NSS
Neutrosophic Sets and Systems, Vol. 73, 2024
University of New Mexico

Enhanced Neutrosophic Set and Machine Learning Approach for Breast Cancer Prediction

Ashika T[1], Hannah Grace[2],* Nivetha Martin[3], and Florentin Smarandache[4]

[1]Vellore Institute of Technology, Department of Mathematics, School of Advanced Sciences, Chennai, India.
E-mail: ashika.t2023@vitstudent.ac.in
[2]Vellore Institute of Technology, Department of Mathematics, School of Advanced Sciences, Chennai, India.
E-mail: hannahgrace.g@vit.ac.in
[3]Arul Anandar College (Autonomous), Karumathur, Madurai, India. E-mail: nivetha.martin710@gmail.com
[4] Emeritus Professor, University of New Mexico, Gallup, United States. E-mail: fsmarandache@gmail.com
* Correspondence: hannahgrace.g@vit.ac.in

Abstract: Breast cancer is the most prevalent type of cancer that affects women worldwide and poses a serious risk to female mortality. In order to lower death rates and enhance treatment results, early detection is critical. Neutrosophic Set Theory (NST) and machine learning (ML) approaches are integrated in this study to provide a novel hybrid methodology (NS-ML) that improves breast cancer diagnosis. Using the Wisconsin Diagnostic Breast Cancer (WDBC) dataset, the research transforms these data into Neutrosophic (N) representations to effectively capture uncertainties. When trained on the N-dataset instead of traditional datasets, ML algorithms such as Decision Tree (DT), Random Forest (RF), and Adaptive Boosting (AdaBoost) perform better. Notably, N-AdaBoost models achieve outstanding results with 99.12% accuracy and 100% precision, highlighting the efficacy of NS in enhancing diagnostic reliability.

Keywords: Neutrosophic sets; Machine Learning; Uncertainity handling; Breast cancer; Classification.

1. Introduction

Women globally have the risk of breast cancer which as a risk factor of cancer due to abnormal growth of cells in breast tissue [1]. Countries like Belgium and the Netherlands reported high rates of incidence, affecting thousands of women, while Barbados and Fiji had notable mortality rates due to the disease [2]. In 2020, it caused over 2 million new cases and led to 6,85,000 deaths globally [3]. By 2030, global cases are expected to rise significantly, impacting approximately 27 million people [4].

Breast cancer involves various tumor types, categorized as malignant or benign, with malignant tumors posing a higher risk due to their rapid spread. Factors contributing to its increasing incidence include lack of awareness, economic disparities, inadequate healthcare access, and screening challenges [5]. Developing independent apps for accurate cancer diagnosis is crucial to overcome these challenges. By analyzing variables and pinpointing the most important elements for a precise diagnosis across several models, ML has shown to be extremely successful in the diagnosis of breast cancer. Several models have been presented in earlier research, using various methods and techniques to identify cancers. ML, particularly integrated with NS to handle data uncertainty, presents an effective approach for enhancing breast cancer models for forecasting.

The main contributions of this study are:

a) Transforming the WDBC into an N-dataset.

b) Model training employing ML classification techniques, such as AdaBoost, DT, and RF.

c) Evaluating the performance of these models.

d) Comparing the performance between the original dataset and the N-dataset.

The remaining sections of the paper are organized as follows: The section 2 summarizes the literature on breast cancer prognosis. The study's materials and procedures are described in depth in Section 3. The study's findings are presented in Section 4. The methodology and results are covered in Section 5, and the paper's conclusions are given in Section 6.

2. Related work

Breast cancer is a serious worldwide health issue, and increasing survival rates requires early identification. Advanced ML techniques, such as Eagle Strategy Optimization (ESO), Gravitational Search Optimization (GSO), and their combined approach, are used to improve classification accuracy on the WDBC dataset. By prioritizing informative features and reducing computational complexity, the approach shows promising results in improving diagnostic precision and efficiency [6]. Breast cancer, characterized by complex development involving various cell types, remains a significant challenge worldwide. Advances in understanding pathogenesis and genetic factors have led to improved prevention and treatment strategies. Effective screening and research into drug-resistant mechanisms have enhanced patient outcomes and quality of life [7]. Over recent decades, significant advancements in breast cancer research have revolutionized treatment approaches, leading to better outcomes. Early detection through improved awareness and screening methods has enabled curative treatments such as surgery and radiation therapy. Ongoing research aims to further enhance diagnostic and therapeutic strategies [8].

