

BIG DATA ANALYSIS OF ALZHEIMER'S DISEASE FOR EARLY PREDICTION

Dissertation submitted in partial fulfillment of the requirement for the degree of

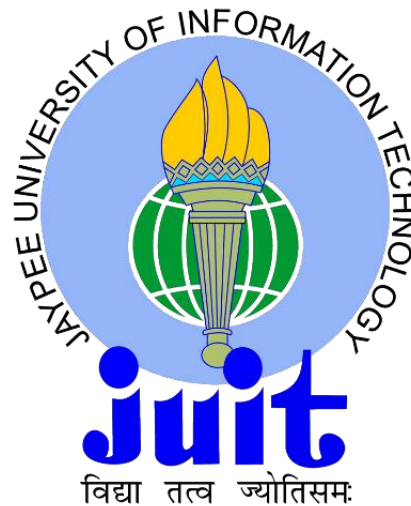
BACHELORS OF TECHNOLOGY IN BIOTECHNOLOGY

By

Nayanika Sharma (171835)

Under the Supervision of

Dr. Rahul Shrivastava



**DEPARTMENT OF BIOTECHNOLOGY AND BIOINFORMATICS
JAYPEE UNIVERSITY OF INFORMATION TECHNOLOGY WAKNAGHAT,
SOLAN (H.P)**

MAY 2021

TABLE OF CONTENTS

DECLARATION.....	5
SUPERVISOR’S CERTIFICATE	6
ACKNOWLEDGEMENT	7
ABSTRACT.....	8
INTRODUCTION.....	9
REVIEW OF LITERATURE	
WHAT IS ALZHEIMER’S DISEASE	10
BELIEVED CAUSES OF AD	13
PARAMETERS INFLUENCING AD	17
MODIFIABLE RISKS	18
UNUSUAL GENETIC MODIFICATIONS.....	22
ALZHEIMER'S DISEASE CONTINUUM	23
BIG DATA	25
BIG DATA CHALLENGES IN AD RESEARCH	27
MATERIALS AND METHODS	29
BIG DATA ANALYTICS FRAMEWORK PROPOSED	30
DATA COLLECTED.....	37
METHODOLOGY FOR LOGICAL REGRESSION.....	41
CODE OF LOGISTIC REGRESSION FOR FINDING TOP TEN FEATURES	42

RESULT AND CONCLUSION..... 47
REFERENCES..... 49

LIST OF SYMBOLS AND ACRONYMS

AD:Alzheimer's Disease

MCI: Mild Cognitive Impairment

TBI: Traumatic Brain Injury

APP:Amyloid precursor protein

EHR: electronic records of health

ADNI:Alzheimer's Disease Neuroimaging Initiative

BDA: Big Data Analytics Tools

ML: Machine learning

HDFS: Hadoop Distributed File System

NHIS-NSC: National Health Insurance Service-National Sample Cohort

HDFS: Hadoop Distributed File System

PCA: Principle Component Analysis

LIST OF FIGURES

S.NO	DESCRIPTION	PAGE NUMBER
1	Alzheimer's Disease Brain Pathology	11
2	Changes in morphology of brain and neurons in AD patients	11
3	Changes in morphology in brain and neurons of AD patients	12
4	Amyloid plaque accumulation in brain of AD patient	12
5	Tau tangles and their effect on brain morphology	15
6	APP Fragmentation with and without AD progression	16
7	AD Continuum	23
8	Six V's of Big Data	26

LIST OF TABLES

S.NO	DESCRIPTION	PAGE NUMBER
1	Alzheimer's Disease Processes Through Distinct Stages	24
2	Databases available for AD research across the globe	36
3	Top 10 logistic regression features and weights	46

DECLARATION

I hereby declare that the literature review work reported in the B.Tech thesis entitled “*BIG DATA ANALYSIS OF ALZHEIMER’S DISEASE USING SORTING MODEL*” submitted at **Jaypee University of Information Technology, Wagnaghat, India**, is an authentic record of work done by me (Nayanika Sharms-171835) carried out under the supervision of **Dr. Rahul Shrivastava** (Associate Professor) Department of Biotechnology and Bioinformatics. I have not submitted this work elsewhere for any other degree or diploma.



Nayanika Sharma (Enrollment number: 171835)

Department of Biotechnology and Bioinformatics

Jaypee University of Information Technology, Solan

Wagnaghat.

Date: 25TH May 2021

SUPERVISOR'S CERTIFICATE

This is to certify that the work reported in the B.Tech thesis entitled “*BIG DATA ANALYSIS OF ALZHEIMER'S DISEASE USING SORTING MODEL*” submitted by **Nayanika Sharma** during their 8th semester in June 2021 in fulfillment for the major project in Biotechnology of Jaypee University of Information Technology, Solan has been carried out under my supervision. This work has not been submitted partially or wholly to any other University or Institute for the award of any degree or appreciation.



(Dr. Rahul Shrivastava)

Dr. Rahul Shrivastava

Associate Professor

Department of Biotechnology and Bioinformatics

Jaypee University of Information Technology (JUIT)

Waknaghat, Solan, India – 173234

Date: 25TH May 2021

ACKNOWLEDGEMENT

We take this opportunity to express our first and foremost gratitude to our “DEPARTMENT OF BIOTECHNOLOGY AND BIOINFORMATICS” for the confidence bestowed upon us and entrusting our project title “*BIG DATA ANALYSIS OF ALZHEIMER’S DISEASE USING SORTING MODEL*”.

At this juncture, with proud privilege and profound sense of gratitude we feel honored in expressing our deepest appreciation to **Dr.Rahul Shrivastava**, for being a lot more than just a supervisor and going beyond the call of duty in our guidance, support, advice, and motivation throughout. He has been the source of inspiration of come what may, these issues cannot bring you down. Sincere thanks for his insightful advice, motivating suggestions, invaluable guidance, help and support in successful completion of this major project and also for his constant encouragement and advice throughout our minor project work.

Special thanks to our parents for their infinite patience and understanding and project partners for the constant support and most importantly God, who in his mysterious ways, always made things work out in the end.

In gratitude,

A handwritten signature in blue ink that reads "Nayanika". The signature is written in a cursive style with a horizontal line underlining the name.

Nayanika Sharma (Enrollment number: 171835)

ABSTRACT

The devastating condition of Alzheimer's Disease (AD) affects up to 50 million of us globally. The biggest risk factor for AD is age and for the first time in history, there are more people over the age of 65, than there are under the age of five, making AD one of the greatest challenges of our time. Dr Alois Alzheimer discovered the disease over 120 years ago nevertheless, the current recovery rate for AD patients is zero percent.

Prevention is the only approach to prevent this disease. There are a large number of health transactions and AD studies generated by the data. Researchers use these patient health data to measure the level of cognitive deterioration in order to determine different stages of dementia. Mild Cognitive Impairment (MCI) might be seen as an AD-to-normal cognition intermediate. In general, people with MCI violate development of AD. This information from MCI patients is crucial to the forecast of AD but is too complex and voluminous for standard approaches to be handled and analysed. Data mining provides methods and technology for turning these complicated data into usable decision-making information.

This research addresses the application of data mining for the production of an automated learning model, based on broad administrative health data in the prediction of AD risk. We sorted the data properly first to establish the functional parameters by using a sorter model to the vast data in the database provided. In order to prevent an AD incident we have trained and tested random forests, a vector machine and logistical regression. Then, logistic regression was used to estimate parameters and dynamically forecast the future result and risk of patient dementia by functional parameters

Keywords: *Alzheimer's Disease, Functional Data analysis, Logistic regression, Prediction model*

INTRODUCTION

Alzheimer's is a complex mix of cognitive and neurological symptoms that cause brain deterioration, overall resulting in death. So with over 15 million people globally suffering from the disease, actually 4.1 million in India alone, there still is no cure. It's in part due to the fact that after billions of dollars and decades of research, we still don't know what exactly caused the disease. We know that clinical signs usually show up after age 65 with changes in the brain maybe years, or in some cases, decades that this makes diagnosis extremely difficult. Patients are funneled through this complicated and convoluted pathway for detection, which can sometimes take years and cost thousands of dollars. Imagine having to go through brain imaging scans cognitive assessment, and in rare cases, even having your spine punctured to confirm a diagnosis assessment, and in rare cases, even having your spine punctured to confirm a diagnosis. These modalities are expensive, they require specialized training, and they're resource intensive. But most importantly, they always happen after symptoms are shown making this overall hurdle to get a diagnosis extremely high. And this is one of the problems with the way we look at the disease because the barrier to obtain a diagnosis is so high, we only start testing for the disease after symptoms show

REVIEW OF LITERATURE

What is Alzheimer's Disease?

