Neutrosophic-based Machine Learning Techniques for Analysis and Diagnosis the Breast Cancer

Rosita Elizabeth O. Torres[1,*], Jhonny Rodríguez Gutiérrez[2], Alex G. Lara Jacome[3]

[1]Docente de la carrera de Medicina de la Universidad Regional Autónoma de los Andes (UNIANDES Ambato), Ecuador

[2]Docente de la carrera de Medicina de la Universidad Regional Autónoma de los Andes (UNIANDES Santo Domingo), Ecuador

[3]Docente de la carrera de Medicina de la Universidad Regional Autónoma de los Andes (UNIANDES Ambato), Ecuador

Email: ua.rositaolivo@uniandes.edu.ec; us.jhonnyrodriguez@uniandes.edu.ec; ua.alexlara@uniandes.edu.ec

Abstract

Approximately one in eight women will get breast cancer in their lifetime. Because of the risks associated with radiation exposure, various women choose to avoid getting detected with breast cancer. Non-invasive breast cancer detection methods have limitations concerning the safety of radiation exposure and the accuracy with which tumors in the breast are diagnosed. Machine learning methods are commonly used to diagnose breast cancer. This paper applied three different machine learning methods like KNN, Naïve Bayes, and ID3. These methods are applied to the Wisconsin Breast Cancer dataset. In the process of categorization, data with unbalanced classes is problematic because methods are more probable to categorize fresh observations to the majority class since the likelihood of cases forming the plurality class is considerably high. So neutrosophic set is used to overcome the vague and uncertain data. This paper used single-valued neutrosophic numbers to evaluate the criteria. This paper used ROC and accuracy to evaluate the methods. The KNN has a 96.7%, Naïve Bayes has a 95.2%, and ID3 has a 95.3% accuracy.

Keywords: Naïve Bayes; Neutrosophic Set; KNN; ID3; Breast Cancer.

1. **Introduction**

Cancer that begins in breast tissue is called breast cancer (BC). A tumor can metastasize to other organs. Breast cancer (BC) is a global illness that mostly affects women between the ages of 25 and 50. The possible increase in the quantity of BC patients in India is concerning because of the misery it might cause. In the United States, patients with BC have a five-year survival rate of over 90%, but in India, it's closer to 60%. In 2020, experts predict that India's number of BC patients might reach 2 million[1], [2].

Medical experts have pinpointed hormonal, lifestyle, and environmental variables that may all play a role in a person's susceptibility to getting BC. It is estimated that between 5 and 6 percent of all cases of BC may be traced back to inherited gene variations. Other causes of BC include being overweight, being older, and experiencing hormonal abnormalities after menopause[3], [4].

Although there is currently no way to prevent BC, the disease may be effectively treated if caught early. In addition, this has the potential to drastically cut down on the overall cost of the therapy. However, atypical manifestations of

DOI: https://doi.org/10.54216/IJNS.210115
Received: February 10, 2023 Revised: April 12, 2023 Accepted: May 11, 2023

cancer symptoms might make for challenging early identification. Mammograms and self-breast examinations are essential for detecting abnormalities early when the tumor has progressed[5]–[7].

Marketers, social scientists, financiers, and medical professionals may all benefit from data mining, which is now a widely used method for information discovery. Classifier techniques have recently been used to medical datasets to do predictive analysis on patients' medical diagnoses. Tumor behavior in breast cancer patients might, for instance, be evaluated using machine learning methods. Since the likelihood of not having this ailment is greater than the risk of having it, there is a disparity in the information regarding training[8], [9].

It is not uncommon for decision factors to be murky. Fuzzy set theory may be used to solve this problem. Atanassov generalizes it to an IFS or an IFS with intuition. After that, Smarandache applies an indeterminacy-membership value to IFS so that it may be applied to a neutrosophic set (NS). Despite the NS theory's ability to deal with uncertainty, it has been challenging to use in practice. Wang et al. presented a revision to NS theory called a single-valued neutrosophic set (SVNS)[10], [11].

The accuracy of three distinct classifiers KNN, Naïve Bayes, and ID3 in the identification of breast cancer is compared in this research. Our goal is to improve the classifier's efficiency by preparing the dataset in a way that accounts for the dataset's inherent inequalities and any missing values. Also, the neutrosophic sets are used to decide what criteria are used as an input of machine learning and exclude others.

2. Literature Review

This section presented previous works on breast cancer using neutrosophic sets (NS). The neutrosophic set has been used in breast cancer as The authors [12] presented a computer-aided diagnosis (CAD) approach for separating normal and abnormal thermograms for breast cancer. Automatic categorization and segmentation are the two primary components of the strategy. Neutrosophic sets (NS) and an optimized Fast Fuzzy c-mean (F-FCM) algorithm were offered as a means of better segmentation during the first stage.

Using neutrosophic sets (NS) and moth-flame optimization (MFO), the authors [13] provided a method for the automated identification of mitosis in histopathology slide imaging. The suggested method involves two basic stages: the extraction of candidates and the categorization of candidates. Histopathological slide images were processed using a Gaussian filter and then projected to the NS domain during the candidate extraction phase. The picture of the truth subset has then undergone morphological procedures to improve it and zero in on cells in mitosis.

A new approach for classifying breast tumors was suggested by the authors [14] which uses a neutrosophic similarity score to make use of both textural and morphologic data. The next step is to use a supervised feature choice method to further condense the available features. Lastly, an SVM classifier is used to validate the suggested features' discriminatory efficacy.

The authors [15] provided a cutting-edge strategy for segmenting ultrasound pictures of the breast. To counteract the speckle noise and tissue-related textures inherent to ultrasound pictures, it employs a blend of region-based active contour and neutrosophic theory.

The authors [16] offered a brand-new method for BUS image division that utilizes the level-set algorithm and the neutrosophic similarity score (NSS). After converting the input BUS picture into the NS area using the three membership subsets T, I, and F, a similarity score NSS is created and used to determine how closely a certain portion of the image corresponds to the actual tumor. Finally, the NSS image's tumor is separated from the surrounding tissue using the level set approach. Several types of clinical BUS pictures have been used in studies.

The authors [17] suggested a neutrosophic connection matrix-based fuzzy clustering technique. Initially, a neutrosophic connection matrix is generated by fuzzifying the information into sets of neutrosophic terms. The neutrosophic equivalency matrix is computed by generating a finite series of neutrosophic connection matrices. The neutrosophic similarity matrix is then lambda-cut to get the final lambda-cut matrix, which is then utilized to identify clusters.

