Proceedings of Engineering and Technology Innovation, vol. 21, 2022, pp. 20-26

Evaluation of the Shapley Additive Explanation Technique for Ensemble Learning Methods

Tsehay Admassu Assegie*

Department of Computer Science, Injibara University, Injibara, Ethiopia

Received 05 December 2021; received in revised form 02 January 2022; accepted 03 January 2022

DOI: https://doi.org/10.46604/peti.2022.9025

Abstract

This study aims to explore the effectiveness of the Shapley additive explanation (SHAP) technique in developing a transparent, interpretable, and explainable ensemble method for heart disease diagnosis using random forest algorithms. Firstly, the features with high impact on the heart disease prediction are selected by SHAP using 1025 heart disease datasets, obtained from a publicly available Kaggle data repository. After that, the features which have the greatest influence on the heart disease prediction are used to develop an interpretable ensemble learning model to automate the heart disease diagnosis by employing the SHAP technique. Finally, the performance of the developed model is evaluated. The SHAP values are used to obtain better performance of heart disease diagnosis. The experimental result shows that 100% prediction accuracy is achieved with the developed model. In addition, the experiment shows that age, chest pain, and maximum heart rate have positive impact on the prediction outcome.

Keywords: model explanation, random forest, ensemble method, interpretability

1. Introduction

Feature visualization and model explanation techniques play an important role to develop interpretable heart disease diagnostic models and improve the diagnostic accuracy of ensemble learning methods. To implement automated ensemble methods for heart disease diagnosis, model explanation techniques are required [1]. Recent research focuses on analyzing the prediction outcome of ensemble learning methods and the interpretation of automated models for creating trust, reliability, and adoption of the ensemble learning methods. The explanation of ensemble learning methods is significant in getting insights into the prediction process of such methods and improving the accuracy of heart disease diagnosis [2]. In addition to improving the prediction accuracy, improving the interpretability of the predicted outcome is important for the adaptability and reliability of such automated methods because the explainability of the automated models can increase the trust and reliability of the prediction outcome. Random forest is one of the most widely implemented methods for heart disease prediction due to high precision [3].

Designing and implementing interpretable ensemble learning methods for heart disease diagnosis has become one of the recent research topics in machine learning [4]. The reason is that, in recent years, the size of heart disease datasets is growing and the complexity of heart disease diagnosis increases the need for ensemble learning methods to obtain better prediction outcome. However, the application of ensemble learning methods to heart disease diagnosis requires additional research effort to develop the automated models that are automatically from the dataset features. Moreover, the prediction results of heart disease diagnosis provided by ensemble methods are more promising than those provided by linear methods such as support

* Corresponding author. E-mail address: tsehayadmassu2006@gmail.com

Tel.: +251-9-21114923

vector machines [5]. Thus, to develop reliable and adaptable ensemble learning methods for heart disease diagnosis, model interpretation is highly required. The reasoning for the prediction outcome made by the ensemble learning methods is important for domain experts to make decisions on whether the outcome is accepted.

This study aims to improve the accuracy of an existing ensemble learning method for heart disease prediction by employing the feature visualization technique and the Shapley additive explanation (SHAP) values with the real-world heart disease datasets obtained from Kaggle data repository. Experiments are conducted to analyze the features with significant influence on the prediction outcome by the SHAP values. Overall, this study aims at achieving the following specific objectives. 1) The study provides a comprehensive and thorough analysis of the literature on the SHAP model explainability technique. 2) The study explores heart disease features that have significant influence on the prediction outcome of an ensemble method using random forest. 3) The study designs and implements a trusted and reasonable ensemble-based heart disease diagnosis method. 4) The study evaluates the performance of the developed ensemble method for heart disease diagnosis and analyzes the effectiveness of the SHAP technique in explaining the ensemble method.

2. Literature Review

Recent research has shown significant improvement in the performance of heart disease diagnosis using ensemble learning methods such as random forest on complex training samples [6]. The evaluation and interpretation of ensemble learning models for heart disease diagnosis have become an active field of research [7]. Model interpretation is gaining much attention and research effort due to the fact that data scientists have noticed the importance of explaining why machine learning models have predicted an instance as a positive or negative class in the medical domain. Interpretability is aimed to increase the trustworthiness and reliability of the predictive outcome made by machine learning models on risky decisions, particularly in the medical domain such as the prediction of diabetes disease [8]. Interpretability has become a major concern in developing robust and trusted machine learning models for heart disease diagnosis. Numerous previous works have investigated the predictive models of heart disease using machine-learning ensemble methods and deep learning algorithms. Literature also shows that high accuracy is attained with the ensemble and deep learning methods [9]. However, the models are not widely adopted due to the lack of interpretability and transparency.

