Engineered Science
DOI: https://dx.doi.org/10.30919/es1736

# Detecting Alzheimer's Disease using Rough Set-based Classification Approach

Fokrul Alom Mazarbhuiya,[1,*] M. Shenify,[2] A.S. Wungreiphi,[1] Janet Kez[1,*] and Rupa Paul[1]

## Abstract

Alzheimer's is a memory deficiency disease that frequently affects elderly persons. Though it cannot be cured or stopped, its progression may be delayed if early diagnosis is possible. Early diagnosis of Alzheimer's disease is one of the most challenging problems for researchers, since it minimizes time, cost, and suffering for the patients, their caregivers, and health institutions. Many methods were already proposed for this purpose, and the rough set-based classification technique is one of them. In this paper, a rough set theory-based classification algorithm is proposed to classify early detection and diagnosis of the disease. The efficacy of the method can be demonstrated with the experiments conducted on a dataset collected from the National Alzheimer Coordination Centre (NACC). To make the classifier compatible with the NACC datasets, it has been customized. The results convincingly show that our method outperforms other known methods.

*Keywords*: Alzheimer's disease, Early diagnosis, Information table, Decision table, α-neighborhood, Neighborhood relation.
Received: 29 December 2024; Revised: 17 June 2025; Accepted: 21 September 2025
Type: Research article.

## 1. Introduction

The rate of occurrence of age-related disorders has increased substantially in recent years due to the great improvement in general life expectancy. An age-related and widespread type of dementia that primarily affects the elderly is Alzheimer's disease (AD).[1] AD is a chronic, degenerative brain illness that gradually impairs cognition, memory, and thinking and remembering skills.[2] After the age of 65, there's a good chance that AD will get worse over the time. Poor judgment, emotional behavioural changes, difficulty in performing familiar jobs, misplacing stuff, difficulties in solving problems, and difficulty in learning new things are some of the warning indicators that precede the actual catastrophic event.[1] Among the risk factors linked to AD are age, diabetes, obesity, smoking, and high blood pressure. The number of AD patients has been rising dramatically over the past few decades, particularly in nations with high life expectancy. Currently, 6.5 million of Americans are affected by AD; among them, 70%, are 75-years-old and older. There are 55 million people suffering from dementia, and 60% to 70% among them are estimated to have AD. The cause of AD is poorly understood, and no treatment can stop or reverse its progression. However, its progression can be delayed up to some extent if symptoms are known in advance. Detecting and diagnosing AD in the earlier stages of its progression can be a challenging task, as it can reduce time, cost, and suffering of the patients, and their caregivers. It can also help health institutions make proper planning to address the problem. Secondly, AD's progression can be delayed using proper medication. There are a couple of approaches tried for this purpose, and machine learning-based approaches are some of them.

Machine learning, since its development, has made inroads significantly in many directions and applied to various fields like data analysis, hand-written text recognition, speech recognition, pattern recognition, search engines, weather forecasting, natural language processing, medical diagnosis, *etc*.[3,4,5] Recently it has been applied in AD and dementia, and many machine learning-based methods have been developed for the early detection of AD.[6] Machine learning, a rapidly growing core area of artificial intelligence, is applied to build classifiers that allow the computer to learn,[7,8] and the aim of it is to develop a classification model.

Classification is a supervised learning approach that learns from the input data and employs this learning to classify new findings,[9,10,11,12,13] and their main focus is on predicting the

[1]*Department of Mathematics, Assam Don Bosco University, Guwahati, Assam, 781017, India*
[2]*College of Computer Science and IT, Albaha University, Albaha, 65799, Saudi Arabia*
[*]*Email*: fokrul.mazarbhuiya@dbuniversity.ac.in (F. A. Mazarbhuiya), DU2023PHD0031@dbuniversity.ac.in (J. Kez)

qualitative response through pattern recognition and analysis of data. In,[14] the authors used a convolution neural network approach to classify AD into four classes. In,[15] the authors proposed a method consisting of Naïve Bayes Classifier (NBC), Support Vector Machine (SVM), and *K*-nearest neighbor (KNN) for efficient classification of AD. A method involving high-dimensionality reduction of Alzheimer's disease datasets using SVM to get better accuracy of classification was introduced in.[16] In,[17] the authors proposed a hybrid model using electroencephalography and fused CT-MRI-based RPCA (principal component analysis) for efficient and early detection of AD. In,[18] a nice tool called rough set theory (RST) to address imprecision, vagueness, and uncertainty arising from real-life problems. Using the equivalence relation of rough set, it was successfully applied to discrete datasets. In order to deal with the high computational cost associated with the continuous data, a couple of algorithms were proposed in.[19,20,21,22] In,[23] authors proposed a neighborhood RST-based model for diagnosis of diseases.

Although various available machine learning-based AD prediction and diagnosis methods achieved the results up to a good extent, there is still more to be done in this area. Since some of the machine-learning-based methods used brain image scanning and others required special devices that needs hospital visits,[2,7] there is need of a comprehensive approach which addresses the aforesaid issues. In this article, we propose an algorithm for the early detection or diagnosis of AD. An Alzheimer's patient can be identified by his/her medical records containing symptoms values which may either be numeric or be categorical. The Alzheimer datasets are the collection of records where symptoms are the attributes with numeric and categorical values. We define a metric,[24] for finding the symptomatic similarity among the patients based on their records. Using this metric, we construct neighborhood relation, which then yields lower and upper approximation spaces and boundary regions. The rules can be extracted from these regions. The proposed algorithm is named as neighborhood rough set-based classification (NRSC) algorithm. The effectiveness of the algorithm is demonstrated using experiment conducted with a real-life collected from National Alzheimer's Coordination Centre (NACC) of University of Washington. The comparative analysis of the proposed algorithm is conducted with well-known classification methods using different performance measures available in the literatures. It is observed from the results that our method outperforms other known methods.

The following is the article's structure. Some recent advancements in this field are discussed in Section 2. The terms used in the paper are explained in Section 3. The approach suggested in this paper is presented in Section 4. The time complexity of the proposed method is discussed in Section 5. The experimental results are discussed in Section 6, and lastly, we conclude our paper with a brief conclusion in Section 7.