Breast cancer is a multifactorial illness with a wide range of subtypes and symptoms. Understanding this diversity is crucial for developing targeted treatment approaches. Research focuses on genetic mutations, micro environmental factors, and epigenetic changes to improve personalized treatment strategies [9]. Technological advancements in mammographic screening and therapeutic interventions have transformed breast cancer management. Innovations in surgical techniques and radiotherapy have improved disease control and cosmetic outcomes. Clinical trials exploring combination therapies and gene-expression profiling aim to enhance treatment selection and patient outcomes [10]. Developing accurate prediction models tailored to specific populations, such as Cuban women, is crucial for effective breast cancer management. ML-based model that achieves high accuracy in estimating breast cancer risk for Cuban women, outperforms existing models. The potential for early diagnosis to enhance patient outcomes and save healthcare expenditures is highlighted by this approach [11]. To select and classify features in multidimensional breast cancer datasets, a new approach called the Rat Swarm Optimizer (RSO) hybridization with Levy Flight-based Cuckoo Search Optimization Algorithm (H-RS-LVCSO) was presented. By merging hybrid adaptive LVCSO with moment invariant wavelet feature extraction, this method greatly improves accuracy, precision, and execution speed. The research contributes to advancing breast cancer classification through innovative Feature Selection (FS) and classification techniques [12].

Evaluating various ML algorithms for breast cancer diagnosis, highlights RF's robustness in handling high-dimensional data and nonlinear decision boundaries. The approach demonstrates high accuracy in distinguishing between healthy individuals and those with breast cancer, showcasing its potential for accurate early detection [13]. The application of ensemble data mining techniques enhances the precision of breast cancer diagnosis by combining Rotation Forest with feature selection based on genetic algorithms. The approach optimizes input variable selection and employs robust classification methods, achieving high accuracy rates and demonstrating the effectiveness of ensemble methods in medical diagnostics [14]. Particle swarm optimization (PSO) was used for FS in data mining techniques to create a predictive model for breast cancer recurrence. The study demonstrates how Particle Swarm Optimization (PSO) enhances classification performance by assessing classifiers such as Naive Bayes (NB), K-Nearest Neighbor (KNN), and the rapid Decision Tree learner (REPTree). The results highlight the effectiveness of feature selection (FS) techniques in optimizing predictive models for breast cancer recurrence [15].

Cardiotocography (CTG) data uncertainty is crucial for classifying fetal heart rate in the biomedical field. The proposed Interval Neutrosophic Rough Neural Network (IN-RNN) framework, utilizing the backpropagation algorithm, enhances RNN's performance through NST. The experimental results indicate exceptional performance, with scores around 95%. Using WEKA application, the framework was compared with algorithms like Neural Network (NN), decision tables, and nearest neighbors, confirming its efficiency. The Receiver Operating Characteristic (ROC) curve displays high and acceptable area-under-curve values for the pathologic, normal, and suspicious states. The IN-RNN framework estimates uncertainty boundaries based on membership, truth, and indeterminacy values, with performance metrics indicating its effectiveness in classifying CTG data [16]. Differentiating COVID-19 from other lung illnesses, like bacterial and viral pneumonia, has become more difficult due to the COVID-19 pandemic. To differentiate between these diseases, a neutrosophic method was put forth, which involved grouping data into sets labeled True (T), False (F), and Indeterminacy (I) to improve feature extraction. Alpha-mean and beta-enhancement preprocessing is applied to chest X-ray pictures in order to decrease indeterminacy and boost opacity detection. Then, in a transfer learning setup, these improved images are examined using ResNet-50, VGG-16, and XGBoost, yielding an accuracy of 97.33% [17].

Decision-making (DM) is naturally challenging because of the uncertain and ambiguous nature of environments, particularly when multiple attributes are considered. The Multi-Polar Interval-Valued Neutrosophic Set (MPIVNS) and the Hypersoft Set (HS) framework were combined to address these issues. New aggregate operators, distance metrics, and similarity measures created especially for MPIVN-HSs are presented. These tools are essential for resolving complex attribute-based decision-making problems. The research utilizes the KNN algorithm to improve decision processes, showcasing practical applications in areas like site selection and beyond. The study significantly advances fields relying on language-based DM, including Artificial Intelligence (AI) and sentiment analysis [18]. The KNN algorithm is a popular non-parametric supervised classifier that assigns class labels to unknown samples based on their nearest neighbors in a training set using distance metrics. While effective, efforts have extended KNN to enhance its accuracy. Neutrosophic KNN, which integrates NST to improve classification. NST computes a final membership $U = T + I - F$ for class labeling, similar to fuzzy KNN, and assigns T, I, and F memberships using a supervised Neutrosophic C-Means (NCM) algorithm. Extensive experiments on synthetic and real-world datasets validate the method's efficacy compared to traditional KNN, fuzzy KNN, and weighted KNN approaches [19].