Model transparency, auditability, and explainability of ensemble learner models are critical to provide a maximally accurate heart disease prediction [10]. Additionally, domain experts, regulators, or policy makers demand medical diagnostic models to be transparent and interpretable. Thus, in medical diagnosis, the very simple predictive models such as logistic regression or decision trees are still widely employed for medical diagnosis. However, the superior predictive power of modern machine learning algorithms in solving complex problems with the growing size of datasets is not leveraged with simple models. Thus, much research effort is needed to develop understandable medical diagnostic models and frameworks for making the “black box” machine learning models transparent, auditable, and explainable. The interpretability of models is gaining much research attention due to three reasons. First, predictive models are widely employed in different areas. Second, simple models are not effective to provide good performance on complex problems. Lastly, there is growing failure of adoption of complex models due to the interpretability and transparency problems.

Interpretable and explainable machine learning is regarded as a process, encompassing three high-level stages [11]. First, model interpretability is critical to understand ensemble learning models and datasets. Second, model interpretability is important to improve the performance of models using explainable artificial intelligence (AI) methods. Third, model interpretability and explanation are applied to refine and optimize machine-learning models. For instance, this study presents an application where interpretability is important. For heart disease prediction, the goal is to identify the patients to be classified into heart disease positive or heart disease negative classes. Specifically, the problem is framed as a binary classification problem where the goal is to predict whether a patient is suffering from heart disease or not by considering the

related factors: heart rate, maximum heart rate achieved (thalach), presence or absence of chest pain (cp), fasting blood sugar level (fbs), cholesterol level (chol), presence or absence of the exercise-induced angina, total resting blood pressure (trestbps), sex, age, slope, resting electrocardiogram (restecg), old peak, and other patient conditions.

The contribution of this study is the use of the SHAP values to explain and interpret the prediction outcome of random forest with the SHAP technique. The literature shows that only few works are published regarding the explainability of ensemble learning methods on heart disease diagnosis [12]. However, the existing models do not explain the prediction outcome. To the best of the author's knowledge, there are no prior published works on the use of interpretable random forest-based heart disease prediction models.

This study explores the application of SHAP in the machine-learning context for medical diagnosis using heart disease datasets as a case study. Overall, the study is designed to answer the following research questions: (1) How to increase the reliability of ensemble learning methods such as a random forest model for heart disease diagnosis? (2) How does the random forest model make predictions with heart disease datasets? (3) How to improve the prediction outcome of an ensemble learning method for heart disease diagnosis? (4) What is the influence of heart disease features on the prediction outcome of the random forest model? Overall, the design and implementation process of the proposed automated heart disease diagnosis system is demonstrated in Fig. 1.

The purpose of this study is to develop a more interpretable and accurate automated heart disease diagnosis model using the random forest model and the SHAP technique. The reliability of the proposed model will be illustrated with the model prototype, and verified with real-world heart disease datasets.

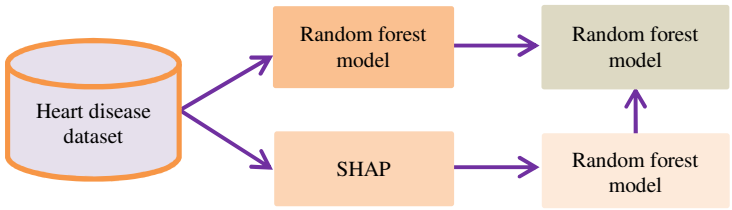

Fig. 1 The process of model explanation with SHAP

3. Design Methodology and Implementation

The complexity of ensemble learning methods and the growing size of real-world heart disease datasets limit the ability to understand what a model has learned or why a given prediction is made, acting as a barrier to the adoption of machine learning [13]. Additionally, from a legal or regulatory perspective, it is required to explain the prediction outcome of a machine-learning model. Machine learning explainability (MLX) is a process of explaining and interpreting an ensemble learning model. To conduct the study, the author collects 1025 real-world heart disease datasets from University of California Irvine (UCI), obtained from a publicly available Kaggle data repository, to the scientific community for experimentation. In this study, SHAP is employed to generate different explainers: objects used to compute the SHAP values for each sample in the datasets. These values represent the impact of certain features on the prediction outcome. Positive values contribute to increasing the final probability, and negative values contribute to decreasing the probability.