3. Research Methodology

In this section, the research methodology will be discussed. The neutrosophic TOPSIS and machine learning are organized in this section as shown in Figure 1.

DOI: https://doi.org/10.54216/IJNS.210115
Received: February 10, 2023 Revised: April 12, 2023 Accepted: May 11, 2023

Zadeh is widely credited with developing the notion of fuzzy sets. In fuzzy set theory, objects may have a level of membership that can be anywhere from 0 to 1, inclusive. It was Smarandache who first developed neutrosophy, the study of neutralities and their relationships to other areas of study and the world at large. Smarandache, Wang, and coworkers devised single-valued neutrosophic sets, which accept values from the interval [0,1], to facilitate their use in practical settings. Thus, a neutrosophic set containing a single value is an example of such a set, and it may be utilized practically to address real-world issues, notably in decision-making systems. An extension of bipolar fuzzy sets, bipolar neutrosophic sets was the topic of discussion in Deli et al[18], [19].

Making the most optimal decision possible from a pool of options defined by numerous, often competing criteria is the goal of multi-criteria decision-making (MCDM). Among the most popular and successful approaches to MCDM problem-solving, TOPSIS was created by Hwang and Yoon. Characteristic quantities and weights are established with precision in traditional MCDM approaches[20], [21]. In 2000, Chen presented the first version of the TOPSIS approach in a fuzzy form, making it possible to work with situations with partial or ambiguous data. This paper used the Bipolar neutrosophic sets.[22], [23]

$$X = \{x, < A_x^+, B_x^+, C_x^+, A_x^-, B_x^-, C_x^- >\} \tag{1}$$

The decision matrix is built as:

$$K = \left[k_{ij}\right]_{m\times n} = \begin{bmatrix} k_{11} & \cdots & k_{1n} \\ \vdots & \ddots & \vdots \\ k_{m1} & \cdots & k_{mn} \end{bmatrix} \tag{2}$$

The weights of criteria can be computed as:

$$w_j = \frac{\sum_{i=1}^{m}\sum_{l=1}^{m}|k_{ik}-k_{lj}|}{\sqrt{\sum_{j=1}^{n}\left(\sum_{i=1}^{m}\sum_{l=1}^{m}|k_{ik}-k_{lj}|\right)^2}} \tag{3}$$

The weights can be normalized as:

$$Nw_j = \frac{\sum_{i=1}^{m}\sum_{l=1}^{m}|k_{ik}-k_{lj}|}{\sum_{j=1}^{n}\left(\sum_{i=1}^{m}\sum_{l=1}^{m}|k_{ik}-k_{lj}|\right)^2} \tag{4}$$

Naive Bayes (NB), a Decision Tree based on the ID3 technique, and KNN were chosen as the three methods for categorization to use. The NB classifier uses the Bayes rule as its basis for probabilistic classification[24], [25]. It works by making the best guess as to the likelihood, for each class value, that a certain instance really does belong to that class. The ID3 method relies on the idea of information entropy, and it operates by breaking down large datasets into smaller ones whose changes in entropy may be more easily analyzed[26], [27].

Predictions of Breast Cancer were made using a fuzzy method based on a function called membership. The authors made an effort to clear up data uncertainty by using the Fuzzy KNN Algorithm. The BC data set was divided into two categories. The dataset was divided into a training and a testing set. After using pre-processing methods, the fuzzy KNN technology was put into action. Accuracy, precision, and the f1 score were only few of the criteria used to evaluate this method. The results showed that the fuzzy KNN classification algorithm was superior to the KNN classifier when it came to accuracy[28], [29], [30].

DOI: https://doi.org/10.54216/IJNS.210115
Received: February 10, 2023 Revised: April 12, 2023 Accepted: May 11, 2023

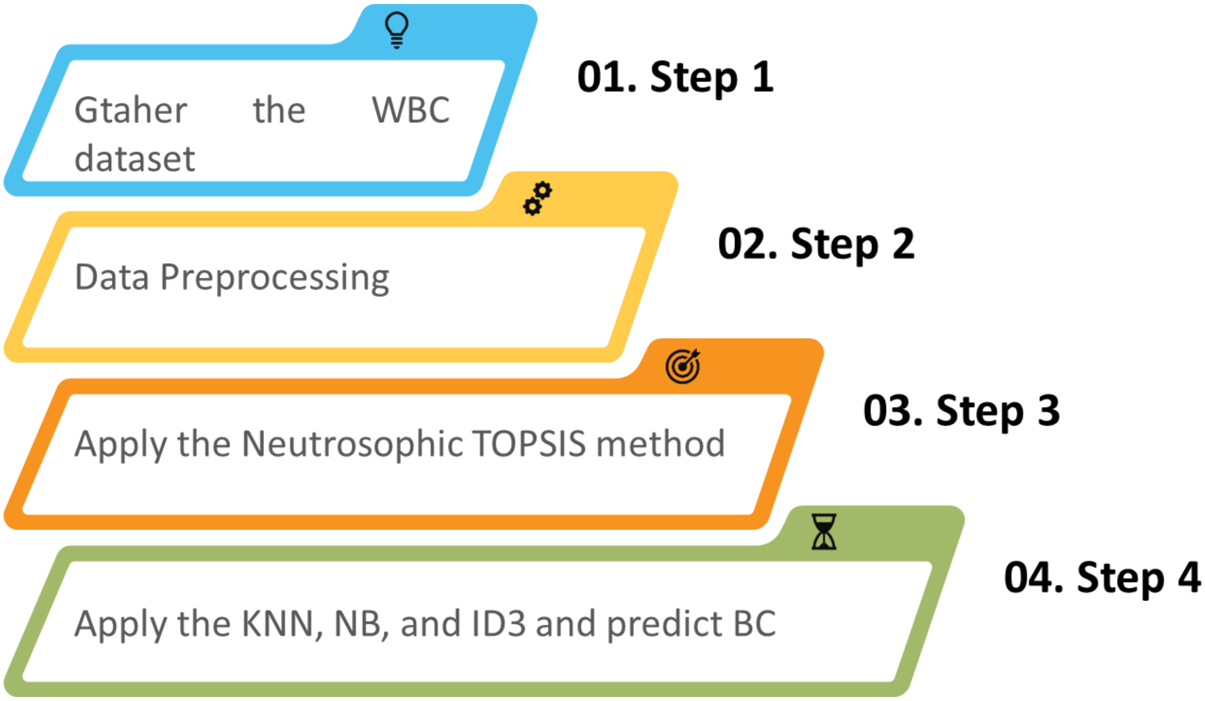

Figure 1: The methodology of neutrosophic TOPSIS and ML methods.