Dr. Alois Alzheimer's, a German psychiatrist, originally documented the symptoms in 1901, after noticing that a particular hospital patient was having some unusual issues, such as difficulty sleeping, slowed memory, abrupt changes, and increased bewilderment. When the patient died, Alzheimer was able to perform an autopsy and test his theory that her symptoms were caused by structural abnormalities in the brain. Under the microscope, he discovered apparent changes in brain tissue, such as misfolded proteins known as plaques and neurofibrillary tangles. The plaques and tangles combine to degrade the brain's structure. Plaques form when a certain enzyme slices another protein in the lipid membrane enclosing nerve cells, resulting in beta amyloid proteins, which are sticky and have a tendency to cluster together. This clumping causes the formation of blocks, limits signalling and hence communication between cells, and appears to stimulate immunological response that cause the elimination of damaged nerve cells. Neurofibrillary tangles are made of a protein called tau in Alzheimer's disease. Nerve cells of the brain contain, among other things, a tube network that acts as a highway for food molecule. The tau protein usually ensures that these tubes are straight, so that molecules are freely transmitted. The protein collapses into twisted strands or pulverisation during Alzheimer's disease. The tubes disintegrate, nutrients are blocked by nervous cell death. In a region called the hippocampus, the destructive pairing of plaques and tangles begins which forms memories. This is why short-term loss of memory is typical for Alzheimer's first symptom. The proteins invade other parts of the brain and create unique changes that signal different stages of the disease.

The proteins on the front of the brain are destructive for the processing of logical thinking. Next, they move to an area that regulates emotions and changes the mood. They lead to paranoia and hallucination at the top of the brain. Once they reach the rear of the brain, they work together to erase the deepest of minds. They work together. In the end, the heart rate and breathing control centres are also being overstretched and cause death.

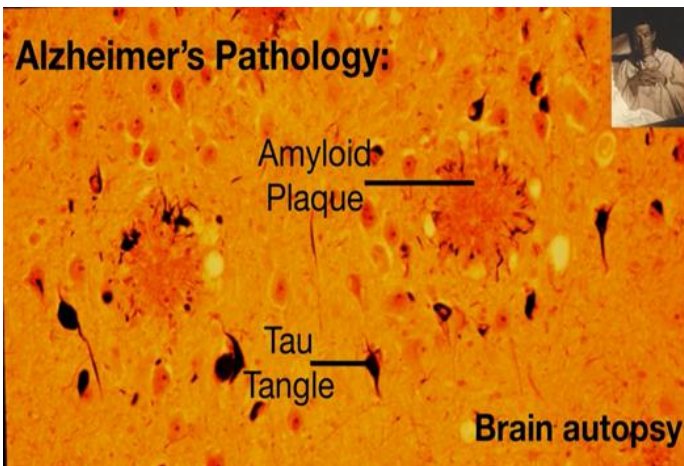


Fig 1: Alzheimer's Disease Brain Pathology(22)

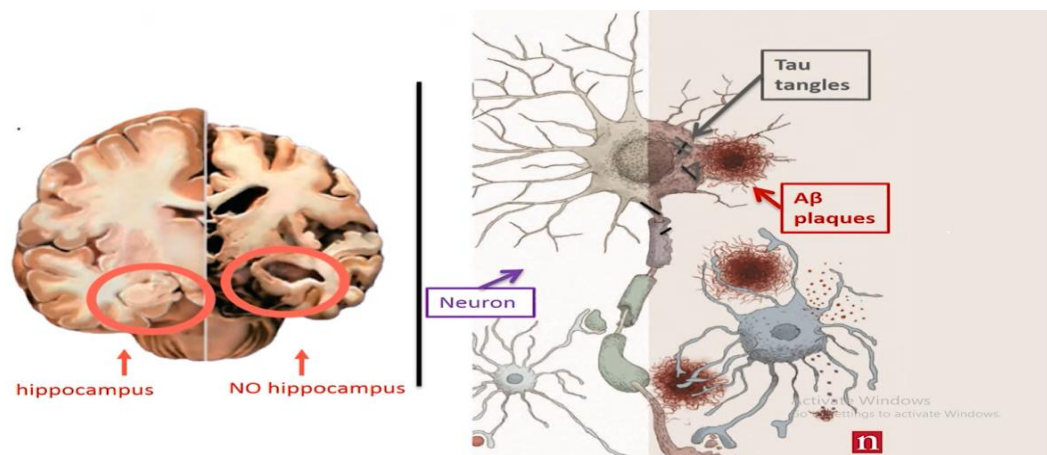


Fig 2: Changes in morphology of brain and neurons in AD patients(22)

Fig: Changes in morphology of brain and neurons in AD patients.

Detect and prevent Alzheimer's disease before memory loss | Bernard Hanseuw |

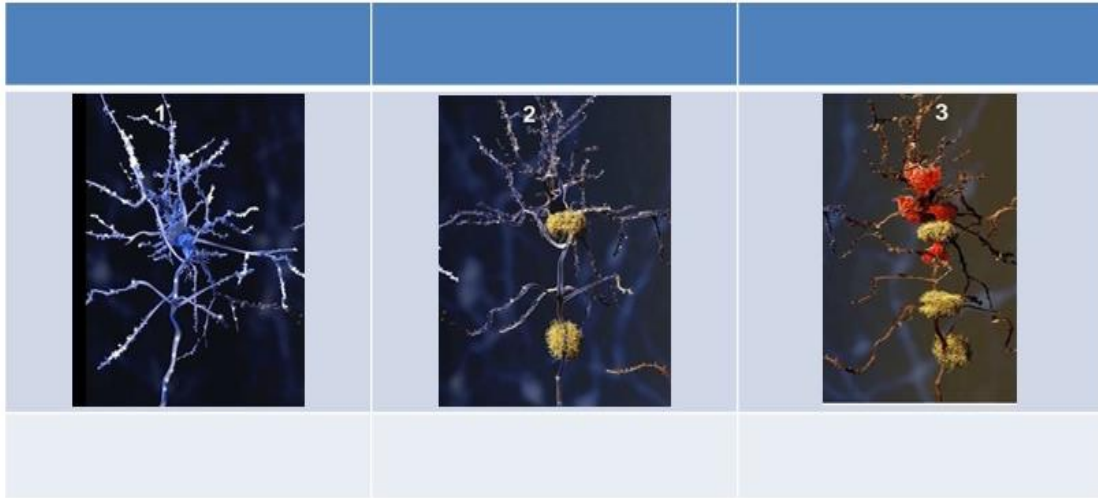


Fig 3: Changes in morphology in brain and neurons of AD patients(19)

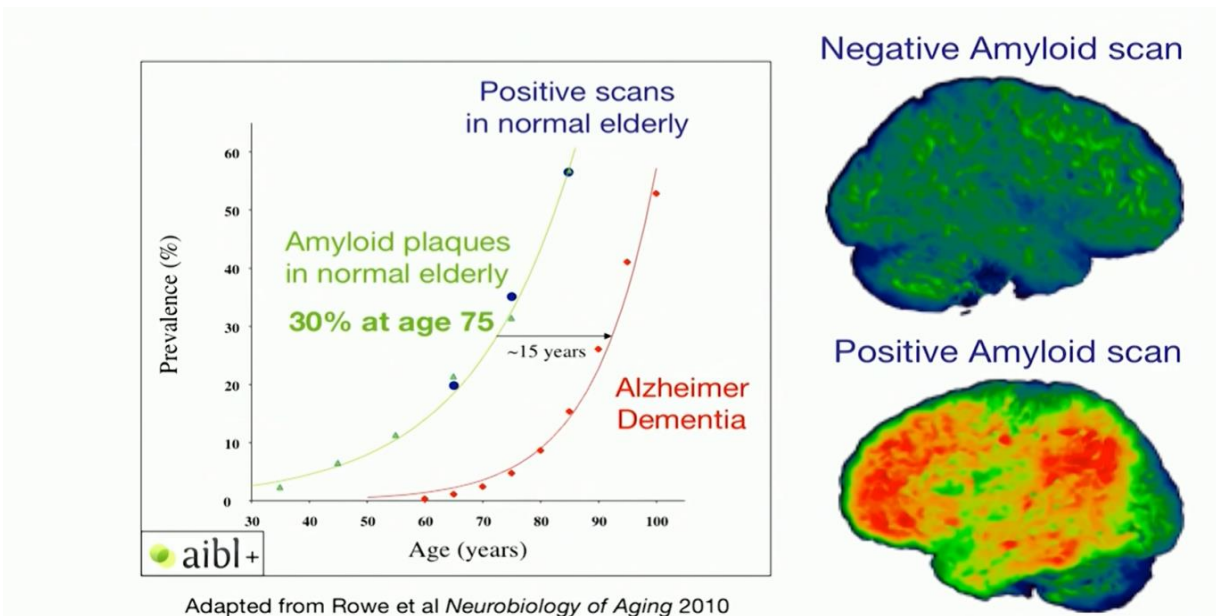


Fig 4: Amyloid plaque accumulation in brain of AD patient(19)

Believed causes of AD:

Let's start with what we now understand about Alzheimer's neurology. The synapse is the connection of neurotransmitters between two neurons. This is where signals or communication occurs. This is where we think, feel, see, hear and remember, and all the synapses are done. Neurons also release a small peptide called amyloid beta during the communication process, in addition to releasing neurotransmitters such as glutamate into the synapse. Normally, amyloid beta is removed and the chemical reasons of Alzheimer's are still discussed by microglia, the janitor cells of our brain. Most neuroscientists believe that the disease begins when amyloid beta begins to build, is released too much or not enough, and synapses containing amyloid beta begin to accumulate. It attaches to itself, which forms sticky masses called amyloid plaques.

