ACSSE
APPLIED COMPUTER SCIENCE AND SOFTWARE ENGINEERING
E-ISSN : 0000000000

Phishing Detection: Comparing Accuracy of Decision Tree, Random Forest, SVM Classifiers

Bahriddin Abapihi1, Halu Oleo University, bahriddin.abapihi@uho.ac.id
Dewi Fortuna2, Amikom University Purwokerto,
dewifortuna202411@gmail.com

ABSTRACT

Phishing attacks have become a serious threat in the realm of cybersecurity, with both direct and indirect impacts on individuals and organizations. This study aims to assess the performance of three machine learning classification algorithms—Decision Tree, Random Forest, and Support Vector Machine—in detecting phishing attempts through the use of suitable datasets. The study focuses on comparing the performance of these algorithms to determine their accuracy and reliability in phishing detection. The experimental findings indicated that Random Forest attained the highest accuracy, reaching 96.675%, followed by Decision Tree (93.57%), while Support Vector Machine showed lower performance (54.724%). Cohen's Kappa also indicated that Random Forest had the highest correlation value. In conclusion, for this dataset, Random Forest and Decision Tree are more effective at detecting phishing attacks compared to Support Vector Machine.

Keywords *: Phishing Detection, Machine Learning Algorithms, Random Forest Accuracy, Decision Tree Accuracy, Support Vector Machine*

1. INTRODUCTION

Phishing attacks are a global threat that leverages social engineering to steal sensitive information. This phenomenon has been rampant with the rapid growth of the internet, attacking various sectors such as online banking, social networking, and healthcare[1]. Despite the existence of detection tools, phishing incidents continue to increase every year, posing a serious threat [2]. The success of these attacks is highly dependent on the human factor, reinforcing the need for effective anti-phishing training and decision support systems[3]. With the impact evenly distributed on individuals, organizations, and economies, the importance of tackling phishing attacks is increasingly evident globally.

New approaches in phishing detection, such as anomaly-based methods, have been developed to block suspicious activity on websites[4]. E-commerce also prioritizes phishing threat analysis to protect online transactions[5]. In addition, the utilization of advanced threat modeling methods such as stide and dread in e-health security promises a better response to phishing attacks [6]. This method provides a flexible solution to combat the continuously changing landscape of phishing threats, underscoring the need for a proactive strategy in risk reduction. A prior study titled 'Comparison of Support Vector Machine and Modified Balanced Random Forest in the Detection of Diabetic Patients' [7], indicates that diabetes results from elevated blood sugar levels due to pancreatic and insulin regulation abnormalities. If left unmanaged, this condition can cause serious health complications, highlighting the necessity for early diagnosis and intervention. According to data from the Indonesian Ministry of Health, diabetes ranks as the third leading cause of death in Indonesia, responsible for 6.7% of total fatalities. In light of this high mortality rate, research has concentrated on developing early detection techniques to mitigate the disease's prevalence and consequences. This study employs Machine Learning to classify data from diabetic patients, focusing on a comparative analysis of the Support Vector Machine (SVM) and the Modified Balanced Random Forest (MBRF), as both algorithms have shown substantial accuracy in previous studies. [8]. The preprocessing stage is applied to prepare the data so that it can be classified.

Both methods underwent identical preprocessing procedures to ensure uniformity in the datasets. The assessment utilized a Confusion Matrix to gauge performance. Results from the experiments indicated that the Support Vector Machine (SVM) achieved a peak performance of 87.94%, whereas the Modified Balanced Random Forest (MBRF) exceeded this with a top accuracy of 97.8% [9]. This research is designed to assess and compare the efficacy of three machine learning algorithms—decision tree, random forest, and support vector machine—in identifying phishing attacks. The objective is to provide comprehensive insights into the strengths and weaknesses of each algorithm and to understand how they address specific challenges in phishing detection

2. METHODOLOGY

Decision Trees are suitable for phishing detection because they provide easy interpretation and do not require data normalization [10]. Random Forest improves the stability of the model with ensemble learning, effectively overcoming overfitting on complex data. Meanwhile, SVM is effective in high dimensions and tolerant of overfitting, providing good performance[11]. In this study, variables such as URL length and HTTPS presence are used as features in model training and testing [12]. The selection of the algorithm should consider the characteristics of the dataset and the purpose of the study, where Decision Tree provides interpretability, Random Forest improves stability, and Support Vector Machine performs well in high dimensions, creating a comprehensive approach to phishing detection [13].

