Multicriteria Algorithms with Applications

Journal Homepage: sciencesforce.com/mawa

Multicriteria. Algo. Appl. Vol. 2 (2024) 80–88

Paper Type: Original Article

Machine Learning with Multi-Criteria Decision Making Model for Thyroid Disease Prediction and Analysis

Ahmed M. Ali [1,*], **and Said Broumi** [2]

[1] Faculty of Computers and Informatics, Zagazig University, Zagazig 44519, Sharqiyah, Egypt; aabdelmonem@fci.zu.edu.eg.

[2] Laboratory of Information Processing, Faculty of Science Ben M'Sik, University of Hassan II, Casablanca, Morocco; s.broumi@flbenmsik.ma.

Received: 17 Oct 2023 **Revised:** 13 Jan 2024 **Accepted:** 25 Jan 2024 **Published:** 30 Jan 2024

Abstract

This study adopted a decision system model that includes machine learning (ML) and multi-criteria decision-making (MCDM) for thyroid prediction and analysis. Many people face thyroid disease, so the early prediction of this disease can aid people around the world in early treatment. This paper integrates the ML algorithms with the MCDM methodology. Three ML algorithms are used in this paper: logistic regression (LR), support vector machine (SVM), and random forest classifier (RF). These algorithms are used to predict and analyse thyroid disease. The results show the RF has the highest accuracy, precision, and F1 score. The RF has 0.95 accuracy. The SVM has a 1.0 recall score. Then, the MCDM methodology is used with various criteria to rank and use the best ML algorithm. The TOPSIS method is used as an MCDM method to rank the ML algorithms. The mean method is used to compute the criteria weights. The results of the MCDM methodology show that RF is the best ML algorithm in this paper, followed by SVM, and the worst ML algorithm is LR.

Keywords: Machine Learning; Artificial Intelligence; Prediction; Multi-Criteria Decision Making, Thyroid Disease.

1 | Introduction

Located in the neck, the thyroid is an endocrine gland that produces and distributes the hormones FT3 and FT4 into the circulation. The body's temperature, heart rate, and, most importantly, metabolism—the process by which the body utilises and absorbs nutrients—are all regulated by thyroid hormones [1]. Major diseases may arise in either scenario where the thyroid gland functions above normal (hyperthyroidism with increased hormones) or below average (hypothyroidism with decreased hormones). The thyroid gland may also become inflamed (thyroiditis) or expand, forming one or more internal swellings (nodules, multinodular goitre). A few of these nodules may have malignant tumors [2].

People have been making choices that affect their everyday lives ever since the beginning of time. Researchers have been interested in analysing how humans do this activity for a long time. To simplify and accurately depict the actual system, we must model the environment in which we operate. This requires that the model be easily understandable and attainable. As a result, we research the options that may be selected and the

Corresponding Author: aabdelmonem@fci.zu.edu.eg

https://doi.org/10.61356/j.mawa.2024.26961

standards by which those options will be judged. This seems straightforward initially, but it is a component of the larger field known as Multiple Criteria Decision Making (MCDM). The MCDM process aims to identify the best option from a group of workable options. A matrix representation of an MCDM issue with options and criteria is possible [3].

Machine learning (ML) technologies have a wide range of possible applications. ML technologies can execute updates and modifications based on real-time data. They can adaptively identify rules and patterns from diverse forms of data with varied features using algorithms without manually creating rules [4, 5]. The checker's programme that Samuel created in the 1950s is where the idea of machine learning originated. The historical context is that artificial intelligence evolved from the reasoning to the knowledge periods, emphasising building knowledge bases via data-driven machine learning without requiring human participation [6, 7].

Since its inception, ML has evolved from logic-based to connectionism, symbolism, and statistical learning, and it can now be applied to various tasks and data sets. Support vector machines, which represent statistical learning, and deep neural networks, which represent connectionism, have each become famous study avenues. Recently, ML has seen the emergence of new concepts such as federated learning, which safeguards data privacy; transfer learning, which guarantees domain knowledge's flexibility; and meta-learning, which can swiftly adjust to new tasks. ML technologies enable data-driven decision-making and are compatible with the information age by combining various MCDM procedures like parameter determination, criteria extraction, and problem definition. They can learn decision information based on accurate data without artificial parameters or rules [8, 9].