The experiment is performed on the Jupyter Notebook in Python 3.7. The datasets are in comma-separated value (CSV) file format. To be used for analysis, the data is imported to the Jupyter Notebook Python environment using Panda's data analysis library in the form of the data frame. The libraries used for data transformation include Pandas, Matplotlib, Scientific learning kit, and SHAP. To develop a model for heart disease diagnosis, the author employs random forest algorithms. The datasets are divided into a training set and a testing set. The training set includes 70% of the datasets or 717 observations, and the testing set includes 30% of the datasets or 308 observations.

4. Existing Design

Fig. 2 shows an existing ensemble learning-based automated heart disease diagnosis system. As can be seen, a model is developed using a feature selection method to improve the prediction outcome. While the application of feature selection improves the prediction accuracy of the system, the reliability and interpretability of the existing model limit the adaptability of such a model in the real-world scenario. Thus, to address the existing gap (i.e., the lack of reasoning) for the prediction outcome of the ensemble learning model on heart disease diagnosis, the study proposes a new model with the help of the SHAP technique to increase the trust of domain experts on using automated models in decision-making. Overall, this study aims to introduce the SHAP technique for model explanation as demonstrated in Fig. 1.

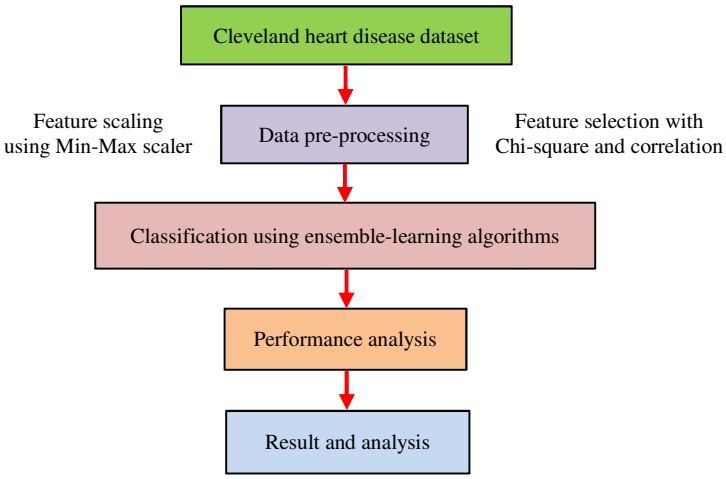

Fig. 2 The existing design

5. Experimental Results and Discussion

This section presents the application of SHAP values for explaining and interpreting the prediction outcome of a random forest model on heart disease diagnosis. The experiment is conducted with SHAP for heart disease prediction with random forest algorithms. To evaluate the proposed random forest model, the author employs SHAP for improving the model performance and interpretability. The experiment is designed to explain the prediction output of the random forest model (i.e., the prediction of heart disease positive or heart disease negative). The author employs the heart disease prediction outcome as a case study to evaluate the interpretability of a random forest model using SHAP. The empirical analysis is conducted to evaluate SHAP-scores using heart disease datasets.

All experiments are implemented in Python 3.7 on Jupyter Notebook environment. The experiment is performed on Intel(R) Core(TM) i7-8565U CPU, 1.80 GHz, 1992 MHz, 4 Core(s), 8 Logical Processor(s) using Microsoft Windows 10 Pro 64-bit version. For data management and optimization, Panda's library is used. The SHAP values of each feature in heart disease datasets are demonstrated in Fig. 3.

As observed from Fig. 3, chest pain (cp), the number of major vessels colored by fluoroscopy (ca), and thalassemia (thal) are the factors that contribute to the model's prediction. Thus, the number of major vessels colored by fluoroscopy (ca), chest pain (cp), and thalassemia (thal) are the features that have the greatest influence on the prediction outcome of the ensemble model. As demonstrated from the experimental result in Fig. 3, a patient with chest pain is most likely predicted as a heart disease patient by the developed random forest model. The number of major vessels colored by fluoroscopy (ca) has the second most impact on the model output. As observed from the plot given in Fig. 3, the features with lower SHAP values include resting electrocardiogram (restecg), fasting blood sugar level (fbs), and slope.