2.1 Problem Identification

In this step, problems are identified and analyzed against previous studies, as well as techniques applied to identify phishing websites [14].

2.2 Data Collection

At the data collection stage, this study utilizes secondary data obtained from the kaggle database [15]. In dataset_phishing consisted of 11430 records with 89 attributes (88 attributes and 1 target attribute). The target attribute has ligmate and phishing. These features will be divided into two categories: 80% allocated for the training dataset and 20% reserved for the testing dataset. The distribution of datasets by class is detailed in Table 1.

Table 1. Data Collection

It	Status	Number of Record datasets
1	legitimate	5715
2	phishing	5715

2.3 Pre-Processing Stage

Pre-processing is a step aimed at selecting data by overcoming missing values, resulting in clean and ready to use data [16]. In this study, missing values are used.

2.4 Use of Decision Tree, Random Forest, and Support Vector Machine Algorithms

After completing the pre-processing phase, the algorithms for Decision Tree, Random Forest, and Support Vector Machine (SVM) were executed. These machine learning models were applied to the prepared dataset to evaluate their performance in classification tasks, with a focus on

accuracy and other relevant metrics.In several metrics that can be used to evaluate classification models, including accuracy [17].

2.5 Drawing conclusions

The process of drawing conclusions seeks to encapsulate the findings of the research by applying Decision Tree, Random Forest, and Support Vector Machine techniques. The goal is to determine which algorithm achieves the greatest accuracy in detecting and categorizing phishing websites based on accuracy metrics [18]. In the context of the classification level in data mining for this algorithm, the assessment scale is applied as follows:

- Excellent classification: 90 - 100
- Good classification: 80 – 89
- Enough classification: 70 - 79
- Low classification: 60 - 69

The following is the framework of the methods used in this study, namely:

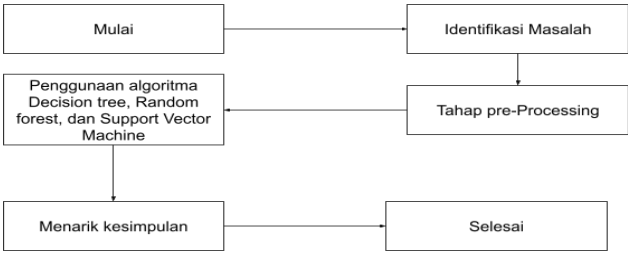


Figure 3.1 Method framework

3. RESULTS AND DISCUSSION

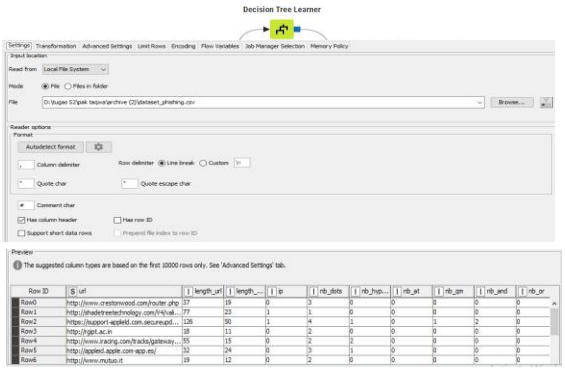
The dataset was obtained from a phishing web survey. Consisting of 89 Columns, the tested attribute is the status attribute.