Scholars must understand the current state of research, potential future directions, and how various ML technologies relate to specific MCDM situations or processes to promote innovative ideas and breakthroughs in ML technology applications to MCDM [10]. This study integrated the MCDM with ML algorithms for thyroid disease prediction.

2 | Dataset

This section describes the dataset of thyroid. This study used the dataset from Kaggle to predict and analyze thyroid disease. This data has 3711 rows and 30 columns; the first five columns show in Table 1. We built some statistical methods of the dataset as shown in Table 2.

Table 1. The first five columns of thyroid dataset.

	Row 0	Row 1	Row 2	Row 3	Row 4
age	0.365591	0.16129	0.430108	0.72043	0.72043
sex	1	1	2	1	1
on thyroxine	0	0	0	1	0
query on thyroxine	0	0	0	0	0
on antithyroid medication	0	0	0	0	0
sick	0	0	0	0	0
pregnant	0	0	0	0	0
thyroid surgery	0	0	0	0	0
I131 treatment	0	0	0	0	0
query hypothyroid	0	0	0	0	0
...	...	...	...	...	...
TT4 measured	1	1	1	1	1
TT4	0.116183	0.012448	0.041494	0.344398	0.834025
T4U measured	1	0	1	0	1
T4U	0.493151	1	0.328767	1	0.30137

FTI measured	1	0	1	0	1
FTI	0.042735	1	0.094017	1	0.850427
TBG measured	0	0	0	0	0
TBG	0	0	0	0	0
referral source	1	4	4	4	3
binaryClass	1	1	1	1	1

Table 2. Some statistical analysis of the dataset.

	count	**mean**	**std**	**min**	**25%**	**50%**	**75%**	**max**
age	3711	0.499935	0.22434	0	0.301075	0.537634	0.677419	1
sex	3711	1.266505	0.52522	0	1	1	2	2
on thyroxine	3711	0.125034	0.330802	0	0	0	0	1
query on thyroxine	3711	0.013473	0.115306	0	0	0	0	1
on antithyroid medication	3711	0.011318	0.105795	0	0	0	0	1
sick	3711	0.039612	0.195072	0	0	0	0	1
pregnant	3711	0.014282	0.118666	0	0	0	0	1
thyroid surgery	3711	0.014282	0.118666	0	0	0	0	1
I131 treatment	3711	0.015899	0.1251	0	0	0	0	1
query hypothyroid	3711	0.063056	0.243096	0	0	0	0	1
...	...	...	...	...	...	...	...	...
TT4 measured	3711	0.953921	0.209685	0	1	1	1	1
TT4	3711	0.494328	0.407627	0	0.087137	0.327801	0.937759	1
T4U measured	3711	0.911884	0.283502	0	1	1	1	1
T4U	3711	0.444211	0.21459	0	0.315068	0.390411	0.486301	1
FTI measured	3711	0.912423	0.282718	0	1	1	1	1
FTI	3711	0.464598	0.414668	0	0.07265	0.239316	0.944444	1
TBG measured	3711	0	0	0	0	0	0	0
TBG	3711	0	0	0	0	0	0	0
referral source	3711	3.267583	1.097079	0	3	4	4	4
binaryClass	3711	0.921584	0.268861	0	1	1	1	1

3 | Machine Learning with Multi-Criteria Decision Making

This section introduces two phases, in the first phase, we apply the machine learning algorithms in thyroid dataset. In the second phase, we test the best model by using the MCDM algorithms. Figure 1 shows the methodology of this paper.

