Improved SVM for Website Phishing Detection Through Recursive Feature Elimination

Feliciano Surya Marcello [1], Farica Perdana Putri [2]
[1,2,3] Department of Informatics, Faculty of Engineering and Informatics, Universitas Multimedia Nusantara, Indonesia
[1]feliciano.marcello@student.umn.ac.id, [2]farica@umn.ac.id

Accepted 09 January 2025
Approved 16 January 2025

***Abstract*— The website has become one of the most widely used media to obtain information, conduct business, transfer data, and others. The development and use of increasingly sophisticated websites also pose a threat to information security, called cybercrime. Website phishing is one type of cybercrime where the perpetrator creates a fake website that mimics the original one in order to steal sensitive user information. The Support Vector Machine (SVM) algorithm is one way that can be implemented in detecting phishing websites by classifying through checking website features. Selecting relevant features for SVM is important to generalize performance and computational efficiency. Thus, the use of Recursive Feature Elimination (RFE) feature selection is conducted to improve the performance of SVM (SVM-RFE). By eliminating irrelevant features, RFE contributes to the increasing accuracy rate for the SVM by 0.15% with an accuracy of 96.09%. In other experiment, the accuracy results of SVM-RFE significantly improved by approximately 15% to 20%.**

***Index Terms*— Feature Selection, Recursive Feature Elimination, Support Vector Machine, Website Phishing Detection.**

I. Introduction

Behind the rapid development of information technology and the internet, there are threats that can harm users of this technology and internet which is commonly called cybercrime [1]. One of the cybercrimes that often occurs today is phishing. A phishing attack is a cybercrime that uses social engineering to deceive users and steal victims' information data, such as personal identity, important information related to finance, and others. Phishing attacks can be carried out in several ways, such as sending fake messages via email or social media platforms, as well as websites [2], [3], [4]. Users who are not aware of this attack will usually be asked to enter information or download files containing malware that can steal the personal data of the victim's device [5].

Website phishing itself is one of the main problems in website security. Phishers usually send Uniform Resource Locator (URL) to victims via email, SMS, and social media [6]. In the second quarter of 2023, the Anti-Phishing Working Group (APWG) noted that there were 1,286,208 phishing attacks and there were 597,789 phishing website attacks detected. The losses resulting from these attacks are considerable, so improvements need to be made.

A phishing website is a replica of a legitimate website. The entire phishing website is not built for phishing, but only a few pages that are specialized by the perpetrator to provide input or download, so that when data is sent, the data is sent to the attacker. There are several methods of cloning websites, which can be done by using special software or by creating manually [7]. The features used to detect phishing websites fall into three categories, namely URL-based features, content-based features, and external service features. URL-based features are features taken directly from the URL of the website, while content-based features refer to characteristics taken from the content of the web page. The content-based features analyze the items displayed on the website or contained within its HTML code to detect trends or abnormalities that could indicate a phishing attempt, such as hyperlink content and abnormal content. External service features are features obtained by querying third-party services or search engines, such as WHOIS, Alexa, Openpagerank, and Google [1].

Support Vector Machine (SVM) is one of the algorithms in machine learning. In SVM, each data item is mapped as a point in n-dimensional space and this algorithm builds a dividing line for the classification of two classes known as hyperplane [8]. Research conducted in [9] detects phishing websites using the SVM algorithm. The research uses six attributes of the URL, including Long URL, Dots, IP Address, SSL Connection, At (@) Symbol, and Dash (-) symbol which are used as the main features in SVM model training. Study in [10], uses Logistic Regression, k-Nearest Neighbors, Decision Tree, Random Forest, SVM, and Gradient Boosting to detect phishing websites. The results show that the proposed SVM algorithm has the lowest performance. The use of kernels in SVM also cannot handle noisy data and overlapping target classes. Therefore, it can be concluded that the SVM algorithm still dealing with high-dimensionality in the feature space [10]. Selecting relevant features for SVM classifiers improves generalization performance, computational efficiency, and feature interpretability [11], [12]. To address the issue of high-dimensional feature space, feature selection methods play a crucial role.

Recursive Feature Elimination (RFE) is one of the feature selection methods that performs model selection based on the learnt model and classification accuracy. RFE sequentially removes unimportant features that can cause a decrease in classification accuracy so that after obtaining the best features, this technique rebuilds a new classification model. The model is trained with the training dataset, feature weights that reflect the importance of each feature are obtained. The features that have been sorted by the highest weight, will be reclassified so that the use of the RFE method based on feature importance can be obtained [13]. This study aims to enhance the performance of SVM algorithm by utilizing RFE feature selection for detecting the web phishing.