3.1 Decision Tree

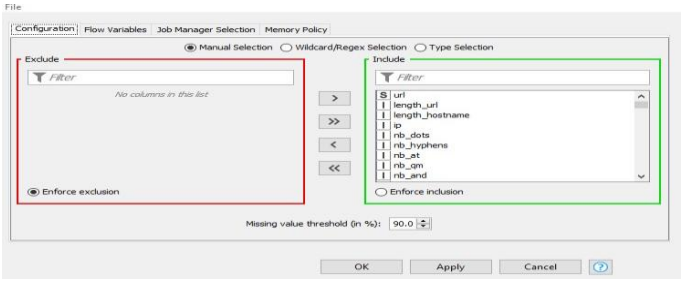
In the image above is the workflow in the training application calculating the decision tree

DOI: https://doi.org/10.63254/acsse.v1i1.12

ACSSE
APPLIED COMPUTER SCIENCE AND SOFTWARE ENGINEERING
E-ISSN : 0000000000

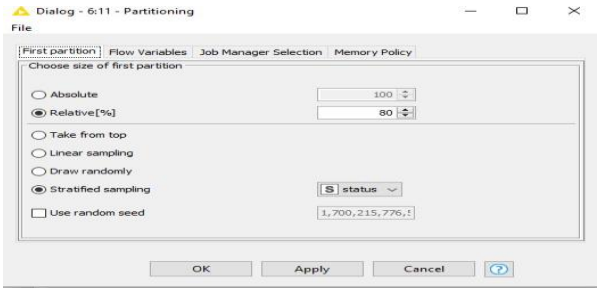


In File Reader, select a dataset to read, the way is by clicking browser



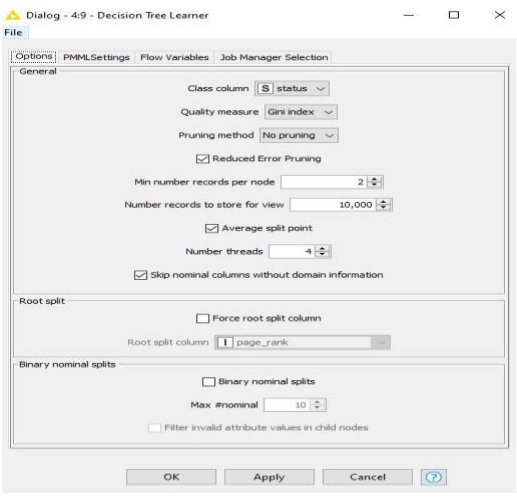
Decision trees are a technique utilized in supervised machine learning, involving the use of labeled input and output data sets for training models. [19]. In the picture is the missing value, here is the explanation:

- Configuration Flow Variables are the names of files, columns, or other data used for system configuration or data projects.
- Missing value threshold (in %): 90.0, the percentage level of the limit that indicates the amount of missing data that can be included in the processing. In this case, the limit percentage rate is 90%.

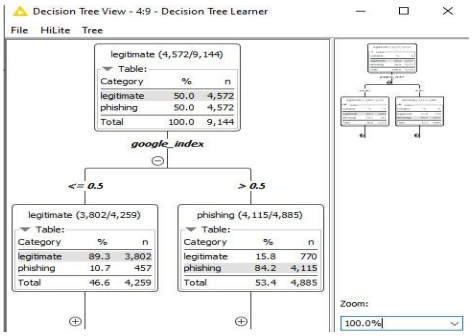


In Relative Partitioning, it is filled with 80 to divide the data into two groups where 80 data will go into one group, and 20 data will go into the other group. For stratified sampling, it is filled with status.

ACSSE
APPLIED COMPUTER SCIENCE AND SOFTWARE ENGINEERING
E-ISSN : 0000000000

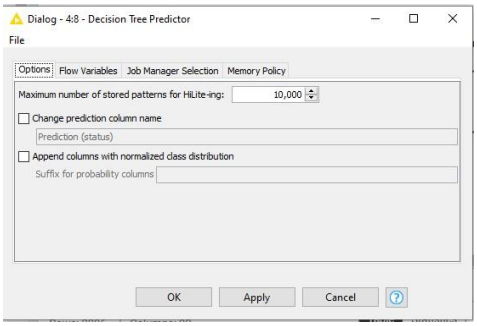


Fill in the Decission Tree Learner as shown in the image, here is the explanation: Decision Tree Learner is one of the nodes or tools used to create a decision tree model [20]. This node enables the analysis and training of decision tree-based models within the KNIME Analytics platform. By using Decision Tree Learner in KNIME, it is possible to solve many problems such as classification and prediction based on the selected features. These nodes make it possible to customize settings such as separation criteria, maximum depth of trees, and many other options according to analysis needs.