4. Experimental Results

The three techniques for categorization were first put to the test on two cancer datasets: WBC. KNN had the most accuracy in the WBC Cancer dataset (96.7%), Naïve Bayes has a 95.2%, and ID3 has a 95.3% accuracy. The dataset has 31 features. Table 1 shows the descriptive statistics on the dataset. Statisticians employ numerical data or numerical values ascribed to qualitative characteristics to illustrate the significance of the data. Descriptive statistics are used to provide a picture of the features of a set of findings, or the information as it is. Scoring factors may be assigned one of many scales in statistical analysis.

Table 1: The descriptive statistics on the breast cancer dataset.

	count	mean	std	min	25%	50%	75%	max
radius_mean	569	14.12729	3.524049	6.981	11.7	13.37	15.78	28.11
texture_mean	569	19.28965	4.301036	9.71	16.17	18.84	21.8	39.28
perimeter_mean	569	91.96903	24.29898	43.79	75.17	86.24	104.1	188.5
area_mean	569	654.8891	351.9141	143.5	420.3	551.1	782.7	2501
smoothness_mean	569	0.09636	0.014064	0.05263	0.08637	0.09587	0.1053	0.1634
compactness_mean	569	0.104341	0.052813	0.01938	0.06492	0.09263	0.1304	0.3454
concavity_mean	569	0.088799	0.07972	0	0.02956	0.06154	0.1307	0.4268
concave points_mean	569	0.048919	0.038803	0	0.02031	0.0335	0.074	0.2012
symmetry_mean	569	0.181162	0.027414	0.106	0.1619	0.1792	0.1957	0.304
fractal_dimension_mean	569	0.062798	0.00706	0.04996	0.0577	0.06154	0.06612	0.09744
...	...	...	...	...	...	...	...	
radius_worst	569	16.26919	4.833242	7.93	13.01	14.97	18.79	36.04
texture_worst	569	25.67722	6.146258	12.02	21.08	25.41	29.72	49.54
perimeter_worst	569	107.2612	33.60254	50.41	84.11	97.66	125.4	251.2
area_worst	569	880.5831	569.357	185.2	515.3	686.5	1084	4254
smoothness_worst	569	0.132369	0.022832	0.07117	0.1166	0.1313	0.146	0.2226
compactness_worst	569	0.254265	0.157336	0.02729	0.1472	0.2119	0.3391	1.058

DOI: https://doi.org/10.54216/IJNS.210115
Received: February 10, 2023 Revised: April 12, 2023 Accepted: May 11, 2023

concavity_worst	569	0.272188	0.208624	0	0.1145	0.2267	0.3829	1.252
concave points_worst	569	0.114606	0.065732	0	0.06493	0.09993	0.1614	0.291
symmetry_worst	569	0.290076	0.061867	0.1565	0.2504	0.2822	0.3179	0.6638
fractal_dimension_worst	569	0.083946	0.018061	0.05504	0.07146	0.08004	0.09208	0.2075

Then introduce some data analysis of the dataset. This analysis shows the relation between the variables and the target variable. For example, the pair plot shows all variables against other as shown in Figure 2. Also, Figure 3 shows the heat map and correlation between variables.

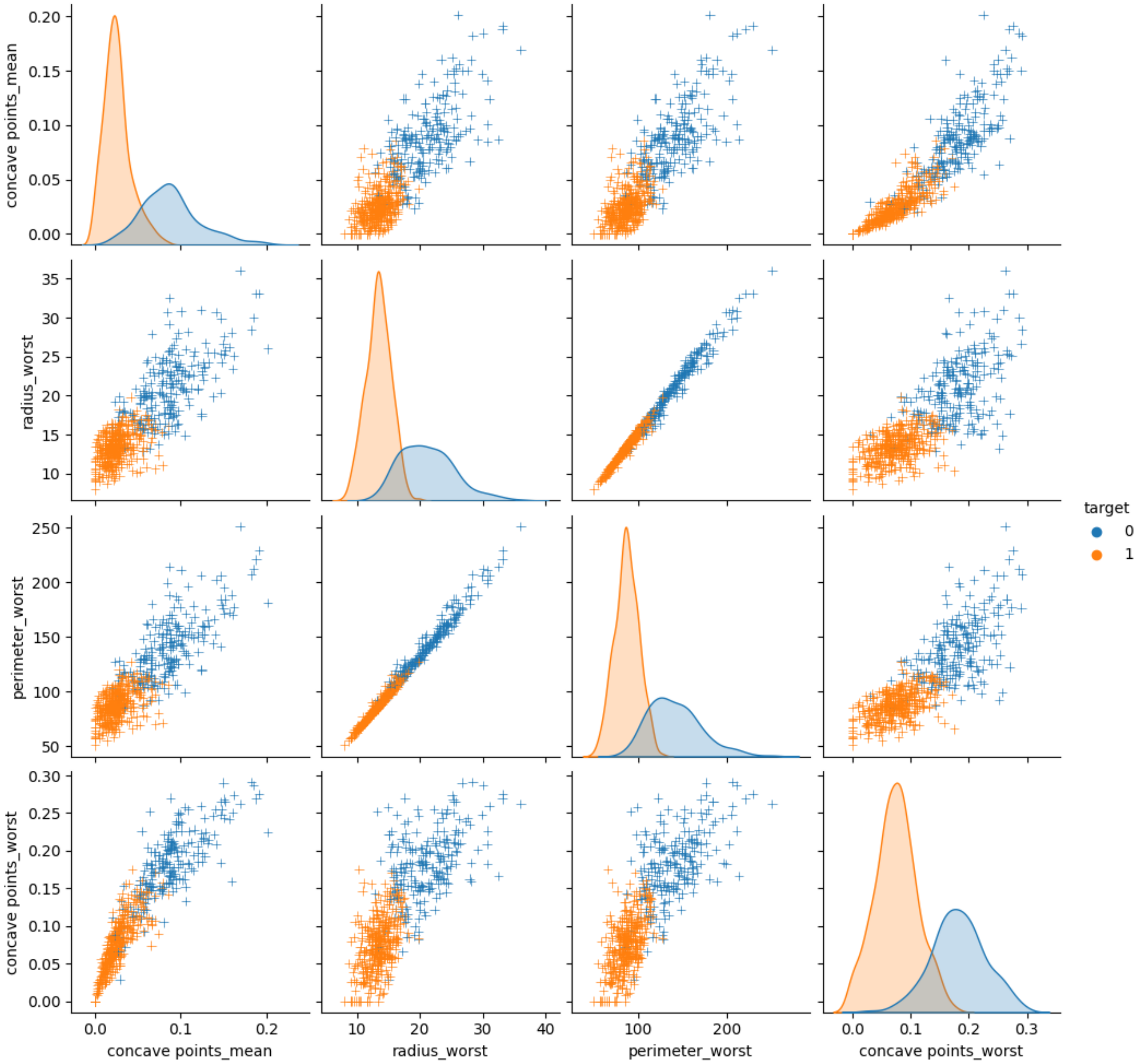

Figure 2: The pair plot of all variables.

