Diagnosing Alzheimer's Disease using a Self-Organising Fuzzy Classifier

Jonathan Stirling[a], Tianhua Chen[a*], Magda Bucholc[b]
For the Alzheimer's Disease Neuroimaging Initiative[1]
[a]Department of Computer Science, School of Computing and Engineering, University of Huddersfield, UK
[b]Intelligent Systems Research Centre, School of Computing, Engineering & Intelligent Systems, Ulster University, UK

*Corresponding Author: Tianhua Chen(T.Chen@hud.ac.uk)

***Abstract*—Dementia is one of the major causes of disability and dependency among older people worldwide. Without treatment currently available to cure dementia or to alter its progressive course, one of the principal goals for dementia care set by the World Health Organisation is the early diagnosis in order to promote early and optimal management. In recognition of the potentials of fuzzy systems in effectively dealing with medical data, this chapter investigates the use of a very recently proposed Self-Organising Fuzzy (SOF) classifier for the prediction of Alzheimer's Disease against Mild Cognitive Impairment and being Cognitively Unimpaired with patient observations provided by the renowned Alzheimer's Disease Neuroimaging Initiative repository. The experimental study demonstrates the effectiveness of SOF, especially in combined use with the Recursive Feature Elimination feature selection.**

I. Introduction

Dementia is a progressive condition with an estimated 50 million cases worldwide in 2018 which is expected to more than triple to 152 million by 2050 [1]. Currently between 60% and 70% of dementia cases are attributed to Alzheimer's Disease (AD) [2] with the remainder consisting of different types including Vascular and Frontotemporal Dementia, each with different causes. Medical research into dementia continues with new types such as Limbic-predominant Age-related TDP-43 Encephalopathy (LATE) still being discovered [3].

Dementia can be generally defined as a condition that impairs the regular cognitive functions of the brain [2]. This impairment affects individuals differently and with varying severity, but typically affects memory, language, behaviour and the ability to carry out day-to-day tasks [4], [2]. Dementia primarily affects older people but can also affect younger people with an estimated 40 thousand cases in the UK alone for age groups under 65 [5]. With increases in life expectancy resulting in larger aged populations, the effect of dementia is expected to have significant implications for economies, healthcare services and society in addition to the substantial physical, psychological and social impact it has on sufferers, their families, friends and carers [2].

Given the impact of and lack of cure for dementia, it is important to diagnose those affected as early as possible and additionally target those at highest risk. Diagnosis allows symptom-slowing medication regimes to be used [6] and for patients in combination with their local healthcare services to prepare care plans to preserve as high a quality of life as possible, for as long as possible [7]. Presently, diagnosis tends to occur late due to: manual diagnosis being time consuming [8]; a lack of practitioners' confidence and / or training to be able to make correct decisions [9]; the limited amount of time in primary care patient interactions [10]; and waiting times up to 18 weeks in the UK [11]. Diagnosis is further complicated by diseases which can show similar symptoms to dementia as well as natural degeneration due to old age [4].

Recent research covers a range of data modalities and machine learning (ML) techniques when considering the diagnosis and pathology of dementia particularly with respect to AD. Of particular relevance for this work are those that consider AD prediction for individual patients based upon data including neuroimaging, neurocognitive assessments and other biomarkers. In general the use of multiple assessments have been shown to be a good indicator of AD [8], [12], while considering assessments individually do not do as well (e.g., when considering only the Mini-Mental State Examination (MMSE) [13], it only lead to less than 0.7 accuracy using SVM and MLP.)

Additionally, there is no global standard for what tests are applied to patients between regions as shown by differences in those available via ADNI and those described in literature [14], [8]. For instance, [8] considers the creation of an aggregated questionnaire, built from selected questions from multiple tests, applying multiple ML techniques as well as FS with the intention to identify a single optimal questionnaire for dementia diagnosis. The simultaneous use of multiple assessments has also been conducted in [12], with SVM models producing their best results on a feature subset consisting of four clinical assessments providing multi-class accuracy of 83% and AUC of 95%, though their work also considers results from other modalities including MRI / PET and CSF biomarker data.

Among recent advances in machine learning that have had

[1]Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf

success in healthcare domains [15], fuzzy systems, which are built on top of fuzzy sets that permit gradual assessment of set elements, enable the tolerance of uncertainty and imprecision that may result from linguistic descriptions while enquiring medical symptoms or noise that may result from inaccurate testing results. Whilst been widely applied in various domains [16], [17], [18], fuzzy techniques have also been intensively utilised in numerous medical applications to tackle challenges raising from healthcare, (e.g., [19], [20]). However, the application of fuzzy systems in diagnosing dementia is relatively limited in the literature. For instance, a recent work has attempted the Fuzzy Logic and Adaptive Neuro-Fuzzy inference Systems [7], however, the dataset used was severely limited in numbers of features and observations leaving questions as to whether the results were flawed.

