Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 227 (2023) 439–445

Procedia Computer Science

www.elsevier.com/locate/procedia

8th International Conference on Computer Science and Computational Intelligence (ICCSCI 2023)

Interpretable Machine Learning Model For Heart Disease Prediction

Putri Sari Asih[a], Yufis Azhar[a], Galih Wasis Wicaksono[a]*, Denar Regata Akbi[a]

[a]*Universitas Muhammadiyah Malang, Jl Raya Tlogomas No. 246, Malang 65144, Indonesia*

Abstract

In the medical industry, accurately predicting a patient's likelihood of heart disease requires a high-performance model and explaining how the model arrived at its conclusion. To address this, a study has proposed a way to interpret machine learning models using SHAP and LIME. Four models have been created: Vector Machine, Random Forest, XGBoost, and k-Nearest Neighbor. The SVM and XGBoost models exhibit the highest f1-score performance, reaching up to 88%. These models can then be utilized during the interpretation stage with the aid of SHAP and LIME. Based on the SHAP visualization results, it is evident that the predictions made include various significant variables. Meanwhile, LIME explains the classification of each data point. Additionally, it confirms that SHAP and LIME are valuable tools for interpreting models.

Keywords: interpretable machine learning; SHAP; LIME

1. Introduction

The principle of equal rights and the pursuit of well-being for all individuals are fundamental to a just and equitable society. Recognizing this, the 1945 Constitution of the Republic of Indonesia, in Article 28I, paragraph 1, affirms that every person has the inherent right to live a life of physical and spiritual prosperity [1]. This encompasses various aspects, including adequate housing, a conducive and healthy environment, and access to essential healthcare services. By guaranteeing these rights, the state assumes the responsibility of ensuring that all Indonesian citizens can enjoy the highest standard of healthcare available.

* Corresponding author. Tel.: +6282142582102.
E-mail address: galih.w.w@umm.ac.id

10.1016/j.procs.2023.10.544

In line with the constitutional mandate, providing quality healthcare services is crucial to safeguarding the population's well-being. The remarkable progress achieved by humanity in the realm of technology has played a pivotal role in advancing healthcare practices and enhancing the lives of individuals. Revolutionary innovations have revolutionized how medical professionals diagnose and treat diseases, significantly improving patient outcomes. A prime example is the advent of Magnetic Resonance Imaging (MRI) machines, which have proven to be highly effective in identifying various medical conditions with exceptional precision. Through the fusion of data processing science and electronic devices, healthcare practitioners now possess a valuable tool to accurately diagnose and address their patients' health concerns.

Technology integration into healthcare has opened up unprecedented possibilities for early detection, precise diagnoses, and tailored treatment plans. By leveraging the power of advanced imaging techniques, such as MRI scans, doctors can identify subtle abnormalities that might otherwise go unnoticed. This early detection allows for timely interventions, potentially saving lives and improving long-term health outcomes. Moreover, the seamless integration of data processing science and electronic devices has resulted in a more streamlined and efficient healthcare system. Patient information can be securely stored and accessed electronically, facilitating comprehensive and coordinated care across healthcare settings.

Machine learning models are being developed to aid in diagnosing diseases based on specific characteristics and criteria [2]. One noteworthy model is used to detect heart disease based on specific criteria and utilizes various methods from different studies to achieve optimal results. These algorithms include Random Forest [3] and hybrid machine learning. These models will assist health workers in treating patients, but from a patient's perspective, they will also expedite the diagnosis process [4].

The models developed to predict heart disease perform fairly well if tested based on classification metrics such as accuracy and precision [5]. One study with the same heart disease dataset managed to achieve an accuracy of 81% precision using the Random Forest algorithm [6]. Other research with similar datasets and methods coupled with the Chi-Square feature selection process achieved an accuracy of 83.7%[7].

This phenomenon is good, as much research is done to get the best machine-learning model with high accuracy. With so many studies being conducted, many models have been produced, but the interpretability of existing models still needs to be studied more [8]. Many machine learning models, in general, are made to predict as many targets as possible correctly. Usually, this is called accuracy. This accuracy is then used as a metric or tool to measure machine learning models' performance [9].

