شده از رمز عبور

رمزعبور: AaBbCc123!!!121	
ویژگی	مقدار
Password encoded	[65, 97, 66, 98, 67, 99, 49, 50, 51, 33, 33, 33, 49, 50, 49]
Length	15
Uppercase	3
Lowercase	3
Digits	6
Special	3
Is-palindrome	False
Is-sequential	False
Has-repeat	True
Num-unique-chars	10

۳-۵. تقسیم داده‌ها و طبقه‌بندی آنها

پس از اینکه برای هر رمزعبور ویژگی‌های معرفی شده استخراج شد، داده‌ها به دو بخش آموزش و آزمایش تقسیم شده‌اند. بطوری‌که ۲۰ درصد از داده‌ها برای آزمایش و ۸۰ درصد برای آموزش در نظرگرفته شده است. در نهایت برای شناسایی قدرت رمزعبورها از روش‌های مختلف یادگیری ماشین مانند رگرسیون لجستیک، نزدیک‌ترین همسایگی، آدابوست، درخت تصمیم، گرادیان بوستینگ، XGBoost و جنگل تصادفی استفاده شده است.

۴. نتایج

در این پژوهش با توجه به شکل (۱) ابتدا بر روی پایگاه داده مورد استفاده پیش‌پردازش‌های لازم اعمال شده است. سپس از

هر رمزعبور ۱۰ ویژگی جدید ذکر شده در جدول (۴) استخراج شده است. در نهایت داده‌ها با نرخ ۸۰ به ۲۰ به دو بخش آموزش[۱] و آزمایش[۲] تقسیم‌بندی شده و طبقه‌بندی قدرت رمزعبورها توسط روش‌های مختلف یادگیری ماشین انجام شده است. در این پژوهش، برای پیاده‌سازی روش پیشنهادی از کتابخانه pandas برای پردازش پایگاه داده و تمیزسازی و از کتابخانه Scikit-learn برای پیاده‌سازی طبقه‌بندی الگوریتم‌های یادگیری ماشین استفاده شده است. اجرای تمام کدها بر روی محیط Colab و با GPU انجام شده است. برای ارزیابی مدل‌ها از معیارهای ارزیابی استاندارد مانند دقت[۳]، صحت[۴]، فراخوان[۵] و معیار F1[۶] [۳۷] استفاده شده است و نتایج به‌دست‌آمده در شکل شماره (۴) خلاصه‌شده است.

با توجه به نتایج به دست آمده می‌توان دریافت که از میان طبقه‌بندهای مختلف یادگیری ماشین، روش جنگل تصادفی و روش XGBoost هر دو بر پایه درخت تصمیم بنا شده‌اند و به دلیل استفاده از یادگیری دسته جمعی[۷] عملکرد بهتری نسبت به دیگر طبقه‌بندها داشته است[۳۸و۳۹]. الگوریتم جنگل تصادفی از تعداد زیادی درخت تصمیم با عمق کم تشکیل شده است. در این الگوریتم هر درخت می‌تواند به صورت تصادفی برخی از ویژگی‌ها را در برداشته و در نتیجه برخی از نمونه‌ها را دسته‌بندی کند. در نهایت دسته‌بندی نهایی بر اساس رای‌گیری[۸] اکثریت بین همه درخت‌ها برای همه داده‌ها انجام می‌شود. عملکرد الگوریتم جنگل تصادفی باعث شده تا در برابر داده‌های نویزدار، بیش‌برازش[۹] و پایگاه‌داده‌های بزرگ همراه با ویژگی‌های گم شده مقاومت و عملکرد بالایی داشته باشد.

