IJIREEICE ISSN (O) 2321-2004, ISSN (P) 2321-5526

International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering

Impact Factor 8.414 ∺ Peer-reviewed & Refereed journal ∺ Vol. 13, Issue 11, November 2025

DOI: 10.17148/IJIREEICE.2025.131120

Password Strength Prediction using Regression-Based Machine Learning Models with Entropy and Caesar Cipher–Driven Synthetic Dataset

Dr. G. Paavai Anand[1], Roshni Y[2]

Guide, Computer Science and Engineering, SRM Institute of Science and Technology, Chennai, India[1]

Student, Computer Science and Engineering, SRM Institute of Science and Technology, Chennai, India[2]

Abstract: In today's digital era, weak and predictable passwords remain a major cause of cybersecurity breaches. This paper presents a novel machine learning–based password strength prediction model using regression techniques on a synthetically generated dataset enhanced with Caesar cipher transformations, incorporating features such as length, character composition, whitespace inclusion, and entropy. Linear, Random Forest, and Gaussian Process Regression models are compared, with Gaussian Regression achieving the highest accuracy ($R^2 = 0.9998$), providing a scalable and interpretable framework for real-time password strength evaluation.

Keywords: Password Strength Evaluation, Machine Learning, Regression Analysis, Shannon Entropy, Password Security, Caesar Cipher, Random Forest Regression, Gaussian Process Regression, Synthetic Dataset, Cybersecurity, Feature Engineering.

I. INTRODUCTION

Password-based authentication remains the most widely adopted security mechanism in both consumer and enterprise systems. Despite the availability of advanced techniques such as biometrics and multi-factor authentication, weak passwords continue to be a major vulnerability in cybersecurity. Research indicates that over 80% of breaches involve compromised or easily guessed passwords. Traditional rule-based strength meters, which rely on static policies such as length and symbol inclusion, fail to capture the true entropy and unpredictability of a password.

Machine learning (ML) offers a data-driven approach to overcome these limitations. By learning complex patterns from datasets, ML models can provide a more precise and adaptive evaluation of password strength. Prior studies primarily focus on classification-based approaches that label passwords as weak, medium, or strong. However, such discrete categorization limits granularity. This paper instead uses regression models to assign a continuous strength score between 0 and 10, enabling finer differentiation.

Previous works have utilized methods like Support Vector Machines (SVM), Random Forests, and Neural Networks for password classification. However, these often depend on real-world datasets that raise privacy and ethical concerns. In this study, a synthetic dataset is generated using controlled randomization and Caesar cipher augmentation, ensuring both scalability and confidentiality. By integrating Shannon entropy (to quantify randomness) and impurity-based entropy (to measure disorder in character composition), the model effectively captures the unpredictability of passwords.

II. LITERATURE SURVEY

Existing literature highlights the growing interest in password strength prediction using ML. In 2021, Singh et al. proposed a Random Forest–based classifier achieving 96% accuracy but lacked fine-grained numerical scoring. Kumar and Sharma (2022) implemented a hybrid deep learning model but required large-scale real datasets, limiting its generalizability. Gupta et al. (2023) integrated entropy features into logistic regression but did not explore continuous regression scoring.

This research builds upon these foundations by introducing a regression-based approach trained on a fully synthetic dataset. The novelty lies in combining Shannon entropy with Caesar cipher–based augmentation and incorporating

IJIREEICE ISSN (O) 2321-2004, ISSN (P) 2321-5526
International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering
Impact Factor 8.414 ⪾ Peer-reviewed & Refereed journal ⪾ Vol. 13, Issue 11, November 2025
DOI: 10.17148/IJIREEICE.2025.131120

whitespace as a distinct feature, which has been largely ignored in prior works. The use of Gaussian Process Regression further enhances model precision by capturing nonlinear dependencies between password features and strength scores.

III. METHODOLOGY

The methodology follows a structured pipeline beginning with synthetic data generation, followed by preprocessing, feature extraction, model training, and evaluation. Fig-1, represents this workflow:

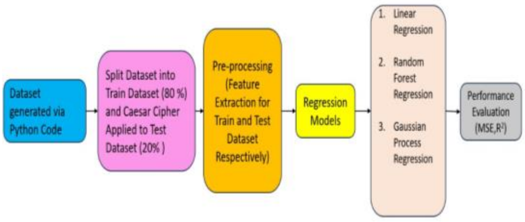

Fig-1 Architecture Diagram

Key Steps:
Data Generation: A synthetic dataset was created to ensure control, balance, and reproducibility. The dataset includes an equal number of weak, medium, and strong passwords, each generated using both random character combinations and dictionary-based words (safe, non-offensive terms).