One issue with machine learning models is that not all aspects can be easily interpreted. The main focus during the model creation process has been maximizing prediction accuracy. However, in the medical field, interpretation is essential for understanding and utilizing the models effectively [10]. To properly diagnose a disease in a patient, it is essential to interpret the model accurately. This requires careful attention to detail.

Efforts should be made to explain the predictive model used for diseases. In this study, we will explore the process of creating a heart disease prediction model that is both highly accurate and easily interpretable. This study will utilize SHAP [11] and LIME [12] to determine the factors contributing to a positive patient diagnosis.

2. Research Methodology

The proposed methods in this study include SHAP and LIME, tools used for interpreting machine learning models. Here is a brief explanation of these two methods.

2.1. Shapley Additive Explanation (SHAP)

A model needs to explain its predictive outcomes regarding interpretable machine learning. For instance, in the context of heart disease prediction, the aim is to enable patients to determine whether or not they are likely to test positive. To achieve this objective, it is helpful to identify the features or variables that have the most significant impact on the predictions [13].

Explainable machine learning models aim to provide insights and justifications for the predictions or decisions made by the model. The SHAP (Shapley Additive Explanations) method is one such approach that offers interpretability in machine learning models. The SHAP method is based on cooperative game theory and borrows the concept of Shapley values, originally developed to distribute payouts among participants in a cooperative game fairly. In machine learning, SHAP values quantify the contribution of each feature in a prediction or decision made by the model. The core idea of SHAP is to assign a value to each feature that represents its importance in the prediction process. These feature importance values are calculated by considering all possible combinations of features and evaluating their impact on the prediction. The SHAP values provide a unified measure of feature importance that considers interactions between features.

The SHAP method considers a reference point, typically a baseline or average prediction, to compute SHAP values. It systematically evaluates the impact of adding or removing each feature from this reference point. By considering all possible subsets of features, the SHAP method calculates the contribution of each feature to the prediction outcome. SHAP can be formulated by equation (1). Where S is a subset of a feature from the existing model, then x is a vector of feature values from the data to be interpreted.

$$\emptyset_j(p) = \sum\nolimits_{S\subseteq\{x_1,\ldots,x_p\}\backslash x_j} \frac{|\mathrm{S}|!\,(p-|S|-1)!}{p!}\left(val\left(S\ \cup\{x_j\}\right) - val(S)\right) \qquad (1)$$

One of the key advantages of the SHAP method is its ability to provide local and global explanations. Locally, SHAP values explain the contribution of each feature for a specific prediction or decision made by the model. Globally, SHAP values provide an overview of feature importance across the entire dataset.

2.2. Local Interpretable Model-Agnostic Explanation (LIME)

The LIME (Local Interpretable Model-agnostic Explanations) method is another approach in explainable machine learning that focuses on providing local explanations for individual predictions made by any black-box machine learning model. The main goal of LIME is to explain the predictions of a complex model by approximating it with a simpler, interpretable model in the local vicinity of the instance being explained. This local approximation allows for better interpretability and understanding of the factors influencing the model's decision.

The LIME method generates perturbations or variations of the instance to be explained. These perturbations are created by randomly modifying or masking some features while keeping the rest of the instances unchanged. The modified instances are then passed through the black-box model, and the predictions are recorded. Next, a simpler, interpretable model is trained to approximate the behavior of the black-box model based on the perturbed instances and their corresponding predictions. This interpretable model can be a linear model, decision tree, or any other interpretable model that can capture the relationships between the features and predictions.

Similar to SHAP, LIME is utilized to uncover the reasoning behind a machine learning model's confident predictions. LIME achieves this by analyzing established models against the given data [13]. The formula for LIME can be expressed as equation (2). In equation (2), the explanatory model aims to minimize loss values L which represents the closeness of the explanation to the predictions of the original model f, while keeping the model complexity Ω (g) at a low value. Therefore, G proves a list of possible explanations of the model generation results and π_π defines how big the value of the data point is around x.

$$explanation\ (x) = arg \min_{\mathrm{g}\,\epsilon\,\mathrm{G}} L(f, g, \pi_\pi) +\ \Omega\ (\mathrm{g}) \qquad (2)$$

The explanations provided by LIME are based on the weights or importance assigned to each feature by the interpretable model. These weights indicate the contribution of each feature towards the prediction and help to understand the decision-making process of the black-box model locally. LIME addresses the local interpretability challenge by focusing on a small neighborhood around the instance of interest. By considering a local context, LIME provides explanations that are more relevant and understandable to users.