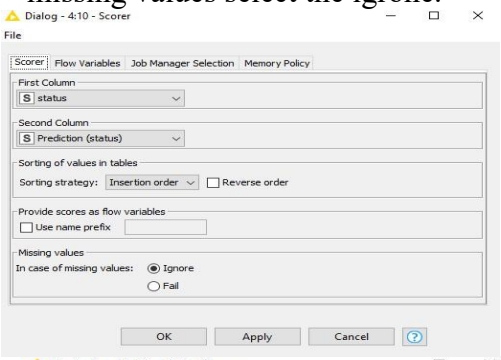


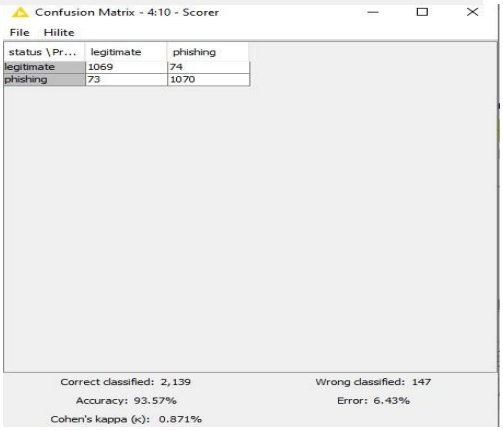
In the figure is a picture of Decision Tree Learner is one of the machine learning algorithms that is often used to calculate the probability of input variables and evaluate how well a model is against training data and testing data. In this case, the Decision Tree Learner algorithm is used to predict whether the received email is phishing or not [21]. In its presentation, the final result is calculated through the amount of training and testing data that is below or above the previously set level of uncertainty (threshold)[22]. Based on the data that has been processed, the Decision Tree Learner algorithm calculates the probabilities for each category at each level of the tree hierarchy. For example, for the legitimate category, the amount of data that is below 0.5 indicates reliable certainty in the prediction. Overall, a tree is needed to predict whether the email is phishing or not. To achieve this, the tree is built by selecting the variable that has the highest probability of recalculating the amount of data on each node.

ACSSE
APPLIED COMPUTER SCIENCE AND SOFTWARE ENGINEERING
E-ISSN : 0000000000



Maximum number of stored patterns for HighLighting -> Sets the number of patterns to be stored for hilite-ing. The value entered in this column will affect the amount of data stored for the hiLite-ing process. Here it is filled with 10,000. In the image Scorer is configured first, the first column is filled with "status" and for the second column is filled with "prediction ", then in case of missing values select the igrone.





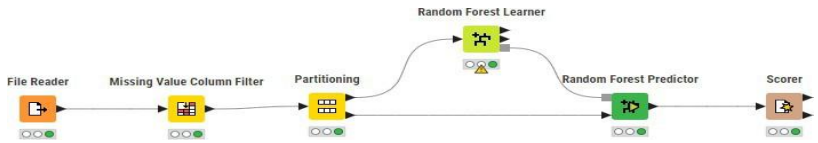
The scorer shows how the score can differentiate between the two classes (in this case, phishing and legitimate) in several different categories. In this matrix, two numbers are displayed for each row and column, the number of examples that have been given to each class and the forecast results that have been generated by the model.

- In the legitimate class and the forecasting results show them as legitimate. This number (1069) indicates the number of instances that the model considers correct.

