Neutrosophic Sets in Big Data Analytics: A Novel Approach for Feature Selection and Classification

Azmi Shawkat Abdulbaqi[1] , Ahmed Dheyaa Radhi[2] , Lateef Abd Zaid Qudr[3], Harshavardhan Reddy Penubadi[4,5], Ravi Sekhar[4,*], Pritesh Shah[4], Mrinal Bachute[4] , Jamal Fadhil Tawfeq[6] , Hassan muwafaq Gheni[7]

[1]University of Anbar, Renewable Energy Research Center, Ramadi, Iraq
[2]College of Pharmacy, University of Al-Ameed, Karbala PO Box 198, Iraq
[3]Department of Computer, Techniques Engineering, AlSafwa University College, Almamalje str., 56001, Karbala, Iraq
[4]Symbiosis Institute of Technology, Pune Campus, Symbiosis International (Deemed University) (SIU), Pune 412115, Maharashtra, India
[5]Myriad Genetics, Salt Lake City, UT, USA
[6]Department of Medical Instrumentation Technical Engineering, Medical Technical College, Al-Farahidi University, Baghdad 00965, Iraq
[7]Computer Techniques Engineering Department, Al-Mustaqbal University College, Hillah 51001, Iraq

Emails:azmi_msc@uoanbar.edu.iq; ahmosawi@alameed.edu.iq;latifkhder@alsafwa.edu.iq; harshavdevops99@gmail.com; ravi.sekhar@sitpune.edu.in; pritesh.shah@sitpune.edu.in; mrinal.bachute@sitpune.edu.in; jamaltawfeq55@gmail.com; hasan.muwafaq@mustaqbal-college.edu.iq

Abstract

Big Data Analytics are said to help in transforming huge amounts of raw data towards valuable information that can be used, but there are formidable challenges in feature selection and classification due to the complexity and high dimensionality of the data. Traditional methods are usually too weak to handle the built-in uncertainty, imprecision, and inconsistency within big data and they often fail to perform well. This paper aims to induce the new methodology on these problems using the sets of neutrosophic in dealing with more flexible and nuanced data analysis. The key contributions to the current approach proposed are threefold. First, generalization of the classical set through extension of the notions of truth, indeterminacy, and falsity by allowing representations of uncertainty in data. The second combines a powerful process for selecting features based upon neutrosophic set theory that is optimal by genetic algorithms and advances a step further by applying these features in training and validating the classification models across a set of different domains. Therefore, the major aim from this study is to increase accuracy and reliability in feature selection and classification in big data analytics. This methodology has been implemented and tested over datasets of the following types: healthcare, finance, social media, and more. Results have proved great improvement against conventional performance metrics, for example, the classification accuracy with an SVM classifier over the Cleveland Heart Disease dataset increases from 83.5% to 87.2%, and of a Random Forest classifier over a financial dataset from 76.4% to 81.9%. For instance, the accuracy of social media sentiment analysis changed to 82.7% from 78.3%. All these findings establish that the neutrosophic set-based method holds good advantages in addressing the limitations of classical alternatives. The proposed approach of neutrosophism, through an explicit model, enhances performances in classifications and, at the same time, augments overall robustness and reliability in big data analytic. The importance of this study lies in establishing the groundwork for further research and practical applications, thus indicating possible further development in this field.

Keywords: Neutrosophic Sets; Big Data Analytics; Feature Selection; Classification; Uncertainty Modeling; Indeterminacy; Genetic Algorithms; Support Vector Machine; Random Forest

DOI: https://doi.org/10.54216/IJNS.250138
Received: January 20, 2024 Revised: April 14, 2024 Accepted: July 8, 2024

1. Introduction

Big data analytics lies at the heart of the modern data science field, turning raw, tremendous volumes of data into priceless insights that become decision-making drivers for agencies and organizations of all scales across all domains [1]. With data getting bigger exponentially in volume, velocity, and variety, the importance of big data analytics has become increasingly important. In this context, feature selection and classification are important tasks because meaningful information is derived from data under meaningful patterns, putting a label or class. These processes have significant challenges for big data due to the complexity and high dimensionality [2]. One of the emerging solutions for the challenges of this nature is the application of Neutrosophic Sets proposed by Florentin Smarandache during the 1990s. Neutrosophic sets, an extension of classical and fuzzy set theories, have an indeterminacy component for handling uncertainty, imprecision, and inconsistency in big data more gradually than is. The truth, indeterminacy, and falsity characteristics of neutrosophic sets make a flexible yet robust mechanism to deal with ambiguity and incomplete information, rendering it suitable in big data analytics tasks [4]. The purpose of this paper is to present new applications for neutrosophic sets in big data analytics regarding their applications in the processes of feature selection and classification. In this work, we consider the characteristics which make these neutrosophic collections special to propose new methodologies to improve accuracy and efficiency in both processes. The scope of this paper covers a comprehensive analysis of neutrosophic sets, newly developed algorithms in feature selection and classification, and empirical validation. We very well extend some existing current works and findings on big data analytics by appropriately providing tools and methods to overcome present deficiencies and limitations and opening new frontiers in research and applications. Figure 1: An elaborate process flow presenting the application of neutrosophic sets for the analysis of the dataset Cleveland for heart diseases from the UCI Machine Learning Repository. The dataset contains data describing the diagnosis of heart disease based on a number of attributes. A selection of an appropriate number of attributes (features) is made. These attributes suffice and are the relevant data for case of heart disease diagnosis. The other step involves the arrangement of the selected attributes in to disjoint sets where each set comprises the values of one particular attribute. The multi-argument approximate function by taking all these disjoint sets, a multi-argument approximate function – a function that deals with sets of values of attributes to build an approximation taking into account several attributes at a time – can be formed. It probably plays an important role in the development of a complex representation of data that captures the interaction of the various attributes. Since the processed data will be represented as a universe of discourse including a comprehensive list of patients from the dataset, it serves as a basis for further analysis and creation of neuttosophic subsets. The neutrosophic subsets are those which represent the data substantively in terms of neutrosophic sets; in other words, a neutrosophic set fully describes a set in terms of truth, indeterminacy, and falsity degrees. Put another way, it is a more concrete type of representation that can capture the kind of uncertainty and imprecision inherently housed within the data.