NST, especially single-valued NS (SVNSs), improves handling of imprecision and uncertainty in medical applications. By integrating NST with fuzzy techniques, more effective solutions for medical image processing, DM, and information fusion are achieved. These methods have shown efficacy in de-noising, clustering, and segmenting medical images. Neutrosophic logic (NL) offers a framework for modeling vagueness and uncertainty, making it ideal for dealing with incomplete or inconsistent information. The importance of NS is emphasized in various medical applications and proposes a framework for leveraging NS to enhance medical image processing and diagnosis [20].

Developing decision support tools for healthcare facility maintenance and asset renewal is challenging due to uncertainties and subjectivity in DM. In order to reduce subjectivity, Neutrosophic Logic (NL), Multi-Attribute Utility Theory (MAUT), and the Analytic Network Process (ANP) were integrated to assess hospital building assets according to their criticality and performance inequalities. ML algorithms, such as DT, KNN, and NB, automate and standardize the prioritization process. Applying the model to healthcare institutions in Canada showed a notable improvement in prediction performance, outperforming the previous model by about 11%. With the help of this framework, hospital asset renewal will be prioritized in a way that is impartial, automated, and consistent, guaranteeing effective resource allocation [21].

A hybrid fuzzy Multi-Criteria Decision-Making (MCDM) methodology utilizing Single-Valued Neutrosophic Fuzzy Sets (SVNFS), Best-Worst Method (BWM), and VIKOR is proposed for assessing cybersecurity risks targeting Connected and Autonomous Vehicles (CAVs). Expert opinions on cyber-attack likelihood and severity are integrated to rank threat-agent categories, identifying insider attackers as posing the greatest risk. This approach addresses subjectivity in opinions and incorporates criteria weights based on the consequences of cyber-attacks, offering a flexible framework applicable beyond CAV cybersecurity to other complex decision contexts with uncertain data [22]. The decision support system (DSS) utilizes the CRITIC and CRADIS models within SVNS to prioritize hydrogen technologies for decarbonizing Iran's oil refining industry. It assesses blue, green, and low-carbon hydrogen technologies across environmental, economic, social, and reliability criteria, identifying solar renewable energy as optimal due to its clean energy conversion and geographical suitability. This study enhances DM under uncertainty, suggesting future research explore broader qualitative factors and stakeholder perspectives [23]. The research focuses on advancing strategic DM in the planning of historic pedestrian bridge remediation through an innovative algorithm based on Rough NS (RNS). This novel approach integrates Rough Sets (RS) and NS theories within a MCDM model. A key contribution is the introduction of a new RN symmetric cross entropy measure and its weighted variant, specifically designed to address uncertainties and the challenge of unknown criteria weights inherent in complex DM processes. By incorporating the VIKOR method, the model enables effective prioritization of bridge remediation efforts by providing robust and reliable rankings. Case studies validate the model's efficacy, demonstrating its practical utility compared to traditional methods in real-world scenarios [24].

A novel approach to Multi-Attribute Decision-making (MADM) was introduced by integrating RS, NS, and Grey System Theory (GST). The RN Grey Relational Analysis (RNGRA) method addresses indeterminate and inconsistent data using RNS, characterized by T, I, and F-membership degrees. Attribute weights are partially known and determined via an information entropy method. The Accumulated Geometric Operator (AGO) converts RN numbers into SVN numbers. The method employs the Hamming distance to calculate the NGR coefficient for assessing reliability and unreliability. Finally, a RN relational degree is established to rank alternatives, with a numerical example provided to demonstrate the method's effectiveness and applicability [25].

A RN TOPSIS method was presented for Multi-Attribute Group DM (MAGDM), effectively handling uncertainty, indeterminacy, and inconsistency in data. By evaluating alternatives and features using RNS, which are distinguished by T, I, and F-membership degrees, the method enhances the conventional TOPSIS technique. Individual opinions are aggregated into a group consensus using the RN weighted averaging operator. The distance between each alternative and the positive and negative Rough Neutrosophic (RN) ideal solutions is estimated using the Euclidean distance. A numerical example demonstrates the method's practicality and efficiency, making it applicable in pattern recognition, AI, and medical diagnosis [26].

3. Materials and Methods

The materials and methods used in the study are described in detail in this section.