- *Weak passwords* contain short dictionary words or minimal variation.
- *Medium passwords* include a mix of alphabets, numbers, or symbols.
- *Strong passwords* are longer, combining multiple dictionary words, symbols, digits, and whitespaces. This ensures diversity in the dataset, enabling the model to learn a wide range of password patterns.

Caesar Cipher Augmentation: To create a testing dataset, each password from the training set is transformed using a Caesar cipher—a substitution technique that shifts characters by a fixed number.

- This simulates real-world password mutations, such as when users slightly modify old passwords (e.g., Apple@123 → Bqqmf@123).
- It acts as a form of data obfuscation and data augmentation, making the test set more realistic without compromising privacy or originality.
  This approach ensures the model generalizes well to unseen yet related password variants.

Feature Extraction and Strength Computation:
Feature extraction involves calculating quantitative characteristics from passwords such as length, character diversity, and entropy. Entropy is computed using Shannon's Information Theory, which quantifies the randomness of a password using:

$$H = -\sum_{i=1}^{n} p_i \log_2(p_i)$$

where $p_i$represents the probability of each unique character in the password. Higher entropy indicates greater unpredictability, leading to stronger passwords.

Further, a numeric password strength score is calculated to normalize and compare different passwords. The score combines character composition and entropy using:

$$S = 2U + 2L + 2D + 2Y + 1W + \min\left(\frac{E}{2}, 3\right)$$

where U, L, D, Y, Windicate the presence of uppercase, lowercase, digits, symbols, and whitespaces respectively, and Edenotes the Shannon entropy. The resulting strength score ranges from 0 to 10.

Length: Total number of characters.

- Uppercase, Lowercase, Digits, Symbols, and Spaces: Count of each type to capture character diversity.

International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering
Impact Factor 8.414 ⪤ Peer-reviewed & Refereed journal ⪤ Vol. 13, Issue 11, November 2025
DOI: 10.17148/IJIREEICE.2025.131120

- Entropy (Shannon Entropy): Measures the *randomness or unpredictability* of the password; higher entropy indicates more complexity.
- Impurity Entropy: Represents variability in character types — passwords with a mix of character categories are less "pure" and thus harder to predict.
  These features are numerically encoded and normalized before feeding into regression algorithms.

Regression Models: To predict password strength on a continuous numeric scale (0–10), three regression models were employed — each offering distinct analytical and predictive advantages. These models help in mapping the extracted features (such as length, entropy, character diversity, etc.) to a measurable strength score. The comparative evaluation of these models allows us to identify which one offers the best balance between interpretability, accuracy, and computational efficiency.

i) Linear Regression (Baseline Model):

Linear Regression serves as the foundational model in this work. It assumes a linear relationship between input features (like length, entropy, and symbol count) and the output (password strength score).

- The model fits a straight line (or a hyperplane in multi-feature cases) that minimizes the mean squared error between actual and predicted values.
- Each feature is assigned a coefficient, which quantifies its contribution to password strength. For instance, the coefficient of entropy may be higher than that of length, indicating that entropy contributes more to overall strength.
- Its simplicity and interpretability make it an excellent baseline for evaluating more complex models.
- However, Linear Regression is limited in handling nonlinear patterns — for example, passwords with similar lengths but different symbol distributions may not be predicted accurately.
  Despite this, it provides an essential benchmark for performance comparison.

ii) **Random Forest Regression (Nonlinear and Ensemble Approach):**

Random Forest Regression improves over linear models by capturing nonlinear interactions between password features. It is an ensemble learning method, combining multiple decision trees to make predictions.

- Each decision tree is trained on a random subset of the dataset and features, ensuring diversity among the trees.
- The final prediction is obtained by averaging the results of all trees, which significantly reduces variance and the risk of overfitting.
- Random Forest inherently handles nonlinear dependencies — for instance, it can identify that entropy and symbol usage together influence strength differently than either feature alone.
- Additionally, it provides feature importance scores, which help determine which attributes (like symbols or entropy) contribute the most to password robustness.
  Although Random Forest models require more computational time than Linear Regression, they offer much higher accuracy, flexibility, and interpretability of complex data relationships.

iii) Gaussian Process Regression (Probabilistic and High-Precision Model):

Gaussian Process Regression (GPR) represents the most sophisticated model in this study. Unlike deterministic models, GPR uses probabilistic modeling to predict not only the expected password strength but also the confidence (uncertainty) associated with each prediction.