This earliest illness phase can already be detected in the brains of forty years of age with the existence of amyloid plaques accumulating. The only way we could be certain of it is with a PET scan, as you are happily ignorant of any memory, language or cognition deficiencies at this moment. It is believed that the amyloid plaque build up takes a minimum of 15 to 20 years and this is a tipping point before a molecular cascade is triggered which produces the disease's medical symptoms. Before the tip, you could recall stuff like I'm coming to this room or what's his name? Or where have I placed my keys? Yeah, now, because I know that half of you did at least one of them last 24 hours, for you all are starting to frighten up again. All of

these are regular forgetfulnesses. In fact, I would argue that these examples may not even involve your memory, because you didn't care where you first put your keys. The memory, language and cognition failures are different following tipping. You can find them at the fridge rather than finally discover your keys in your coat pocket near the door on the table, or locate them something new.

What are they for? What are these?

When amyloid plaques collect to this degree, microglia cells become hyperactivated, and release chemicals causing inflammation and cellular damage, we believe that the synapse itself could begin to clear up. An important brain transport protein called tau becomes hyperphosphorylated and turns into something called tangles that hamper the inner neurons. We have huge inflammation and tangles in the middle of stage AD and all of which induce war and cellular death. Many scientists gambled big on the easiest approach to keep the tipping point of amyloid plaques. In short, drug research focuses mostly on the development of a molecule that prevents or reduces the buildup of amyloid plaque. So it's probably a preventive drug that will treat Alzheimer's. Before the cascade is activated, we will have to take this medication, before leaving our keys inside the refrigerator. We believe that this is why these types of medications in clinical trials have failed to date. Not because science has been established. But since participants had already been symptomatic in these trials. It was too late. It was too late.

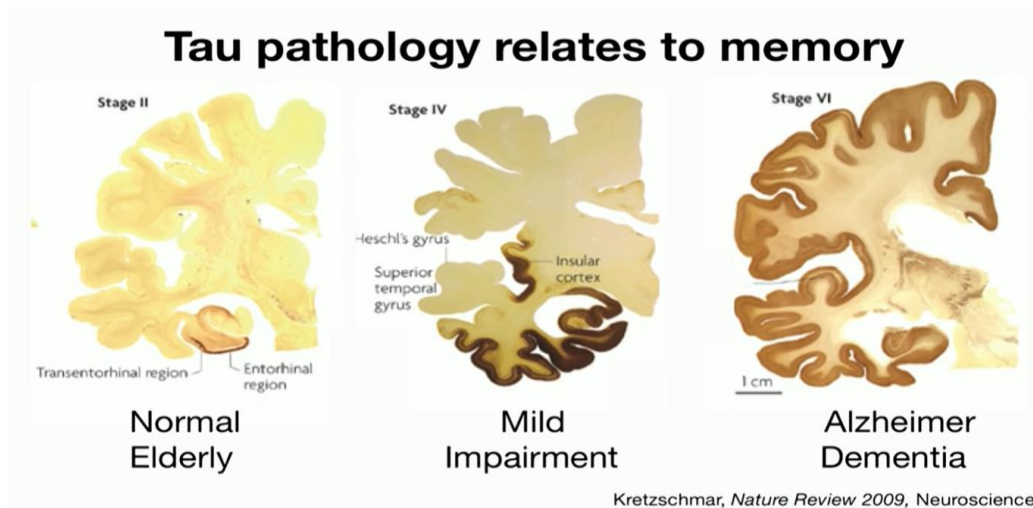


Fig 5: Tau tangles and their effect on brain morphology(20)

Another theory states:

At the heart of AD is a master switch called APP for Amyloid Precursor Protein. APP is a sensor for numerous factors such as hormones, nutrients, inflammation, information and growth factors. When these are optimal, APP is cleaved by molecular scissors at a single site to produce two peptide or protein fragments. sAPP α and α -CTF, these mediate growth and maintenance, you are putting your resources into memory formation and maintenance. On the other hand, in the presence of pathogens, things like viruses, bacteria, fungi, toxins or reduced support APP is cut at 3 sites to produce or 4 different fragments. sAPP β , Amyloid beta (A β) which is the one we classically associated with Alzheimer's disease, Jasp and C31. These mediate protection and retraction. Your brain is now in protection mode, pulling back rather than growth mode. Therefore, what we call Alzheimer's disease is actually a protective

response to numerous different pathogens. It's essentially a scorched earth retreat. And the amyloid villain is not the cause of the disease after all, but actually a protective part of the immune system. So what flips this master Switch, switch? there are dozens of different things. Most patients have more than 10 contributors such as such infections like P gingivalis from our dentition, herpes simplex from our lips, toxins, such as metals, like mercury, or organic toxins like toluene, or bio toxins, which are toxins made by some mold species or reduce support from growth factors like brain derived neurotrophic factors or reduced hormones-estradiol, testosterone, thyroid, or reduced nutrients such as vitamin D, omega three fats, or leakiness of your gut lining, or metabolic abnormalities, such as diabetes and pre diabetes., all of these factors acting on our genetic backgrounds.

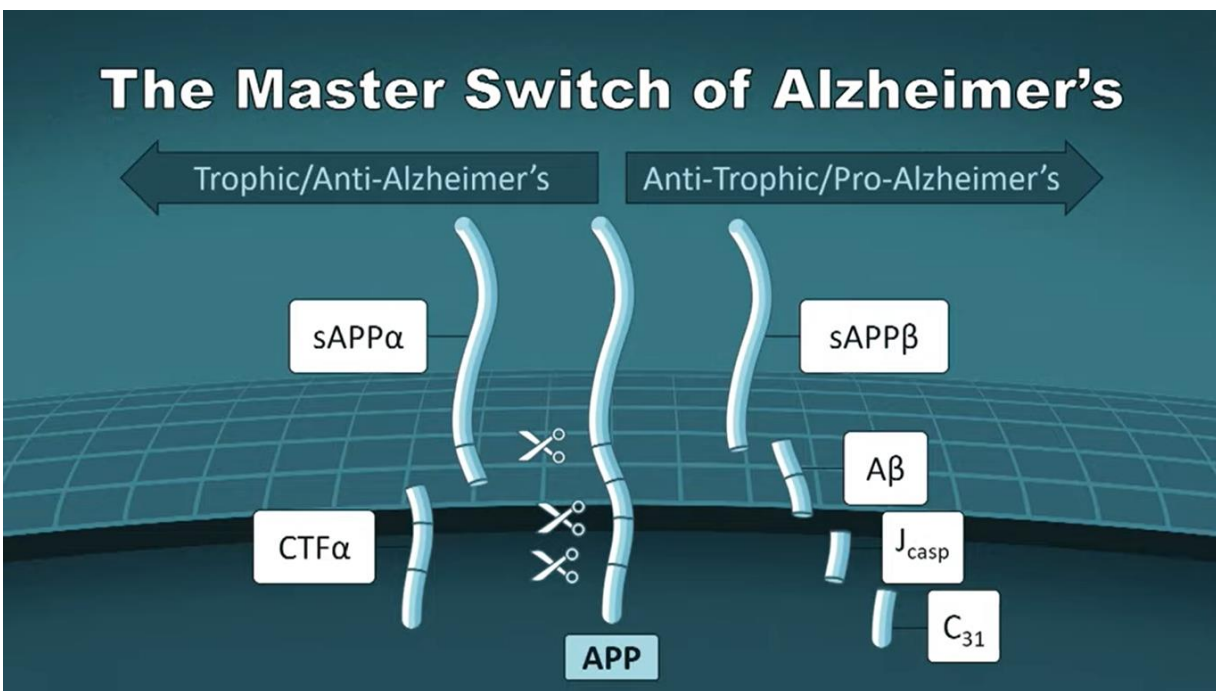


Fig 6: APP Fragmentation with and without AD progression(21)

Parameters Influencing AD

1. **Age-** Age is the biggest of three risk factors, with the vast majority of people living with Alzheimer's being 65 years or older. As shown on the section Prevalence (see page 17) with age 3 percent of people between 65 and 74, 17 percent between 75 and 84 and 32 per cent of those 85 years or older, having Alzheimer's Dementia, the percentage of those with Alzheimer's dementia increases considerably. It is vital to remember that dementia of Alzheimer's is not normal in age, and older age is not enough to develop Alzheimer's disease alone. (16)
2. **Family history-** A family history of Alzheimer's is not required for a person to have the disorder. Persons with an Alzheimer's father, brother or sister are more likely than those with no early-level relative with Alzheimer's to develop the condition. Those who are related to Alzheimer's more than a first degree are at greater risk. In families, inheritance (genetics) and shared environmental and lifestyle factors (e.g. availability to nutritious food and physical behaviour) could be a factor. If an individual has acquired the Risk APOE ϵ 4 gene, the risk related by an Alzheimer's family history is not entirely responsible. (17)
3. **APOE-e4 Gene-** The APOE gene provides the blood cholesterol protein pattern. All are inherited from each parent for one of three different versions of APOE gene e2, e3 or e4. The most popular e3 shape. The most common form is the e4,

whereas the least common form is the e2. With the e4 form, your chance of developing Alzheimer's rise compared to the e3 form, whereas e2 may reduce your risk compared to e3. People with one e4 copy have three times the chance of Alzheimer's being produced in comparison with the ones with the two e3 copies. Those who are inheriting two e4 copies have eight or 12 times the chance of Alzheimer's becoming present. More than with e2 or e3 versions of the APOE gene, it is more likely that Alzheimer develops in a younger age. A meta-analysis of the frequency of e4 for patients diagnosed in the United States with Alzheimer found that five sixty six percent were APOE-e4 gene copied, and eleven percent were two APOE-e4 gene copied (18).