One of the advantages of the LIME method is its model-agnostic nature. It can be applied to any black-box model without requiring knowledge of the model's internal workings. This flexibility makes LIME a widely applicable technique for explaining complex models across various domains.

2.3. Dataset

For this research, we utilized the Cleveland Heart Disease Dataset, which initially contained 76 features. However, only 14 of these features were ultimately released for publication. These features include Age, Sex, CP, Trestbps, Chol, Fbs, Restecg, Thalach, Exang, Oldpeak, Slope, Ca, Thal, and Diagnosis. In this journal, various processes are undertaken before interpreting the model. These include data pre-processing, creating models using different techniques, evaluating the models, and finally, the evaluation stage.

2.4. Data Pre-processing

In this study, two pre-processing stages are carried out: the normalization process [14]. To achieve this, we utilize StandardScaler for assistance. Afterward, we proceed with splitting the data into two categories: training data and test data. For this particular dataset, we allocate 70% of the data for training and 30% for testing.

2.5. Modeling and Evaluation

The following process is to create several machine-learning models[15]. The purpose of this process is to determine the most suitable model for the next stage. Various models have been developed, including Support Vector Machine (SVM), Random Forest, XGBoost, and k-Nearest Neighbor (KNN). The evaluation is conducted based on the f1-score.

SHAP serves as a tool for interpreting global models. Python libraries can be utilized for their implementation. SHAP offers a variety of explainers tailored to different model types. The following are the explainers utilized for various model types.

Table 1. The kernel on each model.

Model	Jenis Explainer
SVM	Kernel Explainer
Random Forest	Tree Explainer
XGBoost	Kernel Explainer
k-NN	Kernel Explainer

To interpret each prediction for every data point, LIME is utilized. Python programming language has a library that can be utilized to implement LIME. Unlike SHAP, Lime Tabular Explainer is the type of explainer every model uses.

3. Result & Discussion

We use the f1-score value to measure model performance, which is an average of the precision and recall values. We have evaluated four models based on this metric, and the results show that the SVM and XGBoost models perform equally well, with Random Forest coming in second place and k-NN in third place. Using SHAP and LIME, we will interpret how these two models work.

Table 2. F1-Score on each model.

Model	Positive (1)	Negative (0)
SVM	88%	85%
Random Forest	87%	82%
XGBoost	88%	85%
k-NN	86%	81%

3.1. SHAP Visualization

The figures labeled as Figure 1 and Figure 2 display the visualization outcomes through SHAP on two models, namely SVM and XGBoost, which include model interpretation. The SHAP value is shown on the x-axis, while the names of the features in the dataset are indicated on the y-axis. The colors red and blue symbolize the positive and negative classes, respectively.

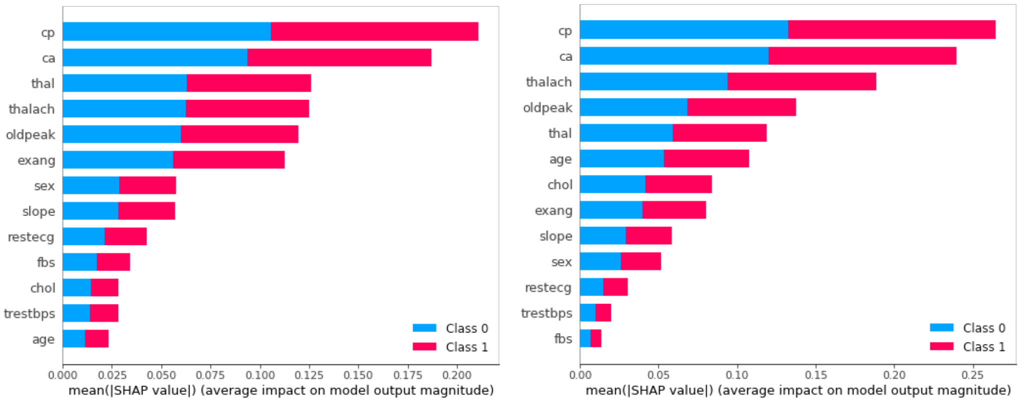


Figure 1. SHAP Visualization on SVM (left) and XGBoost (right).