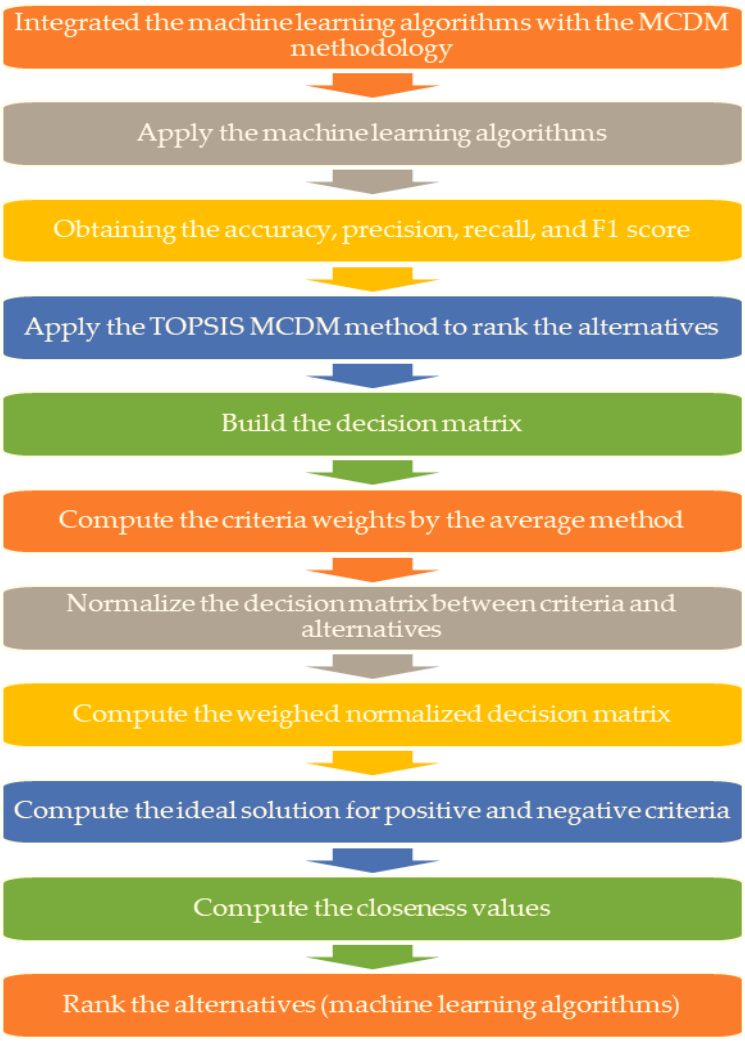

Figure 1. The research framework.

3.1 | Machine Learning Algorithms

This part introduces the ML algorithms used to predict the thyroid disease such as support vector machine, random forest classifier, and logistic regression.

3.1.1 | Support Vector Machine (SVM)

A nonparametric, supervised, kernel-based technique from statistical learning techniques is called Support Vector Machine (SVM). In kernel-based learning, the input data is implicitly mapped into a high-dimensional feature space defined by a kernel function. Stated differently, kernel-based learning performs a back transformation in nonlinear space after using the linear hyperplane as a decision function for nonlinear situations. To get the best answer, SVM uses the Lagrange multiplier to calculate each feature's partial differentiation. As a result, the model narrows down the training data's complexity to a sizable subset of "support vectors."

3.1.2 | Logistic Regression (LR)

Typically, logistic regression (LR) fits the log odds and explanatory factors to determine the class membership likelihood for the two groups.

$$\log\left(\frac{p(Y=1|x)}{1-p(Y=1|x)} = \beta_0 + \beta_1 X1 + \cdots . \beta_N XN\right) \quad (1)$$

Where $\beta = \beta_0, \beta_1, \ldots . \beta_n$ refers to the estimated coefficient regression.

3.1.3 | Random Forest (RF)

In this work, supervised, tree-based ensemble machine learning is accomplished via the application of RF. The random subspace method, which applies the subsampling processes of the parameters without substitute at every division in the tree, and boosting or bootstrap aggregation, which subsamples input samples with substitution, are combined to form the theoretical framework known as RF. The decision trees $\{O_1(X), \dots, O_B(X)\}$, where $X = \{x1, \dots, xp\}$ is a p-dimensional vector of features in thyroid disease.