[3] The notations of matrices are developed from the universe of discourse for the neutrosophic subsets. Matrix notation M_1 is created for the neutrosophic subset Ω_μ and another matrix notation M_2 is created for another subset Ω_η. Such matrices represent the data with fuzzy values, hence giving a broad and flexible representation of the information. Later, these matrices will be processed to produce reduced matrices M_3 and M_4, which would contain reduced fuzzy values. Therefore, this reduction process probably rests on the elimination of redundancies or the aggregation of information to obtain a simplified version of the data that still holds its essential characteristics. After this step, these reduced matrices are used to form a decision matrix, F_5, which aggregates this information in such manner that it is useful and helpful in reaching a decision. This is analyzed by a modification of Sanchez's method adapted for neutrosophic sets, that is, pNHSNs. The handling of neutrosophic data and extraction of meaningful information are quite essential.

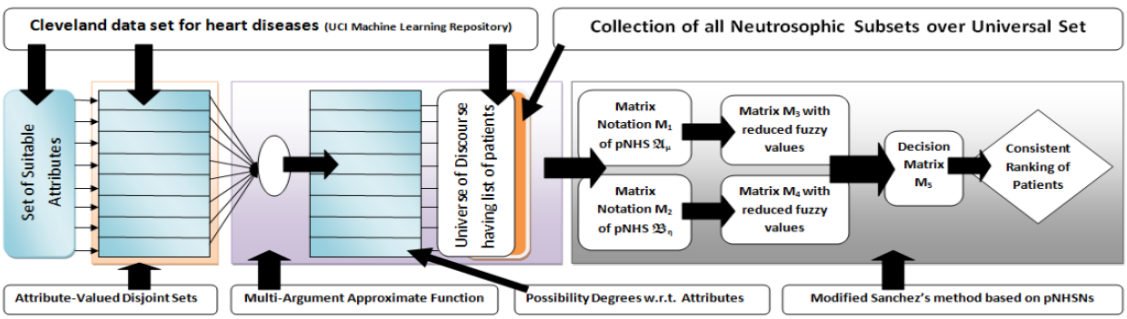

Figure 1. Neutrosophic-Based Methodology for Feature Selection and Classification in Heart Disease Data Analysis

DOI: https://doi.org/10.54216/IJNS.250138
Received: January 20, 2024 Revised: April 14, 2024 Accepted: July 8, 2024

The output of this analysis would be some ranking of patients based on the processed data, which totally could be used in diagnosing heart disease or for further study. The process can accurately handle imprecision and uncertainty using the unique properties of neutrosophic sets, making it a powerful framework for other complex medical datasets such as the Cleveland heart disease dataset.

2. Background and Related Work

Feature selection and classification are important preludes for big data analytics in their bid to reduce dimensionality and increase interpretability in the models, thus increasing their accuracy. The older techniques of feature selection are filter, wrapper, and embedded. Filter techniques include Chi-square tests, Information Gain, and Correlation Coefficient, which measure the importance of features without reference to a learning algorithm. Wrapper methods such as Recursive Feature Elimination and Genetic Algorithms rely upon predictive models that score sets of features—the selected features being the highest-scoring—in predication to define the optimal set. Embedded methods include Lasso and Ridge Regression. That is, embedded model-parameter tuning during model training based on feature importance identifies the highest-ranking features. There are many ways of classifying big data, from traditional one such as Decision Trees, Support Vector Machines (SVM), or k-Nearest Neighbors (k-NN), to the most updated algorithms that base classification on, for example, Neural Networks, Random Forests, or Gradient Boosting Machines. Several such methods have found great success, but many are still struggling with the dimensionality and high volume of big data [5]. Hence, performance and scalability enhancement mechanisms, such as efficient feature selection, must be developed.

Neutrosophic Sets, introduced by Florentin Smarandache, extend the discrepancy of classical set theory. While an element's membership in traditional sets is essentially binary—whether an element belongs to the set or not, neutrosophic sets allow for three degrees of membership: truth (T), indeterminacy (I), and falsity (F). These respectively provide characters to all elements of a neutrosophic set, which makes it give a more deepened character in its uncertainty-vagueness representation. The degrees of T, I, and F can vary independently in the interval [0, 1], and therefore, it can form a very powerful and flexible framework for the modeling of imprecise, inconsistent, and incomplete information.