II. Theory

A. Recursive Feature Elimination

Recursive Feature Elimination (RFE) is a wrapper method in feature selection. It is a method that works by removing redundant and weak features whose removal affects the training error the least and keeping independent and strong features to improve the generalization performance of the model. RFE initially builds a model on the entire feature set and ranks the features based on their importance. After ranking, the lowest ranked feature is removed and the model is rebuilt and re-ranked for the most important feature [14]. Here are the steps in performing the RFE process, namely:

1. Train the model using all features with 10-fold cross-validation.
2. Calculate the mode performance to determine the value (accuracy, precision, and recall) Determine the confusion matrix to use
3. Comparing features with the highest to lowest weights
4. Discard the feature with the lowest weight
5. Perform iterations to calculate the performance of the model until it is finalized:
   (a) Train and test the model on the most recent features
   (b) Recalculate model performance
   (c) Comparing the features with the highest to the lowest weights
   (d) Discard the feature with the lowest weight
6. Use the selected optimal model

B. Support Vector Machine

Support Vector Machine (SVM) is one of the powerful algorithms in machine learning. In SVM, each data item is mapped as a point in an n-dimensional space and the algorithm constructs a dividing line for classification of two classes known as a hyperplane [8]. A hyperplane is a dividing line between classes by maximizing the margin between them. The margin is the distance between the hyperplane and the nearest point in n-dimensional space in each class. The SVM algorithm is included in the ensemble learning method. This is because the learning system of this model uses a hypothesis space in the form of functions from a high-dimensional feature space. In this space, there are many boundaries that can be used to separate the classes, but there is only one boundary that maximizes the margin [15]. The kernel function plays a crucial role in SVM [16] by transforming the input data into a higher-dimensional space where the data becomes linearly separable, including Linear, Radial Basic Function (RBF), and Polynomial kernel.

Linear kernel is the simplest kernel, without using the gamma value (γ) as in (1). x_i is the value of the training data, x_j is the value of the test data, and $k(x_i.x_j)$ is the kernel value.

$$k(x_i.x_j) = x_i^T x_j \tag{1}$$

RBF is a non-linear kernel, using the gamma value parameter (γ>0) as a determinant of the flexibility of this kernel, can be described by (2). This kernel is suitable for data that cannot be solved linearly with a high level of accuracy and precision [17].

$$k(x_i.x_j) = exp\left(-\gamma.\left\|x_i.x_j\right\|^2\right) \tag{2}$$

Polynomial kernel is a non-linear kernel, using the parameter value of the gamma value (γ > 0) and the value of *d* as the coefficient of the penalty degree for flexibility as described in (3).

$$k(x_i.x_j) = \gamma\ (x_i^T.x_j + r)^d \tag{3}$$

A large gamma value will also calculate training data that is far from the decision boundary but will cause a small accuracy value, and the value of *C* is a free parameter, the value of *r* is a bias, and the gamma and *r* parameters have a strong relationship [18].

III. Method

The methodology of this study consists of several steps and is presented in this section. The following steps are described in detail: data collection, data preprocessing, and the proposed method.

A. Data Collection

The dataset used is a dataset from Mendeley data entitled Web page phishing detection from [1]. This dataset contains 11,430 data with 87 attributes. This data has been labelled as phishing and legitimate. All feature values used in this study are integer data types. Only the label is a string data type. Three feature categories included 56 URL-based features, 24 content-based features, and 7 external service features. The labels are equally distributed. Data is divided into 70% testing data and 30% training data.

B. Data Preprocessing

The preprocessing process is divided into three parts, namely data cleaning, standardization, and label encoding. The data cleaning process checks data errors, incorrect data types, duplicates, and empty data. If those exists, the data will be deleted. After data cleaning, the standardization process is carried out to handle outlier data so that it is still in a good scale distribution. Standardization improved model convergence and performance by ensuring features were scaled uniformly. Standardization using the Z-score method as in (4).

$$Z = \frac{x-\mu}{\sigma} \tag{4}$$

where x is a value of data point, μ is mean of data points, and σ is the standard deviation of all data. Z-score standardization transforms features to have a mean of 0 and a standard deviation of 1. This ensures all features contribute equally to the distance-based calculations, particularly for algorithms like SVM.

Labels on imported datasets have a string data type. It is necessary to do a label encoding process to convert category label data into numeric label data since SVM model requires numeric labels to be able to train the model. The LabelEncoder class from scikit-learn is used to transform the status data (label) to change the value of legitimate website to 1 and phishing website to 0.