3.2 | Multi-Criteria Decision Making Algorithms

Many possibilities in MCDM need to be assessed and contrasted based on various factors. The purpose of MCDM is to assist the decision-maker in selecting an option from a range of options. In this sense, several competing requirements often characterise actual issues, and it's possible that no solution can satisfy all the criteria simultaneously. Therefore, the answer is a compromise based on the decision-maker's preferences. Thus, the foundation of TOPSIS is that the selected option should be the closest to the Positive Ideal Solution (PIS) and the furthest from the Negative Ideal Solution (NIS) [11]. The proximity measure is used to determine the final ranking.

The TOPSIS process includes the following phases.

Step 1. Build the decision matrix.

Step 2. Normalize the decision matrix.

$$n_{ij} = \frac{x_{ij}}{\sqrt{\sum_{j=1}^{m}\left(x_{ij}\right)^2}} \; j = 1,2, \dots n; i = 1,2, \dots . m \tag{2}$$

Step 3. Compute the criteria weights.

Step 4. Compute the weighted normalized decision matrix.

$$y_{ij} = w_j n_{ij} \tag{3}$$

Step 5. Compute the negative and positive ideal solutions.

$$A^+ = \{y_1^+, y_2^+, \dots y_n^+\} = \left\{\left(\max y_{ij}\right)\left(\min y_{ij}\right)\right\} \tag{4}$$

$$A^- = \{y_1^-, y_2^-, \dots y_n^-\} = \left\{\left(\min y_{ij}\right)\left(\max y_{ij}\right)\right\} \tag{5}$$

Step 6. Compute the distance matrix.

$$d_i^+ = \left\{\sqrt{\sum_{j=1}^{n}\left(y_{ij} - y_j^+\right)^2}\right\} \tag{6}$$

$$d_i^- = \left\{\sqrt{\sum_{j=1}^{n}\left(y_{ij} - y_j^-\right)^2}\right\} \tag{7}$$

Step 7. Compute the closeness values.

$$T_{ij} = \frac{d_i^-}{d_i^- + d_i^+} \tag{8}$$

4 | Results

In this section, we introduce the results of 3 ML algorithms to predict thyroid disease. Then we use the MCDM methodology to select the best ML algorithm to be used based on several criteria. This study used a dataset from the Kaggle of thyroid disease. Figure 2 shows the accuracy, precision, and recall of the ML algorithms.

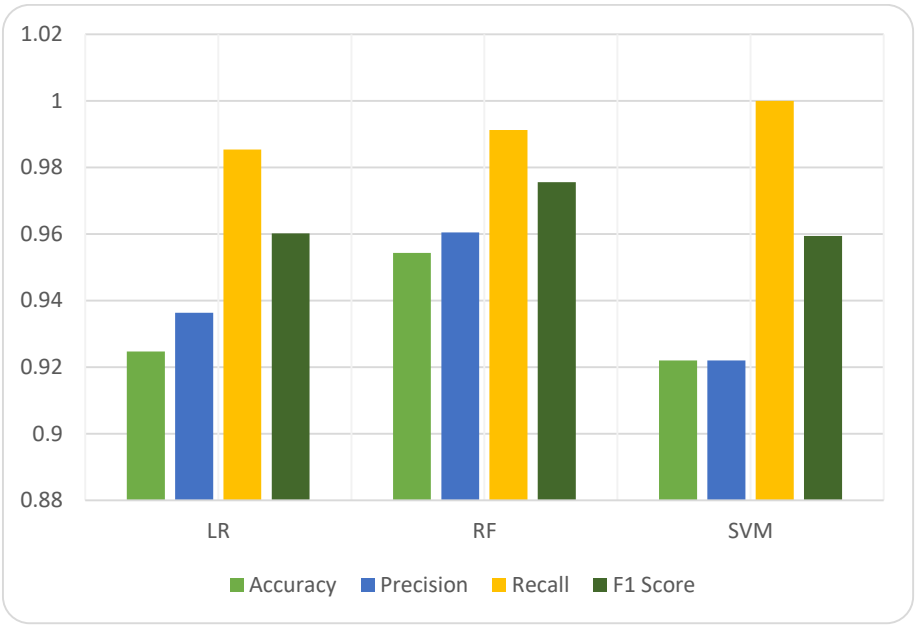

Figure 2. The results of ML algorithms.