As such, in working towards providing assistance for clinicians to conduct effective diagnosis of the Alzhermer's Disease, this chapter therefore proposes to utilise a very recently proposed Self-Organising Fuzzy (SOF) classifier [21] for the prediction of Alzheimer's Disease (AD) against Mild Cognitive Impairment (MCI) and being Cognitively Unimpaired (CU) patient observations. The patients' data used this research comes from the renowned Alzheimer's Disease Neuroimaging Initiative (ADNI) repository. The underlying testing bed is a group of 488 patients and 66 variables, with feature selection methods also applied to explore the effectiveness of selected variables in the experimental study.

The reminder of this chapter is organised as follows. Section II introduces the background of the SOF and the summary of the data set used. Section III describes the proposed pipeline. Section IV presents and discusses the experimental outcomes. Section V concludes the chapter and outlines ideas for further development.

II. Preliminary

A. The Self-Organising Fuzzy Classifier

The Self-Organsing Fuzzy (SOF) [21] is a non-parametric machine learning approach that considers dual phase training made up of offline (initial) and online (run-time) phases and works based upon the creation of computed centres of data clouds (prototypes) and distance measures. This work considers the euclidean distances of points in the offline phase of the technique for initial model development and testing due to the limited amount of data, and time constraints to implement data streaming services.

The offline training of SOF aims to generate 0-order AnYa type fuzzy rules for every unique class in the dataset which take the form of a series of disjunctions of similarities, or fuzzy membership degree between an input vector and "prototypes" of each class. A 0-order AnYa type fuzzy rule has the following form:

$$\begin{aligned} &IF\ (x\ \sim\ p_1^c)\ OR\ ...\ (x\ \sim\ p_n^c) \\ &\quad OR\ \ ...\ (x\ \sim\ p_N^c)\ THEN\ (class) \end{aligned} \tag{1}$$

Where x is the input vector; $\sim$ denotes similarity, which can also be seen as a fuzzy degree of satisfaction; p_n^c represents the n-th prototype for class c;

While other techniques build models for all classes combined, SOF training applies to subsets of the dataset split by the class of each observation, training each set independently with no interference between them. To create these fuzzy rule sets prototypes are derived from the unique samples for each class. For each sample, the multi-modal density is computed and the sample with the maximum density is added to a new list. The remainder of the samples are added to the list recursively, selecting the one with the minimum distance to the sample at the tail end of the list, noting that points cannot appear multiple times on the list.

From each class list, samples with a higher density than those immediately before and after them in the list are added as initial centroids. The items in each class list that were not selected are then used to formed into data clouds around those centroids, with each sample belonging to the one closest to them. The centre of each cloud is determined and the density computed using the number of samples in each cloud as a weighting. Neighbours of each cloud centroid are then computed based on whether the square distance between two of them is within a computed threshold based upon a user-provided granularity and the distances between points, after which the centroids with the highest density within a class neighbourhood are chosen as the prototypes for use in generating the AnYa rules. With respect to the granularity of the technique, the higher the value provided, the more prototypes are expected to be created resulting in a more finely defined area class area.

With respect to classification, once the model has been trained class predictions are made in two parts. First a local decision is made for each class, resulting in an output of the strength of the data point per class by taking the negative square distance between the new observation x and each prototype as exponent to the Euler's constant. The second part selects the maximum strength calculated from all rules to determine the final classification, effectively choosing the single closest prototype to the sample. As such, this work considers whether the use of SOF can predict whether a patient observation can be correctly diagnosed as CU, MCI or AD when considering a subset of ADNI data.

B. Data Summary

Data used in the preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). The ADNI was launched in 2003 as a public-private partnership, led by Principal Investigator Michael W. Weiner, MD. The primary goal of ADNI has been to test whether serial magnetic resonance imaging (MRI), positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment (MCI)) and early Alzheimer's disease (AD). For up-to-date information, see www.adni-info.org.

In this research, aside from the decision variable, the selection of specific predictors from the ADNI repository follows that as described in the recent work by [12], which is briefly summarised as follows.

The dataset used comprises of baseline patient records with any incomplete observations (across all selected variables) having been removed from the set. The resulting dataset comprises of 488 observations split between AD ($n = 76$), MCI ($n = 218$) and CU ($n = 194$). Each observation in the dataset consists of 66 independent variables made up of 26 discrete and 40 continuous items and two dependent variables, one continuous and one discrete. The variables are split by the mode of testing used to extract the values. First, a patient data modality consists of demographic information including age, gender, education level, marital status, ethnicity and race. It also covers 19 items from the patient's medical history (including alcohol / drug abuse, smoking, any cardiovascular or psychological issues) and expands to family history, describing if either parent or any sibling suffers from dementia.

