Neuroscience Patient Identification using Big Data and Fuzzy Logic - An Alzheimer's Disease Case Study

Kamran Munir, email kamran2.munir@uwe.ac.uk[e,*], Alberto de Ramón-Fernández, email aderamon@dtic.ua.es[f], Sohail Iqbal, email sohail.iqbal@seecs.edu.pk[g], Nadeem Javaid, email nadeemjavaid@comsats.edu.pk[h]

[a] *Computer Science and Creative Technologies, University of the West of England, Coldharbour Ln, Bristol BS16 1QY, United Kingdom.*
[b] *Department of Computer Technology, University of Alicante, Ctra. San Vicente del Raspeig s/n, 03690, Alicante, Spain.*
[c] *National University of Sciences and Technology, NUST Campus, H-12, Islamabad, Pakistan.*
[d] *COMSATS Institute of Information Technology, Park Road, Tarlai Kalan, Islamabad, 45550, Pakistan.*

Please cite this article as: K. Munir, A. de Ramón-Fernández and S. Iqbal et al., Neuroscience patient identification using big data and fuzzy logic–An Alzheimer's disease case study, Expert Systems With Applications, https://doi.org/10.1016/j.eswa.2019.06.049

*Corresponding author
Email address: Kamran2.Munir@uwe.ac.uk (Kamran Munir, email kamran2.munir@uwe.ac.uk)

Please cite this article as: K. Munir, A. de Ramón-Fernández and S. Iqbal et al., Neuroscience patient identification using big data and fuzzy logic–An Alzheimer's disease case study, Expert Systems With Applications, https://doi.org/10.1016/j.eswa.2019.06.049

Neuroscience Patient Identification using Big Data and Fuzzy Logic - An Alzheimer's Disease Case Study

Kamran Munir, email kamran2.munir@uwe.ac.uk[e,*], Alberto de Ramón-Fernández, email aderamon@dtic.ua.es[f], Sohail Iqbal, email sohail.iqbal@seecs.edu.pk[g], Nadeem Javaid, email nadeemjavaid@comsats.edu.pk[h]

[e] *Computer Science and Creative Technologies, University of the West of England, Coldharbour Ln, Bristol BS16 1QY, United Kingdom.*

[f] *Department of Computer Technology, University of Alicante, Ctra. San Vicente del Raspeig s/n, 03690, Alicante, Spain.*

[g] *National University of Sciences and Technology, NUST Campus, H-12, Islamabad, Pakistan.*

[h] *COMSATS Institute of Information Technology, Park Road, Tarlai Kalan, Islamabad, 45550, Pakistan.*

Abstract

Modern neuroscience imaging technologies considerably affect diagnostic and prognostic accuracy and facilitate progress towards the cure of brain diseases. The benefits largely depend on the practicalities by which the large-scale imaging and clinical data can be integrated, examined and understood. In EU neuGRID4You (N4U) project, many datasets were generated from research centres and hospitals. In order to perform effective analyses, these datasets and their metadata along with a number of pre-computed parameters are stored in a big data repository. This paper focuses on the patient identification using big data and Fuzzy Logic, which has been achieved through fuzzy processing where a reference number called Alzheimer's Disease Identification Number (ADIN) is calculated. It has enabled patients' sorting for a particular intensity of Alzheimer's disease, short-term estimation of the progression of that disease and context of individual patients with respect to other patients such as appropriate treatment, estimated life expectancy etc. The generated rules define the necessary knowledge base for the inference engine to generate output sets and an aggregate membership function of each rule is formed. Using this function, a most representative value of the total output set is obtained which represents the disease intensity. The implemented system and its evaluation are based on realistic datasets, demonstrators and making use of real-life neuroscience case studies. The presented results of four selected case studies show that this approach have provided sufficient expressiveness in understanding patients' disease information. Finally, a discussion and conclusions are presented on the opportunities offered by the calculation of ADIN to manage Alzheimer's disease along with potential future extensions or applications of this work.

Keywords: Big data, neuroscience, fuzzy processing, brain disease, analysis, Alzheimer

1. Introduction

Exceptional increase in the generation and availability of clinical and neuroimaging data sets has forced the advancements in data processing infrastructures and analysis applications. In med-

*Corresponding author

Email address: Kamran2.Munir@uwe.ac.uk (Kamran Munir, email kamran2.munir@uwe.ac.uk)

Please cite this article as: K. Munir, A. de Ramón-Fernández and S. Iqbal et al., Neuroscience patient identification using big data and fuzzy logic–An Alzheimer's disease case study, Expert Systems With Applications, https://doi.org/10.1016/j.eswa.2019.06.049

ical domains, these massive amounts of heterogeneous data, which are accumulated both in real-time and over decades, are usually extremely critical for diagnostics and decision-making. In the neuroscience domain, various e-infrastructures are offering a suite of services for neuro analyses worldwide. Due to such developments, massive amount of data is being continuously and anonymously shared by the hospitals and research centers to constitute the foundations of brain disease analyses, such as Alzheimer. However, such increase in the volume, variety and velocity of neuroimaging datasets, and ever increasing knowledge complexity in medical research, puts neuroscientists under sever difficulties in data integration, data linking and in performing analyses.

The neuGRID4You project (38) provided an e-infrastructure to neuroscience researchers for image analysis. On one hand such infrastructures are exceptional in providing state of the art neuro imaging and data analysis tools and methodologies. However, on the other hand these technologies depend on the practicalities, speed and reliability with which the neuro imaging datasets along with the clinical data of a substantial population can be analysed and interpreted. The number of datasets considered in neuGRID4You are enormous, these include a variety of datasets generated from several research centres or hospital such as images, internal/external datasets, clinical pipelines, algorithms definitions, patients' clinical and diagnostics information etc.

To effectively analyse neuroscience datasets for decision-making, it requires storage of these datasets and metadata along with several pre-computed parameters in a big data repository before the users can perform any analysis. Moreover, enabling such an end-to-end big data analyses mechanism, and that is also adaptable to other user communities, it requires building of various big datasets storage services to interact with several (heterogeneous) data sets' providing institutes and research centres. These services that can store and index datasets in a big data repository must perform all data ingest, prepare, transform and publish operations to support analyses. Moreover, to achieve the speed and reliability with which such data of substantial populations can be analysed and interpreted, a lot of pre-computations are necessary e.g. to provide quick sorting and retrieval of all patients' population from big data based on several parameters. For example, in the neuGRID4You project alone, thousands of clinical parameters were dealt with. To achieve these necessary pre-computations, fuzzy processing can be applied; for example, to compute the intensity of a disease and to provide quick sorting of all patients based on a disease intensity number. Such fuzzy processing can be very useful while analysing and interpreting neuroimaging data. Moreover, with the help of intensity wise sorted data, it is easy to find out the percentile of the person with disease. This information would eventually help the care providers for the optimisation of their services.

Existing big data approaches provide a limited solution to above mentioned fuzzy neuroscience big data processing problem and three major concerns still remain i.e. (a) the practicalities of transforming and storing neuroscience population's massive amounts of heterogeneous datasets including data dictionaries and images; (b) formulation and enabling run-time disease intensity numbers calculation using fuzzy processing, their formats, storage structure and retrieval; (c) supporting end users in formulating appropriate studies for analysis using clinical datasets, images and pre-computed patients' classifications i.e. based on disease intensity number(s). In this paper, the development of this big data processing system for neuroscience analyses is presented along with selected real life case study and outcomes. The design and development of this system along with the execution of case studies has been based on detailed users' requirements, which

Please cite this article as: K. Munir, A. de Ramón-Fernández and S. Iqbal et al., Neuroscience patient identification using big data and fuzzy logic–An Alzheimer's disease case study, Expert Systems With Applications, https://doi.org/10.1016/j.eswa.2019.06.049

were originally collected through the requirements gathering phases of the EU's neuGRID4You project, and were later extended by collaborating with the neuroscience community. This also includes an enormous number of evolving requirements that appeared while experimenting various datasets, such as volume, increased density and heterogeneity that led further improvements in the implementation.

Fuzzy Logic provides more expressiveness, i.e. explain the fact of being tired or happy in contrast to the traditional Boolean logic the truth value is either '0' or '1'. Therefore, if we are keeping the record of the Alzheimer patient being tired, the answer should be a yes or no. However, the real-life situations are much complex to be described by the Boolean logic. Thus, the more precise answer to if someone is tired may lie somewhere between this dichotomy. Fuzzy logic assigns a truth value from the interval [0,1] to every proposition by using its membership function. Since the interval consist of infinitely many values from 0 to 1, every proposition can be described better using fuzzy logic. Mathematically, fuzzy logic has more power than the Boolean logic, as every Boolean system can be reduced to fuzzy system. This better expressiveness of fuzzy is achieved by using slightly increased computational complexity. But this expressiveness is worth having due to the exponential increase in the computational power of the semi-conductor chips.

Due to enhanced expression of fuzzy logic, it has been a worthy experience of using it to measure the symptoms of Alzheimer's Disease (AD) patient and find out the total intensity of the disease. Having the different parameters of AD recorded as fuzzy membership function values helps to build fuzzy rule base. This fuzzy rules-base eventually provides us the fuzzy inference system. Such fuzzy inference system can be used for making better decisions when data is versatile as in the case of big data.

This paper has been organised as follows: After presenting related literature in Section 2, the architecture for neuroscience analyses is presented in Section 3. The Section 4, presents the challenges of handling large neuroscience datasets. The implementation details of Persistency Service are presented in Section 5. Section 6, presents the implementation of neuroscience detests' analysis mechanism that executes a user defined analysis as Hadoop jobs in MapReduce. In Section 7, the concept, working and details of proposed ADIN are presented. Finally, the conclusions and future work are presented.

2. Related Literature

Management of large volumes of neuroscience data to enable analyses is a key research challenge which neuroscience researchers are facing today. In relation to the neuroscience big data management and analysis system; as presented this paper, fundamental research work have been done in the areas of (a) collection and cataloguing of neuroscience data and resources; (b) managing neuroscience studies data sets, which includes images and patients; and (c) data analysis applications or methods utilising fuzzy processing. The existing efforts related to these are reviewed in this section.

In relation to the collections of neuroscience resources, there are various projects discussed in (19) that discover and integrate neuroscience resources. These projects provide a platform of neuroscience data that is made available to be used by the neuroscience community. The neuroscience community can search for data within these datasets using domain specific keywords. These developed platforms also provide search mechanisms to the neuroscience community to dis-

Please cite this article as: K. Munir, A. de Ramón-Fernández and S. Iqbal et al., Neuroscience patient identification using big data and fuzzy logic–An Alzheimer's disease case study, Expert Systems With Applications, https://doi.org/10.1016/j.eswa.2019.06.049

cover information. One such common example of Web-based catalogue of neuroscience databases is the Neuroscience Database Gateway (NDG) which provides access to over two hundred databases (Sen). Similarly, Neuroscience Information Framework (NIF) (22) provides to access various heterogeneous information resources and is based on resource registry, database mediator and document archiver. The resource integration and data searching of NIP can be done by DISCO (34) web based application. Entrez Neuron (44) uses ontology to provide search within data sets. Unlike Entrez Neuron, we don't provide ontology web language (OWL) based querying and our services interface provides a dynamic analysis building capabilities for both users and other services. The NeuroLOG (37) project provided data retrieval from various sites. In comparison, our system provides big data storage and indexing to search and locate datasets, as well as data sets' analysis services. The linking of all this information and the integration of datasets form various providers are fundamental to support analyses. These existing projects were mainly concerned dealing with heterogeneous neuroscience resources instead of exploring the heterogeneity within the neuro-data. Moreover, the pre-calculation and linking of neuroscience patients' disease identification number using fuzzy processing has been implemented. Some of the other existing work is based on the use of ontologies for querying, which is not related to the focus of this paper.