## 2. Related works

Developed countries with high life expectancy are encountering an exponential rise in age-related diseases. AD is one of the age-related diseases that grows slowly. It is a common form of dementia disorder by which elderly people suffer much.[1] AD affects memory, thinking ability, remembering ability, and reasoning ability[2] slowly and permanently.

Machine learning is a branch of AI that made significant advancements in a wide variety of fields, including data analysis, natural language processing, medical diagnosis, search engines, handwritten text recognition, speech recognition, pattern recognition, and weather forecasting.[3,4,5] It has recently been used in dementia and AD, and numerous machine learning-based techniques have been devised for AD early detection.[6] The goal of machine learning is to create categorization models by designing and developing classifiers that enable computers to learn.[7,8]

Foreseeing the qualitative response through pattern identification and data analysis is the primary goal of classification, a supervised learning technique that learns from the input data and uses this learning to classify fresh results.[9,10,11,12,13] In,[9] the authors made a comprehensive discussion on methods used to classify the breast cancer among women. In,[10] the authors analyzed the most recent ensemble classification techniques used for breast cancer. In,[11] the authors proposed a neighborhood rough set based classification for the anomaly detection in network data. An intuitionistic fuzzy rough set-based approach for the discovery of anomalies in network data was discussed in.[12] In,[13] the authors used a machine learning approach for early prediction of investigated objects over the road network. In,[14] the authors divided AD into four groups using a convolution neural network technique. For the effective classification of AD, the authors in[15] suggested a technique that combines the NBC, SVM, and KNN. In order to improve classification accuracy, a technique utilizing SVM for high-dimensionality reduction of AD datasets was presented in.[16] The authors of[17] suggested a hybrid method for the effective and early identification of AD that combines electroencephalography with fused CT-MRI-based RPCA.

RST is a useful mathematical tool that was presented in[18] to solve ambiguity and uncertainty that arise from real-life problems. It was effectively applied to a discrete datasets using the rough set equivalency relation. In,[19] the authors discussed in detail the innovative models for implementing the rough set in the construction of decision making systems. In,[20] the authors introduced Shannon's entropy to evaluate information quantity in Pawlak's[18] approximation space and then proposed a method to represent Shannon's entropy with a relation matrix. In,[21] the authors presented a review of the existing literature on rough set-based decision analysis by revisiting the probabilistic rough set approximation operators. In,[22] the authors redefined the two basic concepts, the lower approximation and the upper approximation of RST. In,[23] the authors introduced a neighborhood RST-based model for classifying medical diagnosis. In,[24] the authors suggested a metric that leverages both numeric and categorical attributes, which can serve as a valuable similarity measure in various situations. In,[25] the authors proposed to use two machine learning models for the development of early predictive models of AD and related dementias. In,[26] the authors proposed a model based on Synthetic Minority Oversampling Technology-Random Forest for the prediction of AD. In,[26] the authors used a deep generative model for the prediction of preclinical AD using MRI, demographics, and scores. In,[27] the authors implemented a multimodal technique to evaluate the potential of various modalities in predicting the conversion to AD of mild cognitive damaged patients. In,[28] the authors proposed an interpretable deep learning model containing innovative design by incorporating interaction effects and multimodality to improve the prediction accuracy and horizon for MRI to AD progression.

In this article we propose a neighborhood rough set-based classification (NRSC) approach for the early detection of AD. Our proposed approach is much more efficient than most of the existing approaches due to the following facts. AD data often contains incomplete records, overlapping classes or ambiguous symptoms. Most of the existing methods discussed here cannot efficiently handle the uncertainty and vagueness available on real AD data. However, the proposed NRSC algorithm can handle such issues effectively. It can also tolerate uncertainly through the approximation spaces which makes it more reliable for noisy data. The proposed NRSC algorithm can also handle the continuous valued attributes available on any real AD data which most of the aforesaid methods could not. It can detect and eliminate irrelevant attributes through attribute reduction. It does not need any statistical assumption and works good with smaller dataset. Finally, the obtained rules can be interpretable easily.

## 3. Problem Statement

In this article. we present some terms, notations, and definitions used here. Let $U=\{x_1, x_2,\dots x_m\}$ be the records of $m$-AD patients, where each $x_i$ is an $n$-dimensional vector consisting of $k$-numeric and ($n$-$k$)-categorical attributes. The attribute values are either numeric or categorical or any other type of value defined by the domain expert. The attributes are mostly as follows:

i) Cognitive domain: In cognitive domains, the attributes are such as Memory Functions (Mini-Mental State Examination Score: MMSE, Disease Assessment Scale-Cognitive Subscale: ADAS-Cog), Executive Function Score (Trail Making Test: TMT), Language Function Score (Boston Naming Test: BNT), and Visuospatial Ability Score (Clock Drawing Test).

ii) Behavioral and Psychological Domain: In this the categorical attributes are Depression, Anxiety, Aggression and Agitation, Sleep disorder, *etc*.

iii) Functional domain: In this domain, the measures of Activities of Daily Living, Instrumental Activities of Daily Living, Social Functioning are evaluated.

iv) Neuropathological and Biological domain: Here Brain Imaging measure (MRI/PET), CSF Biomarkers, Genetic markers, *etc*. are considered.

v) Motor and Physical domain: Here Motor Function, and Physical Frailty index are considered.

vi) Caregiver and Environmental domain: In this Caregiver Burden is evaluated.

Beside the afore-mentioned attributes, the following numeric attributes are also considered in case of AD patients. For example, Montreal Cognitive Assessment (MoCA) Score, Global Deterioration Scale (GDS) Score, Clinical Dementia Rating (CDR), Brain Imaging Measure, Clock Drawing Test Score, Verbal Fluency Test, Neuropsychological Test Score *etc*.

It is to be mentioned here that each attribute describes the risk factors associated with an AD patient, which are either numeric or categorical.

