International Journal of Scientific Research and Engineering Development-– Volume 8 Issue 2, Mar-Apr 2025
Available at www.ijsred.com

RESEARCH ARTICLE OPEN ACCESS

Enhancing Alzheimer's Disease Stage Prediction with Machine Learning and Multi-Agent Systems

Dr. V. Maniraj
Associate Professor, Research Supervisor, Head of the Department, Department of Computer Science, A.V.V.M.Sri Pushpam College(Autonomous),Poondi,Thanjavur(Dt),Affiliated to Bharathidasan University,Thiruchirappalli,Tamilnadu(Mail Id- manirajv61@gmail.com)
Ms.A. Kamatchi
Research Scholar, Department of Computer Science, A.V.V.M.Sri Pushpam College(Autonomous),Poondi,Thanjavur(Dt),Affiliated to Bharathidasan University,Thiruchirappalli,Tamilnadu(Mail Id- kamatchia06@gmail.com)

1.Introduction

As life expectancy continues to rise in modern society, the prevalence of age-related diseases has also increased. Alzheimer's disease (AD), a common form of dementia that primarily affects older adults, is one such condition. AD is a progressive and irreversible brain disorder that gradually impairs memory, thinking, reasoning, and other cognitive abilities. The likelihood of developing Alzheimer's rises significantly after the age of 65, with the disease's severity worsening over time. Early warning signs of AD include poor judgment, emotional changes, difficulty completing familiar tasks, misplacing items, trouble solving problems, and challenges with learning new things. Key risk factors for AD include smoking, hypertension, diabetes, obesity, and advancing age. In recent years, the number of AD patients has increased dramatically, particularly in developed countries with longer life expectancies.

Machine learning and agent systems have made tremendous strides in various fields, including weather forecasting, robotics, search engines, natural language processing, speech recognition, medical diagnosis, and handwriting recognition. Machine learning, a fundamental component of Artificial Intelligence (AI), is an evolving technology that enables computers to "learn" by developing classifiers. This technology aims to address problems related to inference and prediction using available data, which is essential for decision-making by humans or intelligent systems like agent systems. An agent system refers to a computer system placed in an environment that can perform flexible, autonomous actions to achieve its design goals. One key feature of agent systems is their ability to demonstrate intelligence through interaction. Intelligent agents possess reactive, proactive, and social characteristics. A multi-agent system consists of multiple agents that collaborate or compete to accomplish shared objectives.

Several researchers have proposed different methods for predicting the stages of Alzheimer's Disease (AD) using machine learning applied to various diagnostic techniques. While these studies have made significant progress, some have focused solely on brain imaging scans for diagnosis, while others rely on specialized devices that require patients to visit healthcare centers for evaluation.

1.1.Problem definition

A recent European e-health study reveals that approximately 2.4 million unnecessary visits to healthcare centers occur each year. Among these, elderly patients, particularly those suffering from Alzheimer's disease (AD), are frequent offenders of unnecessary hospital visits. While AD can be diagnosed through a variety of methods, including physical exams, laboratory tests, and brain imaging scans, these procedures typically require the patient's physical presence at the medical facility. This results

in a significant number of AD patients traveling to healthcare centers. The rising number of visits places increased strain on healthcare professionals, leading to longer patient wait times and additional costs for both patients and institutions in terms of time and money. Each visit to the healthcare center involves the registration and storage of large volumes of patient data for future reference. However, this extensive data is often only utilized when there is a need to revisit the patient's medical history. By applying machine learning, this accumulated data on cognitive functions and medical history offers a new opportunity for diagnosing AD patients. This study aims to combine machine learning and agent systems with the vast medical data of 47,000 AD patients to enable diagnosis without the need for unnecessary in-person visits.

1.2. Aim and objectives

The aim of this research is to investigate a machine learning algorithms for Alzheimer disease stage prediction and integrate the algorithm with multi agent system for accuracy improvement.

1.2.1. Objective:

I. Review and assess the current techniques used for diagnosing Alzheimer's Disease (AD).
II. Analyze various machine learning algorithms for their suitability in classification tasks.
III. Determine the most appropriate machine learning classification algorithm for categorizing patients' AD stages using the NACC dataset.
IV. Tailor and implement the chosen machine learning algorithm.

1.3. Research questions

- Which machine learning algorithm performs the best in classifying NACC's medical health and cognitive function data for Alzheimer's Disease diagnosis?
- Does the integration or segregation of medical history and cognitive function data influence the classification accuracy?
- What strategies can be used to enhance classification accuracy through the use of a multi-agent system?