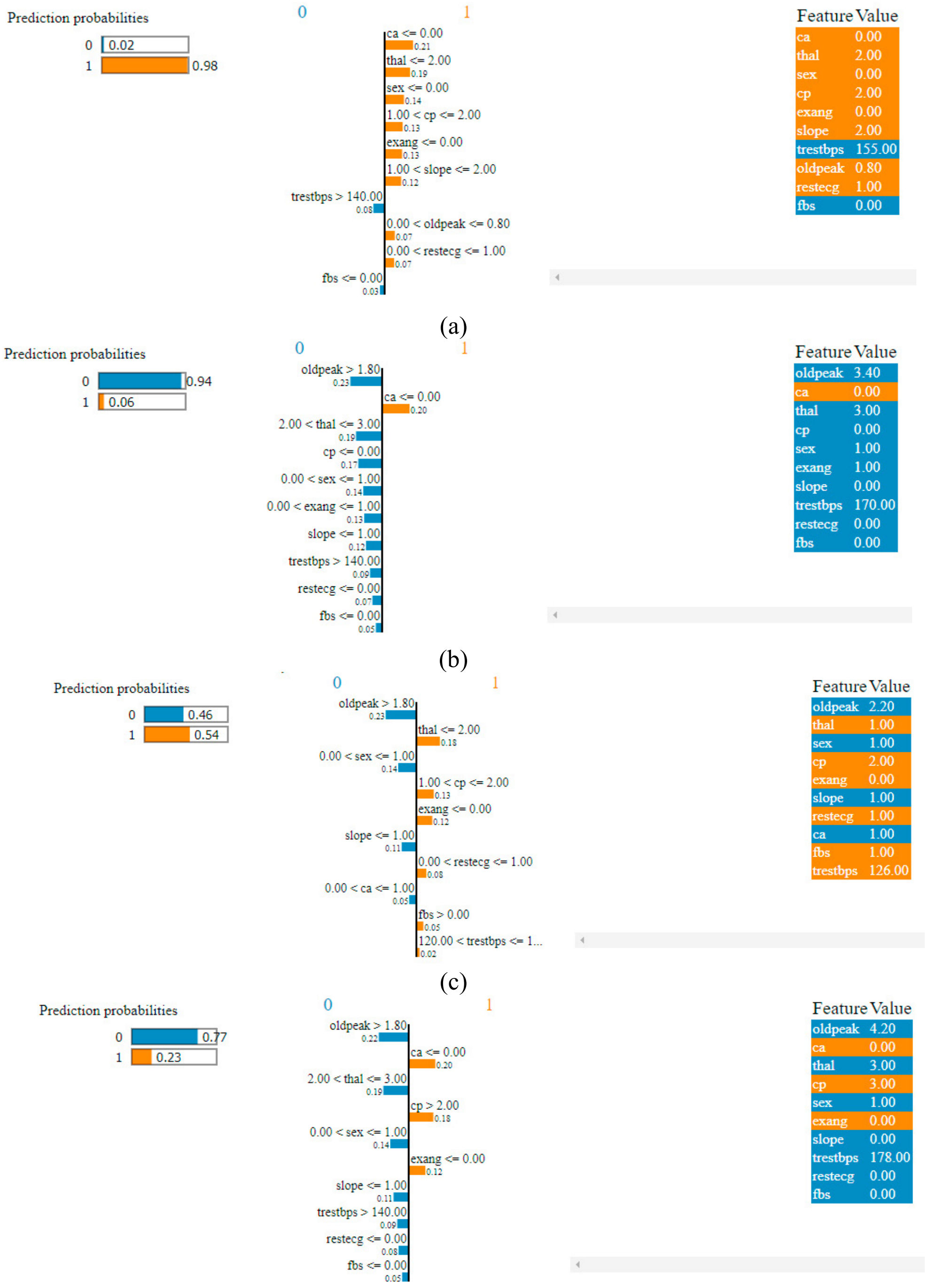
When observing the SVM model visualization, it becomes apparent that variables such as cp, ca, thal, thalach, and oldpeak possess the highest SHAP values, exceeding 0.10 (thal). Interestingly, these five variables exhibit a proportional relationship to the two existing classes, as demonstrated by the size of the colored bars. In contrast, the variables fbs, chol, tresbps, and age possess SHAP values below 0.05. Comparatively, the XGBoost model portrays cp, ca, thalach, oldpeak, and thal as the variables with the highest SHAP values, with thal exhibiting at least 0.10. Conversely, the variables restecg, tresbps, and fbs possess the lowest SHAP values in the XGBoost model, with the highest value being less than 0.05.

