A Comparative Work to Highlight the Superiority of Mouth Brooding Fish (MBF) over the Various ML Techniques in Password Security Classification

Yan Shi, Yue Wang*
Hebei Chemical & Pharmaceutical College,
Shi Jiazhuang 050026, China

***Abstract*—Within the domain of password security classification, the pursuit of practical and dependable methodologies has prompted the examination of both biological and technological paradigms. The present study investigates the efficacy of Mouth Brooding Fish (MBF) as an innovative method in contrast to conventional Machine Learning (ML) approaches for classifying password security. The research approach entails a rigorous examination of the comparative analysis of MBF and ML algorithms, evaluating their effectiveness in password classification using many criteria, including accuracy, robustness, flexibility, and durability against adversarial assaults. The findings suggest that ML approaches have shown significant effectiveness in classifying passwords. However, using methodologies inspired by the minimum Bayes risk framework demonstrates a higher degree of resistance against typical cyber dangers. The intrinsic biological mechanisms of MBF, encompassing adaptive behaviors and inherent protection, play a role in enhancing the resilience and adaptability of the password security categorization system. The results offer significant insights that can inform the evolution of password security systems, integrating biological principles with technical progress to enhance safeguarding measures in digital environments. To emphasize the advantages of the suggested approach, several ML approaches are investigated, such as Support Vector Machines (SVM), AdaBoost, Multilayer Perceptron (MLP), Gaussian Kernel (GK), and Random Forest (RF). The F-score, accuracy, sensitivity, and specificity metrics for MBF exhibit noteworthy performance compared to the other selected models, with values of 100%.**

***Keywords*—Mouth Brooding Fish (MBF); password security; Sber dataset; SVM; Random Forest; AdaBoost**

I. Introduction

The advent of the online society has introduced a user authentication mechanism known as password authentication [1]. The present approach facilitates the registration of a password by the user, followed by the user's authentication by a comparison between the registered password and the input password. Hence, the data that needs safeguarding under this authentication approach is the password provided as input. The process of entering a password typically involves keyboard input, necessitating the implementation of a mechanism to safeguard the data entered via the keyboard [2]. Passwords play a crucial role in ensuring the security of computer systems. While there are several substitutes to passwords for security purposes, passwords remain highly attractive for validating one's identity in a wide range of applications. Digital authentication mechanisms offer a straightforward and efficient approach to safeguarding a system, representing an individual's identity within the system. The inherent weakness of passwords resides in their fundamental characteristics. In contemporary times, individuals are frequently advised on the need to employ robust passwords to safeguard personal information, owing to the proliferation of methods by which unauthorized individuals with limited technological expertise can acquire the passwords of legitimate users. Therefore, businesses must acknowledge the susceptibilities to which passwords are exposed and establish robust policies that control the formulation and utilization of passwords to prevent the exploitation of these vulnerabilities [3].

Over the last twenty years, there has been a significant exponential increase in the production of mobile products by various firms [4]. Nevertheless, despite the continuous advancements in functionality of these gadgets, the security protocols employed to safeguard them have remained essentially identical for the previous twenty years. The significant disparity in growth trajectories observed between devices and their corresponding security measures increasingly exposes a heightened vulnerability, wherein an expanding number of devices are susceptible to infiltration by malicious actors. Building upon prior research in the domain, Pryor et al. [5] investigated several ML methods employed in user authentication systems that incorporate touch dynamics and device mobility. The objective of this paper was to provide a complete examination of the present applications of various ML algorithms commonly employed in user authentication systems that incorporate touch dynamics and device movement. In order to successfully decipher passwords with high levels of complexity, it is imperative to employ a password-cracking methodology that surpasses the limitations of a rule-based dictionary assault.

Consequently, there is a pressing need for extensive research to be conducted in order to advance the creation of such a technique. The subsequent discourse provides an elaborate exposition of the scenarios necessitating the development of password-cracking technologies. One common occurrence is the tendency for individuals to forget their need to remember often. This is especially true when users choose complex passwords that deviate from previously employed patterns. Consequently, an efficient password-cracking technique becomes necessary to address this issue.

Furthermore, it may be necessary for national authorities to decrypt passwords in order to access encrypted criminal evidence or intelligence material. In order to ensure the adequate security of passwords, it is necessary to employ effective password-cracking techniques. The utilization of password-cracking techniques may achieve a realistic estimation of password strength. The zxcvbn approach, as employed in the DropBox system, utilizes straightforward password-cracking techniques to assess the level of password security. Wheeler [6] primarily emphasized enhancing the efficiency of password-cracking techniques rather than assessing the robustness of passwords.

Despite the remarkable advances made in the previous research, there are many limitations regarding the accuracy of the methods proposed for data classification of password security. Accordingly, the current work examined the benefits of MBF over SVM, AdaBoost, MLP, GK, and RF. The results were examined regarding F-score, accuracy, sensitivity, and specificity. The novelty lies in leveraging the unique behavioral traits of MBF to revolutionize data classification within the realm of password security. Drawing inspiration from MBF's instinctive protection mechanisms for their offspring, this approach introduces a fresh perspective to data classification methodologies. This innovative paradigm shift offers a departure from traditional algorithms by integrating biological concepts into the framework of password security, potentially enhancing the resilience and adaptability of data classification systems against cyber threats. Incorporating MBF-inspired strategies introduces a novel avenue for more robust and sophisticated data classification techniques, potentially setting a new standard for safeguarding sensitive information in the digital landscape.

In the subsequent sections of this paper, we delve deeper into the exploration of password security classification methodologies, juxtaposing the innovative MBF approach with conventional ML techniques. Section II provides a literature review of the related works for highlighting the novelty. In Section III, we present a detailed analysis of the performance of each ML approach individually, highlighting their strengths and limitations. Also, the dataset, evaluation criteria, and methodology are illustrated in this section. Section IV focuses on the comparison between MBF and ML methods, showcasing the unique advantages of biological inspiration in password security classification. Finally, Section V concludes the paper.

II. Related Work

In recent years [7], much attention has been devoted to the issues of data classification for password security based on ML techniques [8]. For instance, Saha et al. [9] proposed developing a comprehensive framework for detecting many types of sensitive information, encompassing API keys, asymmetric private keys, client secrets, and generic passwords. ML models were utilized to differentiate between an authentic secret and a spurious detection effectively. Integrating a regular expression-based methodology with ML techniques enabled the detection of many categories of confidential information, particularly generic passwords that were overlooked in previous studies. The proposed method facilitated the minimization of potential instances of inaccurate identification. Huang et al. [10] explored an alternate approach that relies on user keystrokes as a technique. The extraction of touch timings and force characteristics was performed on a piezoelectric force touch panel, which served as an essential component of the hardware system.

Three widely utilized ML classifiers were employed to analyze the gathered dataset, ultimately attaining an Equal Error Rate (EER) of 0.720%. Alswailem et al. [11] presented a sophisticated method to identify and detect fraudulent websites, sometimes called phishing websites. The system served as an auxiliary feature to a web browser, functioning as an extension that autonomously alerts the user upon identifying a phishing website. The system is founded upon a machine learning approach, namely supervised learning. The Random Forest approach was chosen for its strong categorization performance. The primary objective was to enhance the classifier's performance by conducting an in-depth analysis of the characteristics of phishing websites. In another study [12], a novel methodology involved transforming behavioral biometrics data, namely time series, into a three-dimensional picture. The procedure above modification effectively preserved all the inherent attributes of the behavioral signal. No filtering operation was used for the time series in this transformation, and the approach is objective. The performance of the authentication system was assessed using the Equal Error Rate (EER) metric on a substantial dataset, and the efficacy of the suggested technique was demonstrated on a multi-instance system. Murmu et al. [13] proposed a novel ensemble methodology incorporating both a classification algorithm and a guessing technique. The method was based on a bi-directional generative stochastic network for generating individualized passwords. The algorithm was designed to enhance the convergence rate of the password generation process. The proposed method exhibited a higher sample generation rate in a shorter duration when compared to the Generative Adversarial Network (GAN). The one-class SVM was utilized to train a model using both stolen and produced passwords to make predictions about the strength of passwords. The passwords predominantly consist of medium and weak categories, and they exhibited improved performance by establishing a correlation with weak passwords. The LSTM model was optimized to forecast the difficulty associated with cracking a particular test password [7].