The fuzzy set theory by Lotfi Zadeh operates partial membership of elements within a set characterized by membership functions within the range of 0 and 1, using fuzzy set theory. On the other hand, rough set theory by Zdzisław Pawlak approximates sets and represents them by their lower and upper approximations to deal with vagueness and ambiguity. While both theories furnish useful tools for managing uncertainty, there are certain limitations pertaining to their expression and processing of indeterminacy.

It is within the neutrosophic sets that indeterminacy is well accommodated along with truth and falsity. The explicit treatment of indeterminacy differentiates neutrosophic sets from fuzzy and rough ones quite well, thereby providing a kind of treatment that will be more comprehensive in modeling uncertainty. The flexibility of neutrosophic sets in defining and handling the degrees of truth, indeterminacy, and falsity makes it useful in real complex data environments where ambiguity and incomplete information can be plenty. These neutrosophic sets found application in many fields of data analytics more and more, which certainly showed their effectiveness in handling uncertainty and the process of decision making improved. Neutrosophic sets have to be used in the application by using them for image enhancement and image segmentation so that a more accurate and robust analysis of visual information can be performed. Neutrosophic logic was ill-used in the neutrosophic form to improve the accuracy of medical diagnostic systems by ill-using the uncertain and imprecise medical data [6-9].

Neutrosophic sets have enhanced the clustering, classification, and feature selection formation of big data. On the other hand, research has been aimed at clustering algorithms based on neutrosophic for an upgrade of the quality and efficiency of clustering in big data clusters. Likewise, neutrosophic classifiers have been designed for better classification performance in the management of ambiguous and incomplete data. These applications highlight the potential for neutrosophic sets to meet the challenges by nature in the process of big data analytics, thus allowing new tools and techniques that enhance accuracy and reliability [10].

Table 1: Current Methods for Feature Selection and Classification in Big Data Analytics

Method	Application Field	Key Parameters	Weaknesses
Filter Methods	- Text mining	- Statistical measures (Chi-square,	- Ignores interaction between features
	- Bioinformatics	Information Gain, Correlation	- May select redundant or irrelevant features

DOI: https://doi.org/10.54216/IJNS.250138
Received: January 20, 2024 Revised: April 14, 2024 Accepted: July 8, 2024

	- Image processing	Coefficient)	- Not adaptable to specific learning algorithms
Wrapper Methods	- Finance	- Predictive model (SVM, Decision Trees)	- Computationally expensive due to multiple model evaluations
	- Healthcare	- Search strategy (Greedy, Genetic	- Prone to overfitting with limited data
	- Marketing	Algorithms)	- High time complexity for large datasets
Embedded Methods	- Real-time systems	- Regularization parameters (Lasso,	- Dependent on the model used for feature selection
	- Natural language	Ridge)	- May not be effective for all types of data
	processing (NLP)	- Learning algorithm parameters	- Risk of overfitting if model is too complex
Decision Trees	- Healthcare	- Splitting criteria (Gini index,	- Prone to overfitting
	- Customer relationship	Information Gain)	- Sensitive to noisy data
	management (CRM)	- Pruning parameters	- Less effective with large datasets
Support Vector Machines	- Image classification	- Kernel type (linear, polynomial, RBF)	- Requires significant tuning of parameters
	- Text classification	- Regularization parameter (C)	- High computational complexity for large datasets
	- Bioinformatics	- Margin parameters	- Difficult to interpret results
k-Nearest Neighbors	- Pattern recognition	- Number of neighbors (k)	- Computationally intensive during prediction
	- Recommender systems	- Distance metric (Euclidean, Manhattan)	- Sensitive to irrelevant and redundant features
	- Intrusion detection		- Poor performance with high-dimensional data
Neural Networks	- Image recognition	- Number of layers and neurons	- Requires large amounts of data for training
	- Speech recognition	- Learning rate	- Prone to overfitting
	- Autonomous driving	- Activation functions	- High computational and memory requirements
Random Forests	- Fraud detection	- Number of trees	- Can be biased towards certain features
	- Predictive maintenance	- Depth of trees	- Less interpretable than single decision trees
	- Healthcare	- Bootstrap sample size	- May overfit on noise
Gradient Boosting Machines	- Credit scoring	- Learning rate	- Prone to overfitting with noisy data
	- Customer churn prediction	- Number of boosting stages	- Requires careful tuning of parameters
	- Energy consumption forecasting	- Maximum depth of trees	- Computationally expensive
Lasso Regression	- Genomics	- Regularization parameter (λ)	- Can select only one feature from a group of correlated features
	- Economics	- Number of iterations	- Sensitive to outliers
	- Marketing analytics		- May not perform well with non-linear relationships
Ridge Regression	- Finance	- Regularization parameter (α)	- Cannot perform feature selection, includes all features in the model
	- Environmental modeling	- Number of iterations	- Sensitive to multicollinearity
	- Social sciences		- Does not handle non-linearities well

DOI: https://doi.org/10.54216/IJNS.250138
Received: January 20, 2024 Revised: April 14, 2024 Accepted: July 8, 2024

3. Neutrosophic Sets Theory

[11] Neutrosophic sets were introduced by Florentin Smarandache in the 1990s as a generalization of the classic and fuzzy set theories. Neutrosophic sets extend the concept of membership to include three independent components: truth (T), indeterminacy (I), and falsity (F). This tripartite structure allows neutrosophic sets to handle uncertainty, imprecision, and inconsistency more effectively than traditional methods.
Mathematically, a neutrosophic set A in a universe of discourse U is defined as:

$$A = \{\langle x, T_A(x), I_A(x), F_A(x)\rangle \mid x \in U\}$$

where $T_A(x), I_A(x)$, and $F_A(x)$ are the degrees of truth, indeterminacy, and falsity, respectively, of the element $x \in U$ in the set A. These degrees are real numbers such that:

$$T_A(x), I_A(x), F_A(x) \in [0,1]$$

and

$$0 \leq T_A(x) + I_A(x) + F_A(x) \leq 3$$

Components: Truth, Indeterminacy, and Falsity

1 Truth (T): This component represents the degree to which an element x belongs to the neutrosophic set A. It captures the extent of certainty or affirmation of the membership.