The features with lower SHAP values have less impact on the model output as compared to the features with higher SHAP values. The features having high SHAP values impact the model output on the positive class (i.e., heart disease patients). In contrast, the features with lower SHAP values have positive impact on the predictive outcome of heart disease negative (i.e., patients without heart disease). The feature "resting electrocardiogram (restecg)" shows that the lowest SHAP value (approximately 0.4 SHAP value as illustrated in Fig. 3) has positive impact on the predictive outcome of the negative class (class o). Another important plot for analyzing the impact of features on the model output is the SHAP waterfall plot, demonstrated in Fig. 4.

Fig. 4 demonstrates the impact of heart disease features on the random forest model output on heart disease prediction. As observed from Fig. 4, the number of major vessels colored through fluoroscopy (ca) has the highest impact for the class of heart disease positive, but is the second most impactful feature for the positive prediction outcome using the SHAP summary plot demonstrated in Fig. 3. Ranked below the number of major vessels colored by fluoroscopy (ca), chest pain (cp), and thalassemia (thal) are the features which influence the model's prediction towards the class of heart disease positive. In contrast, slope, age, resting electrocardiogram (restecg), exercise-induced angina, and total resting blood pressure (trestbps) influence the model's prediction towards the class of heart disease negative. Features such as maximum heart rate (thalach), fasting blood sugar level (fbs), and cholesterol level (chol) have lower effects on the model's prediction as compared to other heart disease features as demonstrated in Fig. 4.

Fig. 5 demonstrates the interaction plot between maximum heart rate (thalach) and cholesterol level (chol). As illustrated in Fig. 5, maximum heart rate (thalach) and cholesterol level (chol) have a linear relationship with each other, showing that they are highly correlated. The strong correlation between them indicates their dependency on each other.