**Definition 3.1 Metric between two Alzheimer's Disease Records.**

To define the metric function, we use a formula given in.[24] Let $A=\{a_1, a_2, \dots, a_k, a_{k1}, a_{k2}, \dots a_n\}$ be the set of attributes where the first $k$ attributes are numeric and the rest are categorical. Let $x$ and $y$ be the non-empty records of two AD patients. The records $x$ and $y$ are of the form $x=\{a_1(x), a_2(x),\dots a_k(x), a_{k1}(x), a_{k2}(x),\dots a_n(x)\}$ and $y=\{a_1(y), a_2(y),\dots a_k(y), a_{k1}(y), a_{k2}(y),\dots a_n(y)\}$, where $a_i(x)$ is the value of the record $x$ in the attribute $a_i$ and $a_i(y)$ is the value of the record $y$ in the attribute $a_i$. Then the metric between $x$ and $y$ is given by the formula.[24]

$$d(x,y) = \frac{kd_1(x,y) + (n-k)d_2(x,y)}{n} \quad (1)$$

where,

$$d_1(x,y) = \frac{1}{k}\sum_{i=1}^{k} \frac{|a_i(x) - a_i(x)|}{|a_i(x) + a_i(x)|} \quad (2)$$

the metric defined on numeric attributes.

Also,

$$d_2(x,y) = 1 - \frac{|\{a_1(x), a_2(x), \dots a_n(x)\} \cap \{a_1(y), a_2(y), \dots a_n(y)\}|}{|\{a_1(x), a_2(x), \dots a_n(x)\} \cup \{a_1(y), a_2(y), \dots a_n(y)\}|} \quad (3)$$

the metric defined on categorical attributes. Obviously $d(x, y) \in [0, 1]$.

Eq. (2) is derived from Canberra metric[24,29,30,31,32] due to its following features.

**i)** Scale-sensitive normalization: Since each term is divided by the sum of the absolute values of the attribute values $|a_i(x) + a_i(y)|$ of two records, it normalizes the difference of the attribute values. The attributes with small magnitudes, which are diagnostically important, can have larger influence on the Canberra metric. However, the Euclidean distance usually found to be dominated by larger values.

**ii)** High sensitivity near 0: If both $a_i(x)$ and $a_i(y)$ are close to 0, the denominator becomes very small making the distance value large. Because of this, the Canberra metric is susceptible to relative changes, which is helpful when even minor changes are noted for diagnostic purposes.

In case of AD data, each record often includes the attributes such as Cognitive Score, Brain Region Volume, Biomarker Levels, *etc*. Many of these attributes non ony have different assessment scales, but also include small but clinically significant variations. Thus Canberra distance can be suitable for such occasions as it highlights the relative differences. For example, a 0.5-point drop in MMSE is 2 than when it is 28. Additionally, the large-valued attributes that may normally dominate the metric are lessened in importance by the Canberra metric.

Similarly, the Eq. (3) has been derived from Jacard similarity measure,[24,33,34,35] which is effective in comparing the patients with diagnosis code. That's why, it can be useful in finding the similarity in categorical attributes between two AD records.

For defining the metric on the AD data records, the weighted average of Canberra metric defined in Eq. (2) and the similarity measure defined in Eq. (3) is taken so that both the numeric and the categorical attributes of AD data will have proportional contribution on the metric. The metric is then used for generating α-neighborhood relation on the AD datasets.

**Definition 3.2. α-neighborhood.**

For all $x_i \in U$ and $0 \leq \alpha \leq 1$, an α-neighborhood relation $(U, d)$ is as $\alpha(x_i) = \{x \mid d(x_i, x) \leq \alpha\}$.

**Definition 3.3. Neighborhood Decision System (NDS).**

Let $(U, A \cup D)$ be a decision system, where $U$ is the record set (universe) related to AD patients, $A$, the set of conditional attributes, and $D$, the set of decision attributes. For $\alpha < 1$, $A$ generates α-neighborhood relation $N$. Then the α-neighborhood decision system with respect to α thus is represented as NDS = $(U, A \cup D, \alpha)$.

**Definition 3.4.** Let $B$ be a subset of $A$, for any arbitrary $X \subseteq U$, the lower approximation and upper approximation of $X$ in terms of the relation $N$ with respect to $B$ are defined respectively as

$$\underline{N_B}(X) = \{x: \alpha_B(x) \subseteq X, x \in U\} \quad (4)$$

$$\overline{N_B}(X) = \{x: \alpha_B(x) \cap X \neq \emptyset, x \in U\} \quad (5)$$

where,

$$\alpha_B(x) = \{y: d(B(x), B(y)) \geq \alpha, y \in U\} \quad (6)$$

here $B(x)$ is a sub-vector of $A(x)$ having all those values of the attributes belonging to $B \subseteq A$. The boundary region of $D$ with respect to $B$ is defined as

$$Boundary(D) = \overline{N_B}(X) - \underline{N_B}(X) \quad (7)$$

The neighborhood lower approximation is the union of the lower approximation of each $D$ class, which is called the positive region. The boundary region is needed to reduce uncertainty in the decision making process.

## 4. Proposed method

In order to determine the classification rules, we first generate the α-neighbourhood relation by choosing an appropriate value for α using the Eq. (6). The classification rule generation is given as follows: Given a record set associated with $m$ patients, each of which is characterized by $n$-attribute values and represented as an $m \times n$ matrix $[x_{ij}]$, where $x_{ij}$ is the $j$th attribute value for the $i$th patient, $i$=1,2,.. $m$ and $j$=1,2,.. $n$. Generally, supervised datasets can be written as $(U, A \cup D)$, where $U = \{x_1, \dots, x_m\}$, and $A = \{a_1, \dots, a_n\}$, $D = \{d_1, \dots, d_n\}$ are respectively sets of conditional and decision attributes.

The proposed approach begins by calculating the equivalence classes of decision attributes and the neighbourhood relation of the conditional attribute using the formula provided in Definition 3.2. Eq. (1) provides the distance formula for this. After using the "and" operator and decision classes, create neighbourhood approximations to the aforementioned result. The union of each decision class's lower approximation space is known as the NRS lower approximation space of the decision attributes. The boundary region is computed from two or more decision classes using the Eq. (7). The two types of decision rules are generated from the neighborhood approximations, namely possible rules (non-deterministic rules) and certain rules (deterministic rules). Both the certain and the possible rules are generated by using the lower approximation and the upper approximation of the neighborhood rough set given by Eqs. (4) and (5), respectively. The flowchart of the method is given below in Fig. 1.