The current paper addresses several limitations present in previous works within the realm of password security classification. Prior research often focused solely on conventional ML approaches, overlooking the potential insights gleaned from biological paradigms. By introducing the innovative MBF method and juxtaposing it with established ML techniques, this study fills a crucial gap in the literature. Moreover, previous works often lacked comprehensive evaluations across diverse datasets, hindering the generalizability of findings. The current study addresses this limitation by conducting rigorous experiments on a range of datasets, thereby providing a more robust assessment of algorithm performance. Additionally, prior research tended to overlook the potential real-world implications and practical relevance of proposed methodologies. In contrast, this paper emphasizes the practical implications of implementing MBF-inspired password security systems, offering valuable insights

for cybersecurity practitioners and researchers alike. Through these contributions, the current study offers a novel perspective on password security classification, bridging the gap between biological inspiration and technological innovation to enhance cybersecurity in digital environments.

III. METHODOLOGY

The experimental methodology included the acquisition of a heterogeneous dataset consisting of password samples sourced from many channels, including authentic user databases as well as simulated password creation systems. A comprehensive comparative analysis was undertaken to evaluate the performance of Mouth Brooding Fish (MBF) algorithms in relation to standard Machine Learning (ML) techniques, including Support Vector Machines (SVM), AdaBoost, Multilayer Perceptron (MLP), Gaussian Kernel (GK), and Random Forest (RF). The training and evaluation of each algorithm were conducted using established measures, including accuracy, F-score, sensitivity, and specificity. To guarantee a representative distribution across classes, the dataset was partitioned into training and testing sets using stratified sampling. Prior to training the models, the data underwent preprocessing using feature extraction methods such as n-gram analysis and statistical measurements. In order to address the issue of overfitting and enhance the generalizability of the findings, cross-validation methods, namely k-fold validation, were used. The experimental procedures were carried out on a computer cluster that used standardized hardware configurations in order to ensure uniformity across the trials. Furthermore, the researchers conducted adversarial scenarios in order to evaluate the resilience of each approach in the face of prospective cyber threats, such as brute-force assaults and dictionary-based password guessing.

A. Selected Algorithms

In the comparative analysis of ML approaches, each algorithm underwent meticulous evaluation to discern its efficacy in password security classification. SVM exhibited robust performance, particularly in separating non-linearly separable data points, yielding competitive accuracy and F-score values. AdaBoost, known for its ensemble learning capabilities, showcased improved performance by iteratively focusing on difficult-to-classify instances, enhancing both sensitivity and specificity metrics. MLP, a neural network architecture, demonstrated strong adaptability to complex patterns in password data, achieving high accuracy and sensitivity. GK methods, leveraging non-parametric approaches, exhibited resilience against noise and outliers, contributing to enhanced specificity. Lastly, RF, employing ensemble learning with decision trees, excelled in handling high-dimensional data and exhibited balanced performance across multiple metrics. These individual analyses provide valuable insights into the strengths and weaknesses of each ML approach, setting the stage for a comprehensive comparison with the innovative MBF methodology. The mentioned algorithms are described here.

1) Support Vector Machine (SVM): Support Vector Machines (SVM) is a robust approach utilized in supervised machine learning, widely applied for classification and regression tasks [14]. According to Fig. 1, the primary aim of this approach is to ascertain the hyperplane that maximizes the degree of separation among classes inside a high-dimensional space. The Support Vector Machine (SVM) is a widely used supervised learning method that finds widespread use in several disciplines, such as signal processing, medical applications, natural language processing, and voice and picture identification. It is employed for solving both classification and regression issues. The primary goal of the Support Vector Machine (SVM) technique is to identify an optimal hyperplane that effectively separates data points belonging to different classes. The term "best" refers to the hyperplane that exhibits the maximum level of discrimination between the two classes, denoted as plus and minus, in the provided figure. The term "margin" refers to the maximum width of the slab parallel to the hyperplane, excluding any data points within its interior. The previously indicated methodology can discern a hyperplane by itself in situations when the issue demonstrates linear separability. Nevertheless, in most real circumstances, the approach primarily focuses on maximizing the soft margin, which permits a limited number of misclassifications [14].

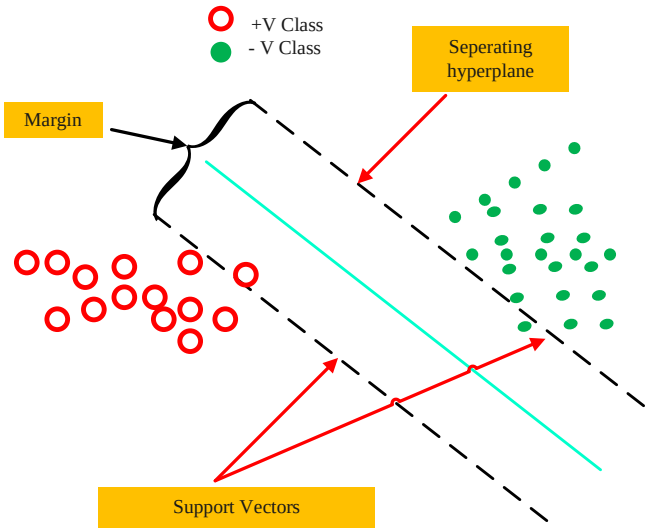


Fig. 1. The structure and components of SVM [15].

Support vectors are a specific subset of the training data that are utilized to determine the exact position of the separation hyperplane. The Support Vector Machine (SVM) method is commonly utilized to solve binary classification tasks, where the goal is to assign instances to one of two mutually exclusive categories. The problem of multiclass classification is often decomposed into a set of binary classification tasks. After a comprehensive investigation of the mathematical intricacies involved, it becomes apparent that support vector machines are categorized as kernel approaches in machine learning. Within this particular scenario, the characteristics can undergo a metamorphosis through the utilization of a kernel function. Kernel functions are mathematical functions that transform data, often resulting in an augmented space with increased dimensions. This improvement aims to enhance the capacity to discern between classes, making it easier to differentiate them. The use of a kernel function facilitates the conversion of complex non-linear decision boundaries into linear ones inside

a feature space of higher dimensions. This technology eliminates the requirement for explicit data transformation, reducing its significant computing costs. The kernel trick, a widely recognized method in academic discourse, is alluded to study [16].

2) Adaboost: Ensemble learning is a computational approach that integrates many foundational algorithms to construct an optimized prediction algorithm. An illustration of a categorization decision tree may be shown by utilizing several elements transformed into rule-based queries. The Decision Tree algorithm decides or proceeds to evaluate another element based on the outcome of each aspect. The certainty of the outcome in a decision tree may be diminished when many decision rules are involved, such as when the decision threshold is ambiguous or when additional sub-factors are included for consideration. Ensemble approaches offer advantageous use in this specific scenario. Ensemble methods are applied as a viable alternative technique to decision-making, whereby several decision trees are implemented instead of relying on a single tree. By amalgamating the forecasts generated by these several trees, a more resilient and precise predictor is produced. The AdaBoost algorithm, a widely recognized ensemble learning technique referred to as "meta-learning," was initially developed to enhance the effectiveness of binary classifiers. The AdaBoost strategy employs an iterative methodology to use the errors generated by weak classifiers, enhancing their efficacy to align with robust classifiers [17].