روش XGBoost الگوریتم قدرتمند دیگری است که بر پایه درخت‌های تصمیم بنا شده و از تکنیک تقویت گرادیان برای دستیابی به دقت و کارایی بالا بهره می‌برد. عملکرد برتر XGBoost را می‌توان به ویژگی‌هایی نظیر استفاده از گرادیان بوستینگ، تکنیک‌های منظم‌سازی برای جلوگیری از بیش‌برازش، قابلیت مدیریت داده‌های گم‌شده و پردازش موازی کارآمد نسبت داد.

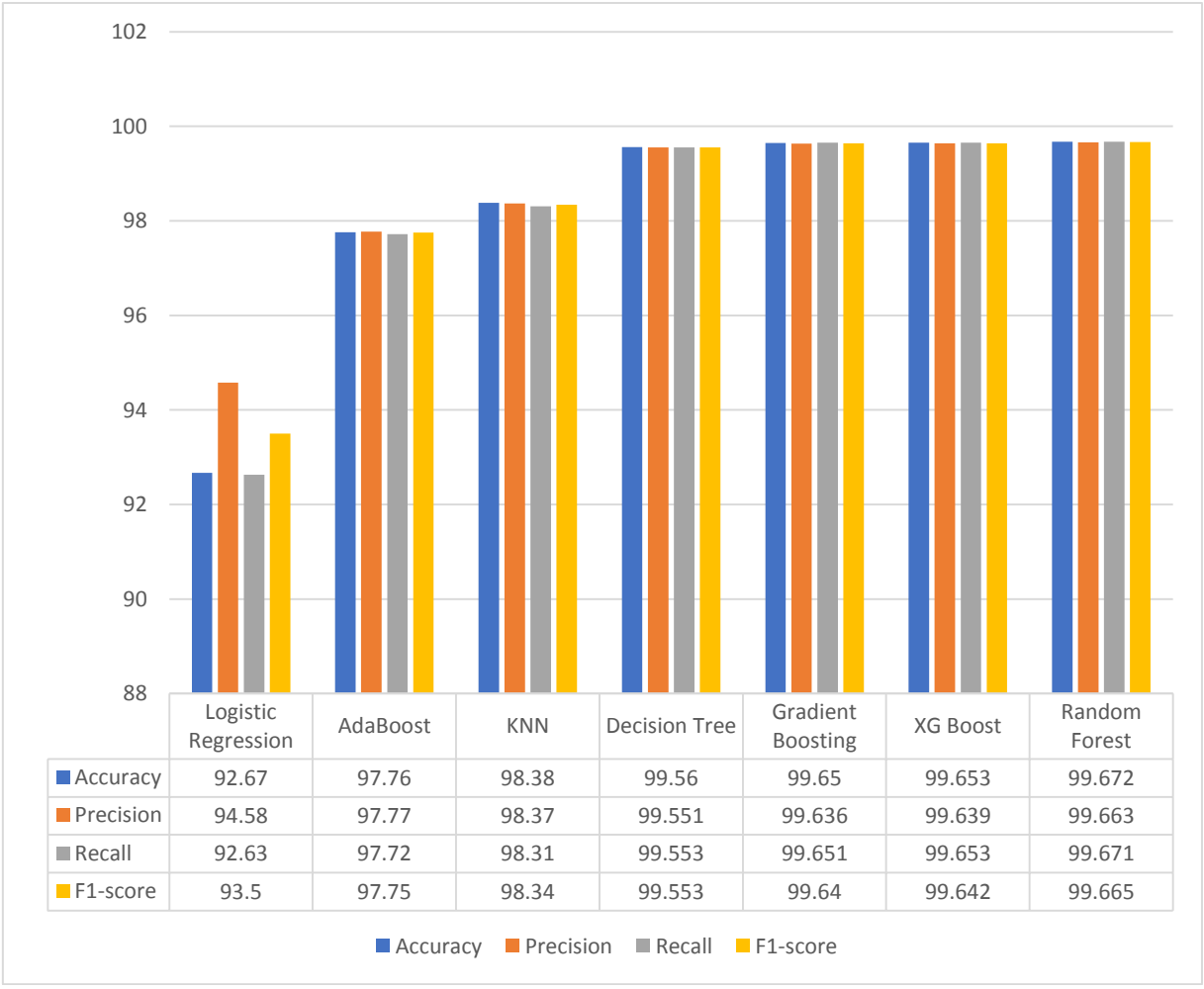
هم‌چنین با توجه به نتایج شکل (۴)، می‌توان دریافت که مدل رگرسیون لاجستیک ضعیف‌ترین عملکرد را در میان مدل‌های بررسی‌شده داشته است. این ضعف عمدتا ناشی از محدودیت‌های ذاتی روش در مدل‌سازی روابط غیرخطی و عدم توانایی در پردازش ساختار پیچیده داده‌های رمزعبور است. هم‌چنین، عملکرد ضعیف آن در مواجهه با داده‌های چندکلاسه، لزوم بهره‌گیری از مدل‌های پیشرفته‌تر و مبتنی بر یادگیری دسته جمعی در زمینه امنیت اطلاعات را برجسته می‌سازد. شکل (۵) ماتریس آشفتگی ۲ مدل برتر آورده شده است.