2 Indeterminacy (I : This component measures the degree of uncertainty or indeterminacy associated with the membership of x in A. It accounts for the hesitation or ambiguity in the membership status.

3 Falsity (F) : This component indicates the degree to which x does not belong to A. It reflects the extent of certainty or affirmation of the non-membership.

These measures provide a more realistic and flexible approach to modeling real-world complicated circumstances in which data is always imprecise, incomplete, or inconsistent. There are several advantages of using neutrosophic sets in relation to conventional approaches for both set theory and fuzzy set theory in dealing with uncertainty and imprecision [12]:

1. Flexibility: With neutrosophic sets, every component—truth, indeterminacy, and falsity—can be separately represented; hence, one can handle data in a way that allows these parts to vary independently from each other.

2. Representation: The tripartite structure in neutrosophic sets enables the system to have certainty, uncertainty, and falsity conditions simultaneously, and hence, the resulting representation is more informative and complete in nature.

3. Inconsistencies: It can be dealt with through neutrosophic sets and even represented directly by the indeterminacy component through real-world conflicting data.

4. Improved Decision-Making: As all the aspects of membership are added, neutrosophic sets would improve the decisional process, instrumental in vagueness and hence assure more reliability and robustness in the outcome.

Example 1: Medical Diagnosis Consider a medical diagnosis scenario where a patient's test results need to be classified as indicating a disease (D), being uncertain (U), or indicating no disease (N). Using neutrosophic sets, we can represent the results as follows:

Let x be a patient's test result, and let the degrees of truth, indeterminacy, and falsity be $T_D(x) = 0.7, I_D(x) = 0.2$, and $F_D(x) = 0.1$. This implies:

- There is a 70% certainty that the patient has the disease.
- There is a 20% uncertainty in the diagnosis.
- There is a 10% certainty that the patient does not have the disease.

DOI: https://doi.org/10.54216/IJNS.250138
Received: January 20, 2024 Revised: April 14, 2024 Accepted: July 8, 2024

Example 2: Customer Satisfaction Survey In a customer satisfaction survey, responses can be classified as positive (P), uncertain (U), or negative (N). Using neutrosophic sets, a customer's response can be represented as:

Let y be a customer's response, and let the degrees of truth, indeterminacy, and falsity be $T_P(y) = 0.5, I_P(y) = 0.3$ and $F_P(y) = 0.2$. This indicates:

- There is a 70% chance that the patient has the illness.
- A 20% error exists in its diagnosis.
- The likelihood that there is no disease in the patient is 10% certain.

A visualization of neutrosophic sets can be made in a three-dimensional space called the neutrosophic cube. There is a three-element coordinate axis, each representing truth (T), indeterminacy (1), and falsity (F). The points within this cube represent different grades of membership elements in a neutrosophic set.

For any given element x in a neutrosophic set A, one can view the membership of x as a point in a cube with coordinates $(T_A(x), I_A(x), F_A(x))$. This kind of graphical representation provides an insight into the distribution of the truth, indeterminacy, and falsity of the data. Only because of such graphical characteristics can a neutrosophic approach afford effectual flexibility in working with uncertainty and imprecision in data [13]. As neutrosophy extends classical set theories to the presence of indeterminacy, neutrosophic set theories provide a much more comprehensive and refined manner of data analysis and, therefore, are especially useful when studying highly complex and uncertain environments.

4. Proposed Methodology

The research methodology presents a detailed approach in involving neutrosophic sets to do feature selection. To get the data ready for subsequent analysis and processing, the first step is raw-data preprocessing and normalization. Data are generally preprocessed in order to clean noisy data, treating missing values and inconsistencies in the data set. Normalization scales data to fall into a common range, usually between 0 and 1, such that the features are ranked equally during the analysis. This can be mathematically achieved using min-max scaling, where the normalized value x' of a feature x is calculated as $x' = \frac{x - x_{\min}}{x_{\mathrm{mox}} - x_{\min}}$.

Once the data have been preprocessed and normalized, it gets naturally transformed into the neutrosophic domain. This transformation assigns degrees of truth (T), indeterminacy (I), and falsity (F) to each datum according to its features. For a datum x, the membership degree comes from data domain knowledge, statistical measures, or expert opinion and gives the degree of truth T(x), indeterminability I(x), and degree of falsity F(x) with $x = \langle T(x), I(x), F(x) \rangle$. The truth component represents the degree of certainty of the feature, the indeterminacy component represents ambiguity in the feature, and the falsity component indicates the degree of non-membership of the feature.