DOI: https://doi.org/10.54216/IJNS.210115
Received: February 10, 2023 Revised: April 12, 2023 Accepted: May 11, 2023

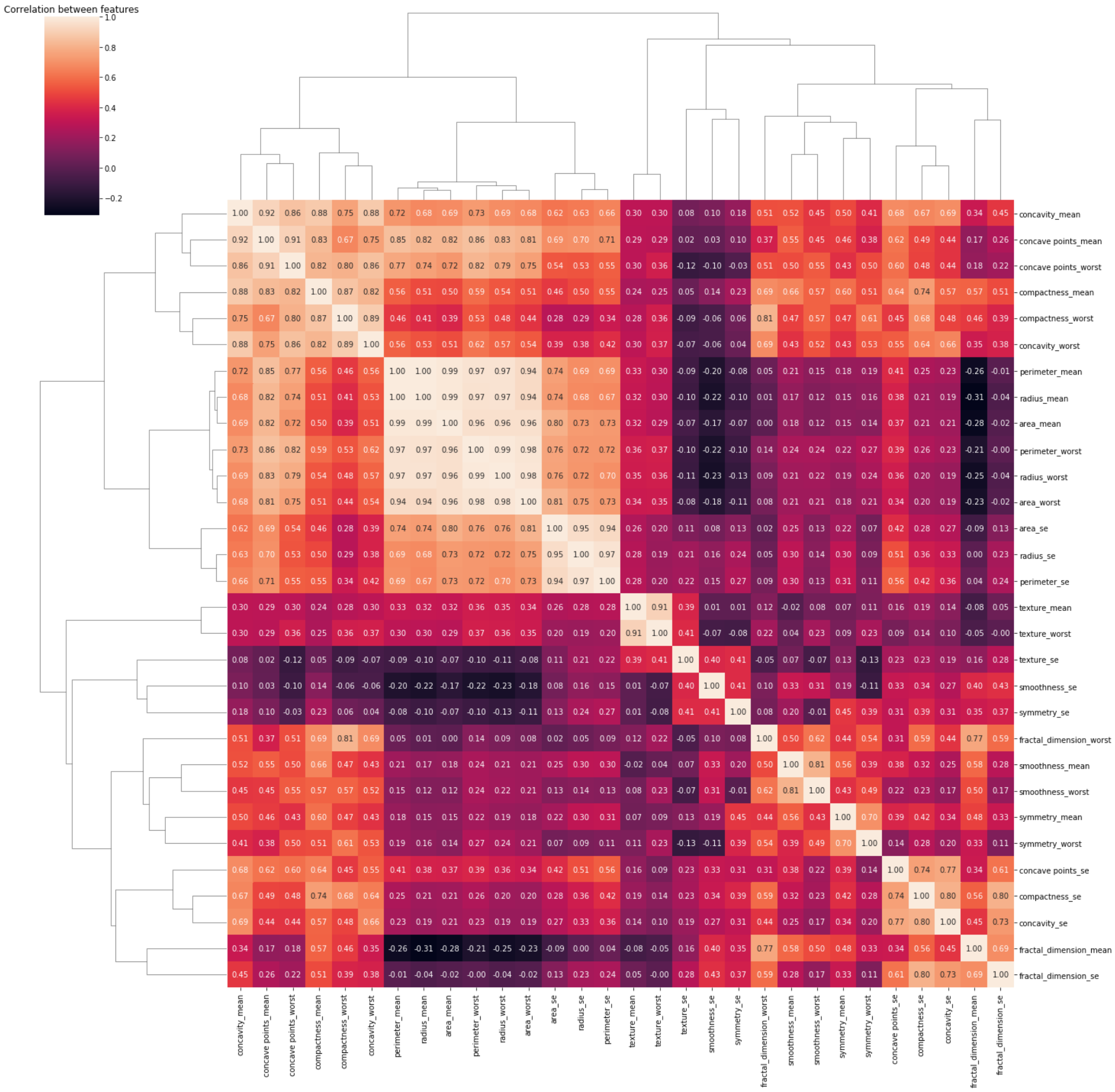

Figure 3: The heat map of all variables in the dataset before excluding some variables.

After applying the steps of the neutrosophic TOPSIS to exclude unrelated variables. The neutrosophic TOPSIS used to compute the weights of the variables in the dataset, the highest weight will be used and others will be excluded. Figure 4 shows the weights of all variables in the dataset. Figure 5 shows the heat map after applying the steps of the neutrosophic TOPSIS method.

DOI: https://doi.org/10.54216/IJNS.210115
Received: February 10, 2023 Revised: April 12, 2023 Accepted: May 11, 2023