در ادامه در جدول شماره (۵) دقت مدل‌های روش پیشنهادی و مدل‌هایی که در مطالعات دیگر به آن‌ها پرداخته‌شده است، نمایش داده‌شده است. این مقایسه به‌منظور ارزیابی و تحلیل عملکرد مدل‌های مختلف در برابر یکدیگر انجام‌شده است. هم‌چنین، نتایج این مقایسه می‌تواند به‌طور واضح نقاط قوت و ضعف هر مدل را نمایان کند و روش پیشنهادی را در زمینه دقت و کارایی با مدل‌های موجود در ادبیات علمی مقایسه کند. در نهایت با مقایسه‌های انجام‌شده در جدول (۵) می‌توان دریافت که که الگوریتم جنگل تصادفی همراه با ویژگی‌های استخراج‌شده جدید با دقت ۹۹/۶۷۲٪ نسبت به الگوریتم‌های استفاده‌شده در مطالعات دیگر، عملکرد بالاتری دارد. روش پیشنهادی برای محیط‌های حوزه امنیت سایبری که نیاز به‌دقت حداکثری دارند، بسیار مناسب به نظر می‌رسند.

با توجه به تمام نتایج به دست آمده در این پژوهش، می‌توان دریافت که استفاده از الگوریتم‌های یادگیری ماشین در ارزیابی قدرت گذرواژه‌ها به دلیل توانایی آن‌ها در تحلیل الگوهای پیچیده، یادگیری از داده‌های واقعی و تعمیم[۱۰] به نمونه‌های جدید، نسبت به روش‌های ساده و شمارشی برتری دارد. در حالی‌که روش‌های سنتی معمولا تنها بر معیارهای محدودی مانند طول یا وجود کاراکتر خاص تکیه دارند، مدل‌های هوش مصنوعی می‌توانند تأثیر هم‌زمان چندین ویژگی را تحلیل کرده و دقت پیش‌بینی بالاتری ارائه دهند. هم‌چنین، این روش‌ها با بهره‌گیری از تکنیک‌هایی مانند استخراج ویژگی، متعادل‌سازی کلاس‌ها و یادگیری از رفتار کاربران، عملکرد پویاتر و قابل اطمینان‌تری در ارزیابی سطح امنیت گذرواژه‌ها ارائه می‌دهند.