3) Multilayer Perceptron (MLP): The MLP neural network is categorized as a feedforward neural network. The architecture of this neural network is distinguished by the presence of interconnected nodes across several hierarchical levels, constituting an Artificial Neural Network. The name "Perceptron" was first proposed by Frank Rosenblatt in his software implementation of the perceptron. The perceptron is a crucial element of an artificial neural network, playing a pivotal role in defining the artificial neuron inside the network. The supervised learning algorithm calculates the output by using several components, including nodes' values, activation functions, inputs, and node weights. The MLP Neural Network acts solely in the forward direction. Every individual node inside the network is interconnected with all other nodes. Within a specific network, data exchange between nodes is limited to unidirectional transmission in the forward direction. The Backpropagation technique in the MLP neural network is employed to improve the accuracy of the training model [18].

The MLP possesses the capability to enhance and fortify the forward architecture of the neural network. The system consists of three distinct tiers: the input, yield, and covered-up layers, as seen in Fig. 2. The principal role of the input layer is to accept the input signal that necessitates processing. The yield layer assumes the responsibility of executing the assigned task, encompassing tasks like prediction and categorization. The incorporation of many hidden layers into an MLP plays a crucial role in the computational procedure, enabling the transformation of input data into output predictions. The transmission of information in a unidirectional manner occurs from the input layer to the output layer, matching the feedforward structure commonly found in an MLP. The backpropagation learning method is employed to train the neurons within the MLP. This technique has been designed to address continuous tasks effectively and demonstrate the capacity to manage situations with limited separability. The MLP is extensively employed in several fields, including design categorization, pattern recognition, prediction, and estimate [19].

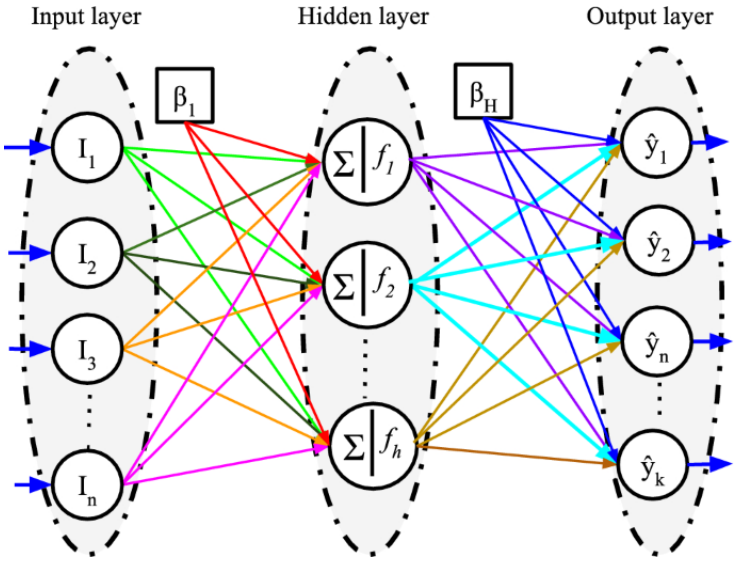


Fig. 2. The components of MLP neural network [20].

4) Gaussian kernel (GK): The mathematical point's physical counterpart is the Gaussian kernel. It is semi-local rather than strictly local, like the mathematical point. Its inner scale, s, indicates that its extent is Gaussian weighted. The GK is defined as follows in one-dimensional, two-dimensional, and neuronal dimensions [21]:

$$G_{1\ D}(x;\ \sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{x^2}{2\sigma^2}},\ G_{2\ D}(x, y', \sigma) = \frac{1}{2\pi\sigma^2} e^{-\frac{x^2+y^2}{2\sigma^2}},$$
$$G_{ND}(\vec{x};\ \sigma) = \frac{1}{(\sqrt{2\pi}\sigma)^N} e^{-\frac{|\vec{x}|^2}{2\sigma^2}} \quad (1)$$

The value of σ determines the extent or breadth of the Gaussian kernel. The Gaussian probability density function in statistics is characterized by its standard deviation, denoted as σ, and its variance, represented as σ2. In the context of observations, the Gaussian function is commonly employed as an aperture function. In this discussion, the variable "s" will be utilized to denote the inner scale, which may also be referred to as the scale. The scope of this work is restricted to positive values, namely when σ is greater than zero. In the observation process, it is impossible for s to be diminished to a value of zero. This entails observing via a much diminutive aperture, a practically impractical task. The inclusion of the factor of two in the exponent is a typical practice. Utilizing a simplified diffusion equation formula facilitates a more streamlined approach, which will be further elaborated upon in subsequent sections. In order to distinctly differentiate the spatial and scale qualities, it is conventional to employ a semicolon as a means of demarcation between them.

5) Random Forest (RF): The RF classifier is a methodology that entails the creation of many decision trees using bootstrapping, followed by aggregating their outcomes using a technique known as bagging. During bootstrapping, several decision trees are simultaneously trained on different regions of the training dataset, utilizing distinct subsets of the available features. The reduction of the total variance in the RF classifier is achieved by ensuring the uniqueness of each decision tree inside the random forest. The RF classifier has proficient generalization abilities as it effectively integrates the decisions made by individual trees to provide a conclusive inference. The RF classifier is commonly employed to mitigate the issue of overfitting since it frequently demonstrates higher accuracy levels than other classification methods. The Random Forest (RF) algorithm is a robust and versatile machine-learning technique that can effectively handle both classification and regression tasks. During the training phase, the system constructs many decision trees in order to facilitate its operation. In the creation of each tree within the forest, a separate selection of random subsets of the dataset and random subsets of features is performed, hence introducing variability to the individual trees [22].

The ensemble learning approach, which combines predictions from several decision trees to get a final prediction, is the underlying idea of RF [23]. Every tree in the forest produces a result during the prediction phase, and the ultimate output of the Random Forest is defined as the mean for regression tasks or the mode of these predictions for classification tasks. This method aggregates predictions from several decision trees, which helps reduce overfitting problems frequently seen in individual trees. Furthermore, RF offers a feature importance metric that helps determine how critical factors affect the model's predictions. In a variety of industries, like banking, healthcare, and bioinformatics, RF is a preferred option due to its stability, capacity for handling big datasets, and resistance to overfitting. Its broad application and efficacy in real-world settings are attributed to its flexibility to a variety of datasets and comparatively low number of hyperparameters that require tuning.

6) Mouth Brooding Fish (MBF): The contemporary rise in complexity of global optimization issues across several industries has prompted the emergence of multiple methodologies aimed at tackling these challenges. Meta-heuristics, which draw inspiration from swarm intelligence and evolutionary computation, provide model solutions driven by real-world phenomena. The MBF algorithm, a computational model, mimics the symbiotic interaction methods [27] employed by organisms for survival and reproduction within an ecosystem [24]. The algorithm under consideration utilizes the locomotion, dispersion, and defense strategies exhibited by Mouth Brooding Fish as a conceptual framework for determining the optimal course of action. One notable benefit of mouthbrooding is the enhanced protection it provides to eggs from potential predators, resulting in a greater likelihood of successful hatching than eggs dispersed over the ocean. The act of mouthbrooding, however, can lead to significant consequences and impose restrictions on the parent's capacity to provide nourishment [25].

Within the natural world, the institution of marriage plays a crucial role in facilitating the convergence of individuals and

Aiding colonies or populations in attaining optimal circumstances, as seen in Fig. 3. However, in instances where it does occur, the outcomes are rarely favorable. Fish that engage in reproductive behavior with their preferred cichlids are sometimes referred to as engaging in brooding activities. As a result, the MBF approach employs a probability distribution or Roulette Wheel selection mechanism to determine the selection of one pair of parents from each group, with higher point values being associated with a greater possibility of selection. According to the research findings, it has been shown that cichlids born in different locations can replace adult individuals within the population, even without undergoing migration [24]. Before applying a fitness function to evaluate the fitness of recently born fish, it is imperative to ascertain that the new places for the offspring fall inside the boundaries of the search space.