While talking about managing neuroscience experiments' data sets, significant research has already been undertaken. Notable examples include: Functional Bioinformatics Research Network (FBIRN) (41), Neuroinformatics Database (NiDB) (15), Alzheimer's Disease Neuroimaging Initiative (ADNI) Datasets (ADN), the Alzheimer's Repository Without Borders (ARWIBO)(ARW), the Northwestern University Schizophrenia Data and Software Tool (NUSDAST) (51)(NUS); the Computational Neuroscience Applications Research Infrastructure (CNARI) framework (46), etc. The Neuroinformatics Database (NiDB) (15) facilitates the storage and manipulation of neuroimaging data sets. SenseLab system (Sen) is driven by metadata to store data. Most of the datasets such as ADNI, FBIRN, ARWIBO, NUSDAST and several others have been included in this work; details are presented in Section 4. In terms of data sets storage and query, aforementioned systems mainly focus on the data sets gathering and storage. The system presented in this paper provides an integrated datasets storage, subject based linking of clinical study data to image datasets, and customisable disease identification number to perform big data analyses. Moreover, unlike the RDBMS, XML or ontology approach adopted in the above systems, we make use a combination of NOSQL, BLOB and Cache storages for data storage and fast retrieval. Furthermore, this work provides generic services to store data generated by external sources and provides a centralised storage view to the end users. While providing this unified view, complete data processing including transformations are handled internally by the system.

The collection and management of large neuroscience data have allowed development of new methods/applications for information analysis. Such as, clinical decision support systems (CDSS), which have played a key role to help experts. These systems, thanks to mathematical rules or statistic and computational methods, can provide important information about any disease (e.g. cluster analysis, clinical value predictions, diagnosis suggestion, etc.). If we focus on the Alzheimer's disease, some good examples of related systems does exist in the literature. For example, in (17) a knowledge base is provided by clinicians along with a validation and reasoning process. This CDSS (17) is able to suggest a probable diagnosis of dementia or Alzheimer. On the same line, the system presented in (45) can evaluate whether the patient has symptoms compatible with Alzheimer's disease or not. If so, an inference system based on Bayesian networks is responsible for assigning the accuracy of the suggested diagnosis. In both of these cases, the inference systems

are based on "if-then-else" rules build up from the different clinical guidelines for the diagnosis of dementia and Alzheimer's disease, as well as from the opinion(s) of experts. In these CDSS, the inference engine is built from a "static set of rules". However, the systems built this way does not have the capacity to evolve or learn from new information. To achieve this, Toro et al.(49) proposed an evolution. The approach presented in (49) is able to learn, discover and suggest new rules using bioinspired techniques and the reasoning capabilities offered by ontologies. The main limitation of this type of decision support systems is the limited flexibility of the "if-then-else" rules that provide output information. Due to this, such systems classify patients within a same level of impairment that are within the same predefined range.

On the contrary, the Fuzzy Logic approach gives more expressiveness. Thanks to the membership functions, fuzzy processing is able to define each state of impairment (fuzzy sets) depending on the symptomatology in a more adjusted way, providing accurate and representative information. Fuzzy processing is used in variety of data analysis applications. These applications include all the important aspects of data analysis such as data cleansing, clustering, integration and interpretability. A good amount of research has been carried out in the area of fuzzy clustering of data. These days, data is coming in huge velocity from various resources such as social media, files sharing services, emails, patients' records, and other electronic devices communicating with each other. We usually first organise such data in clusters to extract the useful information from it. Clustering of data is a crucial step in various area including data cleansing, data mining, computer vision, DNA analysis, stock exchange analysis and medical field. There are some classical clustering techniques upon which fuzzy clustering techniques are based and work efficiently. In clustering, partitioning clustering methods are very important. These methods directly cluster the data points by partitioning the whole data set in mutually exclusive and collectively exhaustive subsets. One of the famous approaches in clustering is to choose the number of clusters first and assign each cluster center arbitrarily. Then, data points are assigned to the nearest cluster center based on Euclidean distance. Over the iterations, the centroids of the data points in the clusters are recomputed and clusters get stabilised. This is the basis idea behind the k-means algorithm (31). Based on this idea, many researchers have generalised this algorithm using fuzzy logic techniques Nanda (29; 40). Similarly, for feature extraction for data mining systems fuzzy c-means clustering has proven itself very effective (28). Interesting results on the telecommunication data are shown depicting the effectiveness of Fuzzy C-means clustering (50). To get inferences from the huge amount of meteorological data fuzzy clustering algorithms are intelligently implemented (30).

Fuzzy logic is not an ultimate solution to all problems. There are certain challenges and issues related to the interpretability of data that arise to rough nature of fuzzy sets (18). But these challenges intrigue researchers to come up with their original ideas (26). There are various techniques that use the potential of fuzzy logic and got patents. One of such patents is by Tanaka (48) that removes various types of noises from data by using a fuzzy smoothing filter. Usage of a 3D elliptic membership function based on fuzzy logic is proposed. Indeed, this can further be generalised to nD elliptic membership function. For unstructured data, it is always practical to have variable membership functions. Haissig et al. (23) has proposed an adaptive fuzzy controller that can modify the member functions in the real time and got their technique patented. Fuzzy data mining techniques demonstrated to be more successful for intrusion detection (16; 42). Moreover, such techniques also show the added advantage of using fuzzy logic in data mining. Fuzzy logic is also used for the fusion of data coming from variety of sources (25). Inspired from this work

Please cite this article as: K. Munir, A. de Ramón-Fernández and S. Iqbal et al., Neuroscience patient identification using big data and fuzzy logic–An Alzheimer's disease case study, Expert Systems With Applications, https://doi.org/10.1016/j.eswa.2019.06.049

several other researches are carried out by taking the advantage of c-means clustering and belief function (21; 20).

3. The Big Data Processing for Neuroscience Experiments and Analysis Architecture

Management of continuously growing large neuroscience datasets, images, pipelines, algorithms and other related entities was the special case of utilising big data technologies for various advantage; for example, to achieve high readiness, unprecedented processing speed with vastly improved research collaborations among neuroscientists in hospitals and research centres. Under normal circumstances, both the neuroscientists and clinicians are confronted with several problems of data processing and heterogeneous access from distributed locations. Indeed, there are numerous new exciting developments in relation to medical imaging, diagnostics and scanning technologies. Previously grid, now cloud computing and various big data technologies have eased various large data sets storage and processing issues. However, when heterogeneous data sets are collected from multiple sources, there are usually no links between their data dictionaries, indexes, images and historical data analysis results. Therefore, it becomes extremely difficult for the neuroscientists to conduct analyses or build further on previously conducted studies. Consequently, there was need to research and provide an intuitive, fast and linked access to all these datasets, tools and information for data analysis.

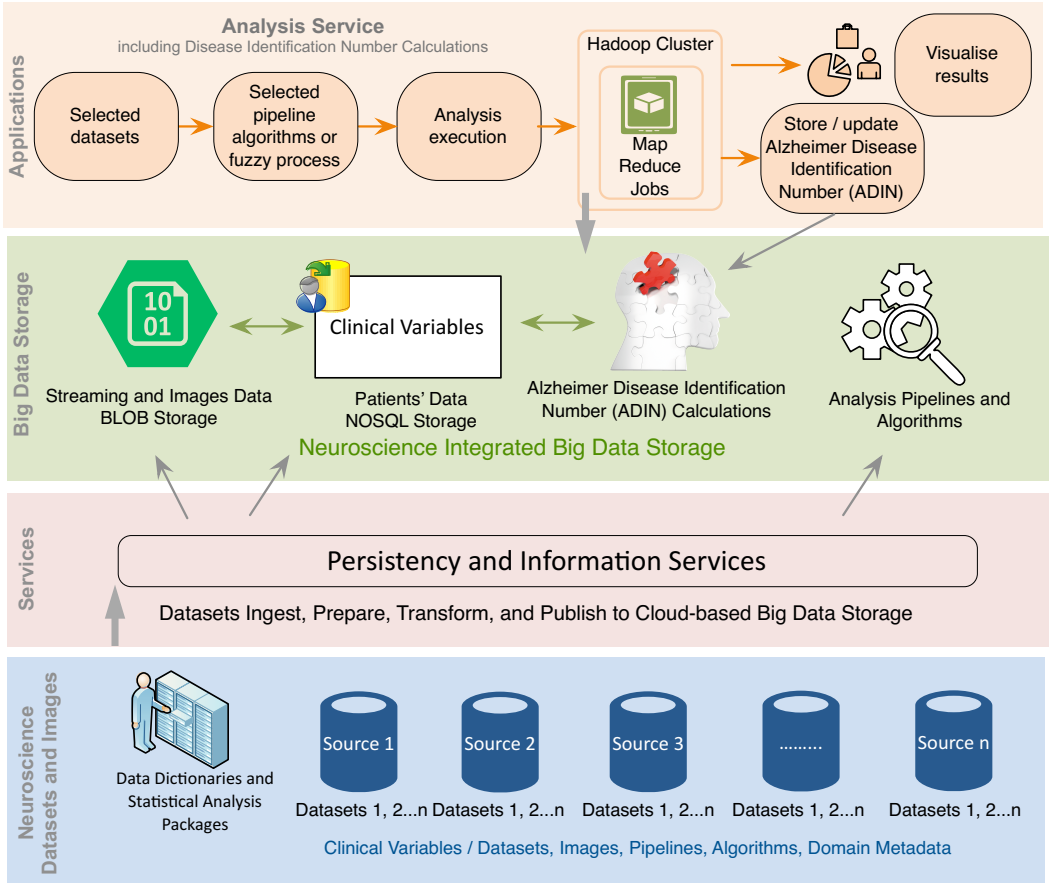

Figure 1: The Big Data Processing for Neuroscience Experiments and Analysis Architecture - with Essential Software Components / Services.

Please cite this article as: K. Munir, A. de Ramón-Fernández and S. Iqbal et al., Neuroscience patient identification using big data and fuzzy logic–An Alzheimer's disease case study, Expert Systems With Applications, https://doi.org/10.1016/j.eswa.2019.06.049

The neuroscience integrated data storage and analysis environment (shown in Figure 1) has been implemented to provide various functionalities and facilities within the neuro-degenerative diseases analysis e-infrastructure, which included (but not limited to): (a) storage and indexing of datasets; (b) store and index/register pipeline definitions; (c) store user-derived and/or pre-calculated datasets and data resulting from pipeline / algorithms executions; (d) provide service interfaces for data import; (e) provide access to stored datasets; and especially, to perform analysis combining datasets and disease identification number calculations. To achieve all these functionalities and facilities, specific data structures including software and methodologies have been implemented. Moreover, indexing, storing and linking the datasets, algorithms and workflows definitions along with data dictionaries have enabled such analyses. Moreover, once the neuroscientists have conduct their analysis or computations, the outcomes and resulting data sets are also made available in the big data storage for reusability purposes. To elaborate these, the architecture components are enclosed into following four layers interlinked by various service interfaces (as shown in Figure 1):

- **Neuroscience Datasets and Images:** These datasets are collected from several data providers that include hospitals and research centres across EU. Each data provider can have any number of heterogeneous datasets and related images including metadata dictionaries and analysis packages / pipelines / algorithms. Moreover, Neuro- and citizen scientists use these datasets for analysis by executing statistical packages, pipelines or algorithms. Thus, a list of statistical packages / pipelines / algorithms including their constraints were also required.

- **Services - Persistency and Information Services:** The services layer contains software components i.e. the persistency and information services, the persistency service provides functionalities such as (a) crawl the datasets; (b) make a model based on the structure of dataset; (c) store both data sets and outcomes, data dictionaries including possible parameters' values, such that these are query-able by other tools and services; and (d) store and index the image files associated with data sets. The information service provides mechanises to directly query datasets and images from the big data repository for querying, indexing and verification purposes.

- **Neuroscience Big Data Storage:** This can store huge amounts of datasets, images, pipelines, and a lot of other pre-computed information. It has for main units i.e. (a) streaming and images data that are stored in BLOB storage;(b) NOSQL storage of patients' datasets that include all clinical variables and their values from all data sources; (c) disease identification number calculations e.g. Alzheimer (ADIN) – for several types of studies / analysis, which are continuously increasing. Here, the links between patients' data and a disease identification numbers are continuously maintained through subject indexing, which further increases the efficiency of data retrieval; and (d) storage of analysis packages, pipelines and algorithms along with their applicability definitions.