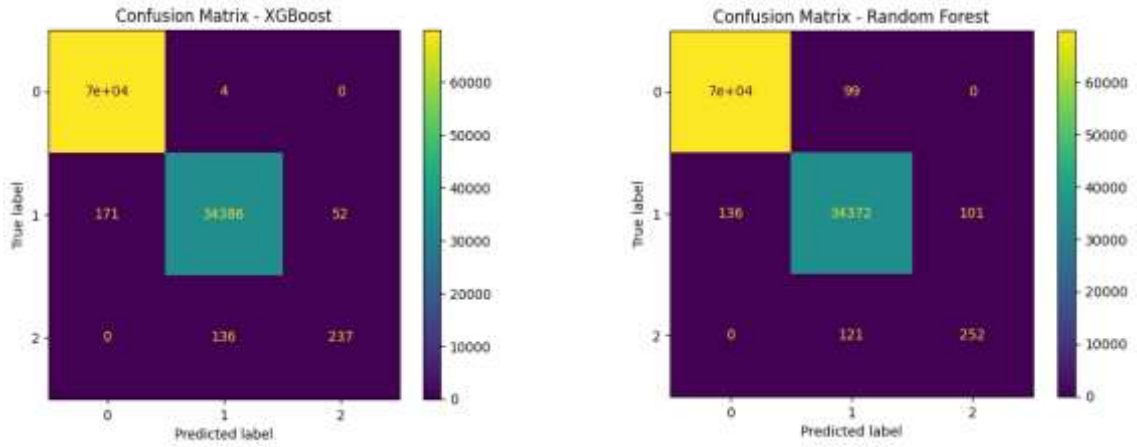
---

[۱] Train
[۲] Test
[۳] Accuracy
[۴] Precision
[۵] Recall
[۶] F1-score
[۷] Ensemble Learning
[۸] Voting
[۹] Overfitting
[۱۰] Generalization



	Logistic Regression	AdaBoost	KNN	Decision Tree	Gradient Boosting	XG Boost	Random Forest
Accuracy	92.67	97.76	98.38	99.56	99.65	99.653	99.672
Precision	94.58	97.77	98.37	99.551	99.636	99.639	99.663
Recall	92.63	97.72	98.31	99.553	99.651	99.653	99.671
F1-score	93.5	97.75	98.34	99.553	99.64	99.642	99.665

شکل (۴). نتایج ارزیابی روش پیشنهادی



شکل (۵): ماتریس آشفتگی دو مدل برتر روش پیشنهادی

جدول (۵). مقایسه روش پیشنهادی با سایر مطالعات

الگوریتم‌های مورد استفاده	دقت	ویژگی‌ها
روش‌های پیشنهادی		
درخت تصمیم	٪۹۹/۵۶	نیازمند تنظیمات بیشتر برای همگرایی بهتر
رگرسیون لجستیک	٪۹۲/۶۷	دقت و کارایی بسیار بالا
XGBoost	٪۹۹/۶۵۳	دقت بالا
جنگل تصادفی	**٪۹۹/۶۷۲**	الگوریتم با دقت مناسب
آدابوست	٪۹۷/۷۶	عملکرد قوی
نزدیک‌ترین همسایه	٪۹۸/۳۸	عملکرد متوسط
گرادیان بوستینگ	٪۹۹/۶۵	دقت بالا
مطالعات دیگر		
استفاده از روش جنگل تصادفی و نایوبیز [۲۳]	٪۸۳/۴	عملکرد متوسط
استفاده ازروش شبکه عصبی بازگشتی [۲۸]	٪۸۳	نیازبه بهبودوترکیب بامدل های دیگر
استفاده از روش جنگل تصادفی و نایوبیز [۲۹]	٪۹۴/۱	دقت و کارایی بسیار بالا
استفاده از روش ماشین بردار پشتیبان ،جنگل تصادفی و نزدیک‌ترین همسایه [۴۰]	٪۸۲	عملکرد متوسط
استفاده از روش جنگل تصادفی و ماشین بردار پشتیبان، رگرسیون لجستیک [۴۱]	٪۸۶	الگوریتم با دقت مناسب
استفاده از روش ماشین بردار پشتیبان،جنگل تصادفی و درخت تصمیم [۴۲]	٪۸۵/۳۴	عملکرد متوسط