Modifiable risk factors:

1. **Risk factors for cardiovascular disease:** The health of the brain and blood arteries is impacted. While the brain is only 2% of the body's weight, it requires 20% of the body's energy supply and oxygen. A healthy heart ensures that sufficient blood is poured into the brain, while healthy blood arteries enable the brain to function properly in oxygen- and nutrient-rich blood. A higher risk of dementia is also connected with many factors increasing the risk of cardiovascular disease. Smoking and diabetes are all factors. Some studies have suggested that the precursor of diabetes (disabled glucose) may potentially raise dementia risk. The age of some risk factors seems to be affecting the risk of dementia. For instance, obesity, hypertension, prehypertension in midlife are related with an increased risk of dementia (systolic

blood pressure from 120 to 139 mm Hg or diastolic pressure from 80 to 89 mm Hg) and high cholesterol. Obesity and high blood pressure after age are, however, connected with lowered dementia risk. Hypertension after 80 years of age can be the body's way of trying to increase cerebral blood when comorbidities, such as vascular disease, impair the blood supply. More research is necessary to understand the implications of some changeable risk variables with age. Researchers have concluded that the elements that protect the heart can also protect the brain and minimise the risk of acquiring Alzheimer's or other dementias by drawing upon the link between heart health and brain health. One of these elements seems to be physical activity. Emergence of information indicates that consumption of a balanced diet can be linked to lower risk of dementia as well as physical exercise. A healthy diet focuses on fruit, vegetables, whole grains, fish, chicken, nuts and legumes while restricting red fats and sugar. Researchers have begun to research health factor mixes and lifestyle behaviours (e.g., blood pressure and physical activity) to see whether mixed factors identify better the risk of Alzheimer's disease and dementia than individual risk factors. They also examine if intervening concurrently on numerous risk variables is likely to reduce the risk more than dealing with a single risk factor.

2. **Education**- Less risk for Alzheimer's and other dementias are education-people with longer years of formal education than people with less than formal schooling.⁸⁵⁻⁸⁹ Some researchers believe that having more years of education builds “cognitive reserve.” Cognitive reserve refers to the brain’s ability to make flexible and efficient use of cognitive networks (networks of neuron-to-neuron connections) to enable a person to continue to carry out cognitive tasks despite damaging brain changes,⁹⁰ such

as beta-amyloid and tau accumulation. The number of years of formal education is not the only determinant of cognitive reserve. Having a mentally stimulating job and engaging in other mentally stimulating activities may also help build cognitive reserve. Some scientists believe factors other than the number of years of formal education may contribute to or explain the increased risk of dementia among those with fewer years of formal education. These factors include an increased likelihood of having occupations that are less mentally stimulating.⁹¹⁻⁹⁴ In addition, having fewer years of formal education is associated with lower socioeconomic status,⁹⁵ which in turn may increase one's likelihood of experiencing poor nutrition and decrease one's ability to afford health care or medical treatments, such as treatments for cardiovascular risk factors. Finally, in the United States, people with fewer years of education tend to have more cardiovascular risk factors for Alzheimer's, including being less physically active⁹⁶ and having a higher risk of diabetes⁹⁷⁻⁹⁹ and cardiovascular disease.¹⁰⁰

- 3. Social and cognitive engagement:** Additional studies suggest that remaining socially and mentally active throughout life may support brain health and possibly reduce the risk of Alzheimer's and other dementias.¹⁰¹⁻¹¹¹ Remaining socially and mentally active may help build cognitive reserve, but the exact mechanism by which this may occur is unknown. More research is needed to better understand how social and cognitive engagement may affect biological processes to reduce risk [\(13\)](#).

4. **Traumatic Brain Injury (TBI) :**

Injury to the traumatic brain (TBI) TBI is the normal brain function disturbance produced by a blow or a blow to the skull or by a skull penetration by an alien item. The main causes of TBIs are falls and the item and engine crashes are affected. The duration of consciousness loss or post-traumatic amnesia¹¹³ and the initial Glasgow Coma scales of 15 points provide two different techniques to assess the severity of TBI. ¹¹⁴

- Mild (TBI), or Glasgow initial scoring of 13 to 15 minutes, is characterised by loss of consciousness or post-traumatic amnesia, and approximately 75% of TBI is mild.¹¹⁵

Moderate TBI is indicated by awareness loss of more than 30 minutes of post-traumatic amnesia but less than 24 hours, or by an initial Glasgow score of 9-12.

- Severe TBIs are known to result in a loss of consciousness of 24 or more hours, or a post-traumatic amnesia of an initial Glasgow score of 8 or less. Solid data shows the risk of various forms of dementia increasing in moderate and severe TBIs. ^{113,116-119} Those who suffer recurrent head traumas (e.g. boxers, soccer players and combat veterans) may have an even higher risk of illness, dementia and cognitive impairment. ¹²⁰⁻¹²⁹. The danger of repetitive head trauma or TBI can be lowered by ensuring that people have a well-lit, safe living environment, wearing seatbelts when on the go and wearing headgear on a bike, a snowmotive or any other open, undisturbed vehicle.

Unusual genetic modifications which increase the risk

1. **Genetic Mutations**-Due to mutations in one of the three genes, minor portions of Alzheimer's cases (estimated to be 1 percent or less) can occur. A genetic change is a change in the genetic structure sequence of the genes. The amyloid precursor protein (APP) and preseniline 1 and preseniline 2 gene are involved in these mutations. The disease is guaranteed for those who inherit Alzheimer's mutation to APP genes or presenilin 1 genes. An Alzheimer mutation to the presenilin 2 gene is 95% likely to acquire the illness. The disease is common in all cases. People with alzheimeric mutations in one of these three genes show signs of late-start disease, which occur at 65 or more years of age, occasionally as young as 30, although the vast majority of Alzheimer's people have a late-start disease. (14)

2. **Trisomy in Down Syndrome**: A copy of chromosome 21, one of the 23 human chromosomes, was born from the syndrome Down. Scientists do not know why persons with Down's syndrome risk more Alzheimer's, although the additional copy of the chromosome 21 may be associated. This gene contains the gene which codes for the creation of APP, which is broken into beta-amyloid fragments accumulating in plaques in persons with Alzheimer's. With an additional copy of chromosome 21 the quantity of betaamyloid fragments in the brain may increase. By the age of 40, most adults with Down syndrome had large levels in their brains of beta-amyloid plaques and tau tangles. As with other adults, a person with Down syndrome is more likely to have Alzheimer's symptoms than any other adult. In the 50s, approximately 30 percent

of adults with Down syndrome had Alzheimer's dementia, according to National Down Syndrome Society. 50% or more of those suffering from Down syndrome will get dementia in their lives (15).

Alzheimer's Disease Continuum

- Alzheimer's disease is known as the continuum of Alzheimer's disease, because of brain changes which are imperceptible by the individual involved in changes in his brain which create memory issues and finally physical incapacity.
- Three broad phases are presented in this continuum; Alzheimer's preclinical disease, mild cognitive impaired (MCI) from Alzheimer's and Alzheimer's dementia.

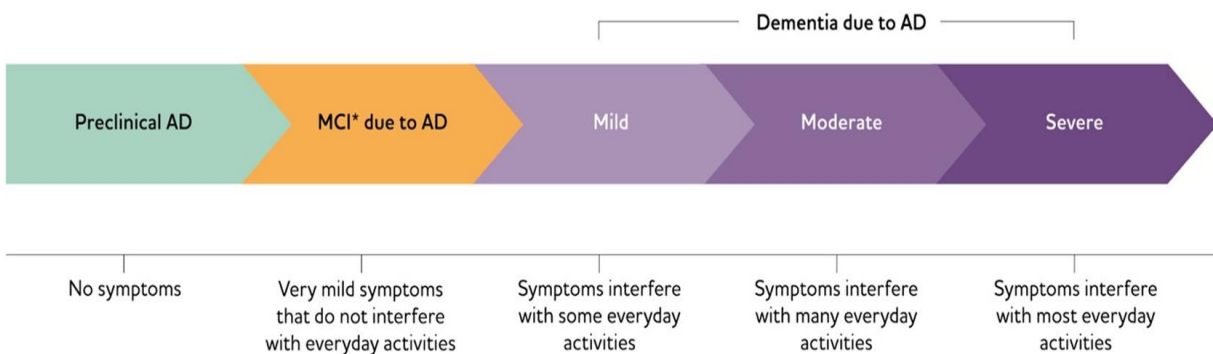


Fig 7:AD Continuum (23).