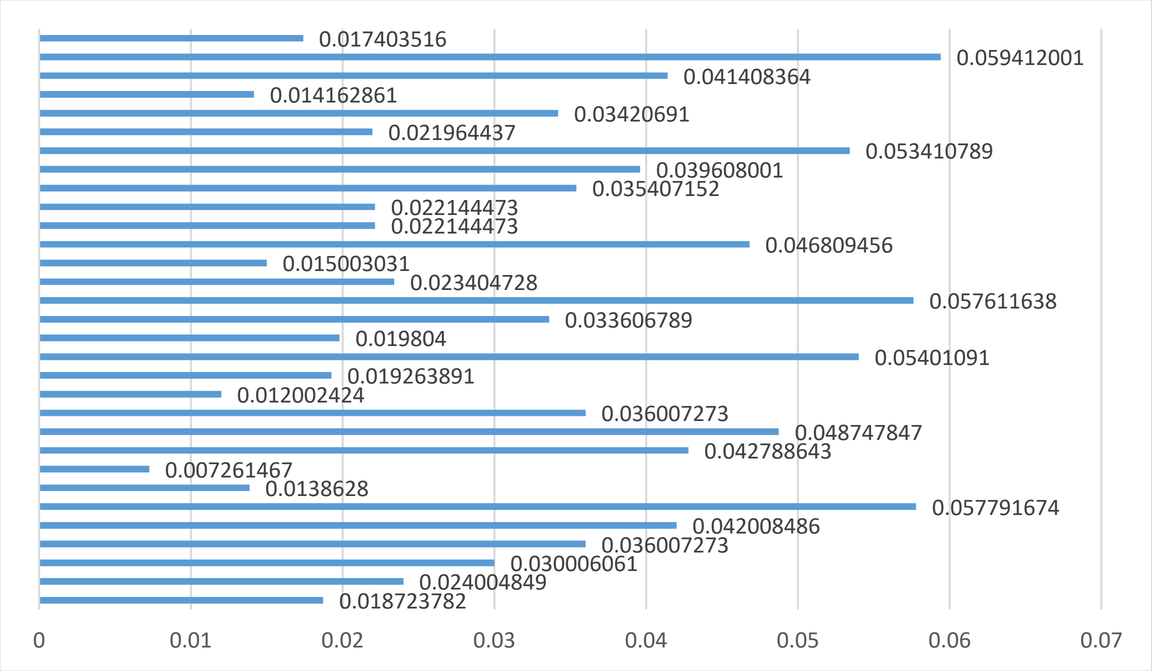

Figure 4: The analysis weights of variables in the dataset

DOI: https://doi.org/10.54216/IJNS.210115
Received: February 10, 2023 Revised: April 12, 2023 Accepted: May 11, 2023

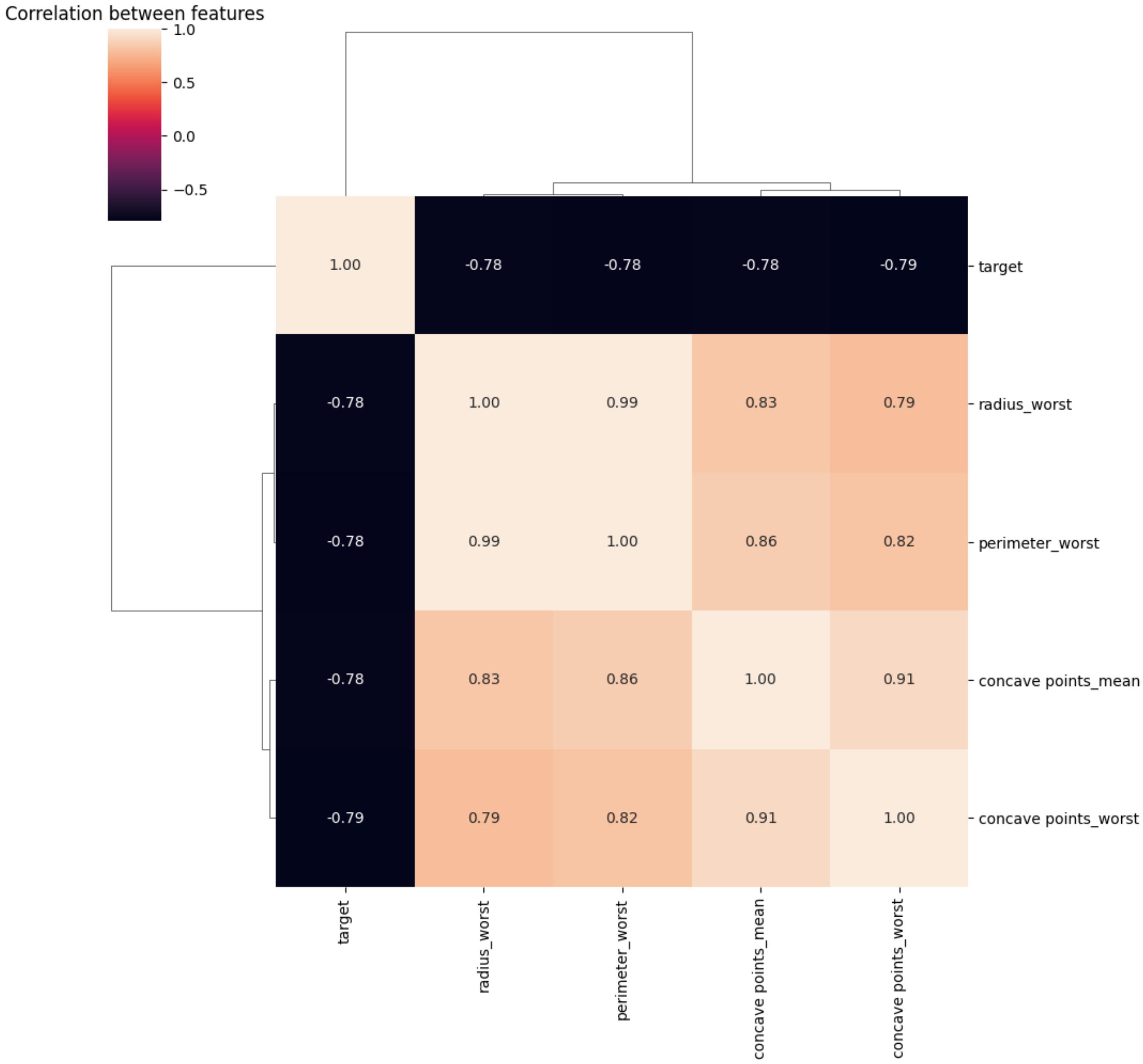

Figure 5: The heat map of all variables in the dataset after excluding some variables.

Then apply the machine learning methods to predict the BC. These methods work on the dataset after applying the neutrosophic TOPSIS method. This paper applied the KNN, Naïve Bayes, and ID3 methods. The accuracy of the three algorithms is shown in Figure 6. After obtaining the accuracy, the ROC is applied for three methods as shown in Figures 7-9.