۵. نتیجه گیری

رمزگذاری به‌عنوان یکی از ابزارهای بنیادین در تضمین محرمانگی داده‌ها، نقش محوری در تأمین امنیت اطلاعات در بسترهای دیجیتال ایفا می‌کند. در پژوهش حاضر، هدف اصلی شناسایی میزان قدرت امنیت گذرواژه‌ها با بهره‌گیری از هوش مصنوعی و الگوریتم‌های یادگیری ماشین است. به‌منظور دستیابی به این مهم در مرحله پیش‌پردازش داده‌ها با دقت بالا پایش شده و نمونه‌های ناقص و گم‌شده حذف شدند. سپس با استفاده از روش نمونه‌برداری افزایشی Smote، عدم توازن کلاس‌ها اصلاح گردیده و تعداد نمونه‌های هر کلاس برابر با تعداد نمونه‌های کلاس اکثریت شدند. پس از اینکه پایگاه داده تمیز شد از هر رمزعبور ۱۰ ویژگی جدید از قبیل لیست کد اسکی کاراکترهای تشکیل‌دهنده رمز عبور، طول رمز عبور، تعداد کاراکترهای حروف بزرگ، تعداد کاراکترهای حروف کوچک، تعداد اعداد به کار رفته در رمز عبور، تعداد نشانه‌های خاص، متقارن بودن یا نبودن رمزعبور، ترتیب دار یا دنباله وار بودن رمزعبور، وجود یا عدم وجود بیش از یک بار کاراکتر یا عبارت خاص در رمزعبور و تعداد کاراکترهای منحصربه‌فرد استخراج شده است. در نهایت پس از تقسیم‌بندی پایگاه داده به دو بخش آموزش و آزمایش، الگوریتم‌های مختلف یادگیری ماشین برای طبقه‌بندی داده‌ها مورد استفاده قرار گرفته است و قدرت رمزعبورها شناسایی شدند. نتایج به‌دست‌آمده نشان داد که الگوریتم جنگل تصادفی و روش XGBoost با تکیه بر مفهوم یادگیری دسته جمعی و ترکیب چندین مدل ضعیف و رای‌گیری اکثریت، بهترین عملکرد را در این پژوهش داشته‌اند. ویژگی بارز این الگوریتم‌ها از قبیل مقاومت بالا در برابر نویز، داده‌های گم‌شده و مشکل بیش‌برازش باعث شده که آنها را به گزینه‌ای مناسب برای کاربرد در محیط‌های واقعی و پیچیده امنیت سایبری تبدل کند. این پژوهش نشان داد استخراج ویژگی‌های جدید و استفاده از الگوریتم‌های یادگیری ماشین به دلیل توانایی بالای آن‌ها در کشف و تحلیل الگوهای غیر خطی و پیچیده، یادگیری از داده‌های واقعی و تعمیم به نمونه‌های جدید، نسبت به روش‌های سنتی برتری دارد. هم‌چنین روش پیشنهادی توانست تأثیر هم‌زمان چندین ویژگی را تحلیل کرده و با دقت بیشتری سطح قدرت و امنیت رمزها را شناسایی کند.

با وجود نتایج مطلوب، این پژوهش با محدودیت‌هایی از جمله وابستگی عملکرد مدل‌ها به کیفیت داده‌های ورودی، عدم تحلیل ساختار معنایی رمزها و تمرکز صرف بر شاخص‌های کمی مانند دقت، صحت و غیره همراه بود. پیشنهاد می‌شود در مطالعات آتی از رویکردهای ترکیبی شامل تحلیل معنایی رمزها، به‌کارگیری شبکه‌های یادگیری عمیق هم در بخش استخراج کننده ویژگی و هم طبقه‌بند و آزمون عملکرد مدل‌ها در محیط‌های عملیاتی واقعی بررسی شود تا جامعیت و کارآمدی مدل‌های پیشنهادی افزایش یابد. در مجموع، این پژوهش گامی مؤثر در جهت بهره‌گیری از الگوریتم‌های یادگیری ماشین در تحلیل و طبقه‌بندی رمزهای عبور به‌شمار می‌رود و می‌تواند مبنای علمی و کاربردی مناسبی برای توسعه سامانه‌های هوشمند شناسایی تهدیدات امنیتی فراهم آورد.