1.4. Research approach

In the previous section, three key research questions have been outlined that need to be explored. To address these questions, a combination of research methods will be employed. A comprehensive literature review and experimental analysis will be conducted to determine the most effective machine learning algorithm for classifying Alzheimer's Disease patients using the NACC dataset. The literature review will also focus on identifying various Alzheimer's diagnosis techniques and potential risk factors. Peer-reviewed articles and academic journals will be carefully examined to gather relevant insights.

The dataset for this study is sourced from the National Alzheimer's Coordinating Center (NACC) at the University of Washington, containing detailed information on over 47,000 patients. Based on the disease risk factors identified during the literature review, relevant data will be extracted from this dataset. Data mining performance can be severely impacted by issues such as noise, missing values, or incomplete and inconsistent data. To mitigate these issues, thorough data preparation will be carried out prior to conducting the experiments. Subsequently, the most effective algorithm will be determined through a combination of literature review and experimentation.

1.5 Research outcome

The anticipated outcomes of this research include:

- Extracting valuable knowledge from the NACC dataset for diagnosing Alzheimer's Disease patients.
- Identifying the most appropriate machine learning classification algorithm for the NACC dataset.
- Developing a multi-agent system model aimed at improving prediction accuracy.

1.6 Structure of the thesis

This thesis is structured as follows: The first chapter introduces Alzheimer's Disease (AD), defines the research problem, presents the research questions, and outlines the aim and objectives of the study. The second chapter provides an in-depth background on various AD diagnosis techniques, machine learning, and agent systems, along with a review of related research in the literature review section. Chapter 3 explains the research methodology used in this study, including diagrams and brief descriptions. In the following chapter, the experimental design, machine learning algorithms, and tools are assessed and selected. Additionally, relevant Alzheimer's disease risk factors are identified from the dataset in this chapter.

2.1 Overview

In developed countries with high life expectancy, the prevalence of age-related diseases increases significantly as the population ages. Alzheimer's Disease (AD) is a slowly progressing, common form of dementia that primarily affects the elderly. It initially impacts areas of the brain responsible for thought, memory, and language. Elderly individuals with AD may struggle to remember recent events, recognize familiar faces, perform everyday tasks, and more. Currently, around 4 million Americans and 5 million Europeans are living with AD. The incidence of AD is projected to increase dramatically, with the number of cases expected to quadruple by 2020. By 2050, it is estimated that someone will develop the disease every 30 seconds. In the United States, AD is one of the leading causes of death among individuals aged 65 and older. Despite ongoing research, there is still no cure for AD, although various diagnostic and treatment approaches are being explored to delay its progression.

Although there is no definitive test to confirm the presence of Alzheimer's Disease (AD), early and accurate diagnosis plays a critical role in tracking the progression of the disease. To differentiate AD from other causes of memory loss, doctors typically rely on a combination of medical history, physical exams, cognitive testing, laboratory studies, and brain imaging techniques[9]. A person's medical history, which includes risk factors for AD such as family history, smoking, alcohol use, diabetes, hypertension, heart disease, obesity (BMI), and gender, is a key part of the assessment process. The physical exam helps ensure that the patient is generally in good health, during which the physician checks vital signs like blood pressure, temperature, pulse, heart rate, and conducts tests on the lungs and heart. Blood or urine samples are also collected for further laboratory analysis. Cognitive or neuropsychological testing is conducted to assess the individual's ability to understand and respond accurately to questions. Common cognitive tests include the Mini-Mental State Examination (MMSE) and the Functional Activity Questionnaire (FAQ). Brain imaging techniques such as Magnetic Resonance Imaging (MRI), Functional MRI (fMRI), Positron Emission Tomography (PET), and Single-Photon Emission Computed Tomography (SPECT) are used to detect abnormalities in the brain. These diagnostic methods help classify individuals as either healthy or

affected by AD. The severity of AD varies among patients and is typically classified into five stages: "No," "Questionable," "Mild," "Moderate," and "Severe."

Supervised learning, also known as classification , focuses on accurately and efficiently classifying new inputs. The classifier is trained on a dataset of labeled instances to predict the class of unlabeled data. Several well-known classification algorithms, including AdaBoost, C4.5, k-Nearest Neighbors (kNN), decision trees, Naïve Bayes, neural networks, and support vector machines (SVM), are widely studied and applied by researchers. Supervised learning finds applications in various fields, such as weather forecasting, predicting disease risk factors in healthcare, classifying network packets in telecommunications, and many other areas.