- **Applications / Analysis Service:** Contains various big data retrieval applications, decision-making and analyses tools. In this paper, we focus on the Analysis Service that allows study definitions and disease identification numbers calculations. This service performs fuzzy processing to compute the intensity of diseases (e.g. Alzheimer disease) and stores the generated values e.g. in the case of Alzheimer it is called "Alzheimer Disease Intensity Number(s) (ADIN)", which are stored in the persistent storage alongside with the patients' data. These

Please cite this article as: K. Munir, A. de Ramón-Fernández and S. Iqbal et al., Neuroscience patient identification using big data and fuzzy logic–An Alzheimer's disease case study, Expert Systems With Applications, https://doi.org/10.1016/j.eswa.2019.06.049

pre-computations are used for the analyses and interpretation of neuroimaging data. For every new dataset, fuzzy processing is applied and new patients'data are sorted in the existing list of millions of patients. This functionality has been provided as a web service that is accessible to both client applications and/or end users (i.e. mainly the neuroscience data administrators). The users or other applications / services interact with is service though a service oriented approach.

The design and methodological details of all the above services along with their implementation outcomes / findings are discussed in the later sections of this paper. In the following section, the obtained neuroscience datasets, metadata and images are described along with their structural description to ascertain the requirement(s) of achieving the above-mentioned data processing approach.

4. Ingestion and Structuring of the Neuroscience Big Datasets - Requirements and Challenges

There are various neuroscience datasets along with images files that have been collected, which are as follows:

- the Alzheimer's Disease Neuroimaging Initiative (ADN). These Alzheimer's Disease Neuroimaging datasets are the most complex and largest datasets in our collections, which also include further dataset's categories i.e., ADNI1, ADNI2 and ADNI GO and their subcategories;
- the Open Access Series of Imaging Studies (OAS), the Open Access Series of Imaging Studies (OASIS) dataset has two subcategories: OASIS Cross-sectional, OASIS Longitudinal;
- the Magnetic Resonance in Multiple Sclerosis dataset (MAG);
- the Schizophrenia Data (51; NUS);
- the 1000 Functional Connectomes Project (FCP);
- the Centre for Biomedical Research Excellence data (COB);
- the Attention Deficit Hyperactivity Disorder dataset (ADH);
- the Diffusion-weighted magnetic resonance imaging dataset (IND);
- the Functional Bioinformatics Research Network (FBI), which has two phases called FBIRN Phase I and FBIRN Phase II;
- the Alzheimer's disease dataset MIRIAD (32);
- the Autism Brain Imaging Data Exchange dataset (ABI);

The functional details of big data storage and services are elaborated using the Alzheimer's disease neuroimaging datasets including its sub-categories. Moreover, these Alzheimer's Disease Neuroimaging datasets and variables are used to demonstrate the practical application of fuzzy processing of big data and calculation of Alzheimer disease identification number calculations

Please cite this article as: K. Munir, A. de Ramón-Fernández and S. Iqbal et al., Neuroscience patient identification using big data and fuzzy logic–An Alzheimer's disease case study, Expert Systems With Applications, https://doi.org/10.1016/j.eswa.2019.06.049

(ADIN) in the later sections of this paper. These datasets include several large data files containing various assessments e.g. ADAS-ADNI, ADAS-ADNIGO2, ADASSCORES, MMSE, MOCA etc., data about bio-specimen, imaging data including parameters, patients' studies and subject characteristics, all in complex directory structure (as show in Figure 2). The storage, indexing and linking of huge datasets, such as ADNI, was a complicated process. The datasets could have been simply stored as files as they were with their providers, but it limits the use of these data sets. This file organisation of storage has insignificant usability due to the unavailability of filtering mechanisms of the dataset contents as well as means to define analysis. For every query, there is need to go through millions and millions of records and there is no linking between different information stored within files containing same subjects.

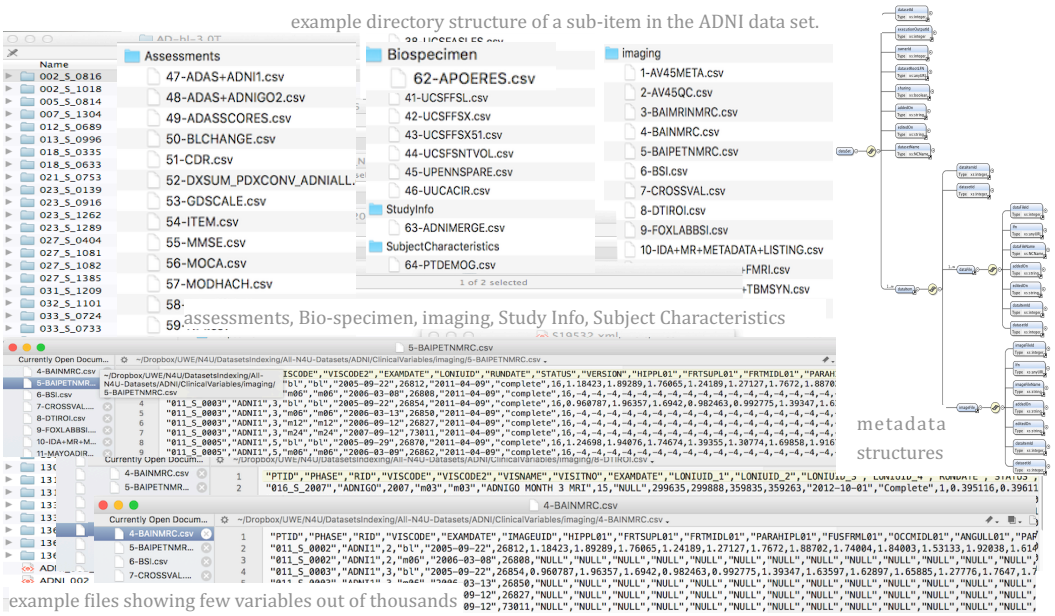

Figure 2: Snapshot of a portion of one neuroscience dataset – showing clinical variables, bio-specimen, imaging directories including parameters, patients' studies and metadata structures.

Moreover, in these existing structures there are either separate metadata files with no link to the datasets files or the metadata files just doesn't exist. In this case, it is impossible for the analysis applications to guide a user about the query parameters and their possible values for precise information retrieval. Nevertheless, the data sets are distributed in various levels of folder structures, which contain outcomes of clinical studies in various phases (as shown in Figure 2). Each stage contains dozens of image outputs (for one subject out of millions) that themselves are stored in various levels of directory structures with no formal metadata associations. In order to deal with all these problems and to provide end users an integrated access to data sets a common meta model has been devised. The administrator is provided with a variables definition wallet for each dataset that is populated and processed once for all the datasets providers. This metadata file is then used to verify (through DatasetCrawler, which is discussed later) each new dataset being made available by the data providers before uploading to the big data repository. While investigating various data sets the following facts / guidelines were obtained i.e. (a) a dataset may contain any number of files containing information about a clinical study performed on a subject; (b) a clinical study information of a subject could be spread into multiple files depending of the sub-types of study; (c) a data set may contain medical images; (d) there could be more than one

Please cite this article as: K. Munir, A. de Ramón-Fernández and S. Iqbal et al., Neuroscience patient identification using big data and fuzzy logic–An Alzheimer's disease case study, Expert Systems With Applications, https://doi.org/10.1016/j.eswa.2019.06.049

image files associated with each subject for each data sets; (e) a data set may contain several files containing information about the subject on which the clinical study was carried out; (f) the data and image files will have inter-relationships that is maintained with subject ID and the image file name which is made of study name, subject id, image id, date and session/visit id. Any further format and naming convention examples are not provided here for privacy reasons.

5. Preparation and Storage of Neuroscience Data Sets' - The Persistency Service

The Persistency Service carries out the preparation and storage of aforementioned neuroscience datasets as show in the Figure 3. It is a web service and an authorised user can invoke its operations. It meets the following requirements: (1) researchers can carry out analysis using one or more stored datasets; (2) it provides access to predefined analyses algorithms executable on different stored datasets; (3) to index the subject associated with each image, the subjects' indexes are stored as exporting the whole datasets images without indexing was inefficient.

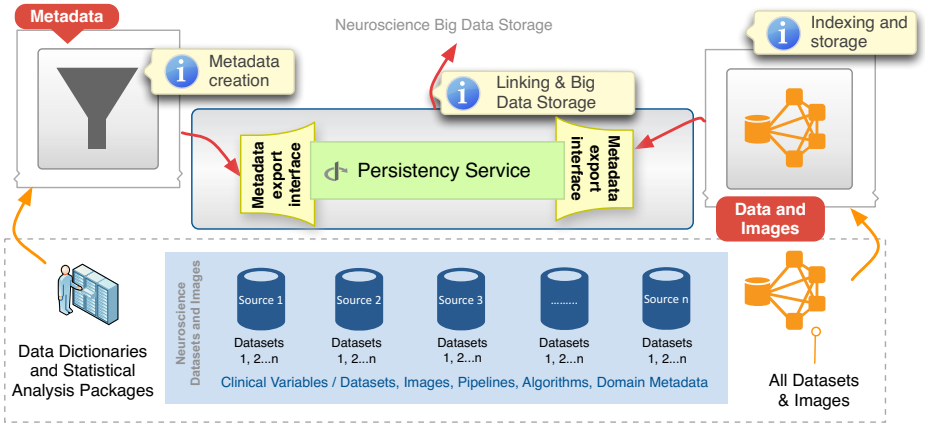

Figure 3: The Persistency Service - Preparation and Storage of Neuroscience Data Sets'.

The Persistency Service (as show in Figure 3) carries out following main tasks:

Statistical Analysis Packages, Pipelines and Algorithms Storage: It stores data analysis packages, pipelines or individual algorithms, along with their indexes.

Clinical Variables and Metadata Storage: Data sets' metadata, clinical variables and their values including patient demographics associated with neuroimages. To store (import) a new dataset, the admin provides the link of the directory containing respective CSV files. The service then generates the records, indexes and metadata. The metadata is then confirmed with the data dictionary of the respective data set and all records along with the indexes are stored.

Image Files Storage and Indexing: The brain images are stored in DICOM (Digital Imaging and Communications in Medicine) format, which is an international standard for medical imaging and related information. Each image ID represents the unique subject that is linked with its other related clinical information.

The functional flow of the Persistency Service is shown as Figure4, which include:

- **Crawl** the datasets given to the Persistency Service.
- **Parse and Ingest** the datasets, variables names and images to import into the persistence storage.

Please cite this article as: K. Munir, A. de Ramón-Fernández and S. Iqbal et al., Neuroscience patient identification using big data and fuzzy logic–An Alzheimer's disease case study, Expert Systems With Applications, https://doi.org/10.1016/j.eswa.2019.06.049

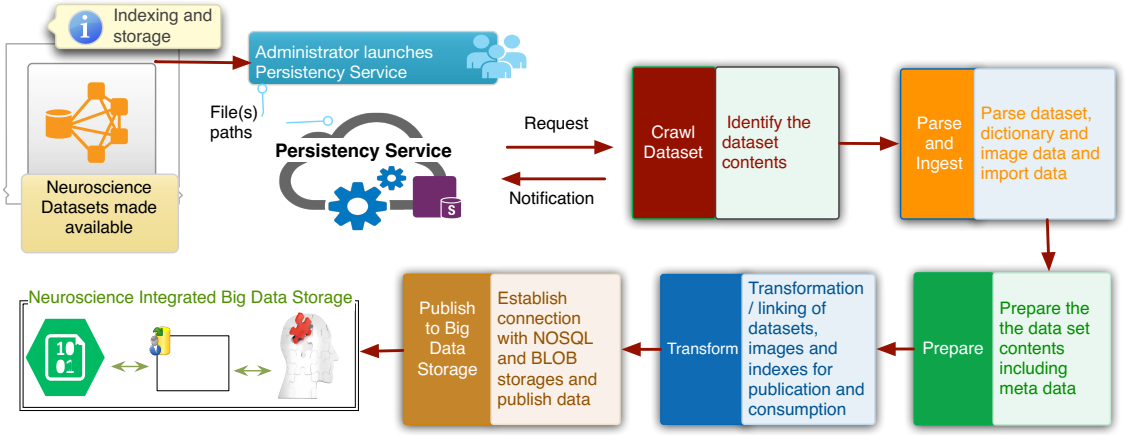

Figure 4: The persistency Service Functional Flow.