۶. مراجع

[1] X. Yan, Y. Liu, and X. Wang, "A survey on password guessing attacks," ACM Computing Surveys (CSUR), vol. 50, no. 3, 2017, DOI 10.48550/arXiv.2212.08796.

[2] H. R. Khodadadi and S. Falsafi, "Improvement of security in wireless communication networks with directional modulation and artificial noise," Sci. J.

Electronic and Cyber Defense, vol. 10, no. 4, 2023, DOI 20.1001.1.23224347.1401.10.4.2.2.

[3] V. Yadegari and A. R. Matinfar, "Detect web denial of service attacks using entropy and support vector machine algorithm," Electronic and Cyber Defense., vol. 6, no. 4, pp. 79–89, 2019, DOI 20.1001.1.23224347.1397.6.4.7.9.

[4] K. Dadashtabar Ahmadi and M. Mahmoudbabouei, "The presentation of an active cyber defense model for application in cyber deception technology," Electronic and Cyber Defense, vol. 9, no. 4, pp. 125–140, 2022, DOI 20.1001.1.23224347.1400.9.4.10.3.

[5] Y. M. Sadeghi, S. M. Agha, and F. Adibnia, "A new approach for static detection of security vulnerabilities in web applications," Electronic and Cyber Defense., vol. 2, no. 4, pp. 65–74, 2020, DOI 20.1001.1.23224347.1393.2.4.21.5.

[6] J. Yan, A. Blackwell, R. Anderson, and A. Grant, "Password memorability and security: Empirical results," IEEE Security & Privacy, vol. 2, no. 5, pp. 25–31, 2004, DOI 10.1109/MSP.2004.81.

[7] J. Tan, L. Bauer, N. Christin, and L. F. Cranor, "Practical recommendations for stronger, more usable passwords combining minimum-strength, minimum-length, and blocklist requirements," in Proc. 2020 ACM SIGSAC Conf. Computer and Communications Security (CCS), pp. 1407–1426, Oct. 2020, DOI 10.1145/3372297.3417882.

[8] C. Herley, and P. Van Oorschot, "A research agenda acknowledging the persistence of passwords," IEEE Security & privacy, vol. 10, no. 1, pp. 28-36, 2011.

[9] M. Antonakakis, T. April, M. Bailey, M. Bernhard, E. Bursztein, J. Cochran, and Y. Zhou, "Understanding the mirai botnet," In 26th USENIX security symposium (USENIX Security 17), pp. 1093-1110, 2017.

[10] N. Lykousas and C. Patsakis, "Decoding developer password patterns: A comparative analysis of password extraction and selection practices," Computers & Security, vol. 145, Art. no. 103974, 2024, DOI 10.1016/j.cose.2024.103974.

[11] A. Constantinides, M. Belk, C. Fidas, R. Beumers, D. Vidal, W. Huang, J. Bowles, T. Webber, A. Silvina, and A. Pitsillides, "Security and usability of a personalized user authentication paradigm: Insights from a longitudinal study with three healthcare organizations," ACM Trans. Comput. Healthcare, vol. 4, no. 1, pp. 1–40, 2023, DOI 10.1145/3564610.

[12] M. Just, and D. Aspinall, "Personal choice and challenge questions: a security and usability assessment," In Proceedings of the 5th Symposium on Usable Privacy and Security, pp. 1-11, 2009, DOI 10.1145/1572532.1572543.

[13] S. Adams and M. A. Sasse, "Users are not the enemy," Commun. ACM, vol. 42, no. 12, pp. 40–46, 1999.