Unsupervised learning, or clustering, involves taking a set of objects as input and identifying patterns or groups within the data, where sequences of similar objects are grouped together. Common unsupervised learning algorithms include k-means, agglomerative clustering, and the Gaussian mixture model. Effective clustering brings similar objects together at the lower levels of the hierarchy and delays merging dissimilar groups until higher up in the hierarchy. Since the potential patterns to be discovered in unsupervised learning are far more extensive, it generally makes unsupervised learning more complex than supervised learning.

An agent system is a computer system designed to perform autonomous actions in its environment to achieve specific objectives. Agents perceive their surroundings through sensors, allowing them to understand what is happening in the environment. Based on this information, they choose and execute appropriate actions using their actuators. Key characteristics of intelligent agents include reactivity, proactivity, and sociability. In addition to these traits, intelligent agents may also incorporate learning behaviors. Agent technology results from the integration of various computer science technologies, such as artificial intelligence and object-oriented programming. A fundamental feature of agents is their ability to provide intelligence through interaction .

A multi-agent system (MAS) involves multiple interacting agents working together, either cooperatively or competitively, to achieve their collective goals. MAS is particularly useful for solving problems that are too complex for a single agent to handle. In other words, MAS is ideal when the knowledge or expertise required to solve a problem is distributed. The interaction among agents goes beyond simple data exchange; it involves cooperation, coordination, and negotiation, much like the interactions in human life.

he Belief-Desire-Intention (BDI) architecture is a widely recognized agent design framework, drawing inspiration from human mental states: Belief, Desire, and Intention. The agent's behavior is modeled around these three components. In this architecture, the interpreter plays a crucial role by continuously updating the agent's beliefs based on information gathered from the environment. Based on the updated beliefs, the agent selects an appropriate desire and plan to execute, and the chosen plan is carried out to achieve the agent’s intention.

The Procedural Reasoning System (PRS) is an agent architecture that explicitly represents the mental states found in the BDI framework. PRS is considered one of the most established and robust implementations of the BDI paradigm. This architecture has been successfully applied to major industrial multi-agent systems, including air traffic control systems like the OASIS system at Sydney Airport and the SPOC business process management system.

The Subsumption architecture, developed by Rodney Brooks, was designed to address the limitations of the previously mentioned agent construction approaches. Unlike others, this architecture does not require the agent to store a world model in memory. Instead, agents react dynamically to events in their environment, making it particularly effective in situations where memory capacity is limited. The agent's behavior in this architecture is organized in hierarchical layers, allowing for easy modification by adding additional behavior layers at the top.

2.2 Literature review

Several studies have investigated the use of various Alzheimer's Disease (AD) diagnostic data, including MRI, fMRI, MMSE, PET, SPECT, demographic information, and behavioral data,the authors aimed to classify individuals as either AD patients or healthy using the Random Forest algorithm. To achieve this, they employed a supervised learning approach with five distinct steps. First, the fMRI data was preprocessed to remove noise, missing values, and errors. Then, the data modeling step followed, after which features distinguishing AD and healthy subjects were extracted. In the fourth step, a subset of the extracted features was selected. Finally, a supervised Random Forest classification algorithm was applied to a dataset of 41 subjects. The proposed method was evaluated on two datasets: one containing healthy young, healthy old, and demented subjects, and the other with only healthy old and demented subjects. In a subsequent study, the same authors applied a similar five-step process on fMRI data, but modified the Random Forest algorithm by introducing majority and weighted voting schemes. In their recent work , they used a comparable method, dataset, and stage, with the aim of providing a supervised method for diagnosing and monitoring the progression of AD using information extracted from fMRI experiments.