## 5. Time complexity of the proposed algorithm

Let $n$ be the number of AD patient records and $m$ be the number of attributes of AD. For finding neighbors with radius $\alpha$, we need to compute the distance between each AD record with others. Such computation uses the distance formula defined in Section 3, which requires $O(n^2 \times m)$ time. For finding lower and upper approximation space based on the neighborhood relation to determine which records belong to the approximations of the decision classes requires $O(n^2)$ time. Finally, rule generation requires $O(n \times r)$ time, where $r$ is the number of rules generated. Total time complexity is $O(n^2 \times m + n^2 + n \times r)$. Since $r < n$ and $n^2 < n^2*$m, the worst-case time complexity of NRSC is $O(n^2 \times m)$. The comparative time complexity of the proposed NRSC algorithm with other well-known algorithms like NBC, KNNC and SVB is given below in a tabular form in Table 1.

**Table1:** Worst-case time complexities of proposed algorithm along with other well-known algorithms.

| Algorithm | NBC | KNNC | SVM | NRSC (proposed) |
|---|---|---|---|---|
| Time complexity | $O(n^2 * m)$ | $O(n*m + \log n)$ | $O(n^2*m)$ | $O(n^2*m)$ |

In practice, the number of attributes is much less than the number of AD patient's record, therefore the proposed algorithm appears to run in quadratic time. Thus, NRSC shows comparable time complexity to other well-known algorithms.

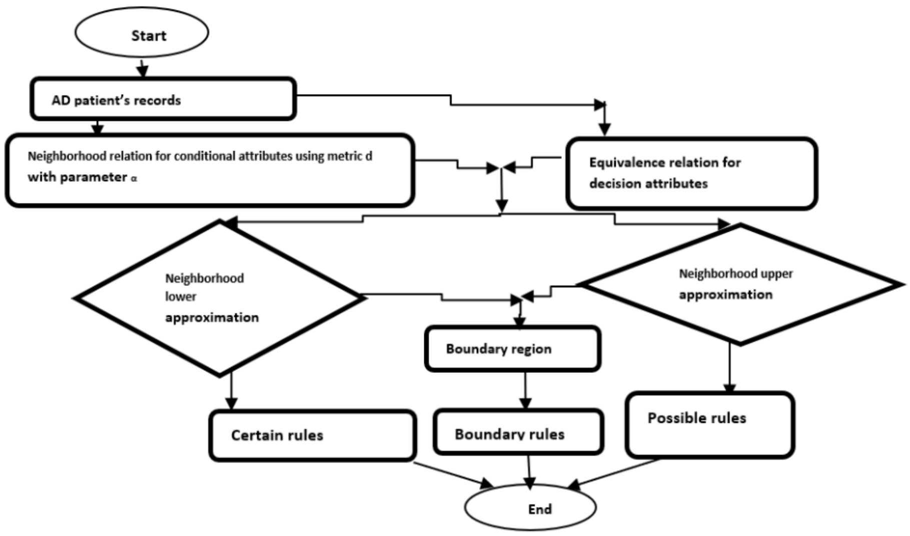


**Fig. 1:** Flowchart of the proposed method.

## 6. Experiments and Discussions

### A. Results and Discussions

In this study, the dataset of NACC of the University of Washington was used for conducting the experiments. The dataset was collected from May 2005 to August 2011. The dataset has 47,000 records and 396 attributes. Some of the attributes are potential risk factors for AD. The risk factors are classified into two classes: medical history and cognitive function. Each record of the dataset is associated with a single person's information that diagnosed AD.

Medical history: A patient's medical history may be obtained by a questionnaire or interview with a doctor. Age, sex, occupation, family history of AD, alcohol and tobacco use, diabetes, high blood pressure, heart disease, obesity, and other behavioural and personal characteristics are all included. This data includes crucial elements that influence AD diagnosis and aid in evaluating AD patients.

Cognitive function: Recently, the physicians are using many tools and methods to assess an AD patient's memory. One such tool is the Mini Mental Status Examination (MMSE), which utilized for measuring the cognitive mental status of a potential AD patient.

Both types of attributes can have numeric or non-numeric (categorical) values. The dataset is preprocessed to make it compatible for the program by removing missing values, noises, and inconsistent values. The proposed algorithm (NRSC) is implemented along with methods such as Naïve Bayesian Classification (NBC), KNN Classification (KNNC), and Support Vector Machine (SVM), using MATLAB with 10,000 refined instances of patient data. Two randomly selected samples of equal size, each having 5,000 records, have been used for training and testing. Then the classifiers are trained and tested 10 times. The value of $\alpha$ is chosen as 0.4. The classifiers were then used to categorize any new record as either AD or not. The effectiveness of the four aforementioned algorithms is assessed using a variety of widely recognized performance metrics, including false positive rate (FPR) and true positive rate (TPR). Figs. 2-7 below show a partial view of the results achieved by the aforementioned algorithms together with a comparative analysis of accuracy metrics such as Precision, Recall, F-measure, Folkes-Mellows Index, Kulczyn ski Index, and Rand Index.