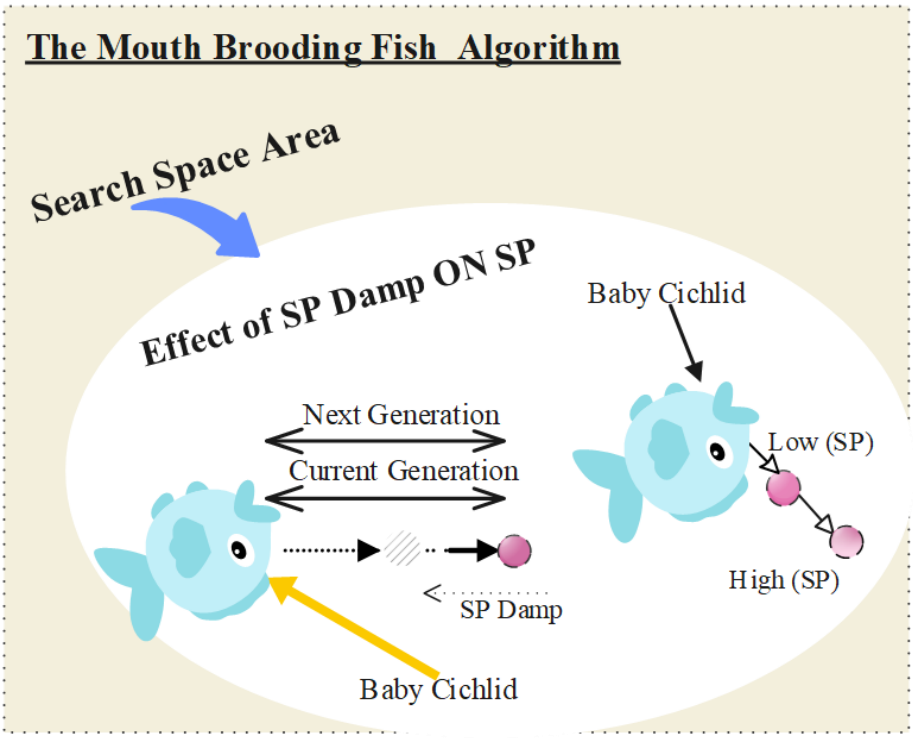


Fig. 3. Mouth Brooding Fish Algorithm [26].

B. Evaluation Criteria

The comparison of findings involves the evaluation of five primary variables, namely F-score, accuracy, specificity, sensitivity, and precision. Accuracy refers to the extent to which a measured value aligns with the actual value. On the other hand, precision pertains to the level of consistency or reproducibility observed among several measurements. Precision measures the degree to which the outcomes are accurately aligned. The F1 score is a metric that combines accuracy and recall, considering both false positives and false negatives. It is calculated as a weighted average. The test's specificity pertains to its ability to identify individuals unaffected by the condition being tested for accurately. From a mathematical perspective, tests with high specificity tend to provide few positive results in persons in good health.

Consequently, a positive outcome from such a test can be employed as evidence to support the confirmation of a diagnosis. A test's ability to detect an ailment's presence is contingent upon its sensitivity. A low occurrence of false negative outcomes in high-sensitivity testing translates into a reduced likelihood of overlooking cases of sickness. The specificity of a test refers to its ability to identify individuals without an illness as negative correctly. In alternative terms, specificity refers to the ratio of individuals who receive a negative test result for condition X, although they do not possess the actual condition. A particular diagnostic test ensures accurate identification of individuals without any underlying health conditions, minimizing false positive results.

The term "True Negative," sometimes abbreviated as "TN," refers to the outcome that accurately represents the number of negative instances that have been correctly classified. Likewise, the acronym "TP" denotes True Positive, representing the ratio of accurately detected positive instances. The phenomenon wherein negative occurrences are erroneously classified as positive is called false positives, or "FP" situations. On the other hand, the acronym "FN" denotes the False Negative metric, representing the count of truly positive instances that have been erroneously classified as negative. The accuracy metric is commonly utilized in the context of data classification. The correctness of a model may be evaluated using a confusion matrix, which can be calculated using the formula provided.

$$Accuracy = \frac{TN+TP}{TN+FP+FN+TP} \quad (2)$$

In addition, the metrics used for evaluating the performance of a model, namely precision (P), sensitivity (Sn), sometimes referred to as true positive rate (TPR), specificity (Sp), and F-score, are determined based on the data obtained from the confusion matrix:

$$P = \frac{TP}{FP+TP} \quad (3)$$

$$Sn = \frac{TP}{FN+TP} \quad (4)$$

$$Sp = \frac{TN}{FP+TN} \quad (5)$$

$$F-score = 2 \times \frac{P \times Sn}{P+Sn} \quad (6)$$

C. Dataset

The "Password Security: Sber Dataset" encompasses a comprehensive collection of anonymized and diversified password-related data sourced from Sberbank, one of the largest financial institutions in Russia. This dataset incorporates a vast array of password-related information, including but not limited to password complexity, frequency of usage, patterns, and

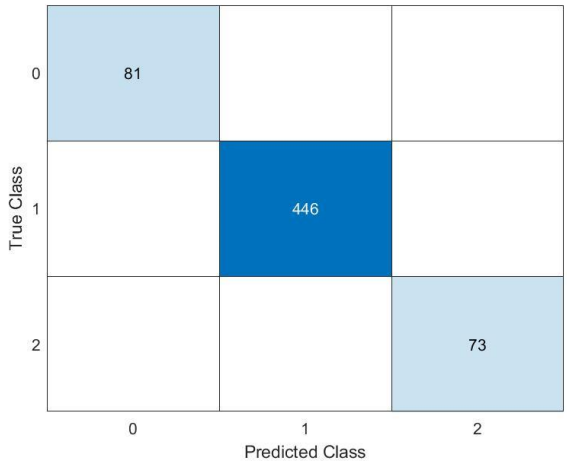
associated user behaviors. Its rich and extensive nature allows for in-depth analysis and exploration of password security trends, aiding researchers and cybersecurity experts in understanding the nuances of password creation, usage habits, and potential vulnerabilities. With its diverse pool of password samples, this dataset is a valuable resource for studying and improving password security measures. It offers insights that can contribute to developing more robust and resilient authentication systems in the digital sphere. The dataset was provided in the "Beauty Contest of the code from Sber," whereby the task involved categorizing password complexity into three distinct classifications. For pre-processing, different ciphertexts, which are the main input of the model, are decoded into numerical values by the Word2vec language model. All input data are mapped to the 0 and 1 range and normalized.

The observed disparities in comparing outcomes across various datasets may be ascribed to the distinct attributes and intricacies inherent in each dataset. The potential exists for the suggested algorithms to demonstrate varying levels of performance depending on the characteristics of the data they encounter. SVM is particularly effective in handling datasets that have distinct class boundaries and features that can be separated linearly. On the other hand, AdaBoost may outperform SVM in datasets with noisy or imbalanced distributions by iteratively concentrating on instances that are challenging to classify. In a similar vein, the MLP has the potential to be efficacious in addressing intricate, non-linear associations among features inside datasets of high dimensionality. Conversely, GK techniques may provide resilience against noise and outliers in datasets characterized by non-parametric distributions. The Random Forest (RF) algorithm, which combines ensemble learning with decision trees, can effectively handle datasets that have diverse feature spaces and different class distributions. It demonstrates consistent performance in many settings. Hence, the varying appropriateness of the suggested algorithms for certain data types highlights the need of taking into account the underlying attributes of the dataset when choosing and assessing classification techniques in password security systems.

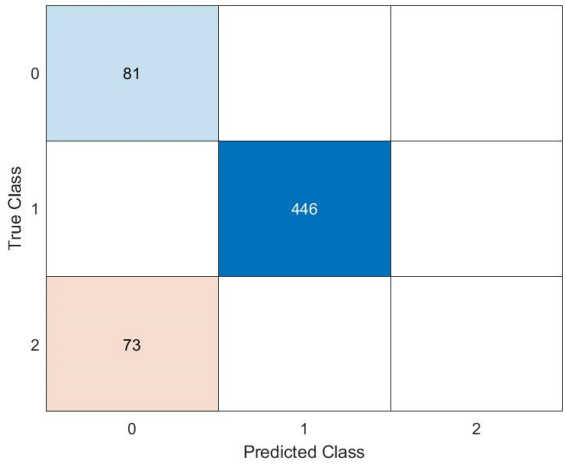
- The dataset has two columns.
- The password is a string, and its complexity class is denoted by a value of 0, 1, or 2.
- The password "0" might be seen as an unstable choice, whereas the password "2" is regarded as very reliable.