Some researchers focus on combining different diagnostic data to improve classification and prediction accuracy. Sandhya Joshi et al, combined MMSE and FAQ tests with four machine learning algorithms on a sample of 860 patients and control subjects to achieve two objectives: improving diagnosis accuracy while saving time and cost, and enhancing sensitivity and specificity. Their dementia classification model consists of four steps: data collection, data preprocessing, feature selection, and classification. The goal of this model is to classify various stages of dementia using selected machine learning methods, including C4.5, Naive Bayes, and Random Forest classifiers. Additionally, neural network methods such as Multilayer Perceptron were used to evaluate the different stages of dementia. In their proposed system, the machine learning algorithms were trained and tested on separate datasets from both MMSE and FAQ tests. The algorithms learned to classify four dementia stages from MMSE data (severe, mild cognitive impairment, moderate AD, and normal), as well as three stages from FAQ data (severe, mild cognitive AD, and normal). These stages were represented by the Clinical Dementia Rating (CDR) score, which is used to determine the presence and severity of dementia. A CDR score of 0 indicates "No Cognitive Impairment," while scores of 0.5, 1, 2, and 3 correspond to "Questionable," "Mild," "Moderate," and "Severe dementia," respectively.

A dataset of MRI data was used, which includes a cross-sectional collection of 218 subjects aged 18 to 59 years and 198 subjects aged 60 to 96 years. The authors proposed a novel automated method for classifying individuals as either Alzheimer's patients or healthy controls using MRI scan data. This method integrates independent component analysis (ICA) with voxel of interest for classification through a five-step process. The steps include preprocessing the MRI data, segmenting the brain's gray matter, applying independent component analysis for decomposition, extracting the voxel of interest, and finally classifying

the data using a support vector machine. After conducting their experiments, the authors found that their proposed method achieved better classification results compared to other related works presented in their study.

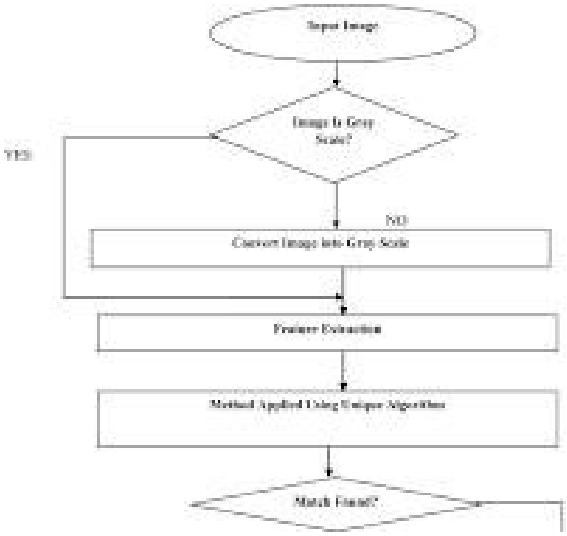
Systems designed to assist Alzheimer's Disease (AD) patients often include reminders for medication or daily activities, as well as memory aids to help trigger recollection of forgotten information. These forms of assistance are commonly implemented in solutions proposed by computer scientists to support AD patients the authors explored the use of Ambient Intelligence (AmI) to aid AD patients. They proposed a mobile application that provides memory triggers to help patients remember the sequence of steps for performing simple, daily tasks. Additionally, the system notifies caregivers if there are significant deviations from the patient's usual routine. The proposed system consists of three main components: the patient terminal, the caregiver station, and the server station. The patient terminal, which operates as a multi-agent system, collects and sends the patient's context information to the web server and provides assistance to the patient based on the acquired data. The web server records, processes, and analyzes the information, while caregivers receive updates about the patient either directly from the terminal or from the server. Overall, the system offers memory trigger assistance to patients and sends reminders for tasks to be completed. Caregivers, on the other hand, benefit from the system by gaining remote communication with the patient, receiving up-to-date information, and receiving alerts in cases of emergency or danger.

3. Research methodology

Research is a structured and organized process aimed at finding solutions to specific problems in a systematic manner. The theoretical framework underlying this process is known as methodology. Methodology refers to "the systematic study of methods that are, can be, or have been applied within a discipline" . In this study, both qualitative and quantitative approaches are employed. The qualitative approach is used to explore and understand previous work, assess various machine learning algorithms, and examine different AD diagnostic techniques. On the other hand, quantitative research is applied to compare the evaluated algorithms and identify the most effective one. The selection of the optimal algorithm is carried out through data collection, feature selection, data preparation, and experimentation using machine learning tools.

3.1 Research process

The research process in this thesis follows a series of steps. In the initial phase, relevant literature is reviewed to help shape the research design. Key risk factors for the disease are identified, and various machine learning classifiers are assessed. The dataset is then prepared for the experiment, ensuring that the data is ready for analysis. In the next phase, the well-known machine learning tool, WEKA , is utilized to compare the evaluated algorithms and select the one with the highest classification accuracy. In the final phase, a Multi-Agent System (MAS) model is developed to further enhance classification accuracy, and its effectiveness is evaluated through simulations. The overall research methodology is outlined in Figure.