– **Prepare** the datasets contents, clinical variables data, images and metadata. Here the identification / linkage between images and clinical data is established through subject ids.

– **Transformation** of CSVs data into records and establishment of associations with patients' images through subject ids are carried out. The indexes of image files on the image names (i.e. subject ids) are also prepared.

– **Publish** / store the image files to the BLOB storage and all prepared records into NOSQL storage along with the indexes.

For further understanding, the Figure 5 illustrates the detailed Persistency Service's functional flow.

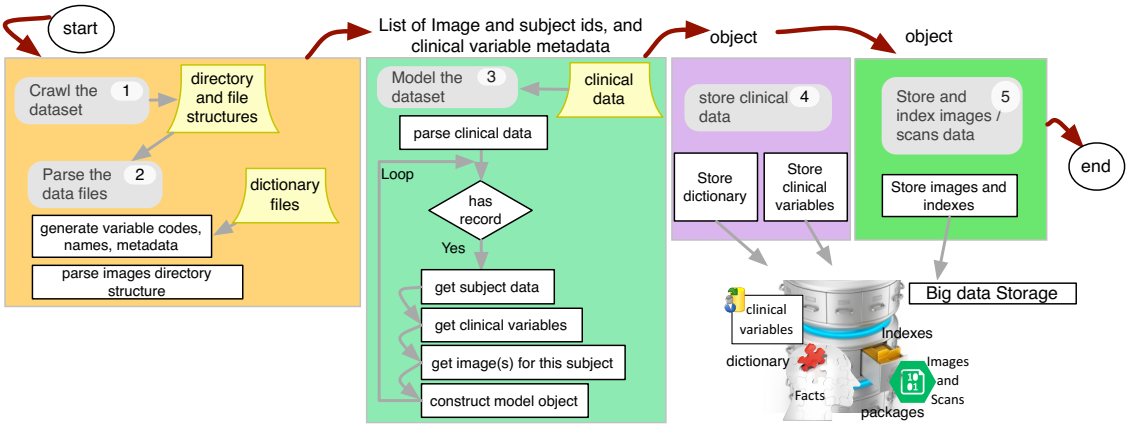

Figure 5: Persistency Service Functional Flow - Detailed Representation.

6. Neuroscience Datasets Analysis and Calculations

To retrieve information from big data storage, various web interfaces (named as Information Services) with basic functionality were initially deployed to fulfil the basic requirements of search and query the data. Through these interfaces, an extensive variety of queries comprising datasets

Please cite this article as: K. Munir, A. de Ramón-Fernández and S. Iqbal et al., Neuroscience patient identification using big data and fuzzy logic–An Alzheimer's disease case study, Expert Systems With Applications, https://doi.org/10.1016/j.eswa.2019.06.049

browsing, viewing of data dictionaries, metadata lookups and neuroimages filtering were all supported. The focus of this paper is not about describing these conventional web / application interfaces. Rather in this section of the paper, the Analysis Service implementation from the big data analysis perspective is presented. In our supported scenario, the neuro or citizen science search requests may arise from end users or services and thus the Analysis Service exposes distinctive interfaces for its clients (as shown in Figure 6). The functional implementation of the neuroscience

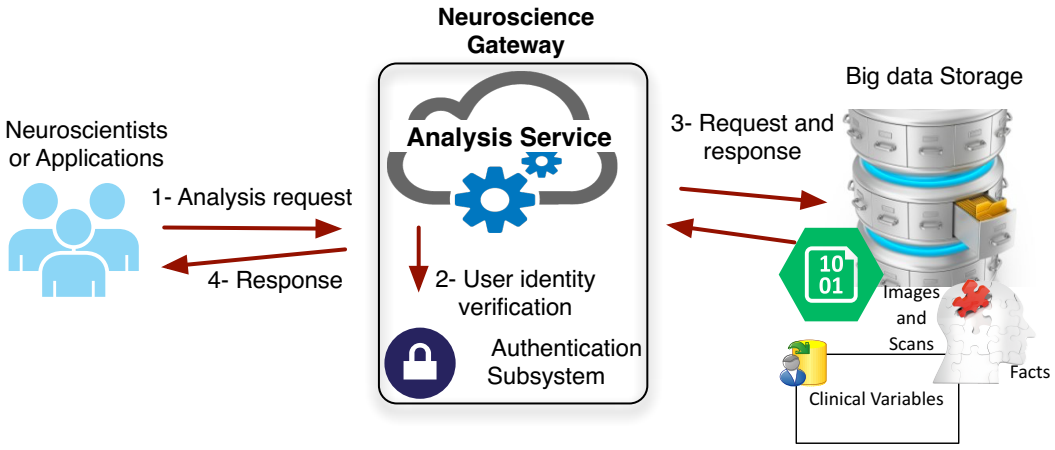

Figure 6: Functional Flow of the Analysis Service.

big data analysis service is shown as Figure 7. Through its big data Analysis Service Interface, the neuroscientists or other services can access the service functionality. The analysis program can be defined using both as Hive (https://hive.apache.org) or Pig (https://pig.apache.org) scripts. The Apache Hive queries facilitate reading large datasets residing in distributed storage using Map Reduce. Pig is a high-level language for expressing data analysis programs and it execute its Hadoop jobs in MapReduce. When a user/client successfully connects to the Analysis Service, the data analysis program is parsed and its Hadoop jobs are executed. Once MapReduce completes the execution the outputs are presented to the user. If the job was to calculate the Disease Identification Number through Fuzzy Processing, then the users are also given the option to store the calculations in the persistency storage for future referencing. In order to test the functionality and outcomes, especially the Fuzzy processing that is also the major focus of this paper, a dedicated work package was designed to specify neuroscientists' analyses using real case studies and the testing was performed in various sessions. In the following sections, we provide details on the fuzzy processing requirements, implementations and data analysis outcomes.

7. The Fuzzy Processing - Disease Identification Number Requirements and Calculations

This section presents the technique used to deal with the Alzheimer's patients. We calculated a reference number called *"Alzheimer's Disease Identification Number (ADIN)"*. Based on the intensity of Alzheimers disease, ADIN is a number $\in[0, 10]$. One of the most interesting reflections is that for a particular intensity of Alzheimer's disease, ADIN sorts a patient in a list of millions of patients. Based on sorting, one can also see how a given intensity evolves in terms of percentile over the years.

Please cite this article as: K. Munir, A. de Ramón-Fernández and S. Iqbal et al., Neuroscience patient identification using big data and fuzzy logic–An Alzheimer's disease case study, Expert Systems With Applications, https://doi.org/10.1016/j.eswa.2019.06.049

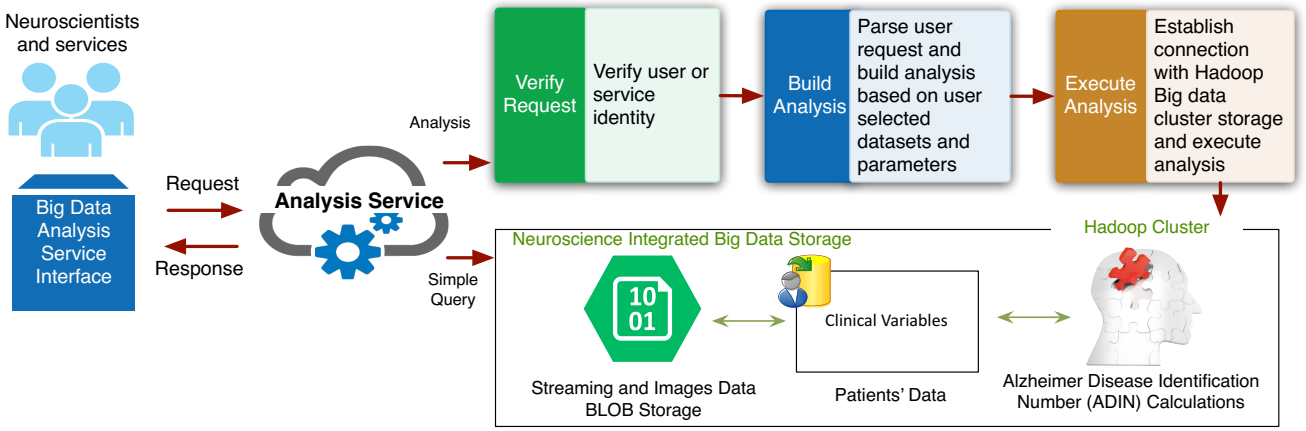

Figure 7: Neuroscience Big Data Analysis Service.

7.1. Alzheimer's diagnosis criteria

One of the most important aspects in the treatment of Alzheimer's is to have an accurate and early diagnosis, so that it shouldn't be confused with another type of dementia or neurodegenerative pathologies. Alzheimer's and dementia are different concepts, although closely related, since Alzheimer's is the main cause of dementia. For this reason, getting complete previous evaluation of a patient's main cognitive areas is essential along with their complete contextual information. Hence, ADIN cannot be calculated if there are missing or incomplete values, as in this case there is a risk of confusing the diagnosis with another type of dementia or with normal cognitive deterioration typical of age.

The Alzheimer's criteria, proposed by the National Institute of Neurological and Communicative Disorders and Stroke and the Alzheimer's Disease and Related Disorders Association (called NINCDS-ADRDA) (36), is among the most used criteria for Alzheimer's disease. Their objective is to give a more accurate approximation of the intensity of an Alzheimer's patient on a probable diagnosis of Alzheimer's according to the NINCDS-ADRDA criterion. It establishes 3 necessary conditions for a probable diagnosis of Alzheimer: (1) progressive worsening of memory; (2) affectation of at least two cognitive areas; and (3) previous absence of systemic or cerebral diseases that may be the cause of the symptoms presented by the patient. These three conditions are considered necessary to establish the diagnosis of probable Alzheimer's disease. In the case of these necessary conditions described are not met, the system will discard the diagnosis. It may be useful to note that the criteria for the diagnosis of Alzheimer's proposed by NINCDS-ADRDA worldwide references for the diagnosis of Alzheimer's disease, do not categorically specify cognitive areas that may be in conflict or that have a greater weight in the diagnosis of the disease. Table 1 shows the main cognitive abilities that may be affected by the disease.

All of the symptoms can be grouped into 8 generic cognitive areas (35):

- **Attention**: It refers to the capacity of an individual to focus its mind on a specific stimulus or task, despite the presence of other environmental stimuli.
- **Memory**: It is the ability of the brain to store information and recover it voluntarily. The information can be stored for a long period of time (long-term memory) or for a short period of time (short-term memory).

Please cite this article as: K. Munir, A. de Ramón-Fernández and S. Iqbal et al., Neuroscience patient identification using big data and fuzzy logic–An Alzheimer's disease case study, Expert Systems With Applications, https://doi.org/10.1016/j.eswa.2019.06.049

Cognitive abilities impaired (35; 36; 43)	
1. *Selective attention*	10. *Perceptual organization*
2. *Sustained attention*	11. *visoconstruction*
3. *Divided attention*	12. *Planning and graphic sequencing*
4. *Information processing speed*	13. *Simple tasks of action and inhibition*
5. *Short-term memory*	14. *Orientation*
6. *Long-term memory*	15. *Motor, verbal and graphic sequence*
7. *Verbal expression*	16. *Capacity of abstraction and reasoning*
8. *Listening comprehension*	17. *Instrumental and regular activities of daily living*
9. *Visual perception*	18. *Changes in behaviour and Social cognition*

Table 1: Parameters for the construction of the membership functions.

- **Language**: The ability to reason with verbal or non-verbal content, establishing among them principles of classification, ordering, relationship and meanings.
- **Visual skills**: The ability to interpret an information that is collected through sight.
- **Praxis**: The ability to plan and carry out motor activities.
- **Problem solving skills (Reasoning and abstraction)**: The ability of reasoning in order to understand a given information and find solutions to problems with verbal or non-verbal content, and establishing principles of classification among them such as ordering, relationship and meanings.
- **Social functions**: About how people process, store, and apply information about other people and social situations.
- **Normal/Instrumental activities of daily living**: These include all daily life activities that have a specific meaning and purpose (e.g. dressing, eating, etc) and it also includes complex activities that require a higher level of personal autonomy (e.g. use the telephone, preparation of the own food, etc).