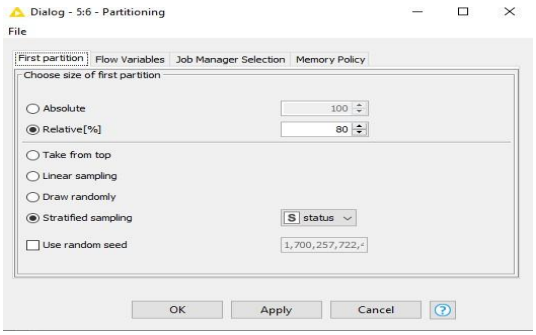
ACSSE
APPLIED COMPUTER SCIENCE AND SOFTWARE ENGINEERING
E-ISSN : 0000000000

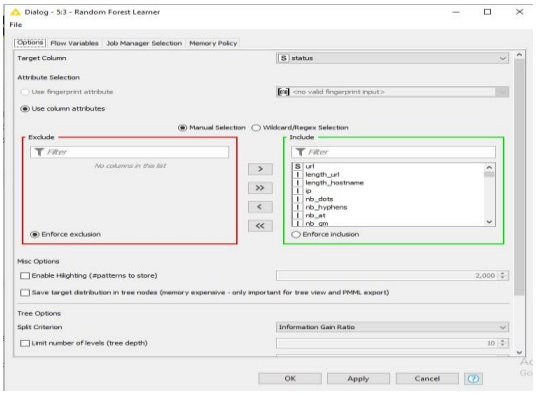
- On legitimate classes and forecasting results that show them as phishing. This number (74) indicates the number of correct instances that the model considers wrong.
- On the phishing class and the forecast results show them as legitimate. This number (73) indicates the number of instances that the model incorrectly considers correct.
- On the phishing class and the forecast results show them as phishing. This number (1070) indicates the number of instances that are considered wrong by the model. Correctly classified: 2,210 Accuracy: 93.57%, Cohen's kappa (K): 0.871%, Wrong classified: 147 Error: 6.43%.

3.2 Random Forest



Random Forest, a popular machine learning technique created by Leo Breiman and Adele Cutler, combines the results from various decision trees to produce a unified prediction [23]. Like the first decision tree is File Reader, then select the missing value column filter then select partitioning. In Relative Partitioning, it is filled with 80 to divide the data into two groups where 80 data will go into one group, and 20 data will go into the other group. For stratified sampling, it is filled with status.

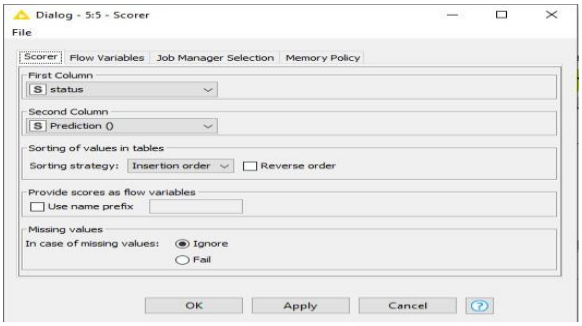




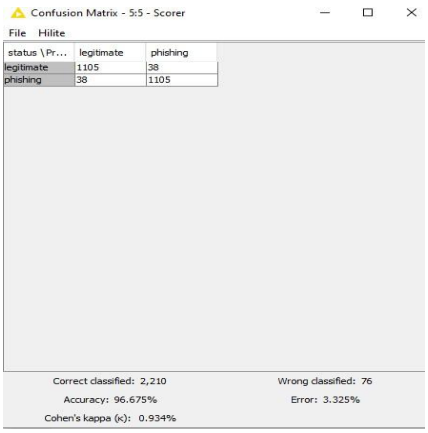
DOI: https://doi.org/10.63254/acsse.v1i1.12

ACSSE
APPLIED COMPUTER SCIENCE AND SOFTWARE ENGINEERING
E-ISSN : 0000000000

To help users understand the forecast results, several options are available in the "Prediction Settings" area. Users can choose to rename prediction columns, add prediction depth (prediction certainty level), or add individual class probabilities. In addition, users can choose to use the soft voting method, which allows for forecasts that occur when classes compete.



In the scorer, the configuration first, namely in the first column is filled with "status" and for the second column is filled with "prediction ()", then in case of missing values select the igrone.



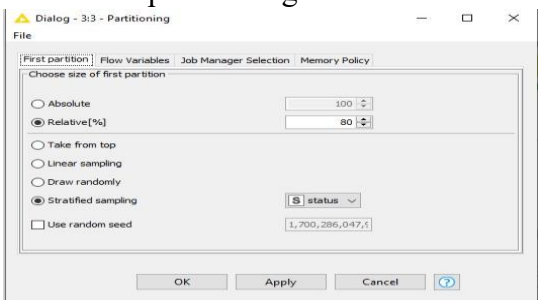
The scorer shows how the score can differentiate between the two classes (in this case, phishing and legitimate) in several different categories. In this matrix, two numbers are displayed for each row and column, denoting the number of examples that have been given to each class and the forecast results that have been generated by the model.