3.2 Problem definition

The increase in a number of visitors in the hospital and medical centers have created a workload on professionals, high number of patients' queues and extra cost in terms of time and money on both the patients and health institutions. There is also a massive stored patients' data at the hospitals and healthcare centers that registered and stored for patients' future follow up and treatment. Usually the data is only used when it is necessary to refer a specific patient's medical history. The collected data at university of Washington can be a good example and there is a need for usage of this NACC dataset to predict AD stage progression, minimize workload, and unnecessary visits.

3.3 Literature review

The literature review involves examining various research papers related to the defined problem area or relevant topics. Several medical and computational studies are reviewed to help formulate the research questions, identify current AD diagnosis techniques and risk factors, and evaluate different machine learning algorithms and tools.

3.4 Research questions

The following three research questions are raised to fill the gap identified in the literature review.

- ✓ Which machine learning algorithm is significantly better to classify NACC's medical health and cognitive function data for AD Diagnosis?
- ✓ Does the combination or separation of medical history and cognitive function data affect the classification accuracy?
- ✓ How to improve the classification accuracy using agent system?
- ✓ In order to answer these research questions literature review and experiments are conducted.

3.5 Hypothesis

Null Hypothesis H0_1: There is no significance difference between all candidate algorithms.
Alternate Hypothesis H1_1: There is a significantly better algorithm in the classification of NACC dataset.
Dependent Variable: accuracy
Independent Variables: NACC dataset, candidate algorithms, and experimental environment

Null Hypothesis H0_2: The separated existing diagnosis techniques , medical history or cognitive function, data are not significantly different in prediction accuracy.
Alternate Hypothesis H1_2: The combined diagnosis techniques, medical history and cognitive function, data gives a better prediction accuracy.
Dependent Variable: accuracy
Independent Variables: NACC dataset, candidate algorithms, and experimental environment.

3.6 Data Preparation

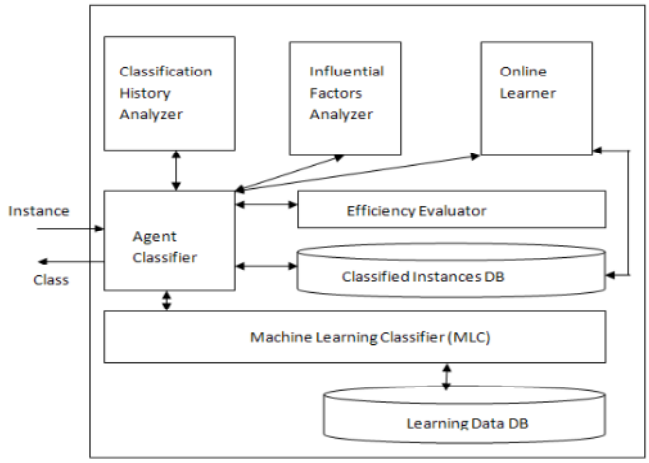
For this study, the data preparation will be carried out in three stages: data collection and overview, attribute evaluation and selection, and data preprocessing. The dataset was obtained from the National Alzheimer's Coordinating Center (NACC) at the University of Washington, covering the period from May 2005 to August 2011. It is provided in CSV (Comma Separated Value) format and contains over 47,000 instances along with 396 attributes, some of which include noise, errors, and missing values that could adversely affect the efficiency of data mining. Therefore, data preprocessing will be applied to address issues such as missing values, noise, and inconsistencies within the data.

3.7 Data preparation and Experimentation

The collected dataset is cleaned and preprocessed using MySQL Workbench and MS Excel VBA to eliminate noise, errors, and missing data that may impact the quality of classification. Following this, two separate experiments are carried out using 10,000 refined instances of patient data. The first experiment aims to identify the most effective algorithm among the candidate algorithms. The second experiment is conducted using the same NACC dataset but focusing solely on the cognitive function data. Simulations are performed in the second experiment to address the final research question.

4.Experimental Design

With the increasing affordability and widespread availability of data storage devices, there is a growing trend of storing large volumes of data for extended periods. Organizations often accumulate substantial amounts of data on their storage systems. For example, companies store their customers' daily transactions, government institutions retain official records, and healthcare organizations keep patient information for future reference. These stored data are used by their respective owners in various ways. To facilitate the retrieval and utilization of this data, various algorithms and software have been developed. Typically, these algorithms are designed for retrieving and presenting the actual recorded data. However, by applying data mining techniques, it is also possible to uncover patterns within the stored data, generating new insights that go beyond the original recorded information.