By analyzing two different models, it is evident that the variable with the highest SHAP value for predicting heart disease remains consistent. These variables include cp, ca, thalach, oldpeak, and thal. Interestingly, both models rank cp and ca as the top two variables. However, in the SVM model, thal is third, while in XGBoost, it is fifth. Similarly, thalach and oldpeak rank fourth and fifth in SVM but third and fourth in XGBoost. These variables play a crucial role in predicting heart disease with varying models.

3.2. 3.2. LIME Visualization

The following report displays the outcomes of implementing LIME on both models' positive and negative prediction samples. LIME's visualization consists of three main components: the left side of the plot shows the probability predictions of each prediction, the second part lists the significant features that contribute to the predictions, and the third part contains the actual values of the predicted data points. Figure 2. (a) and (b) depict the LIME visualization's results, which reveal the differences in positive and negative predictions based on the features that affect the predictions. The negative predictions have fewer features contributing to the negative class than the positive class.

In comparing positive and negative predictions, it is evident that there are numerous influential features in the positive class. While the True Negative prediction only features one ca, that pulls data into the positive class; the other variables enter the negative class. Conversely, the True Positive prediction sees most variables (excluding tresbps and fbs) pulling data into the positive class.



Feature	Value
ca	0.00
thal	2.00
sex	0.00
cp	2.00
exang	0.00
slope	2.00
trestbps	155.00
oldpeak	0.80
restecg	1.00
fbs	0.00

(a)

Feature	Value
oldpeak	3.40
ca	0.00
thal	3.00
cp	0.00
sex	1.00
exang	1.00
slope	0.00
trestbps	170.00
restecg	0.00
fbs	0.00

(b)

Feature	Value
oldpeak	2.20
thal	1.00
sex	1.00
cp	2.00
exang	0.00
slope	1.00
restecg	1.00
ca	1.00
fbs	1.00
trestbps	126.00

(c)

Feature	Value
oldpeak	4.20
ca	0.00
thal	3.00
cp	3.00
sex	1.00
exang	0.00
slope	0.00
trestbps	178.00
restecg	0.00
fbs	0.00

(d)

Figure 2. LIME Visualization on SVM for True Positive (a), True Negative(b), False Positive (c), and False Negative (d)

Figure 2 displays the outcomes of applying LIME to the SVM model, particularly in the case of incorrect predictions (False Positive and False Negative). In the False Positive prediction, the data is drawn towards the positive class due to high values of thal and cp, along with other factors such as exang, restecg, fbs, and tresbps. Conversely, errors arise in False Negative prediction because multiple variables direct the data toward the negative class. The positive class only contains ca, cp, and exang values.

4. Conclusion

In conclusion, our study focused on developing and evaluating multiple models for heart disease prediction, including SVM, Random Forest, XGBoost, and k-NN. Through rigorous assessment, SVM and XGBoost emerged as the top-performing models regarding predictive accuracy. We employed SHAP and LIME techniques to enhance interpretability to unravel the underlying factors driving the models' predictions. Notably, SHAP analysis uncovered a consistent set of top 5 features, namely ca, cp, thal, thalach, and oldpeak, which were pivotal in the prediction process. The visualization capabilities of LIME further facilitated the understanding of each model's interpretability, revealing discrepancies in the variables' impact on positive and negative class data in True Positive and True Negative results. This study contributes valuable insights into the effectiveness of diverse models for heart disease prediction and highlights the applicability of SHAP and LIME methods for comprehending and interpreting these models. Future research can explore the potential of incorporating these techniques into clinical practice to aid in informed decision-making and improve patient outcomes.