C. Proposed Method

The process iterates the training process for each feature using SVM and RFE models. We use 10 cross-validation and calculate the feature ranking based on the accuracy score. After the features are sorted, then remove the features that have the lowest weight. The iteration continues until the features reach convergence. Convergence is a condition when the selected features have found the most optimal optimization and the criteria have been met. In this study, the criteria are the number of features that have been determined.

The SVM model uses parameters previously obtained through the gridsearchCV process. The data training process is carried out by checking convergence. Convergence in SVM refers to SVM training to achieve the most optimal solution or close to the optimal solution. If the convergence has not reached the optimal point, then the training iterates with different points and the point vector will be updated. However, if it has converged, then the accuracy of the hyperparameter will be checked. The prediction results will later be used for system evaluation and determine the performance of the RFE and SVM algorithms

IV. Results and Discussions

In this section, the experiment results are described and analyzed. We also explain the model parameters used in RFE and SVM training, which then become the basis for tuning hyperparameters in the improved-SVM model using RFE, which in this article will be referred to as SVM-RFE.

A. Experimental Setup

The RFECV class of scikit-learn is used with parameters in it, namely, the estimator is SVM as a model, CV is cross validation worth 10, and the weight seen is accuracy. In SVM, the parameters used are C, gamma, and kernel. The values of C are 0.01, 0.1, 1. Gamma values are 0.01, 0.1, 1. Kernels used are linear, polynomial, and RBF. Hyperparameters like the penalty parameter C, kernel type, and kernel-specific parameters control the trade-off between model complexity and accuracy or how the decision boundary is shaped. The hyperparameters values are determined based on the previous studies. Training is carried out on the dataset used, and the results sought are the best accuracy of each parameter.

B. Evaluation Measures

Precision, recall, accuracy, and F1-score are used to evaluate the proposed model. Then the evaluation results are visualized using a heatmap, to see the number of correct and incorrect predictions from the test data using the improved SVM model with RFE. The metrics precision, recall, F1-score, and accuracy are calculated as in (5), (6), (7), and (8), respectively.

$$Precision = \frac{TP}{TP+FP} \tag{5}$$

$$Recall = \frac{TP}{TP+FN} \tag{6}$$

$$F1-score = \frac{2 \times (Recall \times Precision)}{Recall+Precision} \tag{7}$$

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \tag{8}$$

C. Results

The results of finding the best hyperparameters using *gridsearchcv* get results that are C = 1.0, gamma = 0.01, and kernel = 'RBF'. The accuracy of the SVM is 0.96092.

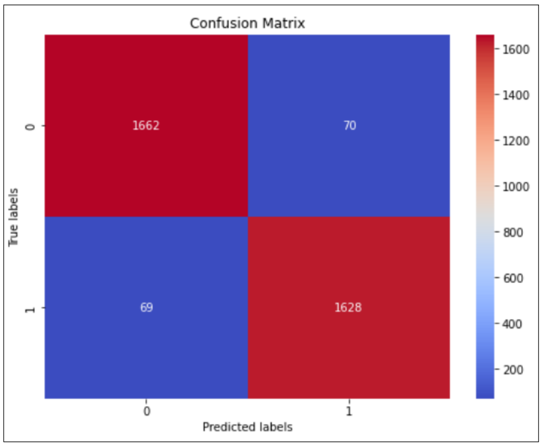

Fig. 1. Confusion matrix without RFE

The evaluation results of the SVM model without RFE can be seen in Figure 1. There are 1662 data for true negative (TN), 69 for false negative (FN), 70 for false positive (FP), and 1628 true positive (TP) data. The evaluation results of the SVM model with RFE are shown in Figure 2. There are 1667 true negative data, 69 false negative data, 65 false positive data, and 1628 true positive data.

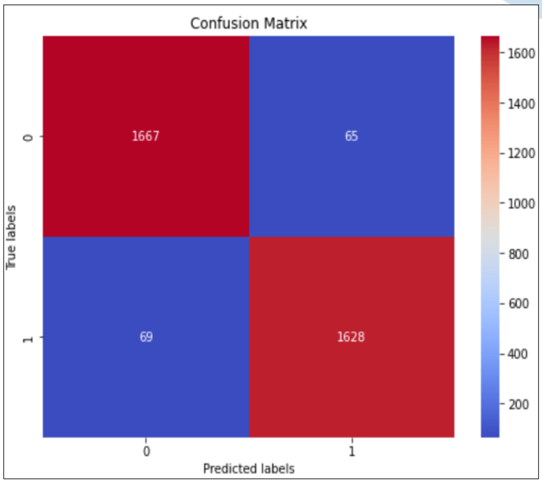

Fig. 2. Confusion matrix with RFE

TABLE I. COMPARISONS OF SVM WITH AND WITHOUT RFE

Method	Precision	Recall	F1-score	Accuracy
Without RFE	0.96	0.96	0.96	0.9594
With RFE	0.96	0.96	0.96	0.9609