3.1. Proposed methodology

The objective is to advance breast cancer prediction by integrating NS with ML algorithms. The WDBC dataset (I_D) was initially retrieved from the UCI ML Repository and then carefully preprocessed (O_{Dpp}) to ensure its quality and consistency.

$$I_D = WDBC \quad (1)$$

$$O_{Dpp} = f_{pp}(I_D) \quad (2)$$

The dataset was then transformed into an N-representation (O_N), where each data point was characterized not only by its specific attributes but also by degrees of T, I, and F. This approach offers a more nuanced depiction of uncertainty and variability inherent in medical datasets.

$$O_N = f_{T,I,F}(O_{Dpp}) \quad (3)$$

Following the transformation, the N-dataset was split ($O_{N_{(s)}}$) into training and testing subsets.

$$O_{N_{(s)}} = s(f_{O_N}) = s\ (X_{train}, X_{test}, y_{train}, y_{test}) \quad (4)$$

The N- dataset was normalized ($O_{N_{(Nor)}}$) to a range of 0 to 1 using Min-Max Scaler.

$$O_{N_{(Nor)}} = f_{Nor}\ (O_{N_{(s)}}) \quad (5)$$

The normalized N - training dataset was employed to train ML classifiers ($O_{N_{(ML)}}$) such as DT, RF, and AdaBoost. These classifiers were selected based on their capacity to handle intricate feature interactions and identify subtle patterns necessary for an accurate diagnosis of breast cancer.

$$O_{N_{(ML)}} = f_{ML}\ (O_{N_{(Nor)}}) = O_{ML}\ (DT_{O_{N_{(Nor)}}}, RF_{O_{N_{(Nor)}}}, AB_{O_{N_{(Nor)}}}) \quad (6)$$

The main performance metrics ($O_{N_{Metrics}}$) for the classifier, which include accuracy, precision, recall, and F1 score, were used to evaluate its performance and provide a thorough examination of its ability to predict.

$$\boldsymbol{O_{N_{Metrics}} = O_{N_{(ML)}}(Metrics_{A_{cc},P_r,R_c,F_1})} \quad (7)$$

Finally, the study conducted a Comparative Analysis (O_{CA}) between the N-dataset and the original dataset to evaluate how integrating NS enhances the accuracy and reliability of breast cancer prediction models.

$$O_{CA} = CA\,(I_D;\, O_N) \quad (8)$$

The workflow is depicted in Figure 1.

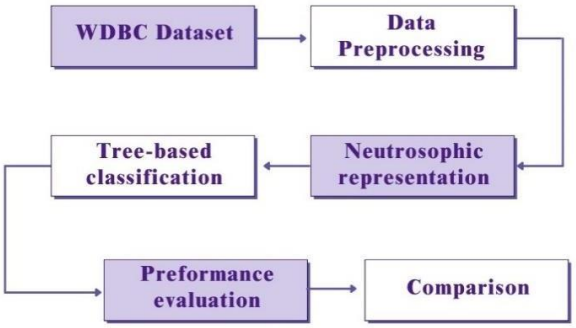

Fig. 1: NS-ML framework for Breast Cancer Prediction

3.2. Dataset description

A popular dataset for ML and statistical analysis, the WDBC dataset was collected by the University of Wisconsin Hospitals in Madison and is particularly useful for predicting and classifying breast cancer. Derived from digital photographs of breast tumors, it has attributes including radius, texture, area, perimeter, smoothness, compactness, concavity, concave spots, symmetry, and fractal dimension. With 212 instances classified as malignant (indicating presence of cancer) and 357 as benign (indicating absence of cancer), this dataset serves as a robust resource for developing accurate models that distinguish between malignant and benign breast cancer cases based on these comprehensive tumor characteristics.

3.3. Preprocessing of data

The breast cancer dataset was preprocessed for effective ML analysis. First, it was divided into features attributes describing tumor and a categorical target variable that distinguishes between benign and malignant cases. This categorical target was encoded into numerical values to make it easier for the models to understand. All features were normalized using Min-Max scaling in order to guarantee accurate predictions. This preprocessing process improved the machine learning models' ability to predict breast cancer diagnosis and maintained data consistency.

3.4. Neutrosophic Sets

Let X be a point or object-containing space, and let x be any element belongs to X. Three unique membership functions define a NS, A within X: T, I and F membership often referred as $T_A(X)$, $I_A(X)$, and $F_A(X)$. These functions assign real values within the interval [0, 1] indicating the degree to which x pertains to the subsets of T, I, and F, respectively:

$$T_A\, maps\, X\, to\, the\, interval\,]\,0^-, 1^+[$$
$$I_A\, maps\, X\, to\, the\, interval\,]\,0^-, 1^+[$$
$$F_A\, maps\, X\, to\, the\, interval\,]\,0^-, 1^+[$$

The sum of $T_A(X)$, $I_A(X)$, $F_A(X)$ for every X, the falls ranges from 0 to 3. This flexibility enables NS to effectively represent and manage uncertainty, ambiguity, and contradiction within sets,

making them valuable in contexts where handling incomplete or uncertain information is crucial [26].