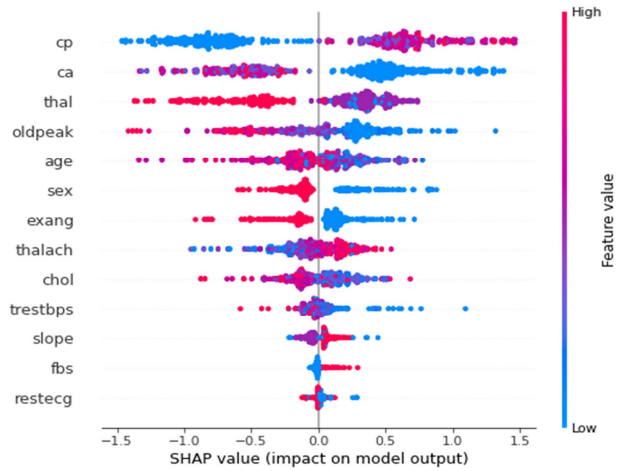

Fig. 3 SHAP values of each feature

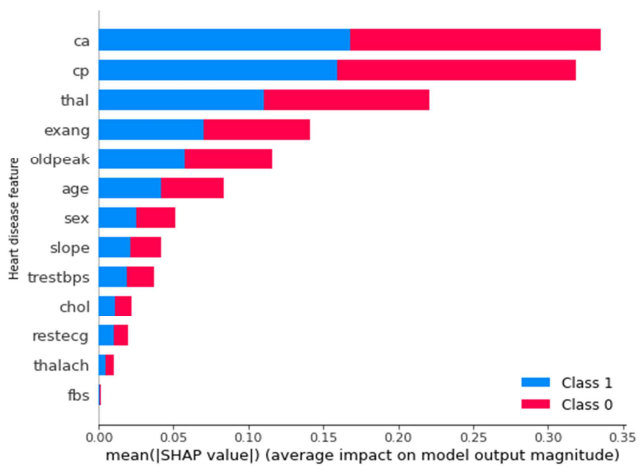

Fig. 4 SHAP waterfall plot

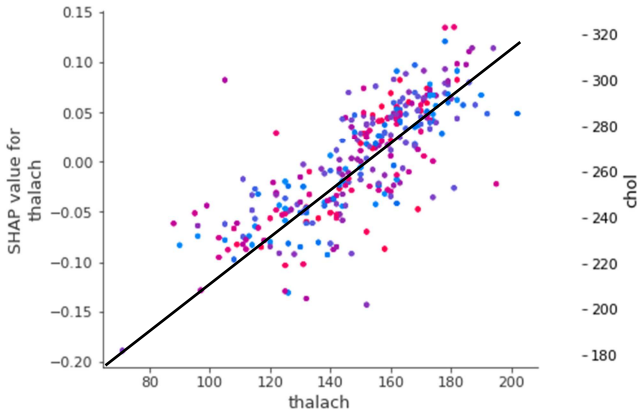

Fig. 5 SHAP interaction plot of maximum heart rate (thalach) vs cholesterol level (chol)

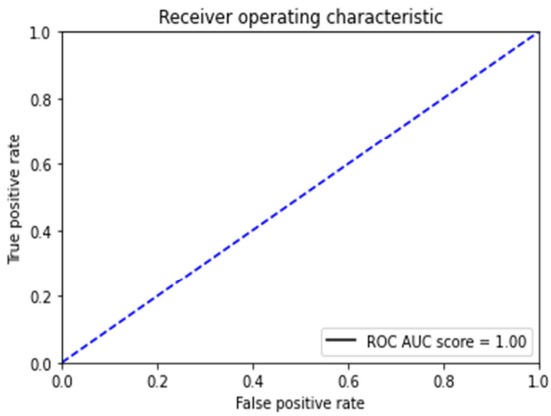

Fig. 6 ROC curve of the developed model

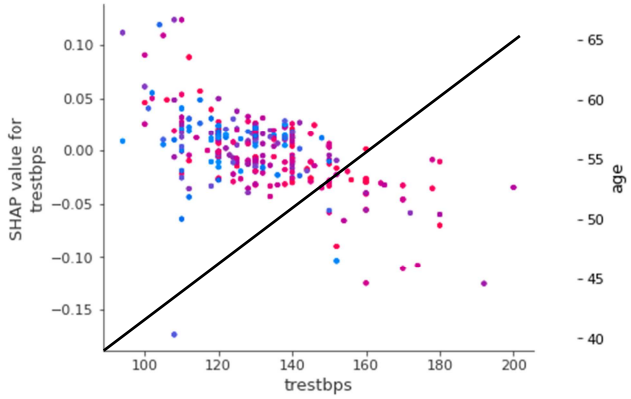

Fig. 7 Interaction plot of total resting blood pressure (trestbps) vs age

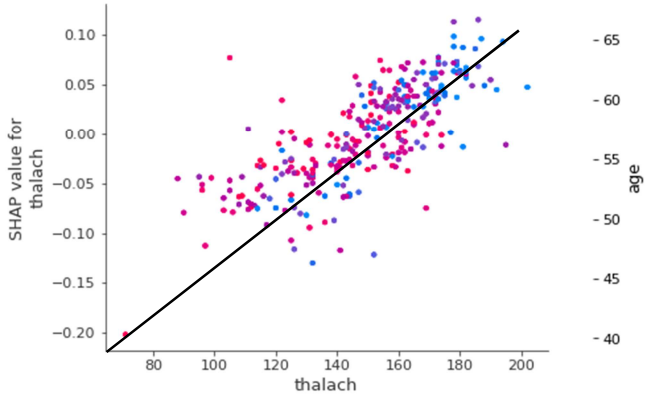

Fig. 8 Interaction plot of maximum heart rate (thalach) vs age

Fig. 6 demonstrates the receiver operating characteristic (ROC) curve of the random forest model for heart disease diagnosis. As can be seen, the area under the ROC curve is 1.00, which shows that the model is perfect for the detection of heart disease. Moreover, the area score or the area under curve (AUC) value of 1.00 evidently proves that the model performs well and does not have the prediction outcome of false positive and false negative classes on the test set conducted in the experiment.

Fig. 7 demonstrates the SHAP interaction plot between total resting blood pressure (trestbps) and age. As demonstrated in Fig. 7, the SHAP interaction values for total resting blood pressure (trestbps) are higher for the patients above the age of 50 years. Thus, the plot evidently proves that there is high positive interaction between age and total resting blood pressure (trestbps). Fig. 8 illustrates the SHAP interaction value for maximum heart rate (thalach) and age. As observed from Fig. 8, the SHAP interaction value for maximum heart rate (thalach) is higher for the patients above the age of 45 years. Thus, the plot appears to prove that there is high positive interaction between age and maximum heart rate (thalach).