Table 1: Alzheimer’s Disease Processes Through Distinct Stages

Mild	Moderate	Severe
Memory loss	Behavioral, personality changes	Gait, incontinence, motor disturbances
Language Problem	Unable to learn/recall new info	Bedridden
Mood swings	Long-term memory affected	Unable to perform ADL
Personality changes	Wandering, agitation, aggression, confusion	Placement in long-term care needed
Diminished judgement	Require assistance w/Adl	

BIG DATA

In several fields the term 'big data' is often used and is always changing in definition. The "big data" defines the initial qualities as the three vs: volume, variety and speed (10). Firstly, the volume of data has dramatically aggregated in the last decade. For instance, in 2013 the US healthcare system already achieved 150 exabytes (10¹⁸). Big information in healthcare will approach the zettabyte (10²¹). (10²⁴). The second feature, variety, refers to the nature of large data heterogeneity. Data from many different sources can be collected such as micro-array data, imagery data, structured data (e.g. drugs, diagnostics) and unstructured data (e.g., clinical notes). The third feature, the speed, is the speed of data generation. The current sequencing techniques, for example, can produce thousands of sequence data every day. Systems with electronic records of health (EHR) may create between millions and billions of medical records every day. Other three features were also evaluated, in addition to those characteristics: variability, truth and value (11). Variability refers to data coherence across time. Veracity is important for large data since uncontrolled data, like unreliable ambulatory measurements, are sometimes available. When the obstacles of the big data analysis can be overcome, the value of big data for health care and patient can be achieved. Despite this new notion of big data, in the field of healthcare research there are no established definitions and attributes of big data. In this survey, we classified large-scale data as complex and heterogeneous, which in traditional ways, including: 1) data sets have been collected from more than one site, for example, AD Neuroimaging Initiatives (ADNI) to combine data. 2) Patients' databases from one or more EHR systems. 3) heterogeneous data sources such as clinical measurements, data on imaging, or computer-based data have been integrated (12) .

The six Vs of big data

Big data is a collection of data from various sources, often characterized by what's become known as the 3Vs: volume, variety and velocity. Over time, other Vs have been added to descriptions of big data:







VOLUME	VARIETY	VELOCITY	VERACITY	VALUE	VARIABILITY
The amount of data from myriad sources.	The types of data: structured, semi-structured, unstructured.	The speed at which big data is generated.	The degree to which big data can be trusted.	The business value of the data collected.	The ways in which the big data can be used and formatted.
					

Fig 8: Six V's of Big Data(21).

Big data are combinations of multi-modal data sets that may be managed separately, but, as a group, are too large for a single machine to handle well and accurately. ML faced the difficulty of effective processing and learning from Big Data with the surge in data output. The development of powerful Big Data Analytics Tools (BDA) is in this context an on-going requirement to handle huge quantities of various data that expand at an extremely quick rate. A broader strategy to adapting speed, volume and variety (19) is needed in AD research (19). Due to their available magnitude and diversity of heterogeneous data gathered with specific protocols using multi-modalities (e.g., MRI, SRMs, and neuropsychological results), the data (19). 'Volume (19) the first feature, has increased. The second characteristic of big data, that is 'Variety' (19) is the heterogeneity of data generated from many data sources. "Velocity" (19) is the third fundamental attribute of the BDA system to be the constantly rising data with

exponential and high processing speed. Due to the ever increasing volume and variety of data, a significant number of diverse data must be stored and processed. Thus, a popular 'Hadoop' ecosystem is built that provides broad, rapid and accurate distributed storage and processing. The Hadoop itself has four modules called "Hadoop common" supporting the other Hadoop modules; "Hadoop Distributed File System" (HDFS) for distributed storage and "Hadoop Yet Another Resourcnegotiator" (YARN) for job planning and cluster resource management.

Within the context of the knowledge of Big Data difficulties in AD research, a new and personalised Hadoop platform involves the administration of clinical data, processing and analytics of numerous multi-modal, neuro-chemical and neuropsychological data. This view is focused on the huge data generated from a variety of processes, including RIM, RMS and neuropsychological tests and targeted the development of a special Big Data Research Framework in response to contemporary scientific concerns. The new strategy is the first step in the early onset and advancement of AD by integrating morphological, metabolism and cognitive modifications.

“Big Data Challenges in AD Research”

In the last few decades extensive research in the ML area for Big Data has been carried out.

However, obstacles remain, including the following:

- **Large data size:** Large data size: Large data investigation in AD has a big challenge in gathering, storing and standarding large-scale information on various and complicated heterogeneous sources for further processing and high-speed analytics.

Data collected from multiple sources require standardised procedures of data collection, nomenclature of data and subsequent data sharing standards.

- **Feature extraction:** Giant aspect dimensionality is a common feature of large data, especially when multimodalities are used. Extraction of the feature is used to minimise data dimension, obtaining information that is important for categorization. Limited literature on the extraction of features from several modalities is currently accessible. The main component analysis and other related techniques can be used to choose features for reduction in dimensionality.

- **Classification:** It is also an arduous process to select the classifiers for certain methods. Validation is therefore important for a correct benchmark classifier.

- **Noise and missing values:** MRS and MRI signals are sometimes loud, and artifact-containing. To detect, assess and delete the data from the analytical pipeline, quality controls should be carried out. Also, there are some missing values to neuropsychological data. The inclusion of noisy information and missing values might result in erroneous models or overfitting.

- **Security:** There is another challenge for data sharing and security at the worldwide level of AD research. Standards for data sharing should be followed rigorously at all levels

MATERIALS AND METHODS

Current AD biomarkers need collection of specimen (for example, sero or Fluid) or imaging data to screen persons at risk for Alzheimer's disease (AD) on the basis of medical performance reports in clinical stage which to provide improved therapeutic options for delaying the emergence of ADC. In contrast, electronic health reports, for example clinical records and administrative data on health, don't take time or effort to obtain data. The amount of such data has also expanded tremendously with the advent of digitalisation The digitalised healthcare database, because it is ubiquitous, cost effective and large, could be an inestimable resource for testing both scalable AD and other diseases predictive models. Despite its enormous potential utility, however, the extent to which massive administrative data on health are beneficial in the prediction of the AD risk is rarely known.

Prediction of risk of DA, prior models often depend on preset medical conditions (sociodemographic (age, sex, education), lifestyle (physical activity), risk factors for the intermediate life of the health system (substantial blood pressure, MIC and total levels of cholesterol); An essential concern remains as to whether the varied etiologies of multifactorial AD in clinical environments may adequately take account of these basic predictive models based on small groups of selected variables. A meta-analysis demonstrates that multifactor models are most likely to predict dementia risks, while single-factor model types do not provide a good indication of the exact AD risk prediction. Here we are testing the extent to which a data-driven machine technique collects important information from a broad medical data encompassing thousands of health trajectory data and predicts AD risk individually.

Machine Learning is the best way to analyse large-scale administrative health data with thousands of descriptors from hundreds of thousands. Studies suggest that machine learning applies successfully in forecasting incident disorders other than AD to the widespread administrative data (diabetes, metabolic syndrome, suicide death, opioid overdose or drug-resistant epilepsy, etc). With machine learning technology recently expanding rapidly, AI technologies are predicted to have a profound impact on medicine when used to clinical predictive modelling. However, we are aware that the data-driven predictive modelling based on national administrative health figures is still to be tested in the AD-risk forecast.

It is inevitable to make employ huge enough data representation of the population in evaluating predictive models. For model performance the amount of the data is important (e.g., accuracy), whereas for model generalizability representativeness is important. In this study, the 1 million persons representing today's South Korean population in the database of the National Health Insurance Service-National Sample Cohort (NHIS-nsc), was utilised in the NHIS-NSC We have created and validated data-driven master learning models for predicting the future incidence of DA using administrative health data, big, longitudinal, thorough, (e.g. insurance claims and health checks) inside this dataset.

Big Data Analytics Framework Proposed (BDA)

The BDA Framework incorporates a wide variety of complicated data structure and structure, storage, processing and analysis. It uses data organisation, parallel computation, spread storage strategies and Machine- Learning-based data-processing algo's that are quick and manageable. The suggested BDA Multi-Modal Data Classification Framework (MCI), Healthy Alters, and AD (HO). The framework proposed can be divided into four main components: (I) data standardisation, (II) data management, (III) data storage, and (IV) data processing. These four components consist of the proposed Hadoop BDA architecture for AD categorization and progression are discussed in detail in this section. The framework makes it possible to accommodate a large number of mixed and humungous data, following data pre-desposition processing, analysis of the data which is processed results and prognosis and diagnostic outcome. The framework provides.