In the next section, the system implementation is presented including the definition of membership functions and generation of inference rules. In addition, selected real life case studies have been described to illustrate the working of the system. Finally, a discussion is presented on the results and possibilities offered by the calculation of ADIN to manage Alzheimer's disease.

7.2. System Development and Processing of Alzheimer's Disease Identification Number

Figure 8 shows an underlying principle of fuzzy logic implementation. First, the intensity of the patient's symptoms is evaluated and recorded as crisp values. Then, these crisp inputs are converted into values obtained from the corresponding membership functions. The rules define the necessary knowledge base for the inference engine to generate an output set for each of them. Next, an aggregate membership function that is composed by the output set of each rule is formed. Using this function, a most representative value of the total output set is obtained which represents the disease intensity.

In general, there are two widespread fuzzy inference systems called Mamdani (33) and Sugeno (47). Mamdani is more intuitive and well suited to human input. In contrast, Sugeno system is

Please cite this article as: K. Munir, A. de Ramón-Fernández and S. Iqbal et al., Neuroscience patient identification using big data and fuzzy logic–An Alzheimer's disease case study, Expert Systems With Applications, https://doi.org/10.1016/j.eswa.2019.06.049

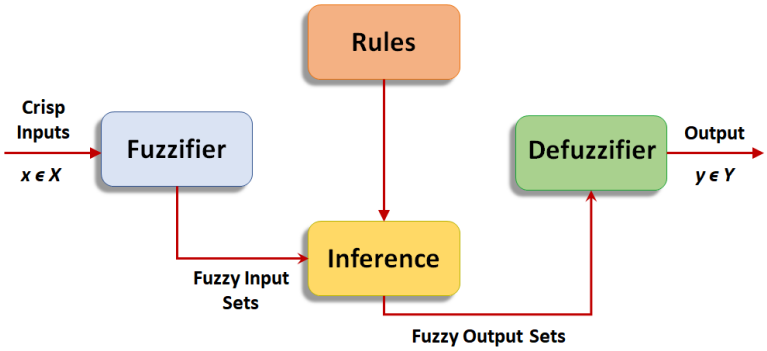

Figure 8: Block diagram of a fuzzy logic system where crisp input could be a scalar value or a vector.

known for being computationally more efficient and well suited to mathematical analysis. The main difference between them is the way crisp output is generated from fuzzy inputs. Sugeno output membership functions are either linear or constant and it calculates final output as a weighted average of all rule outputs. On the contrary, with Mamdani, the output membership functions can be modelled according to ones needs and the final output is calculated through defuzzification. Thus, Sugeno has better processing time since the weighted average replaces the time consuming defuzzification process. However, Mamdani offers a greater expressive power and interpretability (27). Due to the interpretable and intuitive nature of rules, Mamdani matches better to our needs, so we have based the implementation of our system on it (as shown in Figure 9).

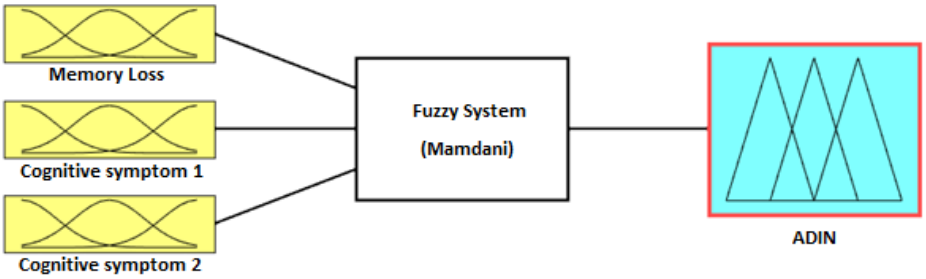

Figure 9: ADIN calculation as a function of 3 cognitive symptoms.

The crisp values correspond with the assessment of the disease symptoms that are modelled though their membership functions. For the fuzzification and defuzzification of input and output values we divide them into four stages: i.e. *normal*, *mild*, *moderate* and *severe*, which follow sigmoidal and normal distributions (as shown in the Figure 10). The sigmoid function allows to describe various natural processes and complex systems. Their progression varies from initially low level to reaching at the maximum levels, which is obtained after a certain time with an intermediate acceleration. Therefore, it is a suitable function for describing the evolution of an individual from a normal (i.e. not affected) status to the appearance of first symptom that causes a progressive impairment. In such a case, it is appropriate to represent the deterioration of a patient from a moderate status to a severe grade of impairment, which will be prolonged over time. The sigmoidal function is defined by its lower limit **a**, upper limit **b** and the **m** value or inflection point, such that $a<m<b$. The slope of the curve increases with the **a - b** distance (as shown in Figure 11a).

A normal or gaussian distribution is commonly used to describe psychological variables such as intelligent quotient and other cognitive parameters. They usually represent a distribution in

Please cite this article as: K. Munir, A. de Ramón-Fernández and S. Iqbal et al., Neuroscience patient identification using big data and fuzzy logic–An Alzheimer's disease case study, Expert Systems With Applications, https://doi.org/10.1016/j.eswa.2019.06.049

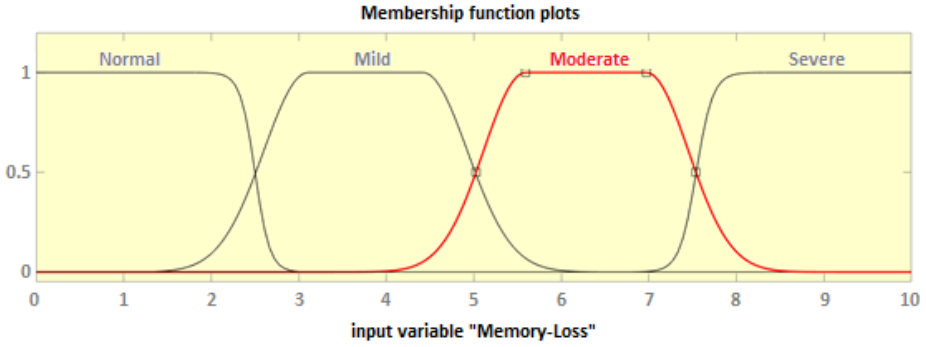

Figure 10: Membership functions for input and output values.

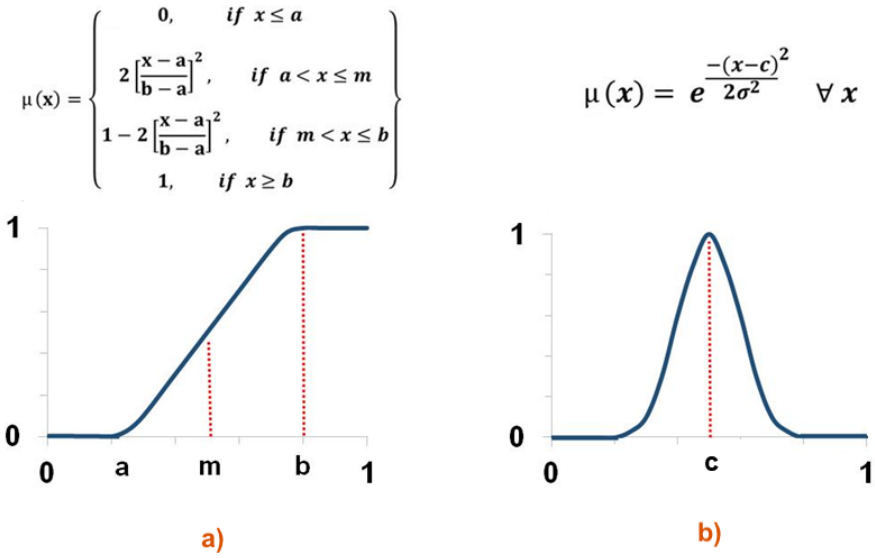

Figure 11: (a) Sigmoidal membership function and (b) Gaussian membership function.

a given set of general population that can be approximated to a normal function. Through this function we can represent intermediate stage(s) of a disease which although can be controlled for a certain time, can still end up evolving towards other greater affectation. It is defined by its average value c and the variance (σ^2) (see Figure 11b). Such membership functions can also be converted to triangular or trapezium membership function, as needed. Table 2 displays empirically selected parameters for the construction of functions. This can be used to replicate the system.

In order to define the fuzzy set, Mamdani inference system needs a set of rules that are usually in the form of implications. In such implications, *input variables* form the antecedents of conditionals and *output variables* constitute the consequents. The greater the number of rules is, the more defined the diffuse set will be. As discussed above, memory deterioration and at-least two cognitive areas are sufficient for the possible diagnosis of Alzheimer. The combination of these three parameters in their different states generates a total of 64 possible cases. These generated cases range from a most favourable case/scenario (i.e. no area of cognition affected) to a most unfavourable scenario (e.g. severe cognitive impact in all of them) and define the fuzzy set (a few examples are listed in Table 3).

Please cite this article as: K. Munir, A. de Ramón-Fernández and S. Iqbal et al., Neuroscience patient identification using big data and fuzzy logic–An Alzheimer's disease case study, Expert Systems With Applications, https://doi.org/10.1016/j.eswa.2019.06.049

Memory Loss & Cognitive Symptoms (Inputs)	**Function parameters**
Normal (N)	μ(x) = sigmf(x,a,c) ⇒ μ(x) = sigmf(x, -11.9, 2.5)}
Mild (M)	μ(x) = Gauss2mf (x,σ1,c1,σ2,c2) ⇒ μ(x) = Gauss2mf (x, 0.506, 3.105, 0.506, 4.42)}
Moderate (Mo)	μ(x) = Gauss2mf (x,σ1,c1,σ2,c2) ⇒ μ(x) = Gauss2mf (x, 0.48, 5.59, 0.48, 6.971)}
Severe (S)	μ(x) = sigmf(x,a,c) ⇒ μ(x) = sigmf(x, 10.5, 7.54)}
ADIN (Output)	**Function parameters**
Normal (N)	μ(x) = sigmf(x,a,c) ⇒ μ(x) = sigmf(x, -11.9, 1.49)}
Mild (M)	μ(x) = Gauss2mf (x,σ1,c1,σ2,c2) ⇒ μ(x) = Gauss2mf (x, 0.506, 2.1, 0.506 4.011)}
Moderate (Mo)	μ(x) = Gauss2mf (x,σ1,c1,σ2,c2) ⇒ μ(x) = Gauss2mf (x, 0.48 5.17 0.48 6.839)}
Severe (S)	μ(x) = sigmf(x,a,c) ⇒ μ(x) = sigmf(x, 10.5, 7.377)}

Table 2: Parameters for the construction of the membership functions.

1.If (ML is N) and (CS1 is N) and (CS2 is N) then (ADIN is N)
2.If (ML is N) and (CS1 is N) and (CS2 is M) then (ADIN is N)
3.If (Memory-Loss is N) and (CS1 is N) and (CS2 is Mo) then (ADIN is N)
. . .
. . .
31.If (ML is M) and (CS1 is S) and (CS2 is Mo) then (ADIN is Mo)
32.If (ML is M) and (CS1 is S) and (CS2 is S) then (ADIN is Mo)
33.If (ML is Mo) and (CS1 is M) and (CS2 is M) then (ADIN is M)
. . .
. . .
62.If (ML is S) and (CS1 is S) and (CS2 is M) then (ADIN is S)
63.If (ML is S) and (CS1 is S) and (CS2 is Mo) then (ADIN is S)
64.If (ML is S) and (CS1 is S) and (CS2 is S) then (ADIN is S)
Table key
ML = Memory Loss; N = Normal; CS1 = Cognitive Symptom 1; CS2 = Cognitive Symptom 2; M= Mild; Mo = Moderate; S = Severe.

Table 3: Example - combination of three clinical parameters in different states generates 64 possible cases

Once we have the crisp values (i.e. the membership functions and the fuzzy inference system), the final step is to find a quantifiable value for the fuzzy output sets by obtaining the corresponding ADIN values. Thus, it is possible to represent the values of ADIN in a response surface by contrasting two of the symptoms and fixing the third in its average value (MAT) (as shown in the Figure 12).