In false positives, there is a difference in the number, namely RFE = 70 and without RFE = 65. In Table 1, the precision, recall, and F1-score values have the same results in RFE and without RFE, there is a difference in the accuracy value, namely with RFE 96.09% and without RFE 95.94%. From the results of this analysis, it can be concluded that the use of RFE on the dataset used in this study has a good impact because it has increased accuracy by 0.15%. The significance of a 0.15% improvement in accuracy may appear insignificant on the surface, but it can have a significant real-world impact, particularly in critical applications such as phishing website detection. Phishing website detection systems are deployed on a huge scale, screening millions or billions of webpages each day. A 0.15% increase in accuracy means properly identifying thousands (or possibly more) of extra phishing websites every day, stopping multiple phishing attacks.

TABLE II. FEATURE SELECTION COMPARISONS WITH [1]

Feature Selection	Accuracy
U + C	81.74%
U + E	76.50%
C + E	80.20%
RFE	**96.04%**

To test the effectiveness of the proposed method, namely SVM-RFE, we also conducted a comparison with the research conducted by Abdelhakim Hannousse [1]. In [1], a combination of features was performed, with the following features: URL-based features (U in Table 2), content-based features (C), and external-based features (E). In Table 2, the accuracy results in this current study which uses RFE have significantly improved compared to previous studies with an increase in accuracy of approximately 15% to 20%.

This paper also uses a different dataset to analyse the proposed SVM-RFE algorithm. In research conducted by Muhammad Hasan [9], the research used a different dataset taken from kaggle which has 32 features, and 11,054 data samples. With the same division of training and testing data: 70% training data and 30% testing data. Study in [9] does not use any selection features and uses the full features in the dataset. The results obtained from the study are also not very good accuracy which is 56.05%. Experiments were also conducted for the dataset used in [9] by comparing with SVM-RFE.

The features selected and removed from the new dataset are nine features and the remaining 22 selected features used for classification. The SVM model was performed using gridsearchCV using a polynomial kernel, gamma of 0.1, and C of 1.0 which are the best parameters. SVM-RFE model obtained an accuracy of

95.23%, while [9] only obtained an accuracy result of 56.05%. This result proves that RFE can provide a great improvement to the performance of SVM classification in detecting phishing websites.

V. Conclusion

Phishing websites are common and cause harm to ordinary people. SVM is one of the machine learning algorithms used to detect phishing websites. However, SVM has disadvantages if the data used has many features. Therefore, feature selection is used to remove irrelevant and redundant features. The use of RFE feature selection coupled with evaluation using 10 cross-validation, makes feature elimination more accurate and better. The use of standardization is also important for the dataset used so that there is no outlier data so that the accuracy generated when predicting training data with the SVM algorithm becomes more accurate and higher.

In this study, the prediction results have an accuracy of 96.09%. In experiments with different datasets with SVM algorithm and RFE feature selection, there is a significant improvement. Without RFE, different datasets only get 56.05% accuracy, while using RFE increases accuracy to 95.23%. Thus, from several trials conducted, it can be concluded that the use of feature selection Recursive Feature Elimination has a good impact in handling many features and helping the Support Vector Machine algorithm in optimizing accuracy for phishing website classification. RFE feature selection is very large due to the repetition of feature checking that is done, necessitating further research on techniques to improve computation complexity.

ACKNOWLEDGMENT

We would like to thank Universitas Multimedia Nusantara (UMN) for supporting this research

References

[1] A. Hannousse and S. Yahiouche, “Towards benchmark datasets for machine learning based website phishing detection: An experimental study,” *Engineering Applications of Artificial Intelligence*, vol. 104, p. 104347, Sep. 2021, doi: 10.1016/j.engappai.2021.104347.

[2] M. Zouina and B. Outtaj, “A novel lightweight URL phishing detection system using SVM and similarity index,” *Hum. Cent. Comput. Inf. Sci.*, vol. 7, no. 1, p. 17, Dec. 2017, doi: 10.1186/s13673-017-0098-1.

[3] A. Safi and S. Singh, “A systematic literature review on phishing website detection techniques,” *Journal of King Saud University - Computer and Information Sciences*, vol. 35, no. 2, pp. 590–611, Feb. 2023, doi: 10.1016/j.jksuci.2023.01.004.