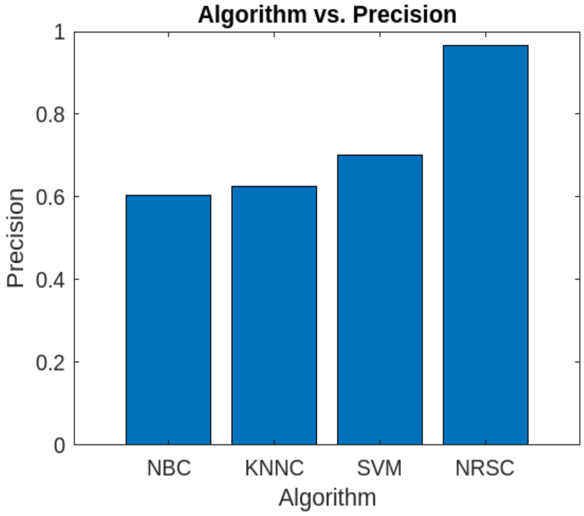


**Fig. 2:** Comparative analysis of Precision.

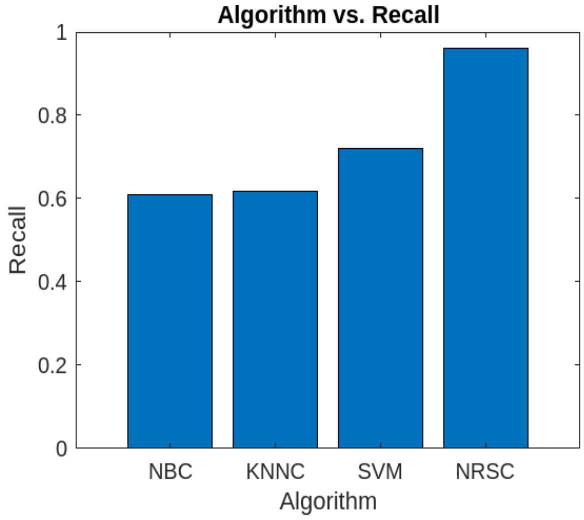


**Fig. 3:** Comparative analysis of Recall.

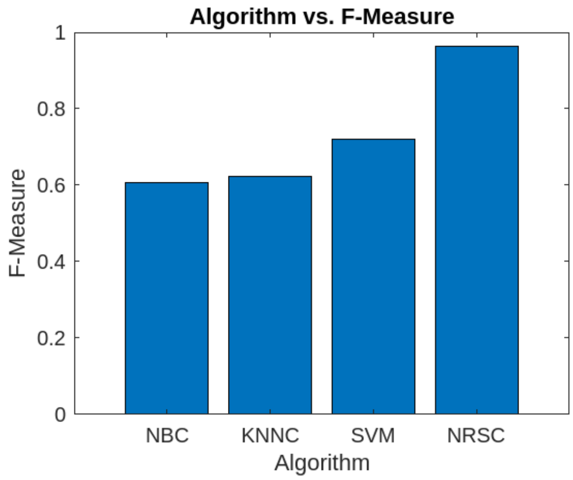


**Fig. 4:** Comparative analysis of F-Measure.

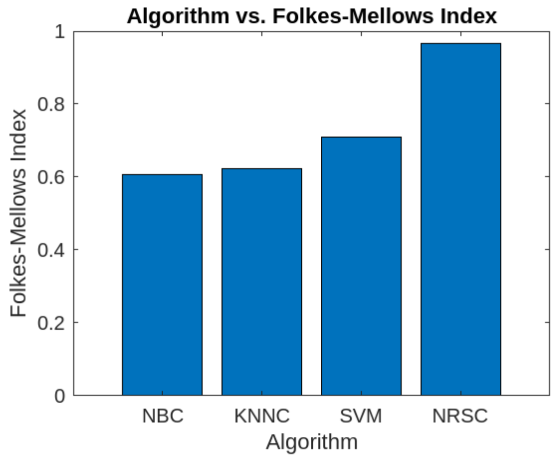


**Fig. 5:** Comparative analysis of Folkes-Mellows Index.

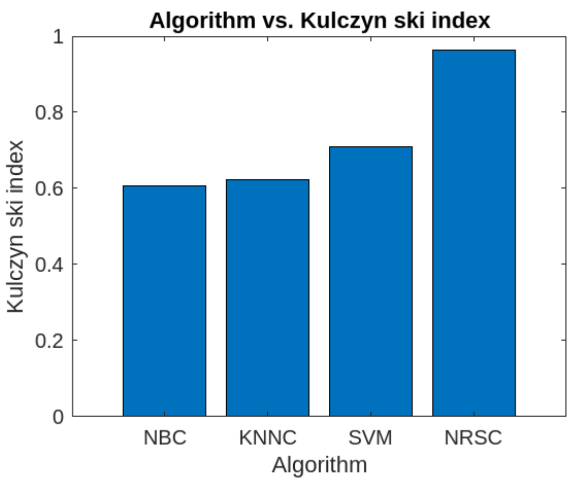


**Fig. 6:** Comparative analysis of Kulczyn ski Index.

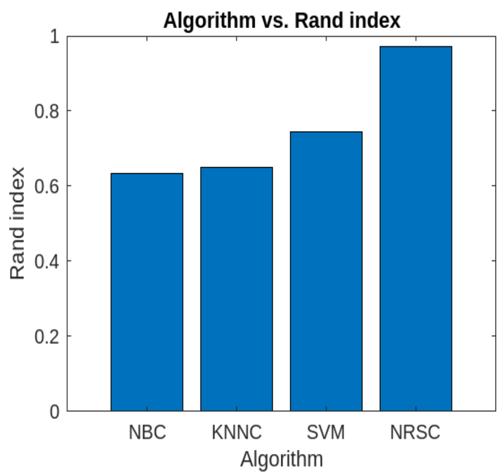


**Fig. 7:** Comparative analysis of Rand Index.

For different values of α such as 0.1, 0.2, 0.3, 0.4, and 0.5, we have implemented the proposed NRSC algorithm with the afore-mentioned datasets and the partial view of the results in terms of performance metrics are presented graphically in the Figs. 8-13 below.