4.1 Algorithm evaluation and selection

Data mining refers to the process of discovering meaningful patterns and insights from data. Various algorithms have been developed for data mining, including those based on visualization, machine learning, and statistical methods. Among these, machine learning-based data mining algorithms are particularly effective in generating easily interpretable patterns. "Machine learning is the study of computational techniques that enhance performance by enabling machines to learn from experience" . Several algorithms have been developed for different types of data mining tasks, such as classification, clustering, and association, the IEEE International Conference on Data Mining (ICDM) identified the top 10 most influential and widely used data mining algorithms. The selection process was carried out in three phases: in the first phase, researchers nominated algorithms they considered most influential and widely used, providing justification and citation references. In the second phase, nominations with fewer than 50 citations were eliminated. The final phase involved researchers voting on the algorithms that successfully passed the previous two phases.

k-means

The k-means algorithm is one of the most straightforward unsupervised learning algorithms. It is an iterative clustering method that groups the given instances into a pre-defined number of clusters . The number of clusters is specified beforehand and remains fixed throughout the process.

SVM

SVM is a statistical set of supervised learning algorithms. In SVM the candidate instance will be classified in to either of two possible classes. The classification is done by dividing examples into two classes which are separated by a clear gap between them. The gap between the categories is known as the optimal separating hyper plane.

Apriori

Identifying associations within data can be challenging and often a complex task. The Apriori algorithm is widely used for mining association rules from datasets. Association rule mining is a technique used to uncover hidden relationships within the data . The process involves first identifying frequent itemsets from the data and then generating the corresponding association rules.

EM

Expectation–Maximization (EM) algorithm is an algorithm which is mostly used to solve data clustering problems. It works by finding iteratively likelihoods between instances in the dataset. EM is a commonly used and more appropriate choice when there is a missing values

PageRank

PageRank is an algorithm used to assess the relative importance of documents on the World Wide Web. Initially, the algorithm was designed to evaluate the importance of a research paper based on how many other papers referenced it. Over time, this concept evolved to focus on measuring the importance of web documents or webpages. In the PageRank algorithm, the value of a web document is determined by the number of links directed toward it .

AdaBoost.

Boosting is a supervised ensemble learning classification algorithm that combines multiple weak learners to form a single strong learner. It aggregates the outputs of different models to classify an input instance. AdaBoost is a widely used ensemble classifier, known for its strong performance due to the variety of weak classifiers it incorporates . Additionally, numerous studies suggest that AdaBoost is less prone to overfitting compared to other algorithms.

KNN (*k*-nearest neighbor algorithm)

KNN (K-Nearest Neighbors) is a classification algorithm that classifies an incoming instance based on the majority vote of its neighboring instances. The algorithm works by storing the entire training dataset and calculating the distance between the input instance and the samples in the training set. It is one of the most widely used algorithms for pattern recognition, exploratory data analysis, and solving data mining problems over a long period of time.

Naïve Bayes

Naïve Bayes algorithm is supervised learning classifier algorithm. It works by calculating the probabilistic likelihood of the given instance with the reference of the training dataset. It assumes the attributes of the datasets are independent to each other

CART

Classification and regression trees (CART) is a classification machine learning algorithm which employees decision tree technique. It generates either regression tree or classification tree based on dependant variable. CART applies recursive partitioning mechanism in order to build the tree.

Among the top algorithms, C4.5, kNN, Naïve Bayes, SVM, CART, and AdaBoost are commonly used for classification in data mining. Classification algorithms are supervised learning methods employed to predict the class of a given instance by analyzing the patterns in attribute values and their corresponding class in the training dataset. Many studies have applied classification algorithms to various aspects of healthcare, such as patient diagnosis, treatment, and support,the mental impairment levels of Alzheimer's Disease (AD) patients can be classified into multiple stages, which represent the severity of the disease. Additionally, the data obtained from the National Alzheimer's Coordinating Center (NACC) at the University of Washington contains a vast amount of real-world diagnostic data of AD patients. The aforementioned algorithms are potential candidates among the top 10 machine learning algorithms. However, due to CART's lower computational efficiency in rule extraction, longer rule lengths, and instability of decision trees, it has been excluded from consideration . To determine the most suitable algorithm for the NACC dataset, further comparisons using data mining tools are necessary.