- In the legitimate class and the forecasting results show them as legitimate. This number (1105) indicates the number of instances that the model considers correct.
- On legitimate classes and forecasting results that show them as phishing. This number (38) indicates the number of correct instances that the model considers wrong.
- On the phishing class and the forecast results show them as legitimate. This number (38) indicates the number of instances that the model incorrectly considers correct.
- On the phishing class and the forecast results show them as phishing. This number (1105) indicates the number of instances that are considered wrong by the model. Correctly classified: 2,210 Accuracy: 96.675%, Cohen's kappa (K): 0.934%, Wrong classified: 76 Error: 3.325%.

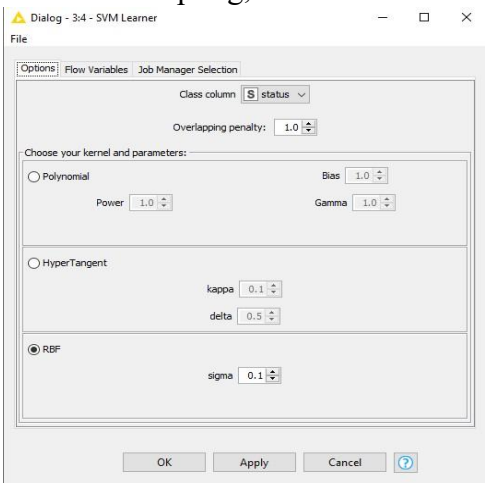
ACSSE
APPLIED COMPUTER SCIENCE AND SOFTWARE ENGINEERING
E-ISSN : 0000000000

3.3 Support Vector Machine

Support Vector Machines (SVMs) were introduced in the early 1990s as classifiers designed to achieve optimal margins, based on Vapnik's theory of statistical learning [24]. Like the decision tree and random forest, the first is File Reader, then select the missing value column filter and then select partitioning. In Relative Partitioning, it is filled with 80 to divide the data into two.



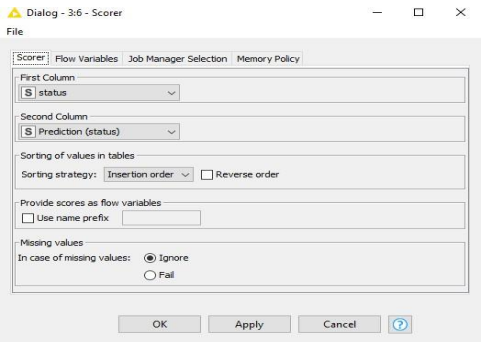
groups where 80 data will go into one group, and 20 data will go into the other group. For stratified sampling, it is filled with status.



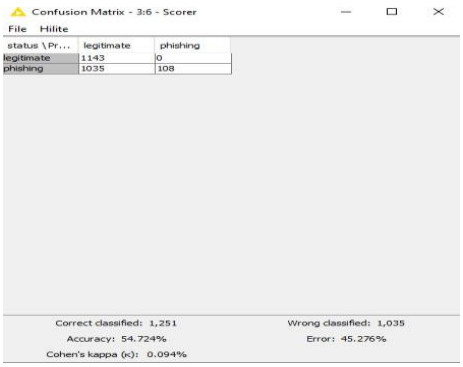
In SVM Learner, use a column class, namely status, then an overlapping penalty, which is 1.0. RBF: This setting uses the Gaussian Radial Basis Function (RBF) kernel which contains sigma 0.1[25].

DOI: https://doi.org/10.63254/acsse.v1i1.12

ACSSE
APPLIED COMPUTER SCIENCE AND SOFTWARE ENGINEERING
E-ISSN : 0000000000



In the scorer, the configuration first, namely in the first column is filled with "status" and for the second column is filled with "prediction ()", then in case of missing values select the igrone.