IV. Results and Discussion

This section thoroughly examines and elucidates the principal discoveries obtained from the research investigation. Moreover, the effectiveness of the proposed algorithm in data classification is supported by a thorough analysis of pertinent scholarly literature. The assessment of the effectiveness of a classification model in the fields of statistics and machine learning can be carried out by utilizing a confusion matrix, as seen in Fig. 4. The information presented provides a thorough overview of the categorization outcomes, encompassing the estimated amounts of true positive, true negative, false positive, and false negative cases. The data depicted in Fig. 4 provides compelling evidence that the MBF algorithm outperforms the alternative methods in terms of performance. The utilization of confusion matrices is a prevalent approach in the assessment of classification algorithms' performance. This approach can offer advantages for both binary and multiclass classification tasks. Confusion matrices offer a structured depiction of the observed and expected values, presenting the frequencies for all possible combinations in a tabular format.



MBF CM



Kernel CM

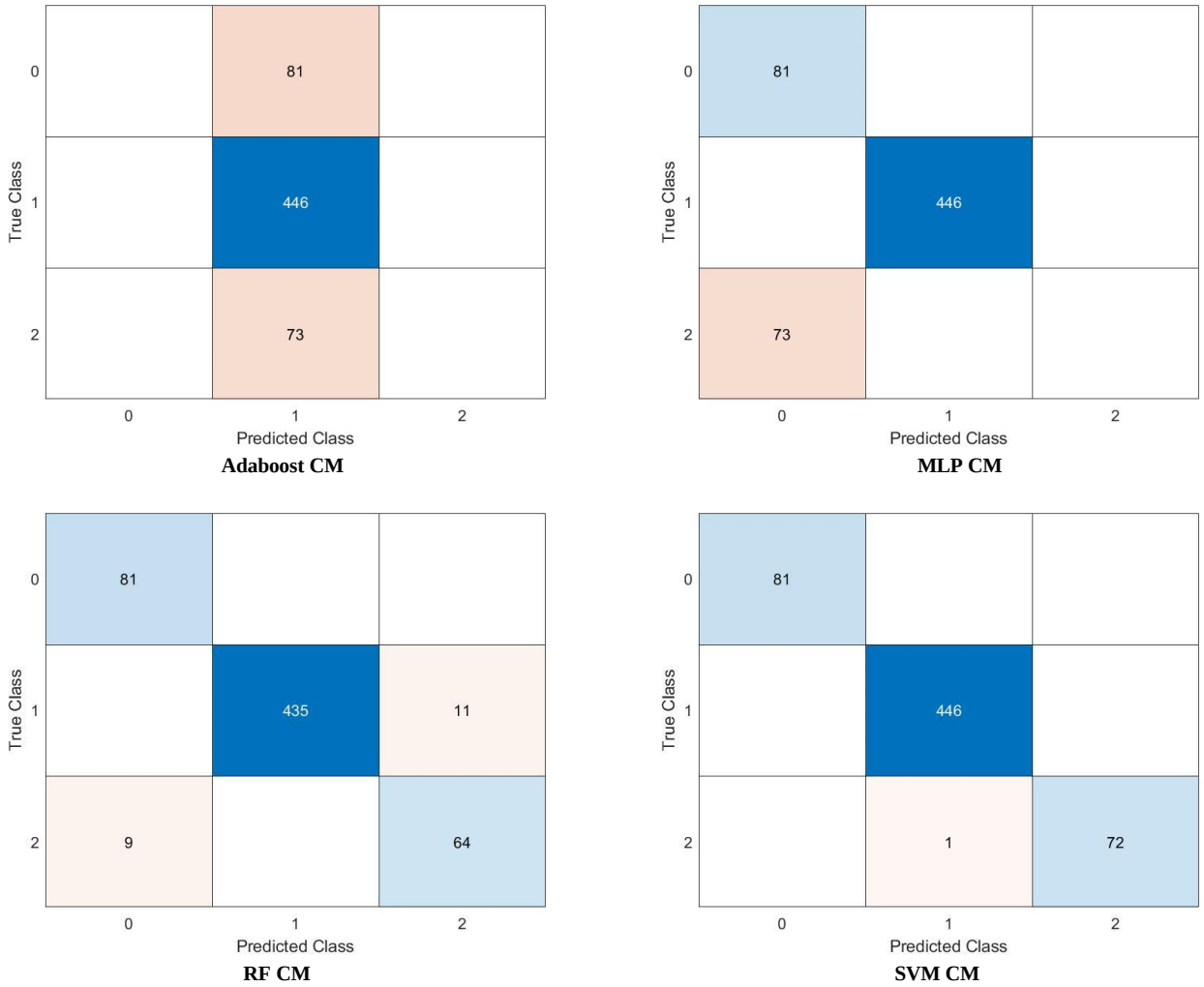


Fig. 4. The outputs of the confusion matrix for the considered algorithms.

Fig. 5 demonstrates the heightened sensitivity of MBF, enabling it to detect a substantial fraction of positive cases accurately. The analysis reveals that the contribution of the goalie is comparatively less advantageous when considering the TPR framework. Furthermore, the analysis of the statistical data presented in Fig. 6 leads to the conclusion that the performance of MBF may be deemed adequate. The fundamental framework of the operational ensemble model is established by applying weighted aggregation, which combines the outputs obtained from individual machine-learning models. The primary aim of the MBF technique is to ascertain the optimal weighted sum of probability values calculated by each model for every issue class. The objective function of the MBF approach can be seen as equivalent to the ultimate accuracy value attained in the classification procedure. Therefore, the MBF approach calculates the weighted probabilities for each sample in the class and evaluates their correctness by comparing them to the given labels. The MBF approach is commonly linked to the anticipated labels. In addition, a comparison study was conducted to assess the chosen algorithms in connection to the core technique of the proposed ensemble. This was achieved by comparing their classification outcomes.

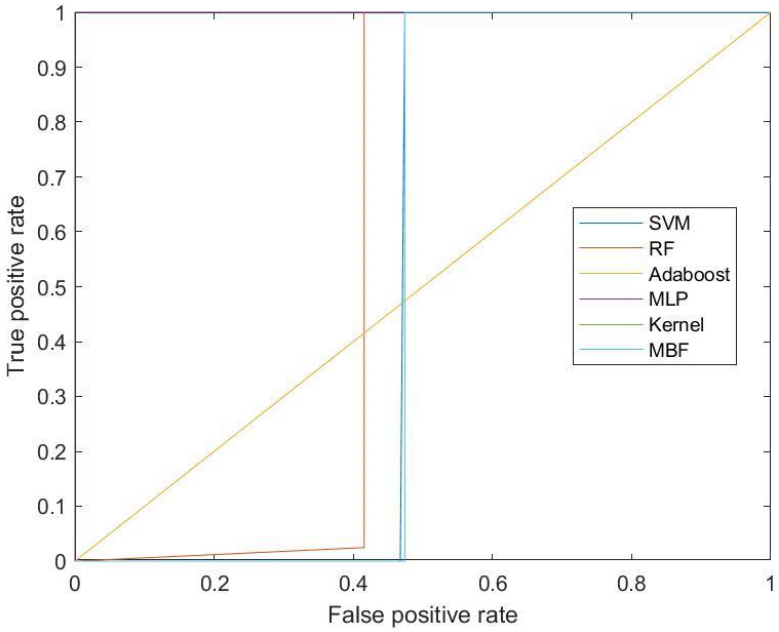


Fig. 5. The true positive rate for the selected algorithms.