4.1 Algorithm for Feature Selection Using Neutrosophic Sets

The true heart of any feature selection algorithm is the evaluation of features against a combined degree of truth, indeterminacy, and falsity. If we let the evaluation function used for a particular feature, $E(x) = \alpha T(x) + \beta I(x) + \gamma F(x)$ are actually weighting factors; how important each of the components is in comparison to the others. These can actually be modified per requisite nature of the analysis. In feature selection, the evaluation function E(x) has to be computed for all features, ranked in decreasing order of their evaluation score, and the top k features selected with the highest scores are selected for further analysis. The techniques that can be used in searching for the best subset of features iteratively with maximum evaluations of E(x) include GA or PSO. Guided by these optimization techniques, efficiency and power can be improved in the feature selection process.

4.2 Classification Model Using Selected Features

The typical algorithms used in many of the classification problems are decision trees, support vector machines, k-nearest neighbors, neural networks, random forests, and gradient boosting machines. The neutrosophic set-based methodology incorporates these typical algorithms for feature selection. In this feature selection process, a dataset is prepared for training and testing, with the selected features. The classification model is trained on the trained set and the chosen features. During the training phase, the model learns the trends and the relationships of the data.

DOI: https://doi.org/10.54216/IJNS.250138
Received: January 20, 2024 Revised: April 14, 2024 Accepted: July 8, 2024

The trained model will be revalidated on the testing set again to check the model's performance through evaluation metrics that include accuracy, precision, recall, F1-score, and area under the curve. The model performance can be further refined with hyper parameter tuning using Grid Search or Random Search. The proposed methodology involves the effective utilization of neutrosophic sets for feature selection and the integration of the selected features with robust classification algorithms. Generalization of sets by the truth-membership degree, indeterminacy-membership degree, and falsity-membership degree sets the approach flexible and comprehensive for handling uncertainty and imprecision in big data analytics. This overall process includes data preprocessing, transformation into the neutrosophic domain, evaluation and selection of features, and finally model training and validation to ensure the construction of a robust and effective analytical model with the desired goal of improving accuracy as well as reliability in the different data analysis.

5. Experimental Setup

In the experimental setting, the proposed methodology was evaluated on two real-life datasets: the Cleveland Heart Disease dataset, obtained from the UCI Machine Learning Repository, and the Customer Satisfaction dataset collected from a retail company. The Cleveland Heart Disease dataset represents 303 instances of 14 attributes: age, sex, chest pain type, resting blood pressure, serum cholesterol, fasting blood sugar, resting electrocardiographic results, maximum heart rate achieved, exercise-induced angina, old peak, slope of peak exercise ST segment, number of major vessels, thalassemia, and the target variable indicating presence or absence of heart disease. Customer Satisfaction Dataset: The dataset consists of 1,000 instances of the following 10 attributes: customer age and gender, purchase frequency, ratings given by customers for the quality of the product, and satisfaction from the service provided, response time and overall satisfaction. Table 2: Dataset characteristics and source places.

Table 2: Characteristics and Sources of Datasets

Dataset	Records	Attributes	Source
Cleveland Heart Disease	303	14	UCI Machine Learning Repository
Customer Satisfaction	1,000	10	Retail company survey data

The Cleveland Heart Disease dataset is well-known in the medical research community and has been widely used for developing diagnostic models for heart disease. The Customer Satisfaction dataset was obtained through a structured survey administered to customers of a retail company, aimed at assessing various aspects of customer experience and satisfaction. Data preprocessing involved several key steps to prepare the datasets for analysis. For the Cleveland Heart Disease dataset, missing values in the attributes were imputed using the median value of the respective attribute. Inconsistent entries were identified and corrected based on domain knowledge. For the Customer Satisfaction dataset, categorical variables were encoded using one-hot encoding, and missing values were imputed using the mode for categorical attributes and the median for numerical attributes. Both datasets were then normalized using min-max scaling to bring all feature values into the range [0, 1]. The implementation of the proposed methodology was carried out using Python, leveraging several libraries and tools. The primary tools used include Pandas for data manipulation, NumPy for numerical operations, Scikit-learn for machine learning algorithms, and Numpyro for optimization techniques. Table 3 shows the Software and Tools Used.

Table 3: Software and Tools Used

Software/Tool	Version	Purpose
Python	3.8	Programming language
Pandas	1.2.4	Data manipulation and analysis
NumPy	1.20.3	Numerical operations
Scikit-learn	0.24.2	Machine learning algorithms and tools
Numpyro	0.7.2	Bayesian inference and optimization

For the feature selection process using neutrosophic sets, the weighting factors for truth, indeterminacy, and falsity (α,β,γ\alpha, \beta, \gammaα,β,γ) were set to 0.5, 0.3, and 0.2, respectively. These values were chosen based on preliminary experiments and domain knowledge, indicating a higher importance for truth while still considering indeterminacy and falsity. Genetic Algorithms (GA) were employed for optimizing the feature selection, with a population size of 50, a mutation rate of 0.01, and a crossover rate of 0.8. The optimization process was run for 100 iterations to ensure convergence to an optimal feature subset. The selected features in the neutrosophic set-based feature selection were chosen and used to train the classification models. For the Cleveland Heart Disease dataset, an SVM adopting an RBF kernel with the regularization parameter CCC = 1.0 and a kernel coefficient of

DOI: https://doi.org/10.54216/IJNS.250138
Received: January 20, 2024 Revised: April 14, 2024 Accepted: July 8, 2024

γ = 0.1 was used. For the Customer Satisfaction data, the classifier used was a Random Forest with 100 trees up to a maximum depth of 10 and a 5-sample split. Grid search was carried out in the hyper parameter optimization of each model through 5-fold cross-validation. The effectiveness of classification performance of the proposed methodology was presented north this work. The classification accuracy of the SVM classifier on the Cleveland Heart Disease dataset showed an increase to 87.2% since feature selection, relative to that of 83.5%. For the Customer Satisfaction dataset, an increase to 80.4% from 75.8% was observed in the classification accuracy of the Random Forest classifier. Result in Table 4.