3.5. Neutrosophic dataset formation

To address the uncertainties inherent in the WDBC dataset for binary classification, An N- dataset was introduced as an inclusive and generalized solution. This dataset goes beyond conventional and high-risk categories by incorporating a degree of neutrality. Thus, the N-dataset is defined as $< T_A, I_A, F_A >$, where each element of the set $X = \{x_1, x_2, \dots, x_n\}$ is specified as follows:

$$\forall x\,(t, i, f) \in < T_A, I_A, F_A >$$

where, *t, i, and f*, respectively, are the real numbers for T, I, and F.

In order to better capture the uncertainty in the data, the model is being developed by adding N-components to the original dataset. The first step in the method is to compute the mean vectors for the training set as a whole (ρ^{all}), the positive class (ρ^{+}), and the negative class (ρ^{-}) which is provided in Eq. (1).

$$\rho^{all} = \sum_{k=1}^{n^{all}} x \;;\; \rho^{+} = \sum_{k=1}^{n^{+}} x \;;\; \rho^{-} = \sum_{k=1}^{n^{-}} x \tag{1}$$

The following Eqs. (2), (3) & (4) are u sed to compute the T, I, and F components for a given sample (x):

$$T = 1 - \frac{||x - \rho^{+}||}{\max(||X_{train} - \rho^{+})} \tag{2}$$

$$I = 1 - \frac{||x - \rho^{all}||}{\max(||X_{train} - \rho^{all})} \tag{3}$$

$$F = 1 - \frac{||x - \rho^{-}||}{\max(||X_{train} - \rho^{-})} \tag{4}$$

These formulas are applied to each sample in the training and testing datasets, yielding features that measure the degrees of T, I, and F. Consequently, an N-dataset is produced, which enhances the ML algorithm's capacity to identify data with inherent uncertainty. This strategy considerably improves the classifier's performance in processing ambiguous and complex biomedical data by utilizing the capabilities of NST. The algorithm for the formation of O_N is provided below.

INPUT: The WDBC dataset (I_D)
OUTPUT: Neutrosophic dataset (O_N)

Step 1: Mean Vector Computation
Step 1.1: Positive class (ρ^{+})

$$\rho^{+} = \sum_{k=1}^{n^{+}} x_k$$

where $y_k = 1$ and n^{+} is the no. of positive samples in X_{train}
Step 1.2: Negative class (ρ^{-})

$$\rho^{-} = \sum_{k=1}^{n^{-}} x_k$$

where $y_k = 0$ and n^{-} is the no. of negative samples in X_{train}
Step 1.3: Overall mean (ρ^{+})

$$\rho^{+} = \sum_{k=1}^{n^{+}} x_k$$

where $y_k = 1$ and n^+ is the no. of positive samples in X_{train}

Step 2: Calculate the N-components (t, i, f) for each sample.

For each sample x_i in both X_{train} and X_{test}

$$T_i = 1 - \frac{||x_i - \rho^+||}{\max(||X_{train} - \rho^+)}$$

$$I_i = 1 - \frac{||x_i - \rho^{all}||}{\max(||X_{train} - \rho^{all})}$$

$$F_i = 1 - \frac{||x_i - \rho^-||}{\max(||X_{train} - \rho^-)}$$

Step 3: Group the N-components

$$f_{T_i, I_i, F_i} = O_N$$

3.6. Classification algorithms

3.6.1. Decision Tree

A hierarchical supervised learning model called a DT makes predictions by gradually dividing the data into groups based on the values of its features. It creates a tree-like structure with internal nodes representing decision points, branches indicating possible outcomes, and leaf nodes delivering final predictions. The tree is built recursively, starting from the root and progressing downwards. At each internal node, the algorithm chooses a feature and threshold that best separates the data, typically using metrics like Gini impurity or entropy. This process continues, refining predictions at each level, until reaching leaf nodes. In order to classify new data, branches are followed depending on feature values as they go from the root of the tree to the leaf. This approach effectively breaks complex decisions into simpler steps, making DT both powerful and interpretable for various prediction and classification tasks [28]. The Gini Impurity and Entropy is calculated using the following Eqs. (5) and (6)

Gini Impurity

$$G = 1 - \sum_{j=1}^{C} {p_j}^2 \tag{5}$$

Entropy:

$$H = -\sum^{C} p_j \log(p_j) \tag{6}$$

where p_j represents the probability of class j in the node, and C is the number of classes