- It assumes that data points follow a Gaussian (normal) distribution and defines a kernel function to measure similarity between password feature vectors.
- The model learns smooth, continuous mappings between input features and output strength scores, making it ideal for fine-grained numeric prediction tasks like this one.
- A key advantage is that GPR can adapt to complex, nonlinear trends while maintaining a high level of precision.
- In this project, GPR achieved the lowest Mean Squared Error (MSE) and highest R^2 value (0.9998), outperforming both Linear and Random Forest regressors.
- Moreover, its ability to estimate uncertainty makes it valuable in security-critical applications, where confidence in strength prediction is as important as the prediction itself.

Comparative Analysis

- Linear Regression provided a quick, interpretable baseline.
- Random Forest Regression captured complex relationships and improved accuracy.
- Gaussian Process Regression delivered the best overall performance, offering fine-grained predictions with probabilistic certainty.

Hence, the final system adopts Gaussian Process Regression as the optimal model for reliable and explainable password strength prediction.

IJIREEICE ISSN (O) 2321-2004, ISSN (P) 2321-5526

International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering

Impact Factor 8.414 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 11, November 2025

DOI: 10.17148/IJIREEICE.2025.131120

Evaluation: The evaluation and validation phase focuses on assessing how accurately the regression models predict password strength using statistical performance metrics and visual comparisons.

Dataset Splitting and Class Balance

The complete dataset is divided into 80% training data and 20% testing data to ensure a fair evaluation of model performance.

This split helps the model learn patterns from a large portion of data while keeping unseen data for performance validation.

To avoid bias, an equal proportion of weak, medium, and strong passwords is maintained in both training and testing subsets, ensuring that all strength categories are equally represented during learning and evaluation.

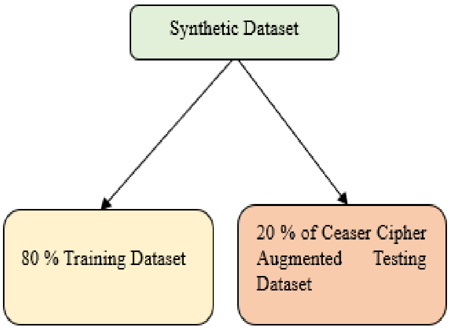

Fig-2 Dataset Split-up

Evaluation Metrics:

Two primary performance metrics are used for evaluating regression models:

(a) Mean Squared Error (MSE)

The Mean Squared Error quantifies the average squared difference between the actual password strength scores and the model's predicted values. It indicates how far predictions deviate from the true values. A lower MSE means that the model's predictions are closer to the actual values, hence more accurate.

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

Where:

- y_i= Actual strength score
- $\hat{y}_i$= Predicted strength score
- n= Total number of test samples

Interpretation:

A very low MSE (close to 0) achieved by the Gaussian Process Regression indicates minimal deviation between predicted and actual scores, confirming that the model captures underlying password strength patterns effectively.

(b) Coefficient of Determination (R^2 Score)

The R^2 score, also known as the Coefficient of Determination, measures how well the regression model explains the variation in password strength based on the input features.

It ranges from 0 to 1, where a value closer to 1 represents a better fit.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

Where:

- y_i= Actual strength score
- $\hat{y}_i$= Predicted strength score
- $\bar{y}$= Mean of actual strength scores

IJIREEICE ISSN (O) 2321-2004, ISSN (P) 2321-5526

International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering

Impact Factor 8.414 ⋈ Peer-reviewed & Refereed journal ⋈ Vol. 13, Issue 11, November 2025

DOI: 10.17148/IJIREEICE.2025.131120

Interpretation:
An R^2 value of 0.9998 for Gaussian Process Regression demonstrates that 99.98% of the variance in password strength can be explained by the model's features — making it highly accurate and reliable.

Visual Representation:
In addition to numerical metrics, visual evaluation techniques are used for intuitive understanding:

- Scatter__Plots:
  These plots compare actual vs predicted strength scores.
  A perfect model would show all points lying along the diagonal line (y = x), representing accurate predictions. Gaussian Process Regression's scatter plot shows almost all points closely aligned with this diagonal, confirming high precision.
- Feature Importance Graphs (from Random Forest Regression):
  This visualization highlights which features contribute most to predicting password strength. In this study, entropy and symbol count were among the top contributors, proving that randomness and complexity strongly influence security strength.