1. Data Normalization

In the suggested BDA framework, multi-modal information from dispersed sources is absorbed into a unified platform (MRI, MRS, and neuropsychological). The data standardisation component is used to organise multi-modal data, which is multi-modal and requires interfaces to provide multiple data on one podium. The proposed design include, MRI DICOM pictures are utilized. Hence, MedCon is used in neuroimaging (nifti) MRI images for medical conversion of the image. DICOM Toolkit is a range of standard libraries and applications for DICOM (DCMTK). It has software for monitoring, constructing, transmitting

and receiving images via a network link. The plugin JMRUI2XML is used to convert and export the MRS data to an XML format for future processing in MRS format. Neuropsychological scores are presented to the file system to be processed in the commonly-split value (csv) file format.

2. Data Management

Includes organisational and user interaction features using a front-end and back-end system. Front-end refers to a user interface and the system accessibility from the back-end. In addition to the front end, the rear end covers raw and processed data storage. It also carries out frontal answers. For interface, coordination and scheduling, the front-end functionality comprises of the user experience Hadoop (Hue)¹, Apache Zookeeper and Oozie. Hadoop (HUE) provides a web browser for Hadoop to access, consult and see data. Hadoop user experience (HUE).

This interface lies between a vast number of storage facilities and tools like HBase, YARN, Oozie etc. It includes HBase and HDFS file browsers, and YARN work browser. It works with Oozie, YARN HBase, HDFS and many other tools for large data. Zookeeper helps maintain sync between distant sources and keeps configuration data. It can even deal with partial failures in the network. Oozie is a Hadoop job workflow planner that defines a sequence of activities and coordinates amongst them in order to accomplish the task. HDFS and HBase are the back-end functions. In addition, the front and back of the website help to construct a whole frame of a web interface for input from a range of data sources and user interactions.

3. Data Storage

In order to organise structured and unstructured data collected in various ways, the data storage component is required. The HDFS and HBase storage facilities are offered. Hadoop Distributed File System (HDFS) is designed to store a wide range of data across many commodity nodes. It has a master-slave architecture consisting of a data node wherein chunks of data are stored, fetched and returned to the node of name (master node) The storing of metadata (data on data) is also a key storage element. Real data is stored in the data node and metadata, including file location, block size, file permission etc is stored in the name (28). It features a built-in fault tolerance mechanism in the event of any node failure. HDFS biggest disadvantage: it uses a writing that reads many patterns (WORM). Therefore, even on one data point, if modifications are required, the full file needs to be updated. HBase is a non-relational database that offers fast random access to a huge amount of structured data called Not just SQL (NoSQL). Furthermore, HBase handles organised, unstructured and semi-structured data. Data in HBase will be saved in columns, sorted by pairs of key values. It also includes encrypted data security software. In the HBase area, HBase provides a library and runtime environment. In the HBase area server HBase provides a library and runtime environment for running user code.

4. Data Processing

The processing of data encompasses, selection ,extraction of features and inclusion of decisions. All of the qualities then are utilised to classify participants in HO, MCI & AD; statistical analysis and verification are then followed. The following is explained to each component of that layer:

a) “Spark, YARN and Map Reduce,”

Data from three different approaches have been grouped into the scope of subject categories from HO, MCI, and AD. These huge databases have difficulties with storage, analysis, and viewing. MapReduce is a reliable and faulty tolerant framework which is used for huge data processing in vast clusters at the same time. ‘MapReduce’ has two functional phases: map and reduction. Map organises raw data in key value pairs and decreases process data simultaneously. Apache Spark's MLlib learning machine library is used to extract, reduce dimensionality, classify, and produce basic statistics. To treat the iterative batch, MLlib uses Spark. Spark promotes iterative processing and increases performance with the use of in-memory calculators. YARN provides employment planning and cluster management. It also organises, plans and supports requests for customer resources.

b) Quality Check Matrix

Matrix for quality control Before multimodality processing began, the quality of the data control was undertaken. In addition, the framework also handles missing data values to prevent overfitting.

c) Feature Extraction

Feature extraction is used to remove from the picture characteristics specific to the condition. For the MRI data structural statistic characteristics consisting of statistical information from AD patients particular areas of interest (ROI) will be retrieved. Statistical elements include entropy, histogram-based characteristics like average and median, AD-related brain area textural information such as hippocampus, frontal cortices, etc. In the MRS, the neurochemical content peak area is the extraction of the spectral characteristics which represent ROI metabolic information. Scores such as MMSE CDR, GDS-SF, HIS, FAQ, TMT-A and TMT-B are present in neuropsychological data.

d) Selection of Features

MRS, MR1, and neuropsychological data extracting features are still high-dimensional categorization data. The ML library's Principle Component Analysis (PCA) is utilised to minimise the feature dimensions by getting the main points. This principle of PCA can be expanded to large space in order to maximise variance by employing the kernel method.

e) Classification and decision-making based ensemble

Our objective[^] for using multimodalities is to improve the precision of the AD patient classification in comparison with the judgement taken using only one data source. Therefore for accurate automated classification, ensemble-based classifiers and their approach to decision-making are offered. The premise of employing a data fusion ensemble strategy is that each modality from distinct sources will be trained by a distinct classifier. In addition, an individual classifier's decisions will be merged in accordance with the relevant mixing rule. To increase diagnostic performance, we can apply the sum rule to obtain data fusion.

f) Statistical Verification

The accuracy of diagnosis of diseases is crucial since it directly affects treatment. Therefore, to validate the outcomes of the proposed framework, the statistical analysis of classification accuracy will be undertaken. Sensitivity, specificity, classification accuracy and operational recipient characteristics (ROC) provide parameters for statistical analysis. Clinicians check diagnostic outcomes.

DATA COLLECTED

Table 2: Databases available for AD research across the globe:

Alzheimer's Disease Data Center, Core or Consortium	Host Institution for Data	Database or Data Characteristics
National Institute of Aging Genetics of Alzheimer's Disease Data Storage Site (NIAGADS)/Alzheimer's Disease Genetics Consortium (ADGC)	National Institute of Aging and the University of Pennsylvania	A repository of national genetic data to promote access to genotypic data to late inception Alzheimer's disease genetics by competent research.
Alzheimer's Disease Neuroimaging Initiative	University of California San Francisco	Consisting of data from 58 North American sites since 2004 on neuropsychological, imaging, genetic, and demographic longitudinal subject
French National Alzheimer's Database (BNA)	Nice University Hospital in France	Registers all medical events carried out in several hundred memory centres in France by memory units and independent professionals.
Integrated NeuroDegenerative Disease database (INDD)	University of Pennsylvania	Integrated database of neurodegenerative disorders connected to ageing, like Alzheimer's, Parkinson's, Lateral Sclerosis Amyotrophic, and lobic lobal degeneration Frontotemporary
AlzPharm and BrainPharm	Yale University	Database(s) of diverse neurological disorders treatment medication. AlzPharm is the BrainPharm "semantic" representation, that means knowledge and relationships are in fact represented in the data.
COMET-AD	Indian University and the State of Indiana	Indiana State-wide record of Alzheimer's adverse events medicines.
neuGRID for you (N4U)	European consortium of multiple partners	European Alzheimer's Data Grid Network
National Alzheimer's Coordinating Center (NACC)	University of Washington	The NACC has built and maintains a huge relativistic database of standardised clinical and

		neuropathological research figures, including data from 27 NIA-funded Alzheimer's Disease Centers (ADCs) around the United States.
Australian Imaging Biomarkers and Lifestyle Flagship Study of Ageing (AIBL)	Commonwealth Scientific and Industrial Research Organization (CSIRO) Australia and partner universities	4.5+year longitudinal prospective cognitive research. Widescale study cohort: more than 1000 students (minimum age 60 years). Alzheimer's (AD) and moderate cognitive impairment (MCI) patients and good workforce. All information is gathered in two centres (40 percent subjects from Perth in Western Australia, 60 percent from Melbourne, Victoria).
European Medical Information Framework (EMIF)	Consortium of European universities, research organizations, pharmaceutical companies, public bodies and non-profit groups	EMIF-AD is the driving force of Alzheimer's disease, which aims at EMIF-AD (1) establishing a large patient data repository to enable biomarker discovery research within the EMIF. (2) Connecting research cohort information to electronic registry data. (3) Identification of new potential targets in pre-symptomatic and prodromal AD for AD medication development using genomics and proteomic techniques.
Framingham Heart Study (FHS)	Boston University	Database includes the Framingham Heart Study genetic, phenotypic, and biomarker data
Brain Health Registry (BHR)	University of California San Francisco and partners	An investigator portal that offers researchers of Alzheimer's disease with access to de-identified brain health registry data is expected to be available in the near future.
Texas Alzheimer's Research	University of Texas	Longitudinal data base for more than 2000