To demonstrate ADIN computation and the working of fuzzy system, this section describes two selected case studies from out patients data repository. Since our Fuzzy System is generic and flexible, it can straightforwardly be extrapolated to the rest of patients in the data repository. This is because ADIN can be calculated independently out of two symptoms besides memory loss (as described in the Alzheimer's clinical guidelines, outlined in Section 7.1):

Selected Example Case study 1: *A 65 year old patient who comes for consultation for the first-time reports changes in mood, slight long-term memory loss and attention problems. The evaluation of cognitive abilities shows affectation in three areas*:

– ***Long-term memory loss***: *The patient is unable to store few memories for a period greater than six (6) months. The patient was evaluated through the* ***California Verbal Learning Test****, which consists of loud reading a list of 16 words to the patient over 5 learning trials. After each trial, the number of remembered words was recorded. The final score in this test*

Please cite this article as: K. Munir, A. de Ramón-Fernández and S. Iqbal et al., Neuroscience patient identification using big data and fuzzy logic–An Alzheimer's disease case study, Expert Systems With Applications, https://doi.org/10.1016/j.eswa.2019.06.049

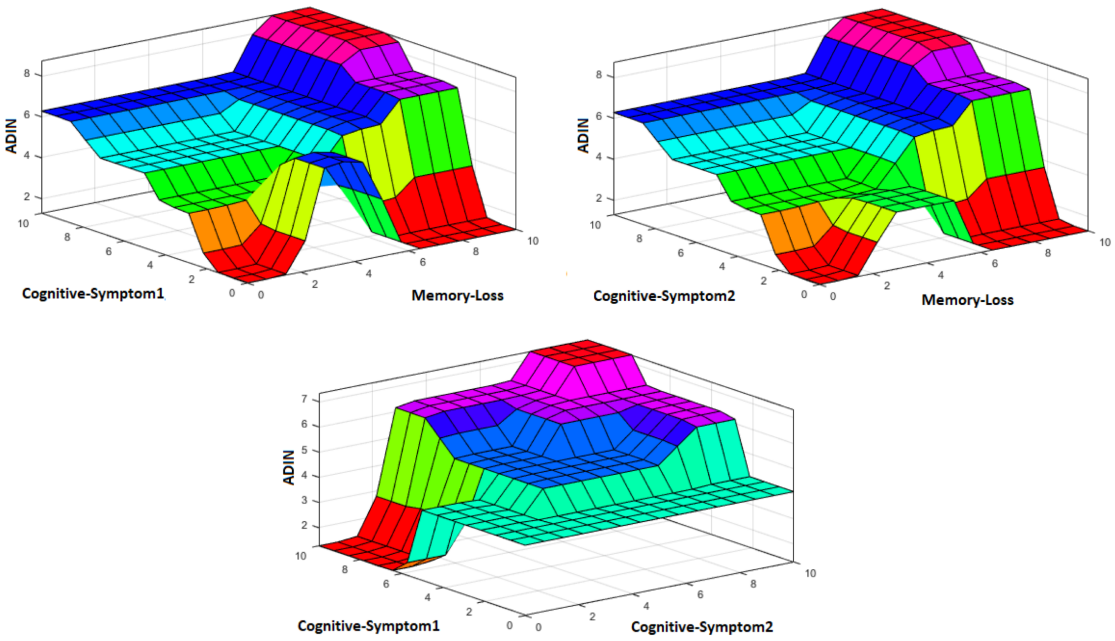

Figure 12: ADIN response surface for every pair of 2 symptoms.

was 3.5, on a scale of 0-10. Here, 0 on a scale of 0-10 means that there is no impairment of the cognitive area and 10 means severe impairment (same is also applicable to other tests).

- ***Sustained attention***: *The patient presents difficulty maintaining an attentional focus over a prolonged period. The* ***Continuous Performance Test****, that evaluates the attentional capacity of the patients, shows a score of 6.*

- ***Visual analysis***: *The perception and visual memory of the patient is evaluated by the* ***Visual Retention test of Benton****, where the patient had to draw a figure that was previously shown to him for a brief period of time. The score obtained was 4.*

For these input values, our Fuzzy System calculates an ADIN value of 3.78, compatible with AD in a mild affectation degree (as shown in Figure 13).

Selected Example Case study 2: *A 76 year old patient who has already been diagnosed with Alzheimer goes to the doctor's office for medical follow-up. The patient has short-term memory loss, difficulty expressing himself, and listening comprehension. In addition, he has a depressive clinical picture and is unable to perform usual daily tasks. The evaluation of cognitive abilities shows affectation in more than 3 areas, the most notable being*:

- ***Short-term memory loss***: *Patient's memory is evaluated from the* ***Digit subtest of WAIS-III*** *scale. This subtest evaluates the patient's ability to memorize and repeat series of numbers in the same order in which they are dictated and in reverse order. The score obtained was 8, on a scale of 0-10. Here, 0 on a scale of 0-10 means that there is no impairment of the cognitive area and 10 means severe impairment (same is also applicable to other tests).*

- ***Alteration of speech***: *The patient's inability to express himself through speech is evaluated. For this, the* ***Boston*** *test was performed in its communication and exposure part. This part*

Please cite this article as: K. Munir, A. de Ramón-Fernández and S. Iqbal et al., Neuroscience patient identification using big data and fuzzy logic–An Alzheimer's disease case study, Expert Systems With Applications, https://doi.org/10.1016/j.eswa.2019.06.049

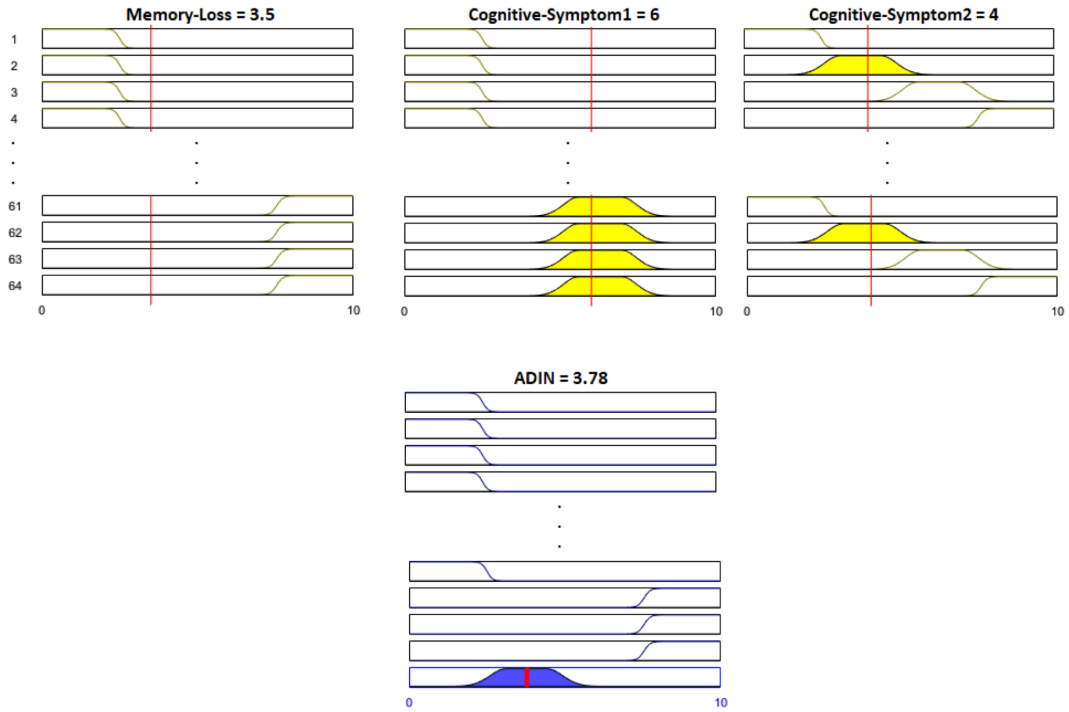

Figure 13: ADIN calculation for a patient with mild affectation degree.

consists of having an informal conversation with the patient and showing him a series of images for describing them. The score obtained was 6.5.

- ***Listening comprehension***: *The patient has difficulty understanding and interpreting orders. The patient was examined through the* ***Token*** *test, which consists of a series of tests where the patient must correctly identify colours, shapes and sizes. The score obtained was 7.*

- ***Instrumental activities of daily living***: *It is confirmed that the patient is unable to perform complex activities where a higher level of personal autonomy is required (e.g. preparation of own food). It is estimated that his degree of affectation is medium (i.e. 6/10).*

When dealing with a patient who has affected more than three areas of cognition, the ADIN is calculated by using the memory loss and the other two symptoms with a higher score (i.e. alteration of speech and listening comprehension). For these three values, our Fuzzy system calculates an ADIN of 8.48, compatible with having AD in a severe affectation degree (as detailed in Figure 14).

Selected Example Case study 3: *A 77 year old woman goes for memory alteration consultation. She also presents easy fatigue, anorexia and dyspnea. Until 9 months ago, she was autonomous for all the basic activities of daily life, but as a result of a respiratory infection associated with anemia and other processes (such as accidental fall), she started getting general fatigue and symptoms of cognitive impairment. After one month of pharmacological treatment, the patient worsened emotionally, got more forgetfulness and no longer goes out on the street. A later evaluation of cognitive abilities showed affectation in multiple areas, being the following ones the most impaired:*

Please cite this article as: K. Munir, A. de Ramón-Fernández and S. Iqbal et al., Neuroscience patient identification using big data and fuzzy logic–An Alzheimer's disease case study, Expert Systems With Applications, https://doi.org/10.1016/j.eswa.2019.06.049

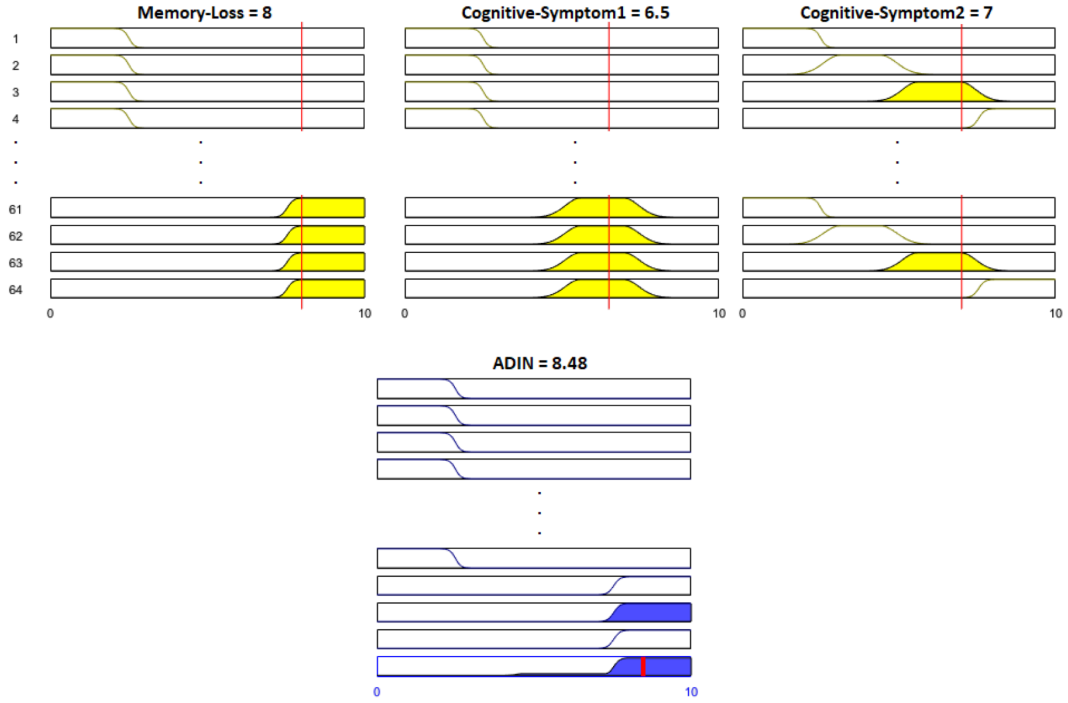

Figure 14: ADIN calculation for a patient with severe affectation degree.