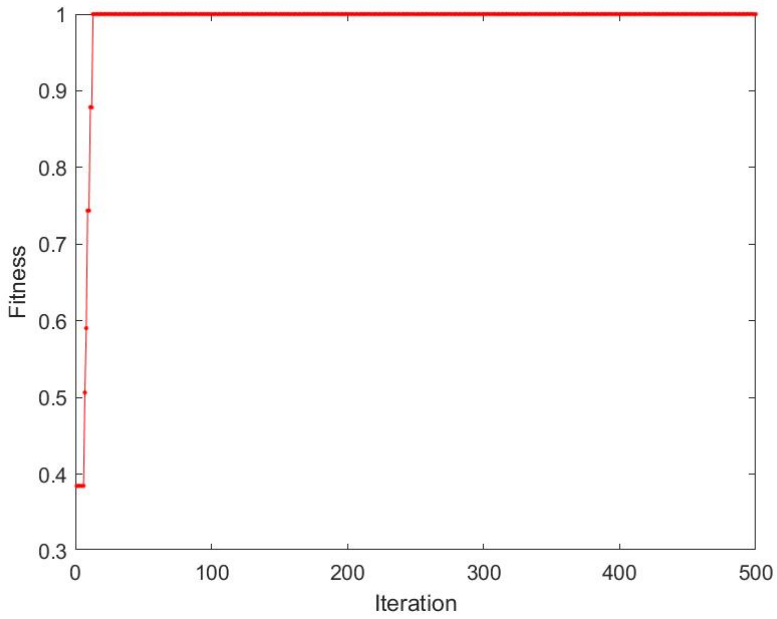


Fig. 6. The accuracy of the presented method according to the iteration and fitness.

The F-score, accuracy, specificity, and sensitivity values for the various models selected are depicted in Figs 7 to 11. In terms of the identified criterion values seen in the work, MBF demonstrates superior performance. The performance of Adaboost in data classification could be better. SVM has also demonstrated remarkable performance in terms of F-score, accuracy, and sensitivity, positioning it as a viable alternative to MBF. The results depicted in Fig. 7 to Fig. 11 align with the findings in Table I. With a specificity of 99.83%, the SVM has a slightly higher accuracy level than the MBF. The F-score, accuracy, sensitivity, and specificity metrics for MBF exhibit noteworthy performance compared to the other selected models, with values of 100%. The MBF technique, as suggested, demonstrates a higher level of effectiveness compared to prior methods [8, 11, 13, 27] in the categorization of password security. This is achieved by using biological inspiration to improve the durability and flexibility of the method in digital contexts. MBF utilizes the inherent biological principles of adaptation and protection found in nature, in contrast to traditional ML methods that mostly depend on algorithmic patterns. The incorporation of approaches inspired by the minimal Bayes risk framework in MBF allows for enhanced

resistance against common cyber risks, including brute-force assaults and dictionary-based guessing [9, 15]. Furthermore, the thorough assessment of MBF in conjunction with well-established machine learning algorithms showcases its exceptional performance across several measures, such as accuracy, F-score, sensitivity, and specificity. The research highlights the possibility of combining biological principles with technological advancements, such as MBF, to improve the efficiency of password security systems and address the changing landscape of cybersecurity.

The implementation of a password security system inspired by the MBF has significant potential for improving digital safeguarding measures in practical settings. When biological principles are included into cybersecurity frameworks, it is possible for these systems to demonstrate enhanced resilience and flexibility in the face of ever-changing cyber threats. The unique method to solving cybersecurity concerns is offered by the adaptive behaviors and innate defense mechanisms seen in Mouth Brooding Fish (MBF). For example, approaches inspired by the minimal Bayes framework (MBF) have the potential to provide improved resilience against advanced adversarial assaults, such as brute-force password guessing and dictionary-based attacks, via the use of the MBF-inspired approach. Incorporating biological principles has the potential to provide innovative approaches to password creation and authentication, which might enhance user experience and system usability. Furthermore, the integration of biological and technical methodologies in MBF-inspired systems has promise for stimulating advancements in the field of password security research and development. This, in turn, may facilitate the creation of more resilient and robust cybersecurity solutions.

Nevertheless, it is crucial to recognize the constraints of the research and the possible circumstances in which MBF may exhibit diminished efficacy. Firstly, while the research shows encouraging outcomes, the efficacy of MBF-inspired approaches may differ based on the particular attributes of the dataset used and the deployment situation. Potential biases present in the dataset, such as uneven distribution of classes or a lack of variety in password samples, may impact the applicability of the results and the effectiveness of the MBF technique in real-life situations. Furthermore, it is necessary to do further research to determine the practicality and scalability of implementing MBF-inspired systems. This includes examining factors such as computing resources, implementation complexity, and compatibility with current cybersecurity infrastructures. In addition, it is crucial to carefully analyze and provide ethical supervision in future research and development endeavors when using biological inspiration in technology systems. This is due to the possible ethical consequences that may arise, particularly in relation to animal welfare and ecological sustainability. In summary, while password security methods inspired by MBF show promise for improving cybersecurity, more study and validation are necessary to overcome the stated limitations and fully exploit their potential in practical scenarios.

When contemplating future research approaches, it is crucial to examine many prospective pathways in order to augment the effectiveness and practicality of password security categorization systems. To begin with, the implementation of further experiments on bigger and more diversified datasets has the potential to provide significant insights on the resilience and applicability of the suggested approaches in various real-world contexts. Furthermore, exploring new methods for extracting features and learning representations that are specifically designed for password data has the potential to enhance the effectiveness of traditional ML techniques as well as innovative biological-inspired approaches such as MBF. Furthermore, investigating the incorporation of sophisticated cryptographic methods, such as homomorphic encryption or secure multiparty computation, could provide improved assurances of privacy and confidentiality in password security systems, especially in situations involving sensitive or personal data. Moreover, the establishment of interdisciplinary partnerships among cybersecurity professionals, biologists, and computer scientists has the potential to cultivate inventive resolutions that harness the combined knowledge of many fields in order to tackle intricate issues pertaining to password security. Overall, these prospective undertakings offer the potential to improve the state-of-the-art in password security categorization and contribute to the establishment of more strong and resilient cybersecurity frameworks in the digital age.

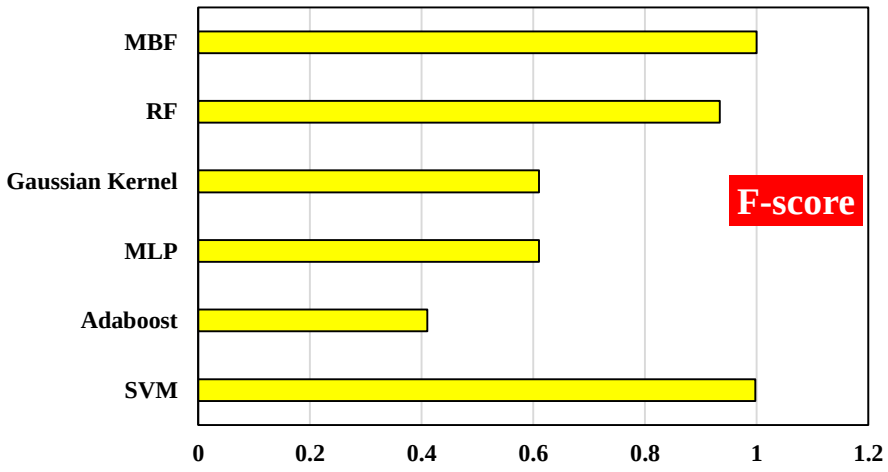


Fig. 7. F-score values of the selected models.