3.6.2. Random Forest

An ensemble learning method called RF combines several DTs to reduce overfitting and improve prediction accuracy. Using a bootstrap sampling of the original dataset, it creates a large number of trees. In order to reduce correlations between the different trees and introduce variety through feature bagging, a random selection of features is chosen at each node for splitting. For classification tasks, the model aggregates predictions by majority voting, while for regression, it uses averaging. By leveraging the collective wisdom of many diverse trees, RF effectively reduces variance and enhances generalization. This method excels in handling complex, high-dimensional datasets and is

widely adopted in ML for its robust performance, ability to capture non-linear relationships, and resilience against overfitting. Additionally, the model's capacity to provide feature importance rankings and handle missing values further contributes to its popularity across various domains [29].

3.6.3. Adaptive Boosting

Boosting is a ML technique that combines multiple weak learners to form a strong predictive model. AdaBoost, developed by Freund and Schapire [30], exemplifies this approach and remains widely used in various fields. In AdaBoost, weak learners are trained iteratively on weighted distributions of training data, with weights adjusted based on their performance. Each weak hypothesis receives a weight αt proportional to its accuracy, thereby minimizing errors. By assigning greater weight to more accurate learners, the final model aggregates these weak learners into a robust overall predictor.

3.7. Performance Evaluation

The proposed methods were compared using metrics including recall, accuracy, precision, and F1-score. The percentage of correctly identified subjects is called accuracy. The precision measure shows the percentage of successfully diagnosed positive subjects out of all predicted positive subjects. The recall metric assesses how well the model can detect positive samples. A fair assessment of the model's performance is provided by the F1-score, which is the harmonic mean of precision and recall.

4. Results

4.1. Experimental setup

Three distinct ML tree-based classifiers were utilized to predict breast cancer using the WDBC dataset. Prior to training, the dataset was transformed into an N-dataset to enhance the models' robustness in handling uncertainties. Several tree-based algorithms were trained on this N-dataset, and standard metrics were employed to evaluate the predictive performance of the methods. Furthermore, comparisons were made between results obtained from the N-dataset and the original dataset. All these experiments were efficiently executed on Google Colab, leveraging GPU acceleration to manage the computational complexity of ML tasks effectively.

4.2. Experimental results

The table provides a comparative analysis of ML algorithm performance using both the N-dataset and the original dataset. N-DT, N-RF, and N-AdaBoost denote models trained with the N-dataset using DT, RF, and AdaBoost algorithms respectively.

Table 2. Comparative evaluation of N-DT and DT model

Metrics	**N-DT**	**DT**
Accuracy	**93.86**	90.35
Precision	**91.66**	84.61
Recall	**93.62**	93.62
F1 score	**92.63**	88.88

Table 3. Comparative evaluation of N-RF and RF model

Metrics	N-RF	RF
Accuracy	**96.49**	96.49
Precision	**95.74**	97.77
Recall	**95.74**	93.61
F1 score	**95.74**	95.65

Table 4. Comparative evaluation of N-AdaBoost and AdaBoost model

Metrics	N-AdaBoost	AdaBoost
Accuracy	**99.12**	98.24
Precision	**100.00**	100.00
Recall	**97.87**	95.74
F1 score	**98.92**	97.82

These metrics provide a detailed comparison of ML algorithms using both the N-dataset and the original dataset. According to Table 2, N-DT demonstrate improvements in accuracy, precision, and F1 score, but slightly lower than RF and AdaBoost in precision and F1 score. Table 3 shows that N-RF exhibits notable enhancements in precision and F1 score, indicating improved capability to accurately classify positive instances while maintaining overall performance metrics. Table 4 highlights performance of N-AdaBoost, achieving 99.12% accuracy, perfect precision of 100.00%, 97.87% recall, and a F1 score of 98.92%. These results underscore AdaBoost's effectiveness in handling uncertainties inherent in biomedical data. In contrast, when using the original dataset, DT, RF, and AdaBoost shows lower precision and F1 scores compared to their performance with the N-dataset. The comparison of the results is displayed in Figure 2. Overall, integrating NS theory enhances the predictive capabilities of these algorithms for breast cancer diagnosis.