The RF has the highest accuracy, precision, and f1 score the SVM has the highest recall. Figure 3 shows the visualization of the RF features. Figure 4 show the ROC curve.

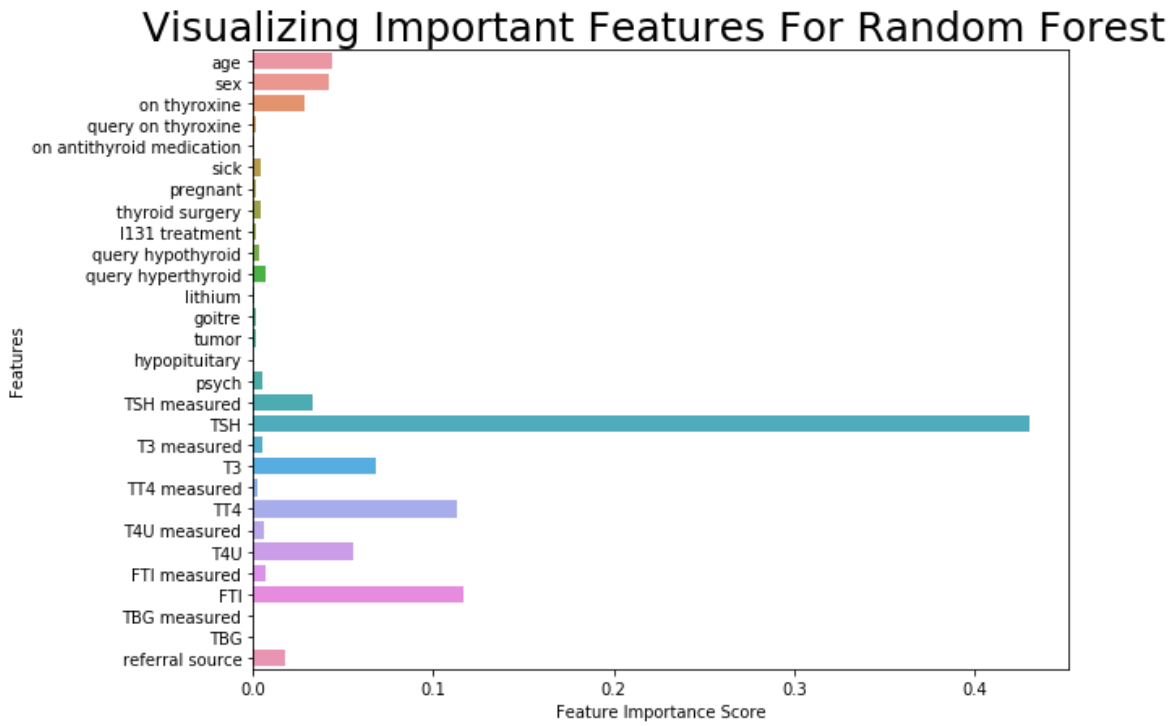

Figure 3. The visualization of RF features.