KNN, Naïve Bayes, SVM, CART and AdaBoost are used for classification in data mining process. Classification algorithms are supervised learning algorithms used for class prediction of a given instances after analyzing the pattern of attribute values and their class in the training dataset. In many researches, classification algorithms used to assist different aspect of health science like patient diagnosis,

treatment and assistance.AD patients' mental impairment level can be classified into number of stages (classes). These stages indicate the severity level of the disease on the patient. On the other hand, the data acquired from National Alzheimer's Coordinating Center University of Washington contains large number of instances contains real data of AD patients' diagnosis and their result. The above mentioned algorithms can be a candidate among those top 10 machine learning algorithms . Nevertheless, since CART has low computation efficiency of extraction rules and long rule length and unstable decision tree excluded from the candidacy . To identify the best algorithm among the candidates with regards to its suitability with NACC dataset, further comparison using data mining tool is required.

4.2 Data mining tools

Data Mining is the process of automatic or semi-automatic extraction of useful unknown information and patterns from real world different sized and complex data. There are 12 popular open source data mining software and application used to perform different data mining tasks such as classification and clustering. These systems, such as ADAM, TANAGRA, WEKA, KNIME, AlphaMiner, Databionic ESOM, Gnome Data Miner, Mining Mart, MLC++, Orange, Rattle and YALE, are developed in either C++, Java, Python, R or C++ and Python programming languages. Except TANAGRA and MLC++, all the other systems can operate on Linux, Mac and Windows platform. The frequency and time of latest updates are low and medium in Gnome Data Miner and ADAM respectively but the rest updates highly. These data mining software are released under GPL (General Public License) except that of KNIME, Mining Mart, TANAGRA, MLC++ and ADAM. Based on their activity, license, language and platform as shown above, most of the systems have similar characteristics. In real world, most of the data sources have different formats. In the view of data source and usability, AlphaMiner, Rattle, WEKA and YALE have more ability to access different application data format with better human interaction and interoperability but Rattle has less capable for data preprocessing. Out of the selected three data mining, (AlphaMiner, WEKA and YALE) WEKA is widely used and popular platform for sharing algorithms

4.3 WEKA

WEKA, (Waikato Environment Knowledge Analysis), is an open source GPL data mining software package written in Java that provides a collection of machine learning algorithms for a data mining task and available in http://www.cs.waikato.ac.nz/ml/weka/.

Since WEKA being placed on Source-Forge in April 2000 until 2007, it was downloaded more than 1.4 million times , but in the 6 months of 2007, there were the average of 21,152 downloads per month. The main WEKA GUI has four different interfaces: the Explorer, Experimenter, Knowledge Flow, and simple CLI (command line) mode to access WEKA.

WEKA offers two main modes: the data exploration mode (Explorer) and the experiment mode (Experimenter). The WEKA Explorer is designed for data exploration and provides features for managing data, such as loading datasets, selecting algorithms, and assigning training and testing data. It also supports classification and reporting of results. The interface consists of six tabs, each serving different purposes: Preprocess (for selecting and modifying datasets), Classify (for training classifiers or performing regression), Cluster (for clustering the data), Associate, Select Attributes, and Visualize.

The WEKA Experimenter allows for the comparison of algorithm performance across various evaluation metrics, helping to identify the most effective classifier using the setup, run, and analyze tabs. The setup tab is used to select the dataset, classifiers, cross-validation methods, and other configurations. The experiment mode facilitates large-scale experiments, storing results in a dataset for further analysis.

However, the results from both the Explorer and Experimenter modes can be influenced by noise, errors, and missing data. Therefore, it is essential to preprocess the NACC dataset before running it through either the Explorer or Experimenter modes to ensure accuracy.

4.4 Data preparation

Data Preparation (DP) is the critical part of predictive algorithm in successful projects. DP is an important part of the mining process and it is 60% more time consuming than the whole data mining process.

- The real world data might be impure due to missing value (empty attribute value), noisy data (having error) and inconsistence data which DP helps to clean it
- DP generates a smaller dataset than original and this can significantly improve the efficiency of data mining by selecting relevant attribute and reducing instances
- DP also generates quality data by filling missed values, correcting errors and others

5.REFERENCES:

[1] S. Joshi, D. Shenoy, G. G. Vibhudendra Simha, P. L. Rrashmi, K. R. Venugopal, and L. M. Patnaik, "Classification of Alzheimer's Disease and Parkinson's Disease by Using Machine Learning and Neural Network Methods," in *Machine Learning and Computing (ICMLC), 2010 Second International Conference on*, 2010, pp. 218 –222.