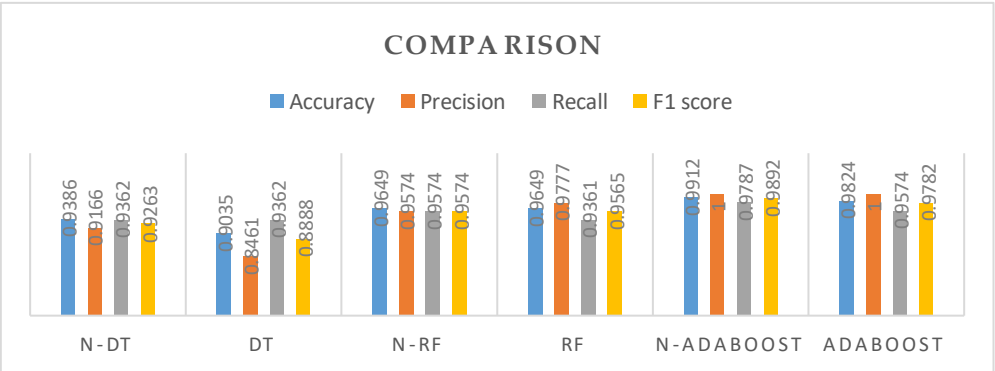

Figure 2. Comparison of N-tree based and conventional ML classifiers

5. Discussion

The proposed hybrid approach, NS-ML aimed at improving the prediction of breast cancer diagnosis. The research findings highlight that N-AdaBoost models, achieve superior accuracy and precision in detecting breast cancer. This underscores the effectiveness of integrating NST into ML models for biomedical applications, particularly in enhancing the reliability and accuracy of breast cancer diagnosis.

Conclusion

This research aims at advancing breast cancer prediction by integrating NS with ML techniques. Dataset was collected from WDBC dataset from the UCI ML Repository. The dataset was transformed into an N-representation, enriching each data point with degrees of T, I, and F to capture the complexities of medical datasets. Subsequently, the N-dataset was partitioned into training and testing subsets for training tree-based classifiers such as DT, RF, and AdaBoost. Predictive performance was measured using evaluation metrics, which demonstrated the models' capacity to recognize trends in breast cancer. Comparative analysis between the N-datasets and original datasets demonstrated improved performance metrics for DT, RF, and AdaBoost with the N-dataset. Notably, N-AdaBoost models demonstrated enhanced reliability of breast cancer diagnosis utilizing NS with scores of 99.12% accuracy, 100.00% precision, 97.87% recall, and an F1 score of 98.92%.

As a future work, the NS-ML approach can be expanded to incorporate Deep Learning (DL) and Neural Network (NN) architectures, aiming to enhance their capability in handling complex biomedical data. This integration will leverage NST to effectively model uncertainties and variability, thereby improving accuracy in tasks like image-based diagnostics and genomic analysis. Developing new neural network structures that integrate N-elements will be crucial for capturing intricate patterns in biomedical data.

References

[1] Panigrahi, L.; Verma, K.; Singh, B. K. Ultrasound image segmentation using a novel multi-scale Gaussian kernel fuzzy clustering and multi-scale vector field convolution. *Expert Syst. Appl.* 2019, *115*, 486–498.

[2] World Cancer Research Fund International, London, breast cancer statistics. Available online: https://www.wcrf.org/cancer-trends/breast-cancer-statistics/.

[3] World Health Organization, Switzerland, statistics on breast cancer. Available online: https://www.who.int/news-room/fact-sheets/detail/breast-cancer (accessed on 9 August 2024).

[4] Kahya, M. A. Classification enhancement of breast cancer histopathological image using penalized logistic regression. *Indones. J. Electr. Eng. Comput. Sci.* 2019, *13*, 1.

[5] Begum, S.A.; Mahmud, T.; Rahman, T.; Zannat, J.; Khatun, F.; Nahar, K.; Towhida, M.; Joarder, M.; Harun, A.; Sharmin, F. Knowledge, attitude and practice of Bangladeshi women towards breast cancer: a cross-sectional study. *Mymensingh Med. J.* 2019, *28*, 96–104.

[6] Singh, L.K.; Khanna, M.; Singh, R. Artificial intelligence-based medical decision support system for early and accurate breast cancer prediction. *Adv. Eng. Softw.* 2023, *175*, 103338.

[7] Sun, Y.S.; Zhao, Z.; Yang, Z.N.; Xu, F.; Lu, H.J.; Zhu, Z.Y.; Zhu, H.P. Risk factors and preventions of breast cancer. *Int. J. Biol. Sci.* 2017, *13*, 1387–1397.

[8] Sharma, G.N.; Dave, R.; Sanadya, J.; Sharma, P.; Sharma, K. Various types and management of breast cancer: an overview. *J. Adv. Pharm. Technol. Res.* 2010, *1*, 109–126.

[9] Polyak, K. Breast cancer: origins and evolution. *J. Clin. Investig.* 2007, *117*, 3155–3163.

[10] Benson, J.R.; Jatoi, I.; Keisch, M.; Esteva, F.J.; Makris, A.; Jordan, V.C. Early breast cancer. *Lancet* 2009,

373, 1463–1479.