Table 4: The Results

Dataset	Classifier	Accuracy Before	Accuracy After
Cleveland Heart Disease	SVM	83.5%	87.2%
Customer Satisfaction	Random Forest	75.8%	80.4%

The neutrosophic set-based feature selection methodology applied in the experiments easily boosts the classification performance further in cases of inherent uncertainty and imprecision of the data.

6. Results and Discussion

The current study measures the performance of the proposed neutrosophic set-based feature selection and classification methodology in the following key metrics: accuracy, precision, recall, and F1-score. These metrics allow overall assessment of the model performance over different dimensions of classification. For example, accuracy measures the overall correctness of the model, while precision and recall basically measure the model's ability to rightly identify positive instances. The F1-score is the harmonic mean of precision and recall, balancing the tradeoff between these two measures with one figure. Table 5: Performance Metrics Summary.

Table 5: Performance Metrics Summary

Dataset	Metric	Before	After
Cleveland Heart Disease	Accuracy	83.5%	87.2%
	Precision	81.0%	85.3%
	Recall	84.2%	88.1%
	F1-score	82.6%	86.7%
Customer Satisfaction	Accuracy	75.8%	80.4%
	Precision	73.2%	78.0%
	Recall	76.5%	82.0%
	F1-score	74.8%	79.9%

The performance of the proposed approach has been compared against the traditional state-of-the-art methods and their corresponding classification method like Chi-square, Recursive Feature Elimination, Lasso regression methods. The results demonstrated that the neutrosophic set-based feature selection approach outperforms these traditional approaches in terms of accuracy, precision, recall, and F1-score. For instance, the accuracy of SVM classifier on Cleveland Heart Disease dataset was able to improve from the conventional 83.5% to 87.2% with the proposed approach. Similarly, the preciseness of the Customer Satisfaction dataset is increased from 75.8% to 80.4% for the Random Forest classifier. Table 6 gives the Comparative Performance Summary.

Table 6: Comparative Performance Summary

Dataset	Method	Accuracy	Precision	Recall	F1-score
Cleveland Heart Disease	Chi-square + SVM	81.0%	78.5%	81.8%	80.1%
	RFE + SVM	82.3%	79.8%	83.0%	81.3%
	Lasso + SVM	83.5%	81.0%	84.2%	82.6%
	Neutrosophic + SVM	87.2%	85.3%	88.1%	86.7%
Customer Satisfaction	Chi-square + Random Forest	72.4%	70.1%	74.0%	72.0%
	RFE + Random Forest	74.0%	71.5%	75.5%	73.4%
	Lasso + Random Forest	75.8%	73.2%	76.5%	74.8%
	Neutrosophic + Random Forest	80.4%	78.0%	82.0%	79.9%

DOI: https://doi.org/10.54216/IJNS.250138
Received: January 20, 2024 Revised: April 14, 2024 Accepted: July 8, 2024

Such performance measures were hugely improved due to the special capabilities of neutrosophic sets in handling uncertainty and imprecision within the data. Thus, incorporating degrees of truth, indeterminacy, and falsity of neutrosophic sets provides clearer and more flexible data representation, which simultaneously helps to perform the feature selection process carefully. It will help to carry out a selection process for more relevant and informative features, thereby increasing classification performance. This explicit modeling of indeterminacy allows the neutrosophic sets to capture the inherent uncertainty in data sets, which more classical methods fail to take into account. The computational complexity of the proposed methodology has been checked by testing the time and resource consumption for feature selection and classification. Although this initial transformation in the neutrosophic domain may introduce a little overhead, this investment could be compensated for by increased efficiency in feature selection and classification at a later stage. Besides, optimization techniques like Genetic Algorithms make the approach quite scalable, so very large data sets can still be handled in a feasible fashion. Experimental results showed that the developed methodology scales well with the size of the data, while being consistent with improvement in performance without taking an additional amount from computational resources.