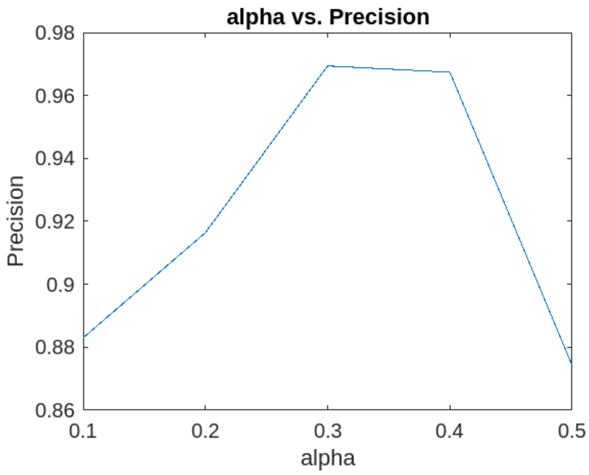


**Fig. 8:** Precision for different values of α

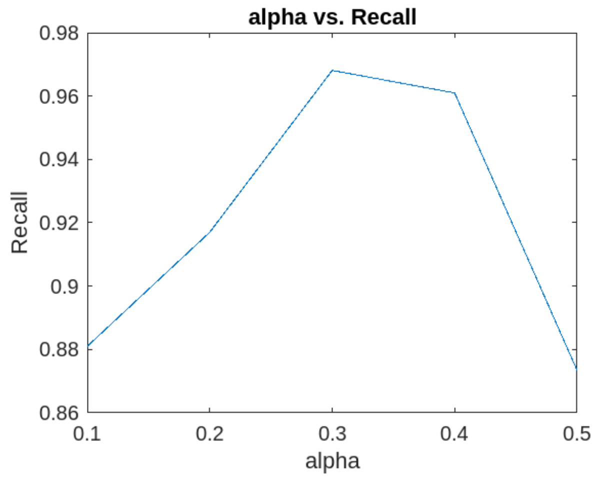


**Fig. 9:** Recall for different values of α.

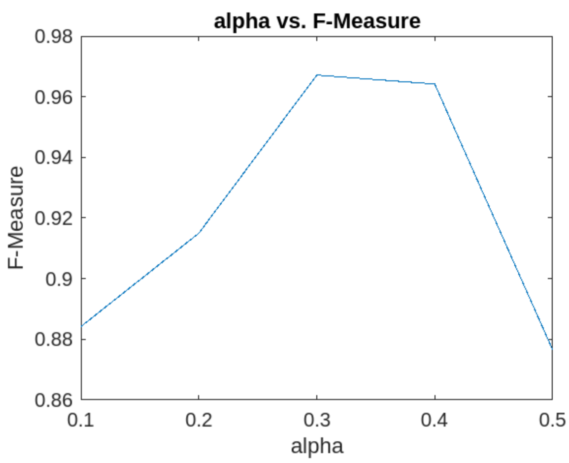


**Fig. 10:** F-Measure for different values of α.

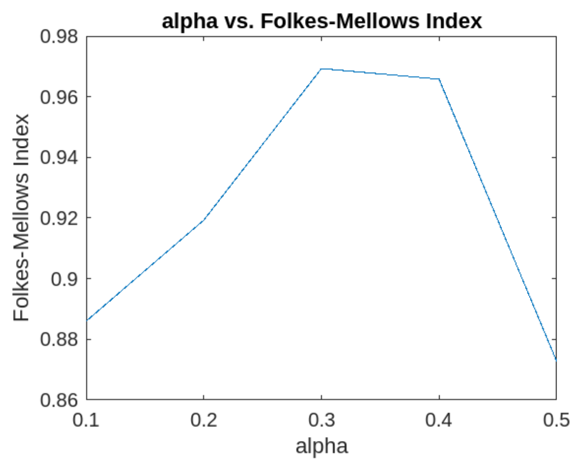


**Fig. 11:** Folkes-Mellows Index for different values of α.

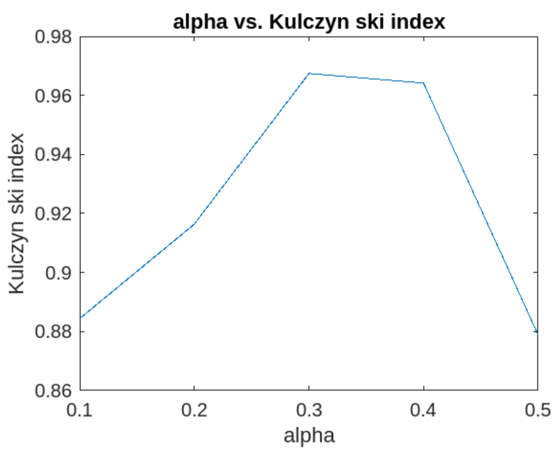


**Fig. 12:** Kulczyn ski index for different values of α.

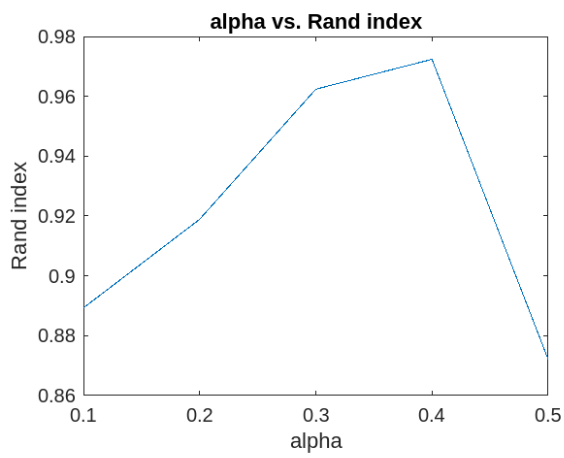


**Fig. 13:** Rand index for different values of α.

From the Figs. 2-7, it has been found that with the dataset of NACC, the proposed NRSC classification algorithm's accuracy is much higher than all other SVM, KNNC, and NBC classification algorithms, which establishes the efficacy of the proposed NRSC classification algorithm over the other aforesaid algorithms. The proposed NRSC algorithm produces almost 96-98% accuracy, whereas SVM, KNNC, and NBC produce 70-75%, 62-65%, and 60-63% accuracy, respectively. Thus, our proposed algorithm outperforms some of the well-known algorithms mentioned in this article.

From the Figs. 8-13, it has been observed that the performance indices namely, Precision, Recall, F-Mesaure, Folkes-Mellows Index, Kulczyn ski index, and Rand Index increase with the increase in the value of $\alpha$ till 0.4, then it start decreasing rapidly. The optimum value of the performance is mostly achieved in the interval [0.3, 0.4]. Therefore, most suitable value of $\alpha$ is taken 0.4 in this implementation of algorithm and comparative studies. Thus for high-dimensional, sparse dataset like Alzheimer's disease, the proposed algorithm works better for a small to moderate value of $\alpha$.

### B. Clinical Interpretations

Since any study concerning classification rules of medical data the clinical interpretation of the obtained results is very crucial, we have attempted to incorporate this issue up to a limited extent here. The following clinical interpretations of the some of the results obtained by NRSC algorithm can be drawn.

The neighbourhood lower approximation space gives certain rules that includes only those AD data records belong to a decision class. Thus such data records represent patients which definitely diagnosed with Alzheimer's. Diseases. An example of such rule is as follows:

For a patient $X$, if hippocampal of < $A$ and MMSE score < $B$, then $X$ has Alzheimer's Disease with high certainty. Similar rules represent strong indicators of the presence or absence AD.