Care Consortium (TARCC)	San Antonio Health Sciences Center and partners.	participants in Texas. Database
Aged Brain Sys Bio (ABSB)	French National Institute of Health	The intention is to provide new scientific resources, including a new open access database, for the ageing research community in Europe.
Alzheimer's Preventative Initiative/Banner Alzheimer's Institute (API/BAI)	Banner Alzheimer's Institute in Phoenix, Arizona	Prevention initiative data from Alzheimer's, including 'Columbia Study'
Women's Healthy Aging Project (WHAP)	University of Melbourne	An epidemiological prospective, longitudinal study of 438 Australian women, covering two decades
The Three City Study (3C)	The consortium of the three cities of Bordeaux, Dijon and Montpellier, France.	Longitudinal population-based studies in adults 65 and older examining the relationship between vascular disorders and dementia. Out of three French cities were recruited 9,294 participants (3,649 men and 5,645 women).
Swedish Dementia Registry	The registry database, SveDem, is maintained at the Uppsala Clinical Research Center in Sweden	Includes the 28, 742 follow-up dementia and demographics from October 2014. There are presently 95 percent of all memory clinics in Sweden currently participating in SveDem.
OPTIMA	University of Oxford	Database contains neuropsychological studies, brain scans, blood samples, CSF, physical test data, and histological information following donation of brain.
Wisconsin Registry for Alzheimer's Prevention (WRAP)	University of Wisconsin	Includes more than 1500 delegates.

Dallas Lifespan Brain Study (DLBS)	University of Texas at Dallas	The research focuses on healthy grown-ups and plays an important role in the understanding of the ageing mind. 350 adults, 50 from 20 to 89 each are comprehensively examined for cognitive, brain structure and function through the whole lifespan of the adult population.
European Alzheimer's Disease Initiative (EADI)		Central data repository on the European Alzheimer's Consortium's seven academic sites (EADC)
Neuroanatomical Database of Normal Japanese Brains	Tohoku University	A data collection has been obtained on 1547 regular individuals aged 16 to 79 years.
Canadian Longitudinal Study on Aging	Consortium of 26 universities across Canada	Consists of 50,000 Canadian women and men aged 45 to 85 years at the time of recruiting, a national, stratified, random sample. At three-year intervals, participants will be subjected to recurrent data collecting rounds and tracked for at least 20 years.
AlzGene	Alzforum, operated by the Biomedical Research Forum (BRF) LLC (in Cambridge, MA)	The AlzGene database is intended to serve as an integrated, impartial, publicly available and frequently updated survey of published studies on AD genetic associations. Data from more than 1600 gene research are included in this database. There have been thorough metaanalysis of more than 300 AD Gene candidates, and the results of more than 40 AD genes are posted publically on the website of AlzGene.org.

Methodology for logical regression

The following are our data pretreatment steps. I Alignment of data: We aligned the data to the initial Alzheimer's prognosis of each subject (event-based sequence). ICD-10 and coding of the drugs: (ii) As we have hierarchical structures ICD-10 for and medicine codes, we have employed the first category of disease codes -FOO[Alzheimer's dentition] including Alzheimer's dentition with prior incipients, FOO.1 ['Alzheimer's dentition with late incipient disease], FOO.2 ['Alzheimer's Dentitis, mixed type, atypical] and FOO.9 [Alzheimer's dentition, unspecified]), and Alzheimer's dentition, respectively, (iii) The study removed rare diseases or drug codes detected in total data fewer than five times (1179 disease and ^362 medicine codes). (iv) We excluded that participant from the test if a participant does not have any health screening data (Lab values, health checkups, personal health along with family health history from the *National Health Screening*) for the past three years (every elder in Korea is required to carry out biannual health testing). 4894 distinct variables utilised in models came out of this preprocessing approach

We utilise the data between 2002 and the year of the AD-n incident for every n-year prediction in the AD Group, because the data required at least n years before the AD incident. We used data from 2002 to 2010 within the non-AD group. For instance, if a patient was diagnosed with AD in 2009, we were using data from 2002 to 2009 for 0 years, 2002 to 2008 for 1 year, 2002 to 2007 for 2 years, 2002 to 2006 for 3 years, 2002 and 2002 to 2005 for 4 years.

We employed the randomly picked balanced data set as well as the whole unbalanced data set for model training, validation and testing. We carried out a bootstrap sample with substitution 10 times for the balanced dataset.

Code of logistic regression for finding top ten features

```
from
sklearn.neural_network
import MLPClassifier

from sklearn.ensemble import RandomForestClassifier
from sklearn import linear_model
from sklearn import svm
from sklearn.svm import SVC
from sklearn.preprocessing import StandardScaler
from sklearn.metrics import
accuracy_score,roc_auc_score,confusion_matrix
from sklearn.model_selection import GridSearchCV,
cross_val_score,cross_val_predict,StratifiedKFold
from sklearn.linear_model import SGDClassifier
from sklearn.kernel_approximation import RBFSampler
from sklearn.ensemble import ExtraTreesClassifier
from sklearn.feature_selection import SelectFromModel
from sklearn.preprocessing import StandardScaler
from sklearn.feature_selection import VarianceThreshold
from sklearn.linear_model import SGDClassifier
from sklearn.pipeline import Pipeline
from sklearn.svm import LinearSVC
import numpy as np
import pickle
from sklearn.metrics import f1_score
import scipy.stats
from sklearn.model_selection import train_test_split
from sklearn.model_selection import StratifiedShuffleSplit
def mean_confidence_interval(data, confidence=0.95):
    a = 1.0 * np.array(data)
    n = len(a)
    m, se = np.mean(a), scipy.stats.sem(a)
    h = se * scipy.stats.t.ppf((1 + confidence) / 2., n - 1)
    return m, m - h, m + h
def getCoeff(X,clf,task):
    diseaseMap = pickle.load(open("diseaseMap_" + task + ".p", "rb"))
    drugMap = pickle.load(open("drugMap_" + task + ".p", "rb"))
    phyExam = pickle.load(open("phyExam.p", "rb"))
    if task == 't1':
        coeff = clf.coef_[0]
        features = np.empty(X.shape[1], dtype=object)
        for code in diseaseMap:
            # print(code,diseaseMap[code])
            features[0] = 'sex'
```

```

        features[1] = 'age_group'
        features[diseaseMap[code] + 2] = code
    for code in drugMap:
        features[len(diseaseMap) + 2 + drugMap[code]] = code
    for peIdx in range(phyExam.shape[1] - 2):
        features[len(diseaseMap) + 2 + len(drugMap) + peIdx] =
'pe' + str(peIdx)
    elif task == 't3':
        features = np.empty(X.shape[1], dtype=object)
        coeff = clf.best_estimator_.coef_[0]
        for code in diseaseMap:
            # print(code,diseaseMap[code])
            features[0] = 'sex'
            features[1] = 'age_group'
            features[2] = 'num_in_patient_visits'
            features[3] = 'num_in_patient_days'
            features[diseaseMap[code] + 4] = code
        for code in drugMap:
            features[len(diseaseMap) + 4 + drugMap[code]] = code
        for peIdx in range(phyExam.shape[1] - 3):
            features[len(diseaseMap) + 4 + len(drugMap) + peIdx] =
'pe' + str(peIdx)
        log_weights = np.transpose(coeff)
        abs_log_weights = np.absolute(log_weights)
        log_srted_ix = np.argsort(abs_log_weights, axis=0)
        srted_array = []
        for i in log_srted_ix:
            line = [[features[i], log_weights[i]]]
            srted_array = srted_array + line
        srted_array = np.array(srted_array)
        pickle.dump(srted_array, open("coeff_" + task + ".p", "wb"))
from sklearn.model_selection import ShuffleSplit
from sklearn.utils import resample
def JP_classify(method,X,y,n_fold):
    """classification
    Args:
        method: classification method (random forest, logistic
regression, SVM)
        X_0: AD patient data
        X_1: normal people data

```

```

        inner_cv = StratifiedKFold(n_splits=n_fold, shuffle=False,
random_state=234)
        outer_cv = StratifiedKFold(n_splits=n_fold, shuffle=False,
random_state=234)
        avg_acc = []
        avg_train_acc = []
        avg_TP = []
        avg_TN = []
        avg_FP = []
        avg_FN = []
        avg_sen = []
        avg_spec = []
        roc_label = []
        roc_pred = []
        roc_prob = []
        outer_loop = 0
        print('loading')
        for train_index, test_index in outer_cv.split(X, y):
            X_train, X_test = X[train_index], X[test_index]
            y_train, y_test = y[train_index], y[test_index]
            print(outer_loop, X_train.shape, X_test.shape)
            outer_loop+=1
            if method == "RF":
                params = {'randomforest__min_samples_leaf': np.arange(1,
51, 5),
                        'randomforest__n_estimators': np.arange(10,
100, 10)}
                pipe = Pipeline([
                    ('featureExtract', VarianceThreshold()),
                    ('scaling', StandardScaler()),
                    ('randomforest',
RandomForestClassifier(random_state=0))
                ])
            elif method == 'SVM':
                params = {'svm__alpha': np.logspace(-4, 7, 12)}
                pipe = Pipeline([
                    ('featureExtract', VarianceThreshold()),
                    ('scaling', StandardScaler()),
                    ("svm", SGDClassifier(max_iter=1000, tol=1e-
5,random_state=0))
                ])
            elif method == 'LR':
                params = {'lr__C': np.logspace(-3, 8, 12)}