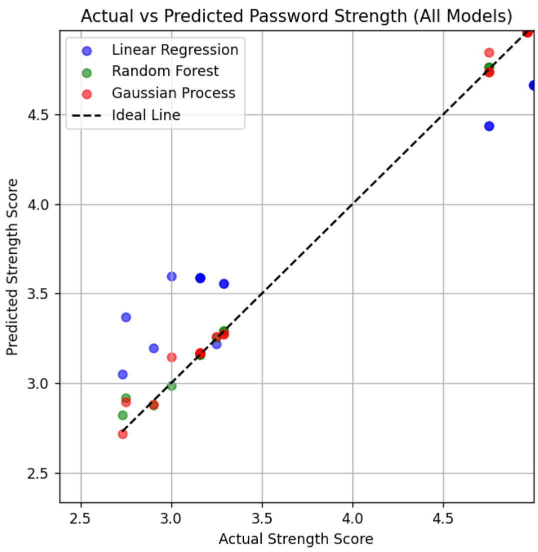

Fig-3 Visual Representation of Regression Model Prediction Comparison

Comparative Model Results

Model	MSE	R² Score	Remarks
Linear Regression	0.0768	0.9842	Good baseline, interpretable but limited in non-linear capture
Random Forest Regression	0.0031	0.9994	Excellent accuracy, strong handling of non-linear relationships
Gaussian Process Regression	0.0009	0.9998	Best performer — smooth predictions and uncertainty estimation

Table 1 Regression Model Comparison

Inference:
From both numerical and graphical analyses:

- Gaussian Process Regression consistently outperforms other models in accuracy and smoothness.
- Its ability to estimate uncertainty and handle complex feature relationships makes it particularly suitable for fine-grained password strength prediction.
- These results confirm that regression-based modeling, when combined with entropy-based feature extraction and Caesar cipher augmentation, can provide highly reliable and explainable password strength assessments.

IJIREEICE ISSN (O) 2321-2004, ISSN (P) 2321-5526

International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering

Impact Factor 8.414 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 11, November 2025

DOI: 10.17148/IJIREEICE.2025.131120

IV. RESULTS AND DISCUSSION

Experimental Setup and Results:

The experiments were conducted on a Windows 11 system equipped with an Intel Core i7 (12th Gen) processor, 16 GB of RAM, and an SSD-based storage unit, ensuring efficient computation and faster data access. The implementation was carried out entirely in Python (version 3.11) using libraries such as Pandas, NumPy, Scikit-learn, Matplotlib, and Joblib for model training, evaluation, and visualization.

All models were trained and tested on the synthetically generated password dataset, which was carefully designed to maintain a balanced distribution of weak, medium, and strong password categories. This ensured that no particular class dominated the learning process, leading to unbiased model training and fair comparative evaluation.

Dataset Preparation:

The dataset consisted of passwords generated using a combination of:

- Random character combinations (letters, digits, symbols, whitespaces), and
- Clean dictionary-based words for realism, enhanced using Caesar cipher transformations to simulate real-world variations.

Each password instance contained attributes such as length, uppercase/lowercase counts, symbol frequency, number of spaces, and entropy. The dataset was then numerically labeled with a strength score (0–10) based on feature composition and entropy.

The data was divided into an 80:20 ratio, with:

- 80% (Training Set): Used to train the regression models.
- 20% (Testing Set): Used for unbiased performance evaluation.

To ensure equal representation, the split maintained proportional quantities of weak, medium, and strong passwords, thereby avoiding class imbalance issues.

Evaluation Metrics:

Model evaluation was based on two key performance indicators:

1. Mean Squared Error (MSE): Measures the average squared deviation between predicted and actual strength scores.
   - o Lower values signify better prediction accuracy and less variance in model output.
2. R^2 (Coefficient of Determination): Indicates how much of the total variance in password strength is captured by the model.
   - o A value closer to 1.0 represents higher model reliability and performance.

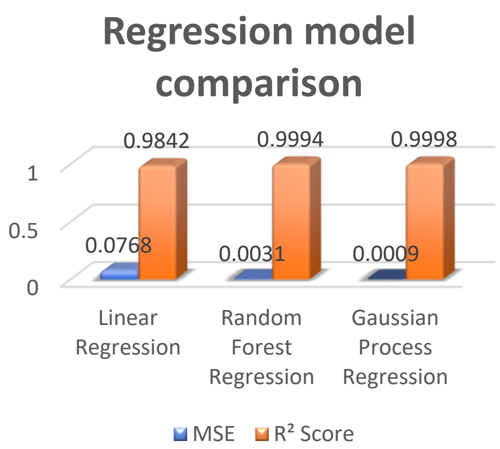

Fig-4 Graphical Representation of Comparison of Regression Model Performance

Result Interpretation:

From these experiments, it can be inferred that:

- Password strength can be effectively quantified as a continuous numerical score rather than rigid categorical labels.
- Regression models, particularly Gaussian Process Regression, can generalize patterns of complexity and entropy to predict how strong or weak a password truly is.