[11] Valencia-Moreno, J.M.; Gonzalez-Fraga, J.A.; Gutierrez-Lopez, E.; Estrada-Senti, V.; Cantero-Ronquillo, H.A.; Kober, V. Breast cancer risk estimation with intelligent algorithms and risk factors for Cuban women. *Comput. Biol. Med.* 2024, *179*, 108818.

[12] Rekha, K.S.; Divya, D.; Amali, M.J.; Yuvaraj, N. Hybrid ML-MDKL feature subset selection and classification technique accompanied with rat swarm optimizer to classify the multidimensional breast cancer mammogram image. *Optik* 2024, *297*, 171574.

[13] Malakouti, S.M.; Menhaj, M.B.; Suratgar, A.A. ML: Early Breast Cancer Diagnosis. *Curr. Probl. Cancer Case Rep.* 2024, *13*, 100278.

[14] Aličković, E.; Subasi, A. Breast cancer diagnosis using GA feature selection and Rotation Forest. *Neural Comput. Appl.* 2017, *28*, 753–763.

[15] Sakri, S.B.; Rashid, N.B.A.; Zain, Z.M. Particle swarm optimization feature selection for breast cancer recurrence prediction. *IEEE Access* 2018, *6*, 29637–29647.

[16] Amin, B.; Salama, A.A.; El-Henawy, I.M.; Mahfouz, K.; Gafar, M.G. Intelligent neutrosophic diagnostic system for cardiotocography data. *Comput. Intell. Neurosci.* 2021, *2021*, 6656770.

[17] Jennifer, J.S.; Sharmila, T.S. A neutrosophic set approach on chest X-rays for automatic lung infection detection. *Inf. Technol. Control* 2023, *52*, 37–52.

[18] Saqlain, M.; Garg, H.; Kumam, P.; Kumam, W. Uncertainty and decision-making with multi-polar interval-valued neutrosophic hypersoft set: A distance, similarity measure and machine learning approach. *Alex. Eng. J.* 2023, *84*, 323–332.

[19] Akbulut, Y.; Sengur, A.; Guo, Y.; Smarandache, F. NS-k-NN: Neutrosophic set-based k-nearest neighbors classifier. *Symmetry* 2017, *9*, 179.

[20] Nguyen, G.N.; Son, L.H.; Ashour, A.S.; Dey, N. A survey of the state-of-the-arts on neutrosophic sets in biomedical diagnoses. *Int. J. Mach. Learn. Cybern.* 2019, *10*, 1–13.

[21] Ahmed, R.; Nasiri, F.; Zayed, T. A novel Neutrosophic-based machine learning approach for maintenance prioritization in healthcare facilities. *J. Build. Eng.* 2021, *42*, 102480.

[22] Tanaji, B.A.; Roychowdhury, S. BWM Integrated VIKOR method using Neutrosophic fuzzy sets for cybersecurity risk assessment of connected and autonomous vehicles. *Appl. Soft Comput.* 2024, *159*, 111628.

[23] Fetanat, A.; Tayebi, M. Sustainability and reliability-based hydrogen technologies prioritization for decarbonization in the oil refining industry: A decision support system under single-valued neutrosophic set. *Int. J. Hydrog. Energy* 2024, *52*, 765–786.

[24] Rogulj, K.; Kilić Pamuković, J.; Ivić, M. Hybrid MCDM based on VIKOR and cross entropy under rough

neutrosophic set theory. *Mathematics* 2021, *9*, 1334.

[25] Mondal, K.; Pramanik, S. Rough neutrosophic multi-attribute decision-making based on rough accuracy score function. *Neutrosophic Sets Syst.* 2015, *8*, 14–21.

[26] Wang, H.; Smarandache, F.; Zhang, Y.; Sunderraman, R. Single valued neutrosophic sets. *Infinite Study*

2010.

[27] Alshikho, M.; Jdid, M.; Broumi, S. Artificial Intelligence and Neutrosophic Machine learning in the Diagnosis and Detection of COVID 19. *J. Prospect. Appl. Math. Data Anal.* 2023, *1*, 2.

[28] Mitchell, T. Decision tree learning. *Mach. Learn.* 1997, *414*, 52–78.

[29] Breiman, L. Random forests. *Mach. Learn.* 2001, *45*, 5–32.

[30] Schapire, R.E. Explaining adaboost. In *Empirical Inference: Festschrift in Honor of Vladimir N. Vapnik*; Springer Berlin Heidelberg: Berlin, Heidelberg, 2013; pp. 37–52.

Received: June 21, 2024. Accepted: August 14, 2024