[2] "2011 Alzheimer's disease facts and figures," *Alzheimer's and Dementia*, vol. 7, no. 2, pp. 208–244, Mar. 2011.

[3] R. Kohavi, G. John, R. Long, D. Manley, and K. Pfleger, "MLC++: a machine learning library in C++," in *Tools with Artificial Intelligence, 1994. Proceedings., Sixth International Conference on*, 1994, pp. 740 – 743.

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

[5] G. Holmes, A. Donkin, and I. H. Witten, "WEKA: a machine learning workbench," in *Proceedings of ANZIIS '94 - Australian New Zealnd Intelligent Information Systems Conference, 29 Nov.-2 Dec. 1994*, New York, NY, USA, 1994, pp. 357–61.

[6] J. Hua, "Study on the application of rough sets theory in machine learning," in *2008 2nd International Symposium on Intelligent Information Technology Application, IITA 2008, December 21, 2008 - December 22, 2008*, Shanghai, China, 2008, vol. 1, pp. 192–196.

[7] N. R. Jennings, K. Sycara, and M. Wooldridge, "A Roadmap of Agent Research and Development," *Autonomous Agents and Multi-Agent Systems*, vol. 1, no. 1, pp. 7–38, Mar. 1998.

[8] J. Tweedale, N. Ichalkaranje, C. Sioutis, B. Jarvis, A. Consoli, and G. Phillips-Wren, "Innovations in multi-agent systems," *Journal of Network and Computer Applications*, vol. 30, no. 3, pp. 1089–1115, Aug. 2007.

[9] S. Joshi, P. Deepa Shenoy, K. R. Venugopal, and L. M. Patnaik, "Evaluation of different stages of dementia employing neuropsychological and machine learning techniques," in *Advanced Computing, 2009. ICAC 2009. First International Conference on*, 2009, pp. 154 –160.
[10] F. Richard and P. Amouyel, "Genetic susceptibility factors for Alzheimer's disease," *European Journal of Pharmacology*, vol. 412, no. 1, pp. 1–12, Jan. 2001.
[11] A. V. Suhanov, P. I. Pilipenko, A. D. Korczyn, A. Hofman, M. I. Voevoda, S. V. Shishkin, G. I. Simonova, Y. P. Nikitin, and V. L. Feigin, "Risk factors for Alzheimer's disease in Russia: a case–control study," *European Journal of Neurology*, vol. 13, no. 9, pp. 990–995, Sep. 2006.
[12] The Duy Bui, Duy Khuong Nguyen, and Tien Dat Ngo, "Supervising an unsupervised neural network," in *2009 First Asian Conference on Intelligent Information and Database Systems, ACIIDS, 1-3 April 2009*, Piscataway, NJ, USA, 2009, pp. 307–12.
[13] S. Chakrabarti, "Data mining for hypertext: a tutorial survey," *SIGKDD Explor. Newsl.*, vol. 1, no. 2, pp. 1–11, Jan. 2000.
[14] M. S. Mouchaweh, "Learning in Dynamic Environments: Application to the Identification of Hybrid Dynamic Systems," in *Machine Learning and Applications (ICMLA), 2010 Ninth International Conference on*, 2010, pp. 555 –560.
[15] R. Nock and F. Nielsen, "Bregman Divergences and Surrogates for Learning," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 31, no. 11, pp. 2048 –2059, Nov. 2009.
[16] B.-F. Zhang, J.-S. Su, and X. Xu, "A Class-Incremental Learning Method for Multi-Class Support Vector Machines in Text Classification," in *Machine Learning and Cybernetics, 2006 International Conference on*, 2006, pp. 2581 –2585.
[17] M.-L. Zhang and Z.-H. Zhou, "Adapting RBF Neural Networks to Multi-Instance Learning," *Neural Processing Letters*, vol. 23, no. 1, pp. 1–26, 2006.
[18] M. Sayed Mouchaweh, "Semi-supervised classification method for dynamic applications," *Fuzzy Sets and Systems*, vol. 161, no. 4, pp. 544–563, Feb. 2010.