DOI: https://doi.org/10.54216/IJNS.210115
Received: February 10, 2023 Revised: April 12, 2023 Accepted: May 11, 2023

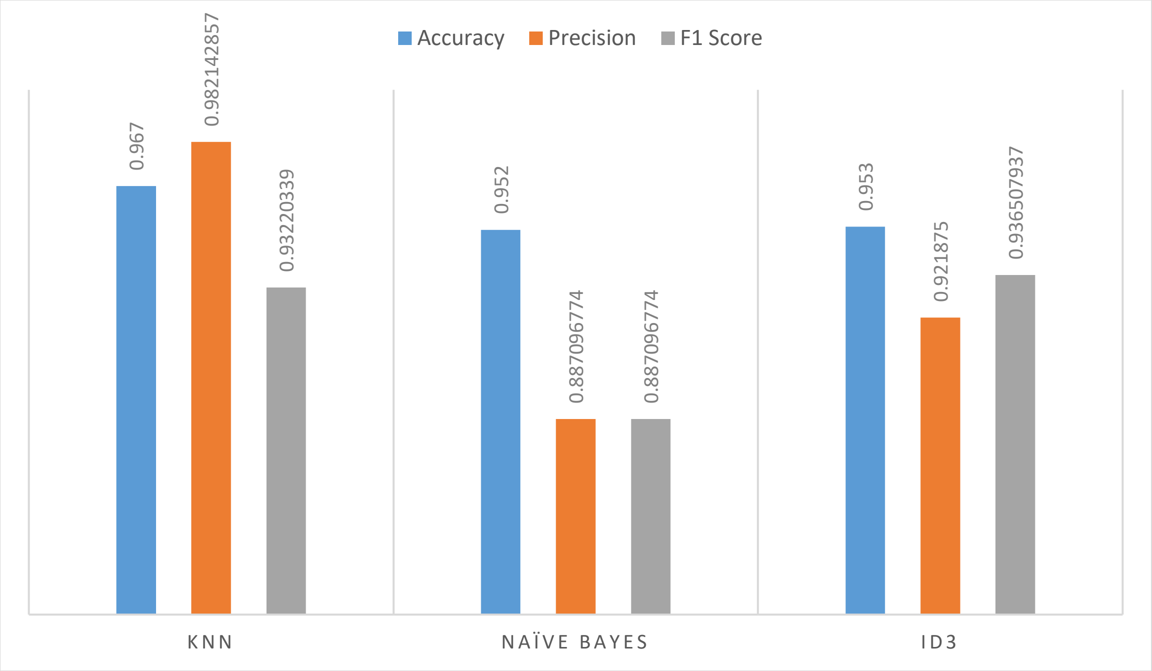

Figure 6: The comparison of three algorithms.

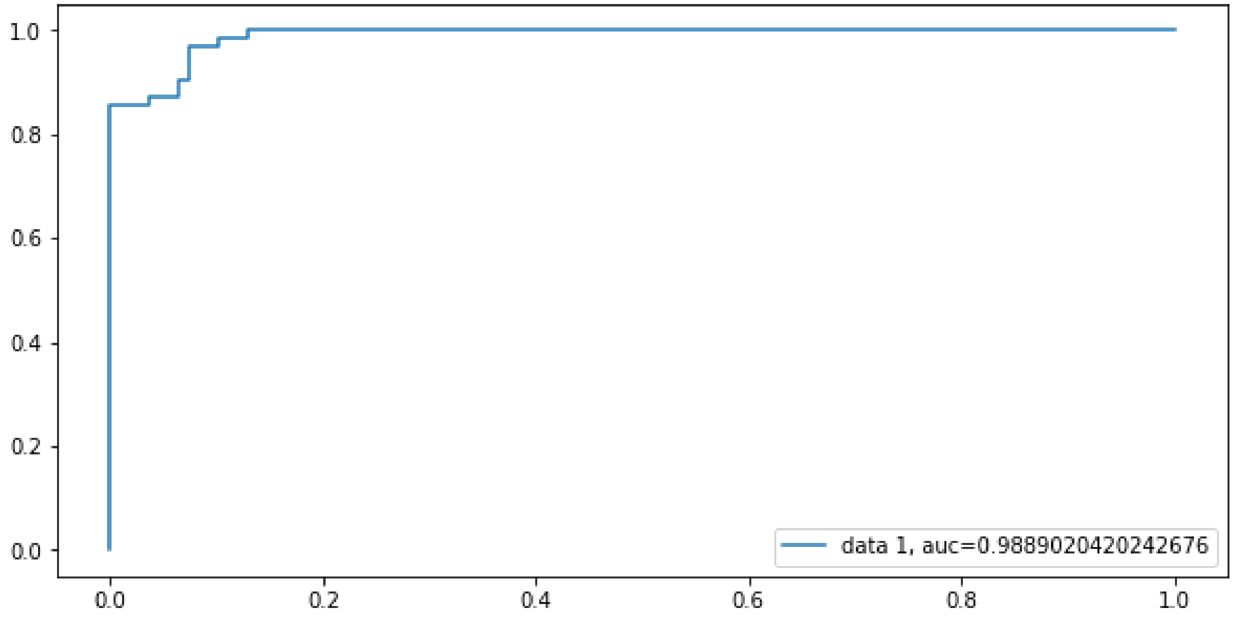

Figure 7: The ROC of the KNN algorithm.

DOI: https://doi.org/10.54216/IJNS.210115
Received: February 10, 2023 Revised: April 12, 2023 Accepted: May 11, 2023