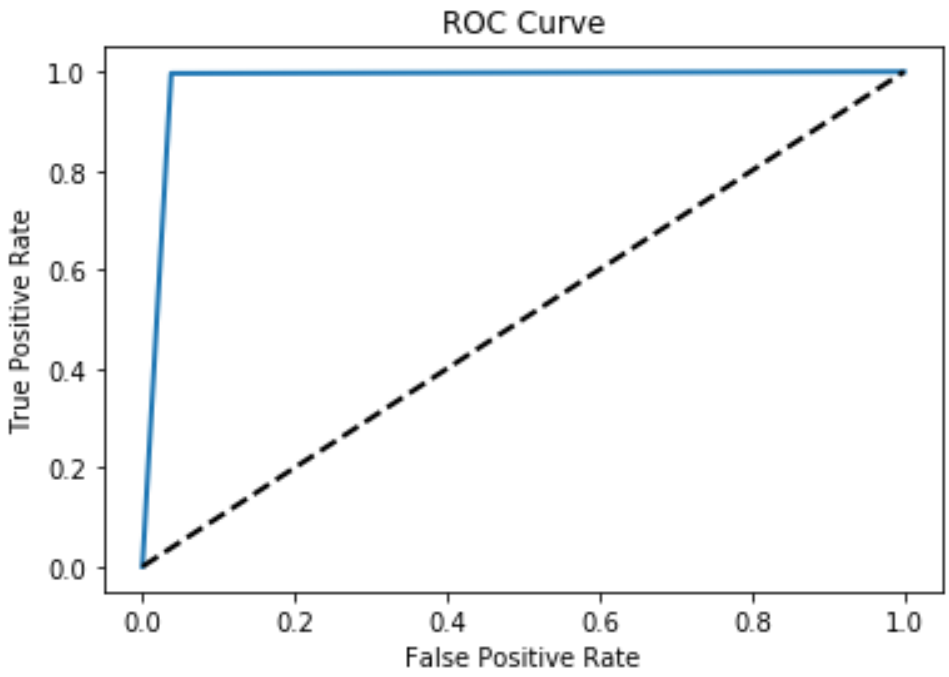

Figure 4. The ROC curve of the random forest.

Then we apply the steps of the MCDM methodology to use the best algorithm.

Step 1. Build the decision matrix between criteria and algorithms.

Step 2. Normalize the decision matrix by using Eq. (2)

Step 3. Compute the criteria weights by using the mean method. The criteria weights show the importance of each feature in the section ML algorithm.

Step 4. Compute the weighted normalized decision matrix by using Eq. (3).

Step 5. Compute the negative and positive ideal solutions by using Eqs. (4) and (5).

Step 6. Compute the distance matrix by using Eqs. (6) and (7)

Step 7. Compute the closeness values by using Eq. (8). The rank of alternatives is shown in Figure 5. The results of MCDM methodology show the RF is the best ML algorithm.

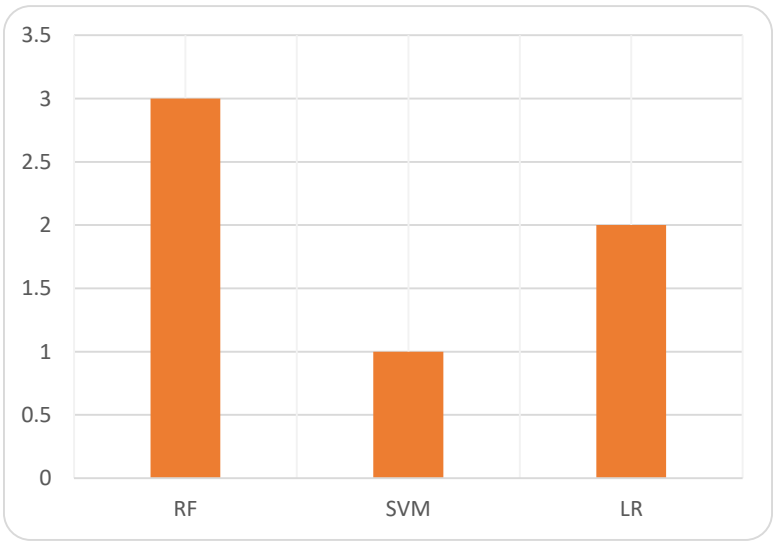

Figure 5. The rank of ML algorithms.

5 | Conclusions

The main goal of this study is to integrate the MCDM methodology with machine learning algorithms to show the best algorithm to predict thyroid disease and analyse it with various criteria. Three ML algorithms are used in this paper to predict thyroid disease with computation of the accuracy, recall score, precision score, and F1 score, such as random forest, logistic regression, and support vector machine. The thyroid disease dataset is used in this study; then, the ML algorithms are trained with this dataset. The results show the random forest has the highest accuracy, precision, and F1 score, but the support vector machine has the highest recall score. The MCDM method is used to select the best algorithms and rank them. The TOPSIS method is an MCDM methodology used to predict thyroid disease based on several criteria. The decision matrix is built based on a set of criteria and alternatives. The criteria weights are computed to determine the most influential of the ML algorithms. The weighted normalized decision matrix is calculated by multiplying the criteria weights by the normalization decision matrix. Then, the positive and negative ideal solutions are computed based on the positive and negative criteria. All requirements are positive. Then, the rank of alternatives is calculated—the results show that the RF has the highest rank, followed by the SVM and the LR.