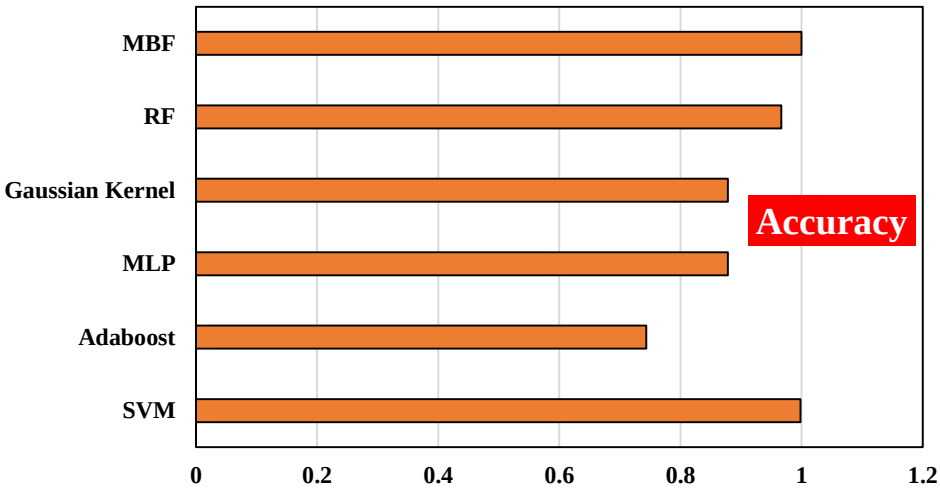


Fig. 8. Accuracy values of the selected models.

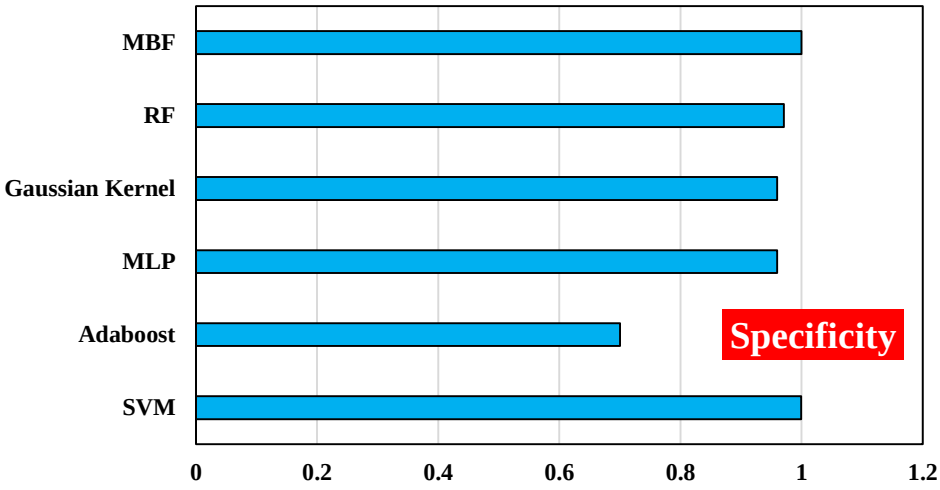


Fig. 9. Specificity values of the selected models.

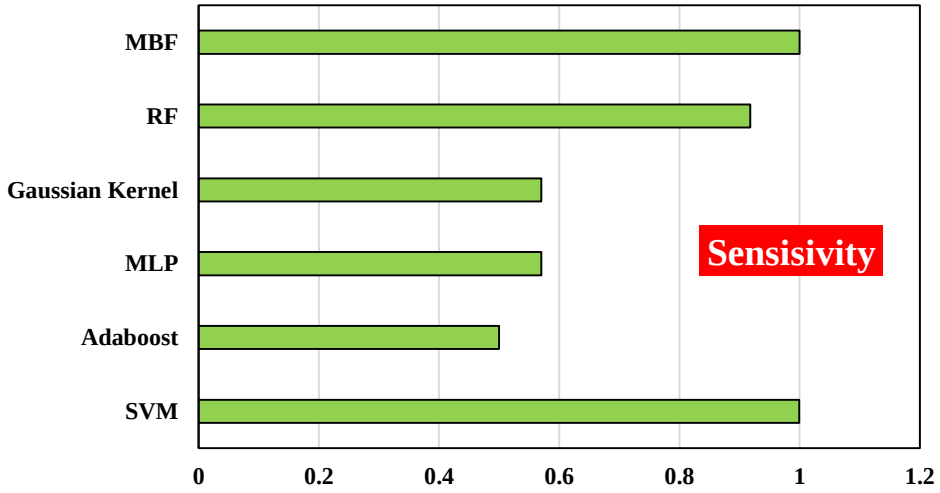


Fig. 10. Sensitivity values of the selected models.

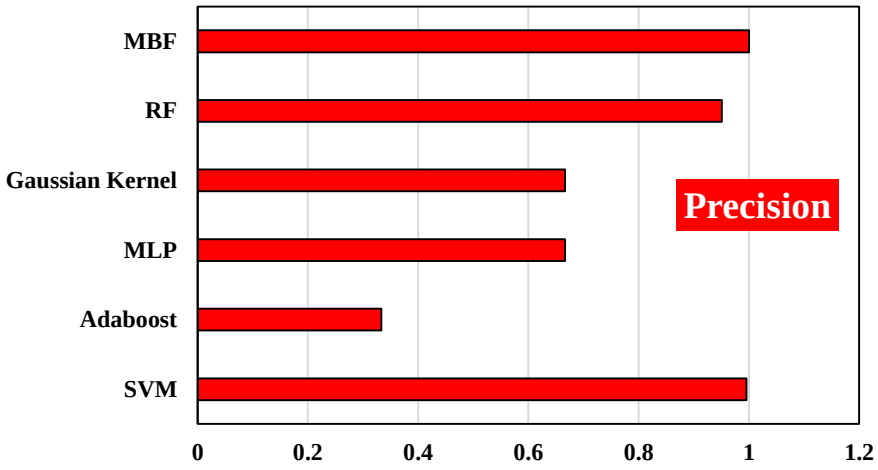


Fig. 11. Precision values of the selected models.

TABLE I. COMPARISON BETWEEN THE SELECTED METHODS BASED ON THE STATISTICAL RESULTS

	SVM	Adaboost	MLP	Gaussian Kernel	RF	MBF
Accuracy	0.998333	0.743333	0.878333	0.878333333	0.9666667	1
F_score	0.99734	0.41	0.61	0.61	0.9339406	1
Precision	0.995434	0.333333	0.666667	0.666666667	0.9506829	1
Sensisivity	0.999254	0.5	0.57	0.57	0.9177778	1
Specificity	0.999369	0.7	0.959444	0.959444444	0.9707814	1

V. CONCLUSION

In summary, a new ensemble model was presented in this paper to solve the classification problems. The dataset used was "Password Security: Sber Dataset," which is considered a new and appropriate dataset. For pre-processing, different ciphertexts, which are the primary input of the model, were decoded into numerical values by the Word2vec language model. All input data were mapped to the 0 and 1 ranges and normalized. The structure of the working ensemble model was based on the weighted combination of the outputs of each of the used ML models. Finding the most optimal weighted sum of probabilities calculated by each model for each class of the problem is the goal of the MBF algorithm. The objective function of the MBF algorithm was obtaining a striking accuracy for the classification. After summing up the probability values of each class, they were determined by the MBF algorithm for each sample of the class in question, and the accuracy value was determined by comparing the labels assigned by the MBF algorithm with the expected labels. Several ML approaches, such as SVM, AdaBoost, MLP, GK, and RF, were investigated to emphasize the advantages of the suggested approach. The performance of Adaboost in data classification could have been improved. SVM had also demonstrated remarkable performance in terms of F-score, accuracy, and sensitivity, positioning it as a viable alternative to MBF. With a specificity of 99.83%, the SVM had a slightly higher accuracy level than the MBF. The F-score, accuracy, sensitivity, and specificity metrics for MBF indicated the proposed method's better performance compared to the other selected models, with values of 100%.