Similarly, the neighbourhood upper approximation space gives possible rules which includes the AD data records that may or may not belong to a decision class. So, it includes both certain and uncertain rules which are more exploratory. An example of such rule is as follows:

For a patient $Y$, if age > 80 and mild memory loss, then $Y$ may have early-stage AD. Such rules help to find out the risky patient who need further medical analysis. These rules are useful for early screening or hypothesis generation.

Again, the boundary region contains the ambiguous rules with patients who have overlapping attribute of multiple conditions. In such cases, further investigations such as advanced imaging, generic testing, *etc*. are required.

## 7. Conclusion

AD is characterized by neuronal degeneration and subsequent brain disorder, resulting in changed behavior and functionality in the affected individual. The existing AD diagnosis is very costly, time consuming, and needs the patient's physical visit to the hospital for regular check-ups and diagnosis. The detection of AD with datasets using machine learning algorithms helps in faster detection. The early prediction can help the AD patients to take required response measures. The cure for the disease is yet to be found, but treatment can be provided to the patients to fight against it. In this article, we propose a neighborhood rough set-based classification method for efficient and early detection of AD patients. Our method classifies the patients into predefined classes; after that, the concern class of patients can be treated together. In this method, first we use a metric defined on both numeric and categorical attributes of AD patients. Then, a suitable value of α is chosen to construct a neighborhood relation. The metric can then be used to find lower approximation space, upper approximation space, and boundary regions. Then the certain rules can be generated from the lower approximation space, the possible rules can be generated from the upper approximation space, and the boundary rules can be extracted from the boundary region. The effectiveness of the algorithm is demonstrated using an experiment conducted with a real-life data collected from the NACC center of the University of Washington. The performance of the proposed method is tested and compared with well-known classification methods not only using complexity analysis but also with the help of different performance measures. It is evaluated using a variety of validation measures, like Precision, Recall, F-Measure, Folkes-Mellows Index, Kulczyn ski Index, and Rand Index. It is observed from the results that our proposed method outperforms other known methods in all the validation measures. In future the following directions of work can be taken.

- Fuzzy can be introduced into the proposed algorithm to enhance accuracy and the user's convenience.
- A Neural network-based approach can be explored

## Acknowledgments

The work is made possible through the assistance and support from almost every colleague and friends. The authors would like to express their heartfelt thanks to the reviewers for any constructive comments and suggestions.

## Conflict of Interest

There is no conflict of interest among the authors

## Supporting Information

Not applicable.

## CRediT Statement

**Fokrul Alom Mazarbhuiya:** Conceptualization, Methodology, Supervision, Project administration. **M. Shenify:** Methodology, Funding acquisition. **A.S. Wungreiphi:** Software, Validation, Writing – Review & editing. **Janet Kez:** Writing – Original draft, Data curation, Formal analysis, Writing—Review and editing, Visualization. **Rupa Paul:** Investigation, Validation.

## References

[1] A. Association, Alzheimer's disease facts and figures, *Alzheimer's & Dementia*, 2011, **7**, 208-244, doi: 10.1016/j.jalz.2011.02.004.

[2] R. Kohavi, G. John, R. Long, D. Manley, K. Pfleger, MLC: a machine learning library in C, *Proceedings Sixth International Conference on Tools with Artificial Intelligence. TAI 94*, New Orleans, LA, USA, November 6-9, 1994, 740-743, doi: 10.1109/TAI.1994.346412.

[3] C. Drummond, "Machine learning as an experimental science," in *AAAI Workshop,* Boston, MA, United states, July 16, 2006 - July 20, 2006, WS-06–06, 1–5.

[4] G. Holmes, A. Donkin, I. H. Witten, WEKA: a machine learning workbench, *Proceedings of ANZIIS '94 - Australian New Zealnd Intelligent Information Systems Conference*, Brisbane, QLD, Australia, November 29 - December 2, 1994, 357-361, doi: 10.1109/ANZIIS.1994.396988.

[5] M. Rana, M. Bhushan, Machine learning and deep learning approach for medical image analysis: diagnosis to detection, *Multimedia Tools and Applications*, 2023, **82**, 26731-26769, doi: 10.1007/s11042-022-14305-w.

[6] S. Joshi, P. Deepa Shenoy, V. K R, L. M. Patnaik, Evaluation of different stages of dementia employing neuropsychological and machine learning techniques, *First International Conference on Advanced Computing*, Chennai, India, December 13-15, 2009,154-160, doi: 10.1109/ICADVC.2009.5378199.

[7] S. Joshi, D. Shenoy, G. G. Vibhudendra Simha, P. L. Rrashmi, K. R. Venugopal, L. M. Patnaik, Classification of Alzheimer's disease and Parkinson's disease by using machine learning and neural network methods, *Second International Conference on Machine Learning and Computing*, Bangalore, India, February 9-11, 2010, 218-222, doi: 10.1109/ICMLC.2010.45.

[8] J. Hua, Study on the application of rough sets theory in machine learning, *Second International Symposium on Intelligent Information Technology Application*, Shanghai, China, December 20-22, 2008, 192-196, doi: 10.1109/IITA.2008.154.

[9] R. Ranjbarzadeh, S. Dorosti, S. Jafarzadeh Ghoushchi, A. Caputo, E. B. Tirkolaee, S. S. Ali, Z. Arshadi, M. Bendechache, Breast tumor localization and segmentation using machine learning techniques: Overview of datasets, findings, and methods, *Computers in Biology and Medicine*, 2023, **152**, 106443, doi: 10.1016/j.compbiomed.2022.106443.

[10] M. Hosni, I. Abnane, A. Idri, J. M. Carrillo de Gea, J. L. Fernández Alemán, Reviewing ensemble classification methods in breast cancer, *Computer Methods and Programs in Biomedicine*, 2019, **177**, 89-112, doi: 10.1016/j.cmpb.2019.05.019.

[11] F. A. Mazarbhuiy, Detecting Anomaly using Neighborhood Rough Set based Classification Approach, *ICIC Express Letters*, 2023, **17**(1), 73-80, doi: 10.24507/icicel.17.01.73.

[12] F. A. Mazarbhuiya, M. Shenify, An intuitionistic fuzzy-rough set-based classification for anomaly detection, *Applied Sciences*, 2023, **13**, 5578, doi: 10.3390/app13095578.