[4] M. H. Alkawaz, S. J. Steven, A. I. Hajamydeen, and R. Ramli, “A Comprehensive Survey on Identification and Analysis of Phishing Website based on Machine Learning Methods,” in *2021 IEEE 11th IEEE Symposium on Computer Applications & Industrial Electronics (ISCAIE)*, Penang, Malaysia: IEEE, Apr. 2021, pp. 82–87. doi: 10.1109/ISCAIE51753.2021.9431794.

[5] S. Asiri, Y. Xiao, S. Alzahrani, S. Li, and T. Li, “A Survey of Intelligent Detection Designs of HTML URL Phishing Attacks,” *IEEE Access*, vol. 11, pp. 6421–6443, 2023, doi: 10.1109/ACCESS.2023.3237798.

[6] H. Abusaimeh, “Detecting the Phishing Website with the Highest Accuracy,” *TEM Journal*, pp. 947–953, May 2021, doi: 10.18421/TEM102-58.

[7] T. Chin, K. Xiong, and C. Hu, “Phishlimiter: A Phishing Detection and Mitigation Approach Using Software-Defined Networking,” *IEEE Access*, vol. 6, pp. 42516–42531, 2018, doi: 10.1109/ACCESS.2018.2837889.

[8] S. Anupam and A. K. Kar, “Phishing website detection using support vector machines and nature-inspired optimization algorithms,” *Telecommun Syst*, vol. 76, no. 1, pp. 17–32, Jan. 2021, doi: 10.1007/s11235-020-00739-w.

[9] A. Abdulwakil, M. A. Aydin, and D. Aksu, “Detecting phishing websites using support vector machine algorithm,” *Pressacademia*, vol. 5, no. 1, pp. 139–142, Jun. 2017, doi: 10.17261/Pressacademia.2017.582.

[10] S. M. Mahamudul Hasan, N. M. Jakilim, Md. Forhad Rabbi, and R. M. S. Rahman Pir, “Determining the Most Effective Machine Learning Techniques for Detecting Phishing Websites,” in *Applications of Artificial Intelligence and Machine Learning*, vol. 925, B. Unhelker, H. M. Pandey, and G. Raj, Eds., in Lecture Notes in Electrical Engineering, vol. 925. , Singapore: Springer Nature Singapore, 2022, pp. 593–603. doi: 10.1007/978-981-19-4831-2_48.

[11] M. H. Nguyen and F. De La Torre, “Optimal feature selection for support vector machines,” *Pattern Recognition*, vol. 43, no. 3, pp. 584–591, Mar. 2010, doi: 10.1016/j.patcog.2009.09.003.

[12] R.-C. Chen, C. Dewi, S.-W. Huang, and R. E. Caraka, “Selecting critical features for data classification based on machine learning methods,” *J Big Data*, vol. 7, no. 1, p. 52, Dec. 2020, doi: 10.1186/s40537-020-00327-4.

[13] H. Jeon and S. Oh, “Hybrid-Recursive Feature Elimination for Efficient Feature Selection,” *Applied Sciences*, vol. 10, no. 9, p. 3211, May 2020, doi: 10.3390/app10093211.

[14] T. E. Mathew, “A Logistic Regression with Recursive Feature Elimination Model for Breast Cancer Diagnosis,” 2019.

[15] R. I. Arumnisaa and A. W. Wijayanto, “Comparison of Ensemble Learning Method: Random Forest, Support Vector Machine, AdaBoost for Classification Human Development Index (HDI),” *SISTEMASI*, vol. 12, no. 1, p. 206, Jan. 2023, doi: 10.32520/stmsi.v12i1.2501.

[16] Jonathan Leander and Arya Wicaksana, “Optimizing a Personalized Movie Recommendation System with Support Vector Machine and Content-Based Filtering,” *JSMS*, vol. 14, no. 1, Dec. 2023, doi: 10.33168/JSMS.2024.0128.

[17] D. K. Choubey, S. Tripathi, P. Kumar, V. Shukla, and V. K. Dhandhania, “Classification of Diabetes by Kernel Based SVM with PSO,” *RACSC*, vol. 14, no. 4, pp. 1242–1255, Jul. 2021, doi: 10.2174/2213275912666190716094836.

[18] K. V. Kamran, B. Feizizadeh, B. Khorrami, and Y. Ebadi, “A comparative approach of support vector machine kernel functions for GIS-based landslide susceptibility mapping,” *Appl Geomat*, vol. 13, no. 4, pp. 837–851, Dec. 2021, doi: 10.1007/s12518-021-00393-0