- ***Long-term and short-term memory loss:*** *The patient has a progressive and accelerated deterioration of memory, the short and long-term memory lapses are becoming more frequent, and she has trouble remembering even with help. Patient's memory is evaluated from the* ***Digit subtest of WAIS-III*** *scale and the* ***California Verbal Learning*** *test. The score obtained was 7 and 8 respectively, on a scale of 0-10.*

- ***Alteration of speech with poor understanding:*** *The speech of the patient is sparse, with multiple nominative mistakes (children and grandchildren names), and continuous use of semantic and phonemic paraphasias. The* ***Boston and Token*** *tests showed a score of 7 out of 10, compatible with a severe affectation.*

- ***Alteration of visuospatial skills:*** *The patient has the visuospatial abilities affected; she even has visual hallucinations (she usually sees a person). Visuoperceptive and Visuoconstructive functions were evaluated through the* ***Clock*** *test. The test is based on two simple exercises; first, the patient must draw a clock with all the hours and the hands dialling a certain time. To do it correctly patient must order the numbers, draw the clock face and place the clock hands in the indicated position. Next, patient is asked to copy a drawn clock as accurately as possible. The total score test was compatible with severe impairment (7/10).*

- ***Instrumental activities of daily living:*** *The patient is highly dependent, being unable to perform either simple or complex activities, where a higher level of personal autonomy is required. She also has urinary and stool incontinence. It is estimated that her degree of affectation is severe (i.e. 7/10).*

For these input values, our Fuzzy System calculates an ADIN value of 8.78, compatible with having AD in a severe affectation degree.

Please cite this article as: K. Munir, A. de Ramón-Fernández and S. Iqbal et al., Neuroscience patient identification using big data and fuzzy logic–An Alzheimer's disease case study, Expert Systems With Applications, https://doi.org/10.1016/j.eswa.2019.06.049

Selected Example Case study 4: *A 70 year old woman with no clinical history of interest comes to the clinic accompanied by her family members because they notice memory deterioration of at least three years of evolution. She begins to suffer affective disorders (feelings of uselessness, frustration, sadness), but the depressive symptoms improved only transiently. Moreover, she is showing more and more carefree in her personal care. The evaluation of cognitive abilities shows affectation in 3 cognitive areas in early stages:*

- ***Short-term memory loss:*** *The patient shows lapses in the reception of messages (she forgets messages that leave her or transmits them repeatedly to the same person). Patient's memory is evaluated from the* ***Digit subtest of WAIS-III scale****. The score obtained was 4.5, on a scale of 0-10.*
- ***Alteration of speech:*** *Her communication skills has been progressively reduced. She finds it difficult to maintain a thread of conversation and to use right words to the context, not being successful with the nomination of objects of common use. For this, the* ***Boston*** *test was performed in its communication and exposure part. The score obtained was 5.*
- ***Alteration of orientation:*** *The patient had been delayed a couple of times upon arriving home because she had had difficulties returning, having been lost and needed help to return on one occasion.* ***Benton Temporal Orientation*** *test was used to assess the orientation in time (weekday, day of the month, current month, current year, time of the day). The score obtained was 5.*

For these input values, our Fuzzy System calculates an ADIN value of 4.96, compatible with AD in a mild-moderate affectation degree.

In all of the above case studies, ADIN has been calculated by using the centroid method as defuzzification method and the following steps were followed:

- ***Step 1:*** In order to find the cut off points with the membership functions, each rule is evaluated individually by entering it in a graph with the corresponding crisp values. Since the rules have been built up from three variables, it can obtain a maximum of three cut off points i.e. there exists a membership function value to each crisp value.
- ***Step 2:*** Since the system is dealing with "*AND*" rules, the minimum value of the cut off points obtained is projected through a horizontal line until cutting the corresponding output membership function. Thus, generating an area under the curve that is usually in a trapezoid shape.
- ***Step 3:*** All the generated trapezoids in Step 3 are superimposed to create a single geometrical figure.
- ***Step 4:*** Finally, the centroid of the graph is calculated, or in other words, its centre of symmetry. The X coordinate of the centroid is the defuzzified value.

As explained above, a complete cognitive assessment of the patient is vital to diagnose a probable case of Alzheimer. This evaluation allows the generation of a specific ADIN for each Alzheimer's patient and evaluate their contextual situation. Using ADIN, it is possible to establish a personalised evolution of each patient and clusters them based on age, sex or any other characteristic.

Please cite this article as: K. Munir, A. de Ramón-Fernández and S. Iqbal et al., Neuroscience patient identification using big data and fuzzy logic–An Alzheimer's disease case study, Expert Systems With Applications, https://doi.org/10.1016/j.eswa.2019.06.049

By having a database with multiple patients evaluated with their corresponding ADIN (see Table 4), it is possible to assess different scenarios: contrast patients with similar characteristics, make trends, warn of extreme cases or monitor whether the evolution of a patient remains stable or, on the contrary, is suffering an anomalous progressive deterioration. As an example, Figure 15 shows the contextual situation of the four selected case studies presented based on their ADIN, age, and the 25th, 50th and 75th percentiles calculated on a database of 300 patients.

	Patient 1	Patient 2	.	Patient case study 1	Patient case study 2	Patient case study 3	Patient case study 4	.	Patient 299	Patient 300
Age	50	52	.	65	76	77	70	.	86	89
Gender	F	F	.	M	M	F	F	.	M	M
Short-term memory	3	3.5	.	2	8	8	4.5	.	7	8.5
Long-term memory	2	3	.	3.5	6	7	3	.	6	9
Selective attention	2	2	.	5	6	6.5	5	.	6	8.5
.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.
Instrumental and regular activities of daily living	2	3	.	2	6	7	2	.	5	8
Changes in behavior and social cognition	1	2	.	3	6	6	2	.	5.5	9
ADIN	**3.77**	**3.77**	.	**3.78**	**8.48**	**8.78**	**4.96**	.	**7.2**	**8.79**

Table 4: Selected example subset of clinical database of Alzhimer's patients

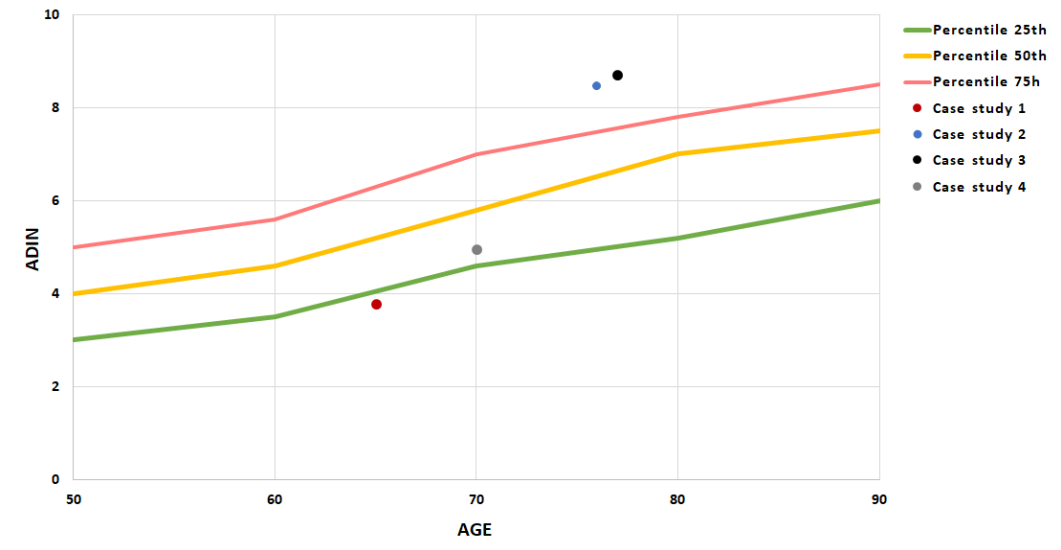

Figure 15: Contextual situation of the selected four study cases

7.3. ***Discussion***

Although Alzheimer's it is a degenerative disease, its correct treatment in the early stages can help to stabilise or stop the progressive deterioration. Determining a true degree of affectation to which a patient is subjected is a crucial aspect to manage Alzheimer's disease. Therefore, the information provided by the ADIN can play a key role in putting each case in context and helping in its interpretation. Moreover, sorting of all the patients is easily done based on ADIN. For every new patient, a tiny amount of fuzzy processing assigns a personal ADIN to that patient, and a new record is sorted in the existing list of millions of patients. In this way, we obtain the status and position of a patient with respect to other patients i.e. a patient is in the 30^{th} percentile is doing better than 70% of the other Alzheimer recorded patients.

Please cite this article as: K. Munir, A. de Ramón-Fernández and S. Iqbal et al., Neuroscience patient identification using big data and fuzzy logic–An Alzheimer's disease case study, Expert Systems With Applications, https://doi.org/10.1016/j.eswa.2019.06.049

Furthermore, an increase of the ADIN in each patient provides information about the evolution of the patient and whether the disease is getting worse or has been stabilised. This increase can be compared with rest of patients in the data repository; and therefore, can assess a positive or negative trend with respect to other patients. Thus, this calculation enables a short-term estimation of the progression of the disease. Knowing the context of each individual with respect to other patients can help to suggest a most appropriate treatment or an estimated life expectancy. Moreover, with ADIN it is possible to clusters patients based on: age group, ethnicity, place of birth or any other characteristic. This makes it possible to study a patient in a more analytical way; for example, (a) the relationship of the ADIN with other diseases (e.g. diabetes, hypertension and narrowing of the carotid artery), and (b) how the risk factors such as smoking, sleep disorder, sedentary lifestyle, family history etc. influence in the development of the disease etc. Nevertheless, since the implemented approach is not fully specific to a particular disease, it can be extended to other diseases or clinical observations.

8. Conclusions

The adoption of big data and latest clinical techniques have eased numerous problems of dealing with large neuroscience data sets including images. However, there is still a need for systems and methodologies that can provide an intuitive analysis mechanism of heterogeneous neuroscience dataset of a very large population. Moreover, such analyses mechanisms require enabling biomedical researchers to conduct neuroscience analyses in a user friendly way. The N4U project and the technique presented in this paper contributes in this direction by enabling storage, indexing and interlinking of neuroimaging datasets, algorithms and data dictionaries etc. One of the main challenges faced and tackled during the implementation of this system were related to the large data sets' volume, diverse structures, dictionaries of datasets and the dissimilar ways in which relationships between data sets and image were conventionally maintained. The implementation of various persistency services, information services and related methodologies helped in simplifying the data analysis process.

With the implementation of ADIN, it became straightforward to identify a patient along with the intensity of Alzheimer's disease. This number is more meaningful when it comes to clustering a large number of (in thousands and millions) of patients. Clusters based on the intensity of disease provide insight about the behaviour of Alzheimer's disease. The implemented system was based on realistic datasets, demonstrators and making use of real-life neuroscience case studies, which also served to evaluate and significantly improve our technique. The testing of system's functionality, outcomes, especially the Fuzzy processing to generate and utilise disease identification numbers have been based on a detailed users requirements gathering process.

There are various possible future extensions or applications of this work, such as studying other ADIN calculation methods based on average values of cognitive symptoms. This can used to compare and study which ones are more representative and can better define the real status of a patient i.e. providing a higher level of confidence. Such a study will be conditioned by (at-least) two factors: (1) the accuracy of the applied diagnostic criteria; and (2) how well the membership functions are adjusted. The definition of these functions should aim to describe a best possible rating system of the tests that evaluate the different cognitive areas according to the scores obtained by the patients. For the study of confidence level, a mid-long-term study on the evolution of patients is needed. That is why, as a continuation of this work, two future research directions are proposed. First, study of the evolution of ADIN in patients for confirmation of

Please cite this article as: K. Munir, A. de Ramón-Fernández and S. Iqbal et al., Neuroscience patient identification using big data and fuzzy logic–An Alzheimer's disease case study, Expert Systems With Applications, https://doi.org/10.1016/j.eswa.2019.06.049

trends and making statistical analysis to validate the ADIN accuracy. Second, an evolution of the designed fuzzy inference system towards a neuro-adaptive fuzzy inference system, which learns about new records so that the output membership function (the one that defines the result) is progressively adjusted to improve confidence of ADIN.