The scorer shows how the score can differentiate between the two classes (in this case, phishing and legitimate) in several different categories. In this matrix, two numbers are displayed for each row and column, denoting the number of examples that have been given to each class and the forecast results that have been generated by the model.

- In the legitimate class and the forecasting results show them as legitimate. This number (1143) indicates the number of instances that the model considers to be correct.
- On legitimate classes and forecasting results that show them as phishing. This number (0) indicates the number of correct instances that the model considers incorrect.
- On the phishing class and the forecast results show them as legitimate. This number (1035) indicates the number of instances that the model incorrectly considers true.
- On the phishing class and the forecast results show them as phishing. This number (108) indicates the number of instances that are considered incorrect by the model. Correctly classified: 1.251 Accuracy: 54.724%, Cohen's kappa (K): 0.094%, Wrong classified: 1,035 Error: 45,276%.

4. CONCLUSION AND RECOMMENDATION

Of the 3 algorithms that have been implemented, Random Forest, and Support Vector Machine provide the highest accuracy results (96.675%), followed by Decision Tree (93.57%), while Support Vector Machine shows lower performance (54.724%). Cohen's Kappa, which provides a correlation value between the observation results and the prediction results adjusted

for the likelihood of success occurring by chance, also shows that Random Forest has the highest value, followed by the Decision Tree and the Support Vector Machine. In terms of the number of errors, Decision Tree and Random Forest produce better results than Support Vector Machine. In conclusion, for this dataset, Random Forest and Decision Tree demonstrate superior performance compared to Support Vector Machine in terms of both accuracy and the correlation between prediction and observation results.

REFERENCES

[1] B. Pranggono and A. Arabo, "COVID-19 pandemic cybersecurity issues," *Internet Technol. Lett.*, vol. 4, no. 2, pp. 4–9, 2021, doi: 10.1002/itl2.247.

[2] A. N. Shaikh, "A Literature Review on Phishing Crime , Prevention Review and Investigation of Gaps," no. March 2018, 2016, doi: 10.1109/SKIMA.2016.7916190.

[3] G. Desolda, L. S. Ferro, A. Marrella, T. Catarci, and M. F. Costabile, "Human Factors in Phishing Attacks: A Systematic Literature Review," *ACM Comput. Surv.*, vol. 54, no. 8, 2022, doi: 10.1145/3469886.

[4] A. Safi and S. Singh, "A systematic literature review on phishing website detection techniques," *J. King Saud Univ. - Comput. Inf. Sci.*, vol. 35, no. 2, pp. 590–611, 2023, doi: 10.1016/j.jksuci.2023.01.004.

[5] M. T. Banday and J. A. Qadri, "Phishing - A Growing Threat to E-Commerce," no. December 2011, 2011, [Online]. Available: http://arxiv.org/abs/1112.5732

[6] M. K. Faridi, I. Riadi, and Y. Prayudi, "Pemodelan Ancaman Sistem Keamanan EHealth menggunakan Metode STRIDE dan DREAD," *Edumatic J. Pendidik. Inform.*, vol. 5, no. 2, pp. 157–166, 2021, doi: 10.29408/edumatic.v5i2.3652.

[7] J. Moedjahedy, A. Setyanto, F. K. Alarfaj, and M. Alreshoodi, "CCrFS: Combine Correlation Features Selection for Detecting Phishing Websites Using Machine Learning," *Futur. Internet*, vol. 14, no. 8, 2022, doi: 10.3390/fi14080229.

[8] M. D. Purbolaksono, M. Irvan Tantowi, A. Imam Hidayat, and A. Adiwijaya, "Perbandingan Support Vector Machine dan Modified Balanced Random Forest dalam Deteksi Pasien Penyakit Diabetes," *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, vol. 5, no. 2, pp. 393–399, 2021, doi: 10.29207/resti.v5i2.3008.

[9] J. Ye *et al.*, "A Chi-MIC Based Adaptive Multi-Branch Decision Tree," *IEEE Access*, vol. 9, pp. 78962–78972, 2021, doi: 10.1109/ACCESS.2021.3077125.

[10] S. Alnemari and M. Alshammari, "Detecting Phishing Domains Using Machine Learning," *Appl. Sci.*, vol. 13, no. 8, 2023, doi: 10.3390/app13084649.