References

[1] MKN, "Undang-Undang Dasar Negara Republik Indonesia Tahun 1945," vol. 105, no. 3, pp. 129–133, 1945, [Online]. Available: https://webcache.googleusercontent.com/search?q=cache:BDsuQOHoCi4J:https://media.neliti.com/media/publications/9138-ID-perlindungan-hukum-terhadap-anak-dari-konten-berbahaya-dalam-media-cetak-dan-ele.pdf+&cd=3&hl=id&ct=clnk&gl=id.

[2] M. M. Ahsan, S. A. Luna, and Z. Siddique, "Machine-Learning-Based Disease Diagnosis : A," *Healthcare*, pp. 1–30, 2022.

[3] N. S. Chandra Reddy, S. Shue Nee, L. Zhi Min, and C. Xin Ying, "Classification and Feature Selection Approaches by Machine Learning Techniques: Heart Disease Prediction," *Int. J. Innov. Comput.*, vol. 9, no. 1, pp. 39–46, 2019, doi: 10.11113/ijic.v9n1.210.

[4] R. H. Scheuermann, W. Ceusters, and B. Smith, "Toward an Ontological Treatment of Disease and Diagnosis Department of Pathology and Division of Biomedical Informatics , University of Texas," *AMIA Summit Transl. Bioinforma.*, pp. 116–120, 2009.

[5] X. Y. Gao, A. Amin Ali, H. Shaban Hassan, and E. M. Anwar, "Improving the Accuracy for Analyzing Heart Diseases Prediction Based on the Ensemble Method," *Complexity*, vol. 2021, 2021, doi: 10.1155/2021/6663455.

[6] A. Rairikar, V. Kulkarni, V. Sabale, H. Kale, and A. Lamgunde, "Heart disease prediction using data mining techniques," *Proc. 2017 Int. Conf. Intell. Comput. Control. I2C2 2017*, vol. 2018-Janua, no. October, pp. 1–8, 2018, doi: 10.1109/I2C2.2017.8321771.

[7] M. A. Jabbar, B. L. Deekshatulu, and P. Chandra, "Intelligent heart disease prediction system using random forest and evolutionary approach," *J. Netw. Innov. Comput.*, vol. 4, no. 1, pp. 175–184, 2016.

[8] A. Vellido, "The importance of interpretability and visualization in machine learning for applications in medicine and health care data overabundance in some of their areas has become the," *Neural Comput. Appl.*, vol. 0123456789, 2019, doi: 10.1007/s00521-019-04051-w.

[9] M. B. Lyons, D. A. Keith, S. R. Phinn, T. J. Mason, and J. Elith, "A comparison of resampling methods for remote sensing classification and accuracy assessment," *Remote Sens. Environ.*, vol. 208, no. February, pp. 145–153, 2018, doi: 10.1016/j.rse.2018.02.026.

[10] C. H. Yoon, R. Torrance, and N. Scheinerman, "Machine learning in medicine : should the pursuit of enhanced interpretability be abandoned ?," pp. 581–585, 2022, doi: 10.1136/medethics-2020-107102.

[11] M. V. García and J. L. Aznarte, "Ecological Informatics Shapley additive explanations for NO 2 forecasting," *Ecol. Inform.*, vol. 56, no. 2, p. 101039, 2020, doi: 10.1016/j.ecoinf.2019.101039.

[12] N. B. Kumarakulasinghe, J. Liu, and A. S. Leao, "Evaluating Local Interpretable Model-Agnostic Explanations on Clinical Machine Learning Classification Models," pp. 7–12, 2020, doi: 10.1109/CBMS49503.2020.00009.

[13] M. Irfan, S. Basuki, and Y. Azhar, "Giving more insight for automatic risk prediction during pregnancy with interpretable machine learning," *Bull. Electr. Eng. Informatics*, vol. 10, no. 3, pp. 1621–1633, 2021, doi: 10.11591/eei.v10i3.2344.

[14] D. Borkin, A. Némethová, G. Michaľčonok, and K. Maiorov, "Impact of Data Normalization on Classification Model Accuracy," *Res. Pap. Fac. Mater. Sci. Technol. Slovak Univ. Technol.*, vol. 27, no. 45, pp. 79–84, 2019, doi: 10.2478/rput-2019-0029.

[15] H. Alves and B. Fonseca, "Experimenting Machine Learning Techniques to Predict Vulnerabilities," 2016, doi: 10.1109/LADC.2016.32.