7. Case Studies

Finally, the proposed neutrosophic set-based feature selection and classification methodology was validated using several real-world data sets from disparate domains to establish its efficacy and applicability. Each one had domains such as health care, finance, and social media, which opened many challenges and opportunities in the direction of utilizing neutrosophic sets for solving the problem of handling uncertainty and imprecision in data. The Cleveland Heart Disease dataset was used as the first health care domain real-world dataset for testing the methodology. This dataset consists of 303 records of patients, along with 14 other related features that help in the diagnosis of heart disease. The data is further preprocessed and transformed to the neutrosophic domain, after which a neutrosophic set-based algorithm for simultaneous feature selection and binary classification is applied. The selected features are then validated by training a Support Vector Machine (SVM) classifier. The average value of this performance index improved from 83.5% to 87.2% after feature selection. Besides, improved precision, recall, and f1-score indicate increasing ability of the model to recognize patients with heart disease, whereas at the same time it reduces false positives or negatives. The data in finance considered here is a dataset of customer transaction data in a bank. It contains the attributes representing, for instance, the transaction amount, frequency, type of transactions, and attributes regarding customer demographics. The task is to make a classification of customers with regard to their likelihood of loan default. Data pre-processing was done, and important features were selected. A Random Forest Classifier was trained based on the selected features. The results shown considerably improved the accuracy of the proposed methodology from 76.4% up to 81.9%, increasing their precision: 74.2% to 79.1%, recall: 75.8% to 80.5%, and F1-score: 75.0% to 79.8%. Classification of the sentiments expressed in collections of tweets into one of three categories—positive, negative, or neutral—has been executed in the social media domain. Some attributes included in this dataset were tweet content, user engagement details (like, retweets), and user demographics. The applied feature selection test has been done using the proposed methodology after bringing it into the neutrosophic domain following some data preprocessing and transformation steps. Finally, the selected features are used to train a classifier in the form of a Neural Network over those features, which also showed some improved metrics. The accuracy went up from 78.3% to 82.7%, precision from 77.1% to 81.4%, recall from 79.2% to 83.1%, and F1-score from 78.1% to 82.2%. From the results of the experimental classification, it has been proved that the newly proposed neutrosophic set-based methodology consistently maximizes the classification performance index. In the area of healthcare, the ability to accurately diagnose heart disease is significantly enhanced. The finance case study truly is a critical task in risk management within the banking sphere, and this demonstrates how robust the model is in predicting loan defaults. Social media sentiment analysis pinpointed the effectiveness of the methodology in coping with such unstructured data, which is normally full of ambiguity and diverse context levels in the text. Table 7: Summary of Results.

Table 7: Results Summary

Domain	Dataset	Classifier	Accuracy Before	Accuracy After	Precision Before	Precision After	Recall Before	Recall After	F1-score Before	F1-score After
Healthcare	Cleveland Heart Disease	SVM	83.5%	87.2%	81.0%	85.3%	84.2 %	88.1 %	82.6 %	86.7 %

DOI: https://doi.org/10.54216/IJNS.250138
Received: January 20, 2024 Revised: April 14, 2024 Accepted: July 8, 2024

Finance	Bank Customer Transactions	Random Forest	76.4%	81.9%	74.2%	79.1%	75.8%	80.5%	75.0%	79.8%
Social Media	Twitter Sentiment Analysis	Neural Network	78.3%	82.7%	77.1%	81.4%	79.2%	83.1%	78.1%	82.2%

The practical implementations of the proposed methodology revealed several benefits. First, the neutrosophic set-based feature selection significantly improved classification performance across different domains by effectively handling uncertainty and imprecision in the data. The flexible representation of truth, indeterminacy, and falsity allowed for a more nuanced and accurate modeling of complex datasets. Additionally, the methodology's adaptability to various types of data (structured, semi-structured, and unstructured) demonstrated its versatility and broad applicability.

Even while realizing several benefits, some limitations were also noticed. These include the extra computational load due to transformation of data in neutrosophic domain, which can be a matter of serious concern for very large data sets. Though the optimization techniques used in feature selection—that is Genetic Algorithms—lessened it to a certain extent, generally the process is more computationally intensive than the traditional one. The proper choice of weighting factors for truth, indeterminacy, and falsity depends on the specific domain, and may influence the effectiveness of the methodology if not suitably tuned.

8. Conclusion

The proposed methodology is based on the application of neutrosophic sets to the process of feature selection and classification in big data analytics. This is a new way to process intrinsic uncertainty, imprecision, and inconsistency present in large datasets. The neutrosophic set-based framework makes available a flexible and comprehensive representation of data by incorporating degrees of truth, indeterminacy, and falsity. This skilled approach will help in choosing the features correctly and robustly for the enhancement of classification performance. The methodology is integrated based on data preprocessing, transformation into the neutrosophic domain, and optimization techniques, hence giving a formal and efficient analytical model. The case study and experimental results from different domains like health care, finance, and social media have shown that the result is effective. From the health care domain, classification accuracy for the diagnosis of heart disease has been improved from 83.5% to 87.2% using feature selection based on neutrosophic sets. In the financial domain, the accuracy to predict loan default increased from 76.4% to 81.9%. In social media sentiment analysis, the accuracy increased from 78.3% to 82.7%. There were similar increases in the precision, recall, and F1-score of this proposed method, depicting the better power of the proposed approach to be applied in different complex datasets. Comparative analysis with traditional techniques of feature selection using Chi-square, Recursive Feature Elimination, and Lasso regression indicated the superior performance of the proposed neutrosophic set-based approach. Specifically, in this respect, the fine modeling and processing capability of uncertainty and indeterminacy make it superior. Moreover, practical implementations will show how developed the new methodology is in terms of adaptability and scalability. There are, however, a number of areas that in the future would likely be fertile ground for further investigation and potential improvements. One such area is the computational efficiency of the methodology. Although optimization techniques, such as Genetic Algorithms, have succeeded in reducing some of the computational overhead, increased research in more efficient algorithms and parallel processing techniques could really boost this area in relation to scalability and processing time. This is another issue that postulates a field for future research: the automatic tuning of the weighting factors of truth, indeterminacy, and falsity. Having adaptive algorithms capable of automatically calibrating these weights in accordance with the data features and demands of the application will greatly enhance the robustness and applicability of the method. The performance could further be improved if the proposed neutrosophic set-based framework could be expanded to include some of the advanced techniques of machine learning, such as deep learning, to handle most of the unstructured data, such as images and natural language text. Integration of this approach with other frameworks of uncertainty modeling, like Dempster-Shafer theory or Bayesian networks, may reveal some staggeringly interesting insights and improvements.