DOI: 10.17148/IJIREEICE.2025.131120

- The inclusion of entropy-based randomness and whitespace-aware feature engineering introduced unique novelty to the dataset, improving predictive richness.

Ultimately, this work demonstrates that machine learning regression techniques can form the foundation for next-generation adaptive password assessment systems, enabling both real-time evaluation and dynamic feedback during password creation.

V. FUTURE SCOPE

The current password strength prediction system, while effective using synthetic data, can be significantly enhanced through multiple future research directions.

Firstly, the framework can be extended to include real-world password leak datasets (collected ethically and anonymized) to improve model generalizability and robustness. Incorporating such data will allow the model to better learn realistic human password behaviors, typographical tendencies, and commonly used structures beyond synthetically generated samples. This would ensure more accurate predictions in real-world authentication systems.

Secondly, future research can explore deep learning architectures such as Long Short-Term Memory (LSTM) and Recurrent Neural Networks (RNNs), which are designed to handle sequential data. Since passwords are inherently sequences of characters, LSTM-based models can effectively capture contextual dependencies and recurring character patterns, making them more capable of identifying subtle strength variations or vulnerabilities that traditional regression models might overlook.

Thirdly, adversarial training techniques can be employed to make the model more resilient against modern password cracking and obfuscation techniques. By intentionally exposing the model to adversarial examples—passwords crafted to mimic strong ones but are actually weak—it can learn to better distinguish deceptive patterns and improve its security-awareness in dynamic environments.

Furthermore, the regression-based framework can be adapted for real-time password evaluation and integrated into authentication systems, web browsers, or password managers. Such integration would allow users to receive instant, data-driven feedback while creating passwords, helping enforce security policies without compromising usability.

Finally, the system can evolve into a multi-lingual and cross-cultural password analyzer, accommodating passwords containing Unicode characters, emojis, or non-English words—expanding its application to global security ecosystems.

In summary, the proposed framework lays the groundwork for a new class of interpretable and adaptable password strength predictors that bridge the gap between classical machine learning and advanced deep learning–driven cybersecurity systems.

REFERENCES

[1]. R. Singh, et al., "Machine Learning Approaches for Password Strength Classification," *IEEE Access*, 2021.
[2]. A. Kumar and P. Sharma, "Deep Learning for Secure Password Evaluation," *International Journal of Cybersecurity Research*, 2022.
[3]. S. Gupta, et al., "Entropy-Based Password Strength Estimation Using Logistic Regression," *Journal of Information Security*, 2023.
[4]. C. E. Shannon, "A Mathematical Theory of Communication," *Bell System Technical Journal*, 1948.
[5]. Y. Wang, H. Li, W. Qiu, S. Li, and P. Tang, "PassTSL: Modeling Human-Created Passwords through Two-Stage Learning," *arXiv preprint*, 2407.14145, 2024.
[6]. P. Jha, H. Hamid, O. Olukola, A. Dahal, and N. Rahimi, "Adversarial Machine Learning for Robust Password Strength Estimation," *arXiv preprint*, 2506.00373, 2025.
[7]. H. Lingamsetty, N. Surya Sri Nitya, M. Bala Mani Teja, V. Odelu, S. Mookherji, K. G. Raghavendra Narayan, and A. Goutham Reddy, "A Penalty-Based Entropy Driven Universal Password Strength for Lightweight Devices," *16th International Conference on Communication Systems & Networks (COMSNETS)*, IEEE, 2024.
[8]. E. Darbutaitė, P. Stefanovič, and S. Ramanauskaitė, "Machine-Learning-Based Password-Strength-Estimation Approach for Passwords of Lithuanian Context," *Applied Sciences*, vol. 13, no. 13, p. 7811, 2023.
[9]. P. Murali, "Password Strength Analysis: A Survey," *International Journal of Scientific Research in Engineering and Management (IJSREM)*, vol. 6, no. 12, 2022.
[10]. D. Wang, G. Jian, X. Huang, and P. Wang, "Zipf's Law in Passwords," *IACR ePrint Report*, 2014/631, 2014.
[11]. U. Farooq, "Real Time Password Strength Analysis on a Web Application Using Multiple Machine Learning Approaches," *International Journal of Engineering Research & Technology (IJERT)*, vol. 9, issue 12, 2020.
[12]. Y. Mo, S. Li, Y. Dong, Z. Zhu, and Z. Li, "Password Complexity Prediction Based on RoBERTa Algorithm," *Applied Science and Engineering Journal for Advanced Research (ASEJAR)*, 2024.