[13] M. Alqarqaz, M. Bani Younes, R. Qaddoura, An object classification approach for autonomous vehicles using machine learning techniques, *World Electric Vehicle Journal*, 2023, **14**, 41, doi: 10.3390/wevj14020041.

[14] N. Raza, A. Naseer, M. Tamoor, K. Zafar, Alzheimer disease classification through transfer learning approach, *Diagnostics*, 2023, **13**, 801, doi: 10.3390/diagnostics13040801.

[15] K. R. Kruthika, Rajeswari, H. D. Maheshappa, Multistage classifier-based approach for Alzheimer's disease prediction and retrieval, *Informatics in Medicine Unlocked*, 2019, **14**, 34-42, doi: 10.1016/j.imu.2018.12.003.

[16] A. S. Alatrany, W. Khan, A. Hussain, H. Kolivand, D. Al-Jumeily, An explainable machine learning approach for Alzheimer's disease classification, *Scientific Reports*, 2024, **14**, 2637, doi: 10.1038/s41598-024-51985-w.

[17] M. Leela, K. Helenprabha, L. Sharmila, Prediction and classification of Alzheimer disease categories using integrated deep transfer learning approach, *Measurement: Sensors*, 2023, **27**, 100749, doi: 10.1016/j.measen.2023.100749.

[18] Z. Pawlak, Rough sets, *International Journal of Computer & Information Sciences*, 1982, **11**, 341-356, doi: 10.1007/BF01001956.

[19] R. R. Nowicki, Rough Set Based Classification Systems, *Studies in Computational Intelligence 802*, Springer, 2019, doi: 10.1007/978-3-030-03895-3.

[20] Q. Hu, D. Yu, Z. Xie, J. Liu, Fuzzy probabilistic approximation spaces and their information measures, *IEEE Transactions on Fuzzy Systems*, 2006, **14**, 191-201, doi: 10.1109/TFUZZ.2005.864086.

[21] Y. Yao, Probabilistic rough set approximations, *International Journal of Approximate Reasoning*, 2008, **49**, 255-271, doi: 10.1016/j.ijar.2007.05.019.

[22] R. Slowinski, D. Vanderpooten, A generalized definition of rough approximations based on similarity, *IEEE Transactions on Knowledge and Data Engineering*, 2000, **12**, 331-336, doi: 10.1109/69.842271.

[23] S. U. Kumar, H. H. Inbarani, A novel neighborhood rough set based classification approach for medical diagnosis, *Procedia Computer Science*, 2015, **47**, 351-359, doi: 10.1016/j.procs.2015.03.216.

[24] F. A. Mazarbhuiya, M. Y. AlZahrani, L. Georgieva, Anomaly detection using agglomerative hierarchical clustering algorithm, *Information Science and Applications*, Springer Singapore, Singapore, 2018, 475-484, doi: 10.1007/978-981-13-1056-0_48.

[25] Q. Li, X. Yang, J. Xu, Y. Guo, X. He, H. Hu, T. Lyu, D. Marra, A. Miller, G. Smith, S. DeKosky, R. D. Boyce, K. Schliep, E. Shenkman, D. Maraganore, Y. Wu, J. Bian, Early prediction of Alzheimer's disease and related dementias using real-world electronic health records, *Alzheimer's & Dementia*, 2023, **19**, 3506-3518, doi: 10.1002/alz.12967.

[25] P. Rani, R. Lamba, R. K. Sachdeva, K. Kumar, C. Iwendi, A machine learning model for Alzheimer's disease prediction, *IET Cyber-Physical Systems: Theory & Applications*, 2024, **9**, 125-134, doi: 10.1049/cps2.12090.

[26] U. Hwang, S.-W. Kim, D. Jung, S. Kim, H. Lee, S. W. Seo, J.-K. Seong, S. Yoon, Real-world prediction of preclinical Alzheimer's disease with a deep generative model, *Artificial Intelligence in Medicine*, 2023, **144**, 102654, doi: 10.1016/j.artmed.2023.102654.

[27] D. Agostinho, M. Simões, M. Castelo-Branco, Predicting conversion from mild cognitive impairment to Alzheimer's disease: a multimodal approach, *Brain Communications*, 2024, **6**, fcae208, doi: 10.1093/braincomms/fcae208.

[28] Y. Wang, R. Gao, T. Wei, L. Johnston, X. Yuan, Y. Zhang, Z. Yu, Alzheimer's Disease Neuroimaging Initiative, Predicting long-term progression of Alzheimer's disease using a multimodal deep learning model incorporating interaction effects, *Journal of Translational Medicine*, 2024, **22**, 265, doi: 10.1186/s12967-024-05025-w.

[29] G. N. Lance, W. T. Williams, Computer programs for hierarchical polythetic classification ("similarity analyses"), *The Computer Journal*, 1966, **9**, 60-64, doi: 10.1093/comjnl/9.1.60.

[30] G. N. Lance, W. T. Williams, Mixed-data classificatory programs I - agglomerative systems, *Australian Computer Journal*, 1967, **1**, 15-20.

[31] T. H. Clifford, and W. Stephenson, "An Introduction to Numerical Classification", *Academic Press*, New York, San Fransisco, London, 1975.

[32] S. M. Emran, and N. Ye, "Robustness of Canberra Metric in Computer Intrusion Detection", *Proceedings of 2001 IEEE Workshop on Information Assurance and Security*, US Military Academy, NY, 5-6 June, 2001, 80-84, ISBN-0-7803-9814-9.

[33] M. Dutta, A. K. Mahanta and M. Mazumder, "An Algorithm for Clustering of Categorical Data using Concept of Neighours", *Proceedings of the 1st National Workshop on Soft Data Mining and Intelligent Systems*, Tezpur University, India, 103-105, 2001.

[34] M. Dutta, and A. K. Mahanta, "An Algorithm for Clustering Large Categorical Databases using a Fuzzy Set based Approach", *Proceedings of the National Workshop on Trends in Advanced Computing*, 2006, Tezpur University, India.

[35] F. A. Mazarbhuiya, and M. Y. AlZahrani, "An Efficient Method for Clustering Periodic Patterns", Computing Conference 2017, *SAI Conference*, London, U.K.