In addition, clinical applications can be developed to interact with patients' datasets such as using domain knowledge ontologies (see (39)), automatic or user-friendly definition of neuroscience analyses (24), patients' profiling using artificial intelligence and information visualisation (14). In relation to the use of domain knowledge, the semantic knowledge can also be used in analyses to increase search performance, creation of reference data and enable reasoning. Moreover, this work can be further enhanced if we correlate ADIN of patients with their social parameter such as income, social interaction, or number of dependents to get further insights. Since our approach is generic, the technique can be extended for other medical studies such as heart or diabetes. However, it will require working through several things e.g. clinical requirements, historical clinical practices, ingestion and structuring of the datasets, disease diagnosis criteria and availability of clinical case studies.

Acknowledgments

We are grateful to European Union in funding the neuGRID and neuGRID4You (N4U) (grant agreement no. 283562) projects that has provided grounding to this work. With special thanks to the partners of the N4U project, the Laboratory of Neuro Imaging (LONI) for Neuro Imaging datasets, researchers of Centre for Complex Corporative Systems (CCCS), researchers from the Clinical Robotics Arm Development lab of National University of Sciences and Technology (NUST), all neuroscience datasets providers, clinical requirements providers, users and testers of various components of the system, and the N4U project consortium. We are also thankful to the Spanish Ministry of Economy and Competitiveness (MINECO) for sponsoring De Ramón-Fernández (Ref. BES-2015-073611) to work on a part of this project at the University of the West of England.

References

Adhd_2016. the attention deficit hyperactivity disorder (adhd-200). `http://fcon_1000.projects.nitrc.org/indi/adhd200`. Accessed: 28 March 2016.

Adibe_2015. autism brain imaging data exchange (abide). `http://fcon_1000.projects.nitrc.org/indi/abide/`. Accessed: 15 September 2015.

Adni_2016. the alzheimer's disease neuroimaging initiative (adni) - adni1, adni2 and adni go. `http://adni.loni.usc.edu/`. Accessed: 25 June 2016.

Arwibo_2015. the alzheimer's repository without borders (arwibo). `http://www.arwibo.it`. Accessed: 10 August 2015.

Cobre_2016. the centre for biomedical research excellence (cobre). `http://fcon_1000.projects.nitrc.org/indi/retro/cobre.html`. Accessed: 06 Janury 2016.

Fbirn_2015. the functional bioinformatics research network (fbirn). `www.fbirnbdr.nbirn.net`. Accessed: 06 March 2015.

Fcp_2016. 1000 functional connectomes project (1000fcp). `http://fcon_1000.projects.nitrc.org/`. Accessed: 18 July 2016.

Indi_dwi_2016. the international neuroimaging data-sharing initiative for diffusion-weighted mri dataset (indi_dwi). `http://fcon_1000.projects.nitrc.org/indi/indi_ack.html`. Accessed: 16 April 2016.

Magnims_2016. the magnetic resonance in multiple sclerosis (magnims). `http://www.magnims.eu`. Accessed: 12 April 2016.

Mat_2018. generate fuzzy inference system output surface (gensurf). `https://es.mathworks.com/help/fuzzy/gensurf.html`. Accessed: 26 June 2018.

Nusdast_2016. the northwestern university schizophrenia data and software tool (nusdast). `http://niacal.northwestern.edu/projects/9`. Accessed: 21 May 2016.

Oasis_2016. open access series of imaging studies (oasis). `http://www.oasis-brains.org`. Accessed: 16 December 2016.

Senselab_2015. `https://senselab.med.yale.edu`. Accessed: 12 September 2015.

Arshad, B., Munir, K., McClatchey, R., Shamdasani, J., and Khan, Z. (2019). Neuroprov: Provenance data visualisation for neuroimaging analyses. *Journal of Computer Languages*, 52:72 – 87.

Book, G. A., Anderson, B. M., Stevens, M. C., Glahn, D. C., Assaf, M., and Pearlson, G. D. (2013). Neuroinformatics database (nidb)–a modular, portable database for the storage, analysis, and sharing of neuroimaging data. *Neuroinformatics*, 11(4):495–505.

Bridges, S. M., Vaughn, R. B., et al. (2000). Fuzzy data mining and genetic algorithms applied to intrusion detection. In *Proceedings of 12th Annual Canadian Information Technology Security Symposium*, pages 109–122.

Carvalho, C. M., Christina, D., Saade, M., Conci, A., Seixas, F. L., and Laks, J. (2017). A clinical decision support system for aiding diagnosis of alzheimer's disease and related disorders in mobile devices. In *Communications (ICC), 2017 IEEE International Conference on*, pages 1–6. IEEE.

Casillas, J., Cordón, O., Triguero, F. H., and Magdalena, L. (2013). *Interpretability issues in fuzzy modeling*, volume 128. Springer.

Cheung, K.-H., Lim, E., Samwald, M., Chen, H., Marenco, L., Holford, M. E., Morse, T. M., Mutalik, P., Shepherd, G. M., and Miller, P. L. (2009). Approaches to neuroscience data integration. *Briefings in bioinformatics*, 10(4):345–353.

Denoeux, T. (2013). Maximum likelihood estimation from uncertain data in the belief function framework. *IEEE Transactions on knowledge and data engineering*, 25(1):119–130.

D'Urso, P. and Giordani, P. (2006). A weighted fuzzy c-means clustering model for fuzzy data. *Computational Statistics & Data Analysis*, 50(6):1496–1523.

Gardner, D., Akil, H., Ascoli, G. A., Bowden, D. M., Bug, W., Donohue, D. E., Goldberg, D. H., Grafstein, B., Grethe, J. S., Gupta, A., et al. (2008). The neuroscience information framework: a data and knowledge environment for neuroscience. *Neuroinformatics*, 6(3):149–160.

Haissig, C. M., Woessner, M. A., and Pirovolou, D. K. (1998). Adaptive fuzzy controller that modifies membership functions. US Patent 5,822,740.

Hasham, K. and Munir, K. (2018). Reproducibility of scientific workflows execution using cloud-aware provenance (recap). *Computing*, 100(12):1299–1333.

Hathaway, R. J., Bezdek, J. C., and Pedrycz, W. (1996). A parametric model for fusing heterogeneous fuzzy data. *IEEE transactions on Fuzzy Systems*, 4(3):270–281.

Iqbal, S. and Boumella, N. (2012). Fuzzy controllers–recent advances in theory and applications.

Kaur, A. and Kaur, A. (2012). Comparison of fuzzy logic and neuro-fuzzy algorithms for air conditioning system. *International journal of soft computing and engineering*, 2(1):417–20.

KG, S., Venugopal, K., and Patnaik, L. (2006). Feature extraction using fuzzy c-means clustering for data mining systems. *IJCSNS*, 6(3A):230.

Li, M. J., Ng, M. K., Cheung, Y.-m., and Huang, J. Z. (2008). Agglomerative fuzzy k-means clustering algorithm with selection of number of clusters. *IEEE transactions on knowledge and data engineering*, 20(11):1519–1534.

Lu, Y., Ma, T., Yin, C., Xie, X., Tian, W., and Zhong, S. (2013). Implementation of the fuzzy c-means clustering algorithm in meteorological data. *International Journal of Database Theory and Application*, 6(6):1–18.

Mac Queen, J. et al. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 281–297. Oakland, CA, USA.

Malone, I., Cash, D., Ridgway, G., MacManus, D., Ourselin, S., Fox, N., and Schott, J. (2013). Miriad (minimal interval resonance imaging in alzheimer's disease).

Mamdani, E. H. and Assilian, S. (1975). An experiment in linguistic synthesis with a fuzzy logic controller. *International journal of man-machine studies*, 7(1):1–13.

Marenco, L., Wang, R., Shepherd, G. M., and Miller, P. L. (2010). The nif disco framework: facilitating automated integration of neuroscience content on the web. *Neuroinformatics*, 8(2):101–112.

McKhann, G., Drachman, D., Folstein, M., Katzman, R., Price, D., and Stadlan, E. M. (1984). Clinical diagnosis of alzheimer's disease report of the nincds-adrda work group under the auspices of department of health and human services task force on alzheimer's disease. *Neurology*, 34(7):939–939.

McKhann, G. M., Knopman, D. S., Chertkow, H., Hyman, B. T., Jack, C. R., Kawas, C. H., Klunk, W. E., Koroshetz, W. J., Manly, J. J., Mayeux, R., et al. (2011). The diagnosis of dementia due to alzheimer's disease: Recommendations from the national institute on aging-alzheimer's association workgroups on diagnostic guidelines for alzheimer's disease. *Alzheimer's & dementia: the journal of the Alzheimer's Association*, 7(3):263–269.

Michel, F., Gaignard, A., Ahmad, F., Barillot, C., Batrancourt, B., Dojat, M., Gibaud, B., Girard, P., Godard, D., Kassel, G., et al. (2010). Grid-wide neuroimaging data federation in the context of the neurolog project. *Studies in Health Technology and Informatics*, 159:112.

Munir, K., Ahmad, K. H., and McClatchey, R. (2015). Development of a large-scale neuroimages and clinical variables data atlas in the neugrid4you (n4u) project. *Journal of biomedical informatics*, 57:245–262.

Munir, K. and Anjum, M. S. (2018). The use of ontologies for effective knowledge modelling and information retrieval. *Applied Computing and Informatics*, 14(2):116 – 126.

Nanda, S. J. and Panda, G. (2014). A survey on nature inspired metaheuristic algorithms for partitional clustering. *Swarm and Evolutionary computation*, 16:1–18.

Ozyurt, I. B., Keator, D. B., Wei, D., Fennema-Notestine, C., Pease, K. R., Bockholt, J., and Grethe, J. S. (2010). Federated web-accessible clinical data management within an extensible neuroimaging database. *Neuroinformatics*, 8(4):231–249.

Patcha, A. and Park, J.-M. (2007). An overview of anomaly detection techniques: Existing solutions and latest technological trends. *Computer networks*, 51(12):3448–3470.

Reisberg, B. (2006). Diagnostic criteria in dementia: a comparison of current criteria, research challenges, and implications for dsm-v. *Journal of geriatric psychiatry and neurology*, 19(3):137–146.

Samwald, M., Lim, E., Masiar, P., Marenco, L., Chen, H., Morse, T., Mutalik, P., Shepherd, G., Miller, P., and Cheung, K.-H. (2009). Entrez neuron rdfa: A pragmatic semantic web application for data integration in neuroscience research. *Studies in health technology and informatics*, 150:317.

Sanchez, E., Toro, C., Carrasco, E., Bonachela, P., Parra, C., Bueno, G., and Guijarro, F. (2011). A knowledge-based clinical decision support system for the diagnosis of alzheimer disease. In *e-Health Networking Applications and Services (Healthcom), 2011 13th IEEE International Conference on*, pages 351–357. IEEE.

Small, S. L., Wilde, M., Kenny, S., Andric, M., and Hasson, U. (2009). Database-managed grid-enabled analysis of neuroimaging data: The cnari framework. *International Journal of Psychophysiology*, 73(1):62–72.

Sugeno, M. (1985). *Industrial applications of fuzzy control*. Elsevier Science Inc.

Tanaka, M. (1995). Fuzzy data processing method and data smoothing filter. US Patent 5,398,303.

Toro, C., Sanchez, E., Carrasco, E., Mancilla-Amaya, L., Sanín, C., Szczerbicki, E., Graña, M., Bonachela, P., Parra, C., Bueno, G., et al. (2012). Using set of experience knowledge structure to extend a rule set of clinical decision support system for alzheimer's disease diagnosis. *Cybernetics and Systems*, 43(2):81–95.

Velmurugan, T. (2014). Performance based analysis between k-means and fuzzy c-means clustering algorithms for connection oriented telecommunication data. *Applied Soft Computing*, 19:134–146.

Wang, L., Kogan, A., Cobia, D., Alpert, K., Kolasny, A., Miller, M. I., and Marcus, D. (2013). Northwestern university schizophrenia data and software tool (nusdast). *Frontiers in neuroinformatics*, 7:25.