[14] S. T. Haque, M. N. Al-Ameen, M. Wright, and S. Scielzo, "Learning system-assigned passwords (up to 56 bits) in a single registration session with cognitive psychology," in Proc. NDSS Symp. (USEC), 2017.

[15] M. N. Al-Ameen, K. Fatema, M. Wright, and S. Scielzo, "The impact of cues and user interaction on the memorability of system-assigned recognition-based graphical passwords," in Proc. 11th SOUPS, pp. 185–196, 2015.

[16] B. Ur, "Do users' perceptions of password security match reality?" in Proc. 2016 CHI Conf. Human Factors in Computing Systems, pp. 3748–3760, 2016, DOI 10.1145/2858036.2858546.

[17] D. Florencio and C. Herley, "A large-scale study of web password habits," in Proc. 16th Int. Conf. World Wide Web, pp. 657–666, 2007, DOI 10.1145/1242572.1242661.

[18] "Generating Strong Passwords with Deep Learning," [Online]. Available: [URL Not Provided].

[19] O. Fierro, N. Grandi, and J. Oliva, "Superradiance of charged black holes in Einstein–Gauss–Bonnet gravity," Classical Quantum Gravity, vol. 35, no. 10, 2018, DOI 10.1088/1361-6382/aab3f6.

[20] W. Han, M. Xu, J. Zhang, C. Wang, K. Zhang, and X. S. Wang, "TransPCFG: Transferring the grammars from short passwords to guess long passwords effectively," IEEE Trans. Inf. Forensics Security, vol. 16, pp. 451–465, 2020, DOI 10.1109/TIFS.2020.3003696.

[21] B. Hitaj, "Passgan: A deep learning approach for password guessing," in Proc. ACNS 2019, Springer, 2019, DOI 10.48550/arXiv.1709.00440.

[22] S. H. M. Zarch, H. Soltani, and M. S. Yazdani, "Enhace the security of password by fuzzy controller," in Proc. Iranian Conf. Intelligent Systems (ICIS), pp. 1–5, 2014, DOI 10.1109/IranianCIS.2014.6802541.

[23] M. Jiao, "Application of Random Forest Algorithm in Network Intrusion Detection of Government Affairs Departments," Int. J. Comput. Intell. Appl., vol. 24, no. 4, 2024, DOI 10.1142/S1469026823420038.

[24] Y. Abdrabou, "'Your Eyes Tell You Have Used This Password Before': Identifying Password Reuse from Gaze and Keystroke Dynamics," in Proc. 2022 CHI Conf. Human Factors in Computing Systems, pp. 1–16, 2022, DOI 10.1145/3491102.3517531.

[25] S. Parkinson, "Password policy characteristics and keystroke biometric authentication," IET Biometrics, vol. 10, no. 2, pp. 163–178, 2021, DOI 10.1049/bme2.12017.

[26] B. Suruthi, "Efficient handwritten passwords to overcome spyware attacks," Sci. Technol., vol. 3, pp. 1–9, 2021.

[27] M. Ishak, "Correlation impact by random forest towards prediction of phishing website," in IOP Conf. Ser.: Mater. Sci. Eng., vol. 917, no. 1, 2020, DOI 10.1088/1757-899X/917/1/012043.

[28] A. Nosenko, Y. Cheng, and H. Chen, "Learning password modification patterns with recurrent neural networks," in Int. Conf. Secure Knowl. Manage. AI Era, Springer, pp. 110–129, 2021, DOI 10.1007/978-3-030-97532-6_7.

[29] A. Demenongo and A. Iorshase, "Ensemble model for the detection of phishing URLs," Ilorin J. Comput. Sci. Inf. Technol., vol. 7, no. 1, pp. 1–25, 2024.