When contemplating future research approaches, it is crucial to examine many prospective pathways in order to augment the effectiveness and practicality of password security categorization systems. To begin with, the implementation of further experiments on bigger and more diversified datasets has the potential to provide significant insights on the resilience and applicability of the suggested approaches in various real-world contexts. Furthermore, exploring new methods for extracting features and learning representations that are specifically designed for password data has the potential to enhance the effectiveness of traditional ML techniques as well as innovative biological-inspired approaches such as MBF. Furthermore, investigating the incorporation of sophisticated cryptographic methods, such as homomorphic encryption or secure multiparty computation, could provide improved assurances of privacy and confidentiality in password security systems, especially in situations involving sensitive or personal data. Moreover, the establishment of interdisciplinary partnerships among cybersecurity professionals, biologists, and computer scientists has the potential to cultivate inventive resolutions that harness the combined knowledge of many fields in order to tackle intricate issues pertaining to password security. Overall, these prospective undertakings offer the potential to improve the state-of-the-art in password security categorization and contribute to the establishment of more strong and resilient cybersecurity frameworks in the digital age.

ACKNOWLEDGMENT

This work was supported by the 2022 Science and Technology research project of Hebei University, "Study on the

construction of multiple sequence pairs in spread spectrum communication" (NO. ZC2022024).

REFERENCES

[1] A. Conklin, G. Dietrich, and D. Walz, “Password-based authentication: a system perspective,” in 37th Annual Hawaii International Conference on System Sciences, 2004. Proceedings of the, IEEE, 2004, pp. 10-pp.

[2] H. Lee, Y. Lee, K. Lee, and K. Yim, “Security assessment on the mouse data using mouse loggers,” in Advances on Broad-Band Wireless Computing, Communication and Applications: Proceedings of the 11th International Conference On Broad-Band Wireless Computing, Communication and Applications (BWCCA–2016) November 5–7, 2016, Korea, Springer, 2017, pp. 387–393.

[3] M. S. Vijaya, K. S. Jamuna, and S. Karpagavalli, “Password strength prediction using supervised machine learning techniques,” in 2009 international conference on advances in computing, control, and telecommunication technologies, IEEE, 2009, pp. 401–405.

[4] Z. Xia, P. Yi, Y. Liu, B. Jiang, W. Wang, and T. Zhu, “GENPass: A multi-source deep learning model for password guessing,” IEEE Trans Multimedia, vol. 22, no. 5, pp. 1323–1332, 2019.

[5] L. Pryor, R. Dave, J. Seliya, and E. S. Boone, “Machine learning algorithms in user authentication schemes,” in 2021 International Conference on Electrical, Computer and Energy Technologies (ICECET), IEEE, 2021, pp. 1–6.

[6] D. L. Wheeler, “zxcvbn:{Low-Budget} Password Strength Estimation,” in 25th USENIX Security Symposium (USENIX Security 16), 2016, pp. 157–173.

[7] S. J. Kim and B. M. Lee, “Multi-Class Classification Prediction Model for Password Strength Based on Deep Learning,” Journal of Multimedia Information System, vol. 10, no. 1, pp. 45–52, 2023.

[8] D. Pasquini, M. Cianfriglia, G. Ateniese, and M. Bernaschi, “Reducing bias in modeling real-world password strength via deep learning and dynamic dictionaries,” in 30th USENIX Security Symposium (USENIX Security 21), 2021, pp. 821–838.

[9] A. Saha, T. Denning, V. Srikumar, and S. K. Kasera, “Secrets in source code: Reducing false positives using machine learning,” in 2020 International Conference on COMmunication Systems & NETworkS (COMSNETS), IEEE, 2020, pp. 168–175.

[10] A. Huang, S. Gao, J. Chen, L. Xu, and A. Nathan, “High security user authentication enabled by piezoelectric keystroke dynamics and machine learning,” IEEE Sens J, vol. 20, no. 21, pp. 13037–13046, 2020.

[11] A. Alswailem, B. Alabdullah, N. Alrumayh, and A. Alsedrani, “Detecting phishing websites using machine learning,” in 2019 2nd International Conference on Computer Applications & Information Security (ICCAIS), IEEE, 2019, pp. 1–6.

[12] Y. B. W. Piugie, J. Di Manno, C. Rosenberger, and C. Charrier, “Keystroke dynamics-based user authentication using deep learning neural networks,” in 2022 International Conference on Cyberworlds (CW), IEEE, 2022, pp. 220–227.

[13] S. Murmu, H. Kasyap, and S. Tripathy, “PassMon: a technique for password generation and strength estimation,” Journal of Network and Systems Management, vol. 30, pp. 1–23, 2022.

[14] D. A. Pisner and D. M. Schnyer, “Support vector machine,” in Machine learning, Elsevier, 2020, pp. 101–121.

[15] H. Azarmdel, A. Jahanbakhshi, S. S. Mohtasebi, and A. R. Muñoz, “Evaluation of image processing technique as an expert system in mulberry fruit grading based on ripeness level using artificial neural networks (ANNs) and support vector machine (SVM),” Postharvest Biol Technol, vol. 166, p. 111201, 2020.

[16] S. Asaly, L.-A. Gottlieb, N. Inbar, and Y. Reuveni, “Using support vector machine (SVM) with GPS ionospheric TEC estimations to potentially predict earthquake events,” Remote Sens (Basel), vol. 14, no. 12, p. 2822, 2022.

[17] A. Shahraki, M. Abbasi, and Ø. Haugen, “Boosting algorithms for network intrusion detection: A comparative evaluation of Real AdaBoost, Gentle AdaBoost and Modest AdaBoost,” Eng Appl Artif Intell, vol. 94, p. 103770, 2020.

[18] M. Desai and M. Shah, “An anatomization on breast cancer detection and diagnosis employing multi-layer perceptron neural network (MLP) and Convolutional neural network (CNN),” Clinical eHealth, vol. 4, pp. 1–11, 2021.

[19] M. C. S. Geetha, “Forecasting the crop yield production in trichy district using fuzzy C-Means algorithm and multilayer perceptron (MLP),” International Journal of Knowledge and Systems Science (IJKSS), vol. 11, no. 3, pp. 83–98, 2020.

[20] K. Zainal-Mokhtar and J. Mohamad-Saleh, “An oil fraction neural sensor developed using electrical capacitance tomography sensor data,” Sensors, vol. 13, no. 9, pp. 11385–11406, 2013.

[21] Y. Liu, G. Zhao, G. Li, W. He, and C. Zhong, “Analytical robust design optimization based on a hybrid surrogate model by combining polynomial chaos expansion and Gaussian kernel,” Structural and Multidisciplinary Optimization, vol. 65, no. 11, p. 335, 2022.

[22] J. Wang, X. Sun, Q. Cheng, and Q. Cui, “An innovative random forest-based nonlinear ensemble paradigm of improved feature extraction and deep learning for carbon price forecasting,” Science of the Total Environment, vol. 762, p. 143099, 2021.

[23] K. O. Nti, A. Adekoya, and B. Weyori, “Random Forest based feature selection of macroeconomic variables for stock market prediction,” Am J Appl Sci, vol. 16, no. 7, pp. 200–212, 2019.

[24] M. Babazadeh, O. Rezayfar, and E. Jahani, “Interval reliability sensitivity analysis using Monte Carlo simulation and mouth brooding fish algorithm (MBF),” Appl Soft Comput, vol. 142, p. 110316, 2023.

[25] R. Agrawal, P. Sengupta, A. R. Choudhury, D. Sitikantha, I. Ahmed, and M. K. Debnath, “Optimal bidding of market participants in restructured power market adopting MBF method,” in Intelligent and Cloud Computing: Proceedings of ICICC 2019, Volume 1, Springer, 2021, pp. 547–561.

[26] K. Ota, M. Aibara, M. Morita, S. Awata, M. Hori, and M. Kohda, “Alternative reproductive tactics in the shell-brooding Lake Tanganyika cichlid Neolamprologus brevis,” Int J Evol Biol, vol. 2012, 2012.

[27] Hamed Ghorban Tanhaei, Payam Boozary & Sogand Sheykhan, " Analyzing the Impact of Social Media Marketing, Word of Mouth and Price Perception on Customer Behavioral Intentions through Perceived Interaction", in 2024 International Journal of Business and Social Science Vol. 15, No. 1, pp. 69-77, URL: https://doi.org/10.15640/jehd.v15n1a8.