```

```

        pipe = Pipeline([
            ('featureExtract', VarianceThreshold()),
            ('scaling', StandardScaler()),
            ('lr',
linear_model.LogisticRegression(random_state=0))
        ])
        clf = GridSearchCV(estimator=pipe, param_grid=params,
cv=inner_cv, scoring='f1', n_jobs=-1)
        clf.fit(X_train, y_train)
        fs = clf.best_estimator_.named_steps['featureExtract']
        y_pred = clf.predict(X_test)
        if method == 'SVM':
            y_prob = clf.decision_function(X_test)
        else:
            y_prob = clf.predict_proba(X_test)
        y_pred_train = clf.predict(X_train)
        acc = accuracy_score(y_test, y_pred)
        train_acc = accuracy_score(y_train, y_pred_train)
        if method == 'SVM':
            auc = roc_auc_score(y_test, y_prob)
        else:
            auc = roc_auc_score(y_test, y_prob[:, 1])
        f1 = f1_score(y_test, y_pred, average='weighted')
        roc_label = np.append(roc_label, y_test)
        roc_pred = np.append(roc_pred, y_pred)
        if method == 'SVM':
            roc_prob = np.append(roc_prob, y_prob)
        else:
            roc_prob = np.append(roc_prob, y_prob[:, 1])
        TN, FP, FN, TP = confusion_matrix(y_test, y_pred).ravel()
        avg_TP = np.append(avg_TP, TP)
        avg_TN = np.append(avg_TN, TN)
        avg_FP = np.append(avg_FP, FP)
        avg_FN = np.append(avg_FN, FN)
        avg_acc = np.append(avg_acc, acc)
        avg_train_acc = np.append(avg_train_acc, train_acc)
        print(TP, FP, FN, TN)
        sen = TP / (TP + FN)
        spec = TN / (TN + FP)
        avg_sen = np.append(avg_sen, sen)
        avg_spec = np.append(avg_spec, spec)
        print('Accuracy:{},AUC:{},F1:{}'.format(acc, auc,f1))
        print('Train Accuracy:{}'.format(train_acc))

```

```

        print('Sensitivity:{},Specificity:{}'.format(sen, spec))
    print("Train Accuracy Avg: {}".format(np.mean(avg_train_acc)))
    print("Accuracy Avg: {}
({})".format(np.mean(avg_acc),np.std(avg_acc)))
    m, m_h1, m_h2 = mean_confidence_interval(avg_acc)
    print('Accuracy: {},{},{}'.format(m, m_h1, m_h2))
    print("Sensitivity Avg: {}
({})".format(np.mean(avg_sen),np.std(avg_sen)))
    m, m_h1, m_h2 = mean_confidence_interval(avg_sen)
    print('Sensitivity: {},{},{}'.format(m, m_h1, m_h2))
    print("Specificity Avg:
{}({})".format(np.mean(avg_spec),np.std(avg_spec)))
    m, m_h1, m_h2 = mean_confidence_interval(avg_spec)
    print('Specificity: {},{},{}'.format(m , m_h1 , m_h2))
    if method == 'RF' or method == 'SVM':
        return roc_label,roc_pred,roc_prob,clf
    elif method == 'LR':
        return roc_label, roc_pred, roc_prob,clf

```

RESULT AND CONCLUSION

Table 3: Top 10 logistic regression features and weights

Data Type	Name	b value	95% CI	Odd ratio	p-value
Health checkup	Hemoglobin (g/dL)	-0.90 2	-0.903/-0.901	0.405	<0.001
Demography	Age	0.689	0.687/0.690	1.991	<0.001
Control of health	Urine protein ^a	0.303	0.300/0.306	1.353	<0.001
Medicine	Zotepine (antipsychotic drug)	0.303	0.280/0.325	1.353	<0.001
Medicine	Nicametate Citrate (vasodilator)	-0.29 7	-0.298/-0.295	0.743	<0.001
Code of Disease	Other degenerative disorders of nervous system in diseases classified elsewhere	-0.29 2	-0.309/-0.274	0.746	<0.001

Code of Disease	External ear disorders in diseases otherwise categorised	-0.27 4	-0.328/-0. 220	0.760	<0.001
Medicine	Tolfenamic acid 200 mg (pain killer)	-0.26 6	-0.279/-0. 254	0.766	<0.001
Code of Disease	Respiratory distress syndrome for adults	-0.25 9	-0.282/-0. 236	0.771	<0.001
Medicine	Eperisone Hydrochloride (antispasmodic drug)	0.255	0.237/0.272	1.290	<0.001

References

1. Sharma, Ankita, et al. "BHARAT: an integrated big data analytic model for early diagnostic biomarker of Alzheimer's disease." *Frontiers in neurology* 10 (2019): 9.
2. Zhang, Rui, Gyorgy Simon, and Fang Yu. "Advancing Alzheimer's research: A review of big data promises." *International journal of medical informatics* 106 (2017): 48-56.
3. Ricciarelli, Roberta, and Ernesto Fedele. "The amyloid cascade hypothesis in Alzheimer's disease: it's time to change our mind." *Current neuropharmacology* 15.6 (2017): 926-935.
4. Wang, Chun, et al. "Statistical methods and computing for big data." *Statistics and its interface* 9.4 (2016): 399.
5. Krumholz, Harlan M. "Big data and new knowledge in medicine: the thinking, training, and tools needed for a learning health system." *Health Affairs* 33.7 (2014): 1163-1170.
6. Toga, Arthur W., and Naveen Ashish Priya Bhatt. "Global data sharing in Alzheimer's disease research." *Alzheimer disease and associated disorders* 30.2 (2016): 160.
7. Alzheimer's Disease International. Dementia statistics. 2015 Available at: <http://www.alz.co.uk/research/statistics>.
8. LaFerla, F. M. & Green, K. N. Animal models of Alzheimer disease. *Cold Spring Harb. Perspect. Med.* 2, a006320 (2012).
9. Giacobini, E. & Becker, R. E. One hundred years after discovery of Alzheimer's disease. A turning point for therapy? *J. Alzheimers Dis.*
10. (Laney, 2001). (2001). 3D data management: Controlling data volume, velocity and variety. *META group research note*, 6(70), 1.

11. Cottle M, Hoover W, Kanwal S, Kohn M, Strome T, Treister NT. Transforming Health Care through Big Data
12. Ross MK, Wei W, Ohno-Machado L. “Big Data” and the Electronic Health Record. *Yearb Med Inform.* 2014;9(1):97–104.
13. Wang, H. X., Xu, W., & Pei, J. J. (2012). Leisure activities, cognition and dementia. *Biochimica et Biophysica Acta (BBA)-Molecular Basis of Disease*, 1822(3), 482-491.
14. Bekris, L. M., Yu, C. E., Bird, T. D., & Tsuang, D. W. (2010). Genetics of Alzheimer disease. *Journal of geriatric psychiatry and neurology*, 23(4), 213-227.
15. Lott, I. T., & Dierssen, M. (2010). Cognitive deficits and associated neurological complications in individuals with Down's syndrome. *The Lancet Neurology*, 9(6), 623-633.
16. Hebert, L. E., Weuve, J., Scherr, P. A., & Evans, D. A. (2013). Alzheimer disease in the United States (2010–2050) estimated using the 2010 census. *Neurology*, 80(19), 1778-1783.
17. Green, R. C., Cupples, L. A., Go, R., Benke, K. S., Edeki, T., Griffith, P. A., ... & Farrer, L. A. (2002). Risk of dementia among white and African American relatives of patients with Alzheimer disease. *Jama*, 287(3), 329-336.
18. Mayeux, R., Saunders, A. M., Shea, S., Mirra, S., Evans, D., Roses, A. D., ... & Phelps, C. H. (1998). Utility of the apolipoprotein E genotype in the diagnosis of Alzheimer's disease. *New England Journal of Medicine*, 338(8), 506-511.
19. A precision approach to end Alzheimer's Disease | Dale Bredesen (video)

20. Preventing Alzheimer's Disease Using Groundbreaking Diagnostics | Gillian Coughlan
(video)

21. <https://searchdatamanagement.techtarget.com/definition/big-data>

22. <https://biovox.eu/alzheimer-s-can-be-predicted-early/>

23. Downing, A. M., Yaari, R., Ball, D. E., Selzler, K. J., & Devous Sr, M. D. (2016).

Bridging the gap between research and clinical practice in asymptomatic Alzheimer's disease. *J Prev Alz Dis*, 3(1), 30-42.