[30] R. Mohammed, J. Rawashdeh, and M. Abdullah, "Machine learning with oversampling and undersampling techniques: overview study and experimental results," in Proc. 2020 11th Int. Conf. Inf. Commun. Syst. (ICICS), pp. 243–248, 2020, DOI 10.1109/ICICS49469.2020.239556.

[31] S. Wang, "Research on expansion and classification of imbalanced data based on SMOTE algorithm,"

Sci. Rep., vol. 11, no. 1, 2021, DOI 10.1038/s41598-021-03430-5.

[32] S. Das Guptta, "Modeling hybrid feature-based phishing websites detection using machine learning techniques," Ann. Data Sci., vol. 11, no. 1, pp. 217–242, 2024, DOI 10.1007/s40745-022-00379-8.

[33] J. Z. Ahmadabadi, F. Z. Mehrjardi, M. Ghanbary, and M. Mirzaei, "Identification of Effective Factors and Prediction of Ischemic Heart Disease Using Machine Learning Methods and Data from the Yazd Health Study (YaHS)," Journal of Shahid Sadoughi University of Medical Sciences, vol. 32, no. 7, pp. 8067-8079, 2024, DOI 10.18502/ssu.v32i7.16571.

[34] M. Akbari Podineh, F. Zare Mehrjardi, and M. Sardari Zarchi, "Multimodal analysis of ECG signals for cardiac arrhythmia detection using machine learning and deep learning methods," Applied and basic Machine intelligence research, vol. 3, no. 1, pp. 17-34, 2025, DOI 10.22034/abmir.2025.22930.1118.

[35] M. R. Esmaeili Noroozi, and F. Zare Mehrjardi, "Optimization of Steel Alloy Composition to Maximize Yield Strength Using a Machine Learning Model and the Cuckoo Optimization Algorithm (COA)," Engineering Management and Soft Computing, vol. 12, no. 1, pp. 131-143, 2026, DOI 10.22091/jemsc.2026.13746.1299.

[36] R. Torkashvan, S. Parsa, and B. Vaziri, "Fault Proness Estimation of Software Modules Using Machine learning," Electronic and Cyber Defense, vol. 11, no. 4, pp. 45-59, 2024, DOI 20.1001.1.23224347.1402.11.4.4.1

[37] F. Z. Mehrjardi, A. M. Latif and M. S. Zarchi, "Forgery detection in digital images using the hybrid deep learning method," Electronic and Cyber Defense, vol. 11, no. 4, pp. 99-116, 2025. (in Persian), DOI 20.1001.1.23224347.1402.11.4.9.6

[38] R. torkashvan, S. Parsa and B. vaziri, "Fault Proness Estimation of Software Modules Using Machine learning," Electronic and Cyber Defense, vol. 11, no. 4, pp. 44-59, 2025. (in Persian), DOI 20.1001.1.23224347.1402.11.4.4.1.

[39] A. Karimi and M. R. Khosravi Farsani, "Improving the accuracy of code smell identification using the gray wolf algorithm based on machine learning techniques and majority voting," Electronic and Cyber Defense, vol. 12, no. 1, pp. 108-122, 2024. (in Persian), DOI 20.1001.1.23224347.1403.12.1.9.7.

[40] J. Mallet, "Hold on and swipe: a touch-movement based continuous authentication schema based on machine learning," in Proc. 2022 Asia Conf. Algorithms, Comput. Mach. Learn. (CACML), pp. 442–447, IEEE, 2022, DOI 10.1109/CACML55074.2022.00081.

[41] L. Pryor, "Evaluation of a User Authentication Schema Using Behavioral Biometrics and Machine Learning," arXiv preprint arXiv:2205.08371, 2022, DOI 10.48550/arXiv.2205.08371.

[42] M. Anwer, "Attack detection in IoT using machine learning," Eng., Technol. Appl. Sci. Res., vol. 11, no. 3, pp. 7273–7278, 2021, DOI 10.48084/etasr.4202.