Acknowledgments

The author is grateful to the editorial and reviewers, as well as the correspondent author, who offered assistance in the form of advice, assessment, and checking during the study period.

Author Contributaion

All authors contributed equally to this work.

Funding

This research has no funding source.

Data Availability

The datasets generated during and/or analyzed during the current study are not publicly available due to the privacy-preserving nature of the data but are available from the corresponding author upon reasonable request.

Conflicts of Interest

The authors declare that there is no conflict of interest in the research.

Ethical Approval

This article does not contain any studies with human participants or animals performed by any of the authors.

References

[1] Sinha, B. B., Ahsan, M., & Dhanalakshmi, R. (2023). LightGBM empowered by whale optimization for thyroid disease detection. International Journal of Information Technology, 1-10. https://doi.org/10.1007/s41870-023-01261-3

[2] Priya, V. V., Subashini, R., & Priya, S. H. (2023, February). Thyroid Disease Prediction using Random Forest Algorithm. In 2023 7th International Conference on Computing Methodologies and Communication (ICCMC) (pp. 794-799). IEEE. https://doi.org/10.1109/ICCMC56507.2023.10083592

[3] Sultana, A., & Islam, R. (2023). Machine learning framework with feature selection approaches for thyroid disease classification and associated risk factors identification. Journal of Electrical Systems and Information Technology, 10(1), 1-23. https://doi.org/10.1186/s43067-023-00101-5

[4] Botlagunta, M., Botlagunta, M. D., Myneni, M. B., Lakshmi, D., Nayyar, A., Gullapalli, J. S., & Shah, M. A. (2023). Classification and diagnostic prediction of breast cancer metastasis on clinical data using machine learning algorithms. Scientific Reports, 13(1), 485. https://doi.org/10.1038/s41598-023-27548-w

[5] Yang, Y., Zhou, W., Jiskani, I. M., Lu, X., Wang, Z., & Luan, B. (2023). Slope Stability Prediction Method Based on Intelligent Optimization and Machine Learning Algorithms. Sustainability, 15(2), 1169. https://doi.org/10.3390/su15021169

[6] Peng, Y., & Unluer, C. (2023). Modeling the mechanical properties of recycled aggregate concrete using hybrid machine learning algorithms. Resources, Conservation and Recycling, 190, 106812. https://doi.org/10.1016/j.resconrec.2022.106812

[7] Nancy Noella, R. S., & Priyadarshini, J. (2023). Machine learning algorithms for the diagnosis of Alzheimer and Parkinson disease. Journal of Medical Engineering & Technology, 47(1), 35-43. https://doi.org/10.1080/03091902.2022.2097326

[8] Tian, G., Lu, W., Zhang, X., Zhan, M., Dulebenets, M. A., Aleksandrov, A., ... & Ivanov, M. (2023). A survey of multi-criteria decision-making techniques for green logistics and low-carbon transportation systems. Environmental Science and Pollution Research, 30(20), 57279-57301. https://doi.org/10.1007/s11356-023-26577-2

[9] Méndez, M., Merayo, M. G., & Núñez, M. (2023). Machine learning algorithms to forecast air quality: a survey. Artificial Intelligence Review, 1-36. https://doi.org/10.1007/s10462-023-10424-4

[10] Herm, L. V., Heinrich, K., Wanner, J., & Janiesch, C. (2023). Stop ordering machine learning algorithms by their explainability! A user-centered investigation of performance and explainability. International Journal of Information Management, 69, 102538. https://doi.org/10.1016/j.ijinfomgt.2022.102538

[11] Ayan, B., Abacıoğlu, S., & Basilio, M. P. (2023). A Comprehensive Review of the Novel Weighting Methods for Multi-Criteria Decision-Making. Information, 14(5), 285. https://doi.org/10.3390/info14050285