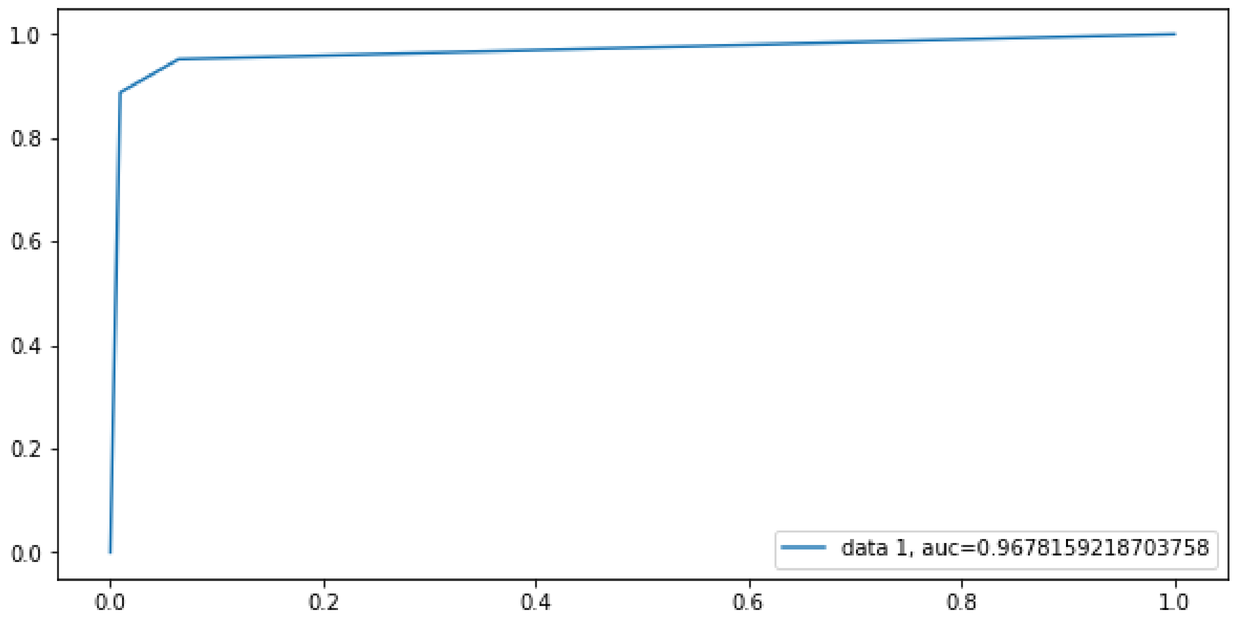

Figure 8: The ROC of the Naïve Bayes algorithm.

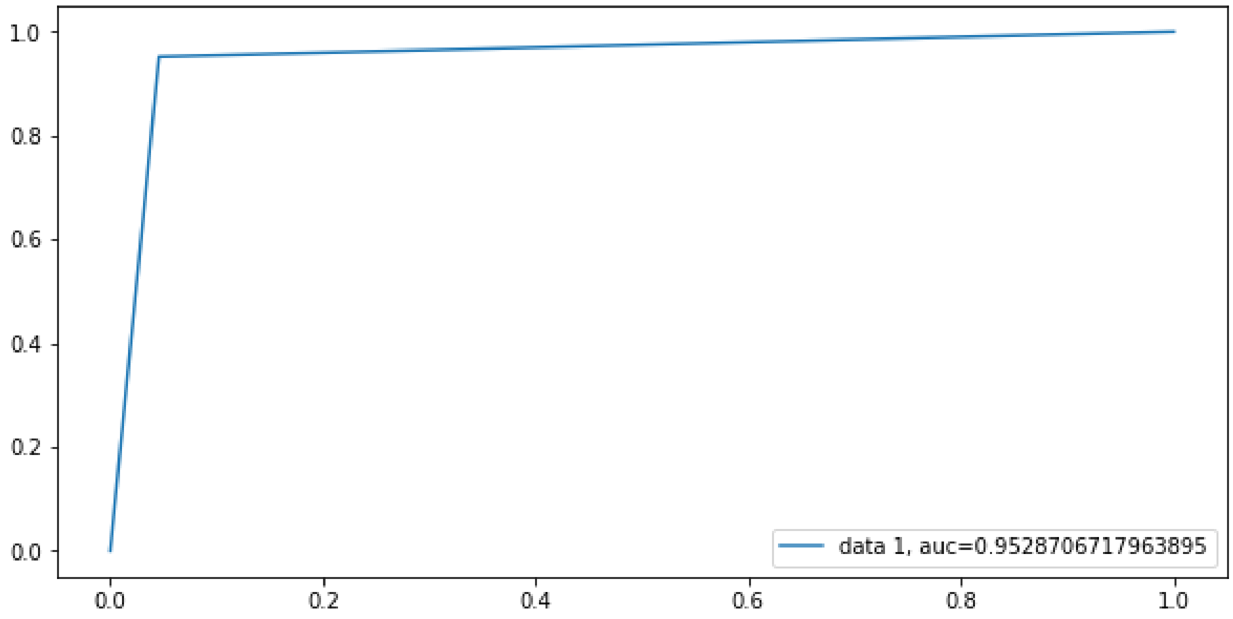

Figure 9: The ROC of the ID3 algorithm.

5. Conclusion

Among the leading killers of women, breast cancer is high. The key to saving women's lives from breast cancer is finding it early. Advanced machine learning techniques may aid in breast cancer screening. This paper used the WBC dataset. But this dataset has many features unreliable and contain vague information so, the neutrosophic TOPSIS method is used to solve this problem. After reducing the features of the dataset, this paper applied machine learning algorithms to predict breast cancer. This paper applied the KNN, Naïve Bayes, and ID3 methods. This paper introduced the accuracy, precision, and F1 score. The KNN algorithm has the highest machine learning accuracy

DOI: https://doi.org/10.54216/IJNS.210115
Received: February 10, 2023 Revised: April 12, 2023 Accepted: May 11, 2023

followed by the Naïve Bayes and then the ID3. This paper built a comparison between the accuracy, precision, and f1 score. Also, this paper introduced some descriptive statistics on the dataset like mean, standard deviation, minimum, and maximum values.

References

[1] W. Yue, Z. Wang, H. Chen, A. Payne, and X. Liu, “Machine learning with applications in breast cancer diagnosis and prognosis,” *Designs*, vol. 2, no. 2, p. 13, 2018.

[2] M. Amrane, S. Oukid, I. Gagaoua, and T. Ensari, “Breast cancer classification using machine learning,” in *2018 electric electronics, computer science, biomedical engineerings’ meeting (EBBT)*, IEEE, 2018, pp. 1–4.

[3] H. Asri, H. Mousannif, H. Al Moatassime, and T. Noel, “Using machine learning algorithms for breast cancer risk prediction and diagnosis,” *Procedia Comput. Sci.*, vol. 83, pp. 1064–1069, 2016.

[4] M. D. Ganggayah, N. A. Taib, Y. C. Har, P. Lio, and S. K. Dhillon, “Predicting factors for survival of breast cancer patients using machine learning techniques,” *BMC Med. Inform. Decis. Mak.*, vol. 19, pp. 1–17, 2019.

[5] D. A. Omondiagbe, S. Veeramani, and A. S. Sidhu, “Machine learning classification techniques for breast cancer diagnosis,” in *IOP Conference Series: Materials Science and Engineering*, IOP Publishing, 2019, p. 12033.

[6] E. H. Houssein, M. M. Emam, A. A. Ali, and P. N. Suganthan, “Deep and machine learning techniques for medical imaging-based breast cancer: A comprehensive review,” *Expert Syst. Appl.*, vol. 167, p. 114161, 2021.

[7] A. R. Vaka, B. Soni, and S. Reddy, “Breast cancer detection by leveraging Machine Learning,” *Ict Express*, vol. 6, no. 4, pp. 320–324, 2020.