[11] M. Awad and R. Khanna, "Efficient learning machines: Theories, concepts, and applications for engineers and system designers," *Effic. Learn. Mach. Theor. Concepts, Appl. Eng. Syst. Des.*, no. April 2015, pp. 1–248, 2015, doi: 10.1007/978-1-4302-59909.

[12] H. Le, Q. Pham, D. Sahoo, and S. C. H. Hoi, "URLNet: Learning a URL Representation with Deep Learning for Malicious URL Detection," no. i, 2018, [Online]. Available: http://arxiv.org/abs/1802.03162

[13] I. D. Mienye and N. Jere, "A Survey of Decision Trees: Concepts, Algorithms, and Applications," *IEEE Access*, vol. 12, pp. 86716–86727, 2024, doi: 10.1109/ACCESS.2024.3416838.

[14] A. K. Dutta, "Detecting phishing websites using machine learning technique," *PLoS One*, vol. 16, no. 10 October, pp. 1–17, 2021, doi: 10.1371/journal.pone.0258361.
[15] D. Finfgeld-Connett and E. D. Johnson, "Data Collection and Sampling," *A Guid. to Qual. Meta-Synthesis*, no. January, pp. 18–30, 2018, doi: 10.4324/9781351212793-3.
[16] S. Roy, P. Sharma, K. Nath, D. K. Bhattacharyya, and J. K. Kalita, "Pre-processing: A data preparation step," *Encycl. Bioinforma. Comput. Biol. ABC Bioinforma.*, vol. 1–3, no. January, pp. 463–471, 2018, doi: 10.1016/B978-0-12-809633-8.20457-3.
[17] "Course Title : Foundation to Data Science (70620) Instructor : Dr . Usman Arif Group 11 : Rafay Mahar (24559) Hassan Naeem (24318) Assignment : 2," no. 70620.
[18] A. Rahmah, N. Sepriyanti, M. H. Zikri, I. Ambarani, and M. Y. bin Shahar, "Implementation of Support Vector Machine and Random Forest for Heart Failure Disease Classification," *Public Res. J. Eng. Data Technol. Comput. Sci.*, vol. 1, no. 1, pp. 34–40, 2023, doi: 10.57152/predatecs.v1i1.816.
[19] S. Cycles, "Chapter 9 Chapter 9," *Cycle*, vol. 1897, no. Figure 1, pp. 44–45, 1989, doi: 10.1007/0-387-25465-X.
[20] T. Elomaa, *Tools and techniques for decision tree learning*. 1996. [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.45.9025&rep=rep1&type=pdf
[21] S. H. Abdulrahman and M. Salim, "Using Decision Tree Algorithms in Detecting Spam Emails Written in Malay: A Comparison Study," *ITM Web Conf.*, vol. 42, p. 01001, 2022, doi: 10.1051/itmconf/20224201001.
[22] L. D. Maxim, R. Niebo, and M. J. Utell, "Screening tests : a review with examples," vol. 8378, no. 13, pp. 811–828, 2014, doi: 10.3109/08958378.2014.955932.
[23] A. Cutler, D. R. Cutler, and J. R. Stevens, "Ensemble Machine Learning," *Ensemble Mach. Learn.*, no. February 2014, 2012, doi: 10.1007/978-1-4419-9326-7.
[24] Y. X. Chu, X. G. Liu, and C. H. Gao, "Multiscale models on time series of silicon content in blast furnace hot metal based on Hilbert-Huang transform," *Proc. 2011 Chinese Control Decis. Conf. CCDC 2011*, pp. 842–847, 2011, doi: 10.1109/CCDC.2011.5968300.
[25] Lensa Rosdiana Safitri, Nur Chamidah, Toha Saifudin, and Gaos Tipki Alpandi, "Comparison Of Kernel Support Vector Machine In Stroke Risk Classification (Case Study:IFLS data)," *Proc. Int. Conf. Data Sci. Off. Stat.*, vol. 2023, no. 1, pp. 309–316, 2023, doi: 10.34123/icdsos.v2023i1.381.