6. Conclusions

This study presents an explainable and automated heart disease diagnosis model using the SHAP model explanation technique and random forest-based ensemble learning method. With the SHAP technique, an interpretable random forest model with 100% prediction accuracy is successfully developed. The SHAP technique is effective for providing the intuition and the reasoning behind the predictive outcome of ensemble learning methods such as random forest for heart disease diagnosis. For future work, it is recommended to use various model explanation techniques and machine learning algorithms, such as extreme boosting and deep learning, to identify whether a patient is suffering from heart disease or not. The results of future work will facilitate the trusted and reliable automated intelligent heart disease diagnosis models that would become alternatives to the healthcare experts.

Conflicts of Interest

The author has no conflicts of interest.

References

[1] S. Liu, X. Wang, M. Liu, and J. Zhu, "Towards Better Analysis of Machine Learning Models: A Visual Analytics Perspective," Visual Informatics, vol. 1, no. 1, pp. 48-56, March 2017.

[2] K. Aas, M. Jullum, and A. Løland, "Explaining Individual Predictions When Features Are Dependent: More Accurate Approximations to Shapley Values," Artificial Intelligence, vol. 298, Article no. 103502, September 2021.

[3] A. Chatzimparmpas, R. M. Martins, I. Jusufi, and A. Kerren, "A Survey of Surveys on the Use of Visualization for Interpreting Machine-Learning Models," Information Visualization, vol. 19, no. 3, pp. 207-233, July 2020.

[4] J. Zhou, A. H. Gandomi, F. Chen, and A. Holzinger, "Evaluating the Quality of Machine Learning Explanations: A Survey on Methods and Metrics," Electronics, vol. 10, no. 5, Article no. 593, March 2021.
[5] P. Linardatos, V. Papastefanopoulos, and S. Kotsiantis, "Explainable AI: A Review of Machine Learning Interpretability Methods," Entropy, vol. 23, no. 1, Article no. 18, January 2021.
[6] H. S. Yan, M. C. Tsai, and M. H. Hsu, "An Experimental Study of the Effects of Cam Speeds on Cam-Follower Systems," Mechanism and Machine Theory, vol. 31, no. 4, pp. 397-412, May 1996.
[7] D. Farrugia, C. Zerafa, T. Cini, B. Kuasney, and K. Livori, "A Real-Time Prescriptive Solution for Explainable Cyber-Fraud Detection within the iGaming Industry," SN Computer Science, vol. 2, no. 3, Article no. 215, May 2021.
[8] K. Futagami, Y. Fukazawa, N. Kapoor, and T. Kito, "Pairwise Acquisition Prediction with SHAP Value Interpretation," The Journal of Finance and Data Science, vol. 7, pp. 22- 44, November 2021.
[9] M. Chaibi, E. M. Benghoulam, L. Tarik, M. Berrada, and A. E. Hmaidi, "An Interpretable Machine Learning Model for Daily Global Solar Radiation Prediction," Energies, vol. 14, no. 21, Article no. 7367, November 2021.
[10] P Csókaa, F. Illés, and T. Solymosi, "On the Shapley Value of Liability Games," European Journal of Operational Research, in press.
[11] C. M. Viana, M. Santos, D. Freire, P. Abrantes, and J. Rocha, "Evaluation of the Factors Explaining the Use of Agricultural Land: A Machine Learning and Model-Agnostic Approach," Ecological Indicators, vol. 131, Article no. 108200, November 2021.
[12] S. N. Payrovnaziri, Z. Chen, P. Rengifo-Moreno, T. Miller, J. Bian, J. H. Chen, et al., "Explainable Artificial Intelligence Models Using Real-World Electronic Health Record Data: A Systematic Scoping Review," Journal of the American Medical Informatics Association, vol. 27, no. 7, pp. 1173-1185, 2020.
[13] K. Dissanayake and M. G. M. Johar, "Comparative Study on Heart Disease Prediction Using Feature Selection Techniques on Classification Algorithms," Applied Computational Intelligence and Soft Computing, vol. 2021, Article no. 5581806, 2021.