[8] J. Wu and C. Hicks, “Breast cancer type classification using machine learning,” *J. Pers. Med.*, vol. 11, no. 2, p. 61, 2021.

[9] E. A. Bayrak, P. Kırcı, and T. Ensari, “Comparison of machine learning methods for breast cancer diagnosis,” in *2019 Scientific meeting on electrical-electronics & biomedical engineering and computer science (EBBT)*, IEEE, 2019, pp. 1–3.

[10] P. Biswas, S. Pramanik, and B. C. Giri, “Neutrosophic TOPSIS with group decision making,” *fuzzy multi-criteria Decis. using neutrosophic sets*, pp. 543–585, 2019.

[11] R. M. Zulqarnain, X. L. Xin, M. Saqlain, F. Smarandache, and M. I. Ahamad, “An integrated model of neutrosophic TOPSIS with application in multi-criteria decision-making problem,” *Neutrosophic Sets Syst.*, vol. 40, pp. 253–269, 2021.

[12] T. Gaber *et al.*, “Thermogram breast cancer prediction approach based on neutrosophic sets and fuzzy c-means algorithm,” in *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, Ieee, 2015, pp. 4254–4257.

[13] G. I. Sayed and A. E. Hassanien, “Moth-flame swarm optimization with neutrosophic sets for automatic mitosis detection in breast cancer histology images,” *Appl. Intell.*, vol. 47, pp. 397–408, 2017.

[14] K. M. Amin, A. I. Shahin, and Y. Guo, “A novel breast tumor classification algorithm using neutrosophic score features,” *Measurement*, vol. 81, pp. 210–220, 2016.

[15] M. Lotfollahi, M. Gity, J. Y. Ye, and A. Mahlooji Far, “Segmentation of breast ultrasound images based on active contours using neutrosophic theory,” *J. Med. Ultrason.*, vol. 45, pp. 205–212, 2018.

[16] Nada A. Nabeeh , Alshaimaa A. Tantawy, A Neutrosophic Model for Blockchain Platform Selection based on SWARA and WSM, Neutrosophic and Information Fusion, Vol. 1 , No. 2 , (2023) : 29-43 (Doi : https://doi.org/10.54216/NIF.010204)

[17] Mona Mohamed , Nissreen El Saber, Prioritization Thermochemical Materials based on Neutrosophic sets

DOI: https://doi.org/10.54216/IJNS.210115
Received: February 10, 2023 Revised: April 12, 2023 Accepted: May 11, 2023

Hybrid MULTIMOORA Ranker Method, Neutrosophic and Information Fusion, Vol. 2 , No. 1 , (2023) : 08-22 (Doi : https://doi.org/10.54216/NIF.020101)

[18] Abedallah abualkishik , Rasha Almajed , Watson Thompson, Improving the perfoamnce of Fog-assisted Internet of Things Networks using Bipolar Trapezoidal Neutrosophic sets, International Journal of Wireless and Ad Hoc Communication, Vol. 6 , No. 1 , (2023) : 30-37 (Doi : https://doi.org/10.54216/IJWAC.060103)

[19] H. Sharma, A. Tandon, P. K. Kapur, and A. G. Aggarwal, “Ranking hotels using aspect ratings based sentiment classification and interval-valued neutrosophic TOPSIS,” *Int. J. Syst. Assur. Eng. Manag.*, vol. 10, pp. 973–983, 2019.

[20] Ahmed Abdelhafeez , Hoda K Mohamed, Enhance the Performance of Bus Rapid Transit (BRT) through the Evaluation of Alternatives under an Integrated MCDM Neutrosophic Environment, Neutrosophic and Information Fusion, Vol. 1 , No. 2 , (2023) : 08-15 (Doi : https://doi.org/10.54216/NIF.010201)

[21] Abdullah Ali Salamai, Evaluation and Selection of Cloud Service: A neutrosophic model, Neutrosophic and Information Fusion, Vol. 1 , No. 2 , (2023) : 16-25 (Doi : https://doi.org/10.54216/NIF.010202)

[22] Ahmed Abdelhafeez , Hoda K. Mohamed, Skin Cancer Detection using Neutrosophic c-means and Fuzzy c-means Clustering Algorithms, Journal of Intelligent Systems and Internet of Things, Vol. 8 , No. 1 , (2023) : 33-42 (Doi : https://doi.org/10.54216/JISIoT.080103)

[23] Gopal Chaudhary , Manju Khari , Amena Mahmoud, Intelligent Video Moving Target Detection Based on Multi-Attribute Single Value Medium Neutrosophic Method, Journal of Intelligent Systems and Internet of Things, Vol. 5 , No. 1 , (2021) : 49-59 (Doi : https://doi.org/10.54216/JISIoT.050105)

[24] S. Chen, G. I. Webb, L. Liu, and X. Ma, “A novel selective naïve Bayes algorithm,” *Knowledge-Based Syst.*, vol. 192, p. 105361, 2020.

[25] M. El Kourdi, A. Bensaid, and T. Rachidi, “Automatic Arabic document categorization based on the Naïve Bayes algorithm,” in *proceedings of the Workshop on Computational Approaches to Arabic Script-based Languages*, 2004, pp. 51–58.

[26] R. Bhardwaj and S. Vatta, “Implementation of ID3 algorithm,” *Int. J. Adv. Res. Comput. Sci. Softw. Eng.*, vol. 3, no. 6, 2013.

[27] M. Slocum, “Decision making using id3 algorithm,” *Insight River Acad. J*, vol. 8, no. 2, 2012.

[28] S. Zhang, D. Cheng, Z. Deng, M. Zong, and X. Deng, “A novel kNN algorithm with data-driven k parameter computation,” *Pattern Recognit. Lett.*, vol. 109, pp. 44–54, 2018.

[29] A. Moldagulova and R. B. Sulaiman, “Using KNN algorithm for classification of textual documents,” in *2017 8th international conference on information technology (ICIT)*, IEEE, 2017, pp. 665–671.

[30] Gómez, Gustavo Adolfo Álvarez, Maikel Yelandi Leyva Vázquez, Jesús Estupiñán Ricardo. "Application of Neutrosophy to the Analysis of Open Government, its Implementation and Contribution to the Ecuadorian Judicial System." Neutrosophic Sets and Systems vol 52, pp.215-224., 2022.

DOI: https://doi.org/10.54216/IJNS.210115
Received: February 10, 2023 Revised: April 12, 2023 Accepted: May 11, 2023