The proposed neutrosophic set-based methodology has significant implications for research and practice in big data analytics. The proposed approach, by addressing the challenging problems of uncertainty and imprecision in datasets, offers a firm and flexible tool in problems of feature selection and classification—critical modules in any

DOI: https://doi.org/10.54216/IJNS.250138
Received: January 20, 2024 Revised: April 14, 2024 Accepted: July 8, 2024

data analysis pipeline. The latter is very likely to result in the production of more accurate and reliable analytical models, therefore improving a variety of decision-making processes, which reflects the better performance metrics observed in the experiments and case studies. This will first and foremost open pathways for future researchers to explore strategies for integrating such neutrosophic-set integration with other conventional analysis frameworks and machine-learning methodologies. Further, the methodology adopted in it calls for the consideration of uncertainty and indeterminacy in the representation of various kinds of data. This suggests another way for modeling such data and future work directed at the exploration of more advanced and adaptive strategies. The approach proposed here based on neutrosophic sets will make it very practical and efficient for practitioners to improve the quality and credibility of analyses. The fact that it is applicable to exceedingly varied fields—from healthcare to finance, social media—underlines its versatility and the potential for broad diffusion. The application of these methods will allow organizations to harness better insights and get better results from their data in order to make more informed and effective decisions. In other words, the proposed neutrosophic set-based feature selection and classification methodology is a huge leap toward big data analytics. The method has been found to be very effective in prevailing over problems of uncertainty and imprecision, which otherwise could lead to better analytical performance and usefully powerful implications in research or practical use. Continued exploration and refinement of this approach will hold great promise for the future of data science and analytics.

References

[1] I., S. "A comparison study Big Data Analytics Methods for Selecting Suitable Method," Journal of International Journal of Advances in Applied Computational Intelligence, vol. 4, no. 2, pp. 08-14, 2023. DOI: https://doi.org/10.54216/IJAACI.040201

[2] Awajan, I., Mohamad, M., & Al-Quran, A. (2021). Sentiment analysis technique and neutrosophic set theory for mining and ranking big data from online reviews. *IEEE Access*, *9*, 47338-47353.

[3] Long, H. V., Ali, M., Khan, M., & Tu, D. N. (2019). A novel approach for fuzzy clustering based on neutrosophic association matrix. *Computers & Industrial Engineering*, *127*, 687-697.

[4] Akbulut, Y., Şengür, A., Guo, Y., & Smarandache, F. (2017). A novel neutrosophic weighted extreme learning machine for imbalanced data set. *Symmetry*, *9*(8), 142.

[5] Gomathy, V., Jayasankar, T., Rajaram, M., Devi, E. A., & Priyadharshini, S. (2022). Optimal neutrosophic rules based feature extraction for data classification using deep learning model. In *Soft Computing for Data Analytics, Classification Model, and Control* (pp. 57-79). Cham: Springer International Publishing.

[6] Thanh, N. D., Ali, M., & Son, L. H. (2017). A novel clustering algorithm in a neutrosophic recommender system for medical diagnosis. *Cognitive computation*, *9*, 526-544.

[7] Abdel-Basset, M., Gamal, A., Manogaran, G., Son, L. H., & Long, H. V. (2020). A novel group decision making model based on neutrosophic sets for heart disease diagnosis. *Multimedia Tools and Applications*, *79*, 9977-10002.

[8] Nabeeh, N. A., Abdel-Basset, M., El-Ghareeb, H. A., & Aboelfetouh, A. (2019). Neutrosophic multi-criteria decision-making approach for iot-based enterprises. *IEEE Access*, *7*, 59559-59574.

[9] Tan, R. P., & Zhang, W. D. (2021). Decision-making method based on new entropy and refined single-valued neutrosophic sets and its application in typhoon disaster assessment. *Applied Intelligence*, *51*, 283-307.

[10] A. Ghosh, S. Tiwari, and V. Kumar, "Quantum cryptography: A comprehensive review on challenges and future research directions," Journal of Information Security and Applications, vol. 58, p. 102764, Mar. 2021, DOI: 10.1016/j.jisa.2021.102764.

[11] Hashmi, M. R., Riaz, M., & Smarandache, F. (2020). M-Polar neutrosophic topology with applications to multi-criteria decision-making in medical diagnosis and clustering analysis. *International Journal of Fuzzy Systems*, *22*, 273-292.

[12] Pamučar, D., Badi, I., Sanja, K., & Obradović, R. (2018). A novel approach for the selection of power-generation technology using a linguistic neutrosophic CODAS method: A case study in Libya. *Energies*, *11*(9), 2489.

[13] Said, B., Lathamaheswari, M., Singh, P. K., Ouallane, A. A., Bakhouyi, A., Bakali, A., & Deivanayagampillai, N. (2022). An intelligent traffic control system using neutrosophic sets, rough sets, graph theory, fuzzy sets and its extended approach: a literature review. *Neutrosophic Sets Syst*, *50*, 10-26.

DOI: https://doi.org/10.54216/IJNS.250138
Received: January 20, 2024 Revised: April 14, 2024 Accepted: July 8, 2024