A Clinical Measures (CM) modality covers the results of various test suites designed to determine issues with cognitive functions such as memory, learning and language. Specifically, this dataset collates the results from Alzheimer's Disease Assessment Scale 13 (ADAS-13), MMSE, three RAVLT trials, Functional Assessment Questionnaire (FAQ), Montreal Cognitive Assessment (MoCA) and two logical memory tests for immediate (LIMM) and delayed recall (LDEL), each of which tests specific areas of brain function.

MRI and PET scans make up a neuroimaging modality covering 23 of the variables provided in the datset. The MRI measures describe the volumes of seven areas of the brain along with the volumes of white and grey matter, white matter hyperintensities, cerebrospinal fluid and intercranial volume. MRI data is also used to extract boundary shift integral values for the whole brain and ventricles. PET scans, depending on the tracer compound used, measure: glucose metabolism across five regions of the brain and over all five; the mean uptake of tracer; and the sums of pixels Z-scores two or three standard deviations from 0. MRI and PET data can be considered as independent modalities within neuroimaging.

Finally, CSF extraction as an independent and invasive procedure is considered as an independent modality. The measures extracted from the fluid consist of the concentrations of certain proteins and the ratios between them.

With respect to dependent variables, the clinical decision for baseline observations is provided from ADNI, consisting of three classes describing CU, MCI and AD. It is worth noting that [12] made use of the Clinical Dementia Rating Sum of Boxes (CDRSB) continuous value as the dependent for regression and created their own classification variable based on CDRSB using two thresholds for the classification task. Whereas this work only looks at classification rather than regression problem, and to keep the feature parity with the original work, the CDRSB variable is ignored as an independent feature while the categorical diagnosis alone is used as the ground truth for each observation.

III. Methodology

This work makes use of a standard machine learning pipeline as shown in Figure 1. The following sub-sections discuss each individual component of the pipeline in further detail.

A. Data Pre-Processing

Built on top the pre-processing done in [12], further processing was applied to the dataset for this work that included applying min-max normalisation of all continuous variables, scaling them to between 0 and 1 to remove any feature domination issues caused by discrepancies in ranges [22] and to bring them into a normal distribution, allowing them to be considered equally. Discrete data with a cardinality of 2 were scaled to 0 and 1 values for use as binary features, while those with high cardinality (>2) were one-hot encoded to separate the categories into independent features. The use of one-hot encoding removes ordinality implications between values that could affect results when considering distances.

B. Modelling Approach

The processed data was then split randomly into stratified k-folds ($k = 10$), maintaining the class ratios between each fold after which $k-1$ folds were collated for use as training data for model development and the remaining fold kept separate for evaluating each trained instance. Given the imbalance between each of the data classes caused primarily by the limited number of AD samples, the training sets are modified to include generated observations created using the Synthetic Minority Oversampling TEchnique (SMOTE) to reduce bias towards the majority class [23]. The k-fold cross-validation process is repeated 5 times with randomised splits of the data upon each iteration to get a better estimate of the performance of the models across different combinations of data.

Feature Selection (FS) is applied as part of the pipeline to reduce the dimensionality of the data to sets that provide the most relevant information with respect to the classification task. Selection occurs for each k-th training set to reduce noise within the data [24], improving generalisation of model instances and computational efficiency [12]. For each set of derived features, the SOF model is developed using the oversampled training set and then evaluated against the left-out test fold. The results across all model evaluations are collated according to the metrics described in the evaluation section.

C. Feature Selection

While previous works considered univariate FS approaches, their use only considers how much a single variable can discriminate between prediction classes in isolation, leaving room for features to be removed that may provide greater value when used in combination with others. This work considers three FS options including: no FS, as a base comparison; Binary Particle Swarm Optimisation (BPSO) [25], [26]; and Recursive Feature Elimination (RFE) [27], with the latter two being examples of multi-variate selection methods. BPSO and RFE are wrapper techniques that determine a final feature subset

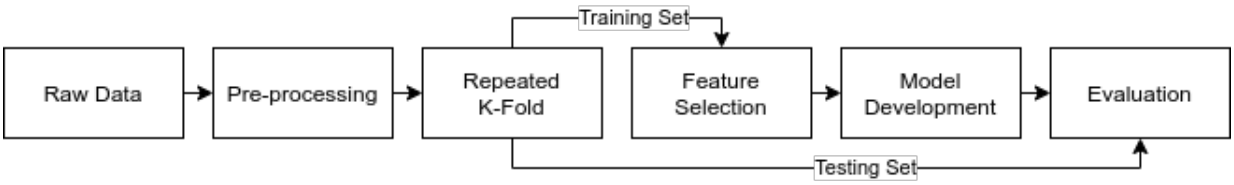

Fig. 1. ML Pipeline

by building and evaluating models separately to the rest of the pipeline, returning the subset that produces the best evaluation measure across the multiple sets that are attempted. The use of wrapper methods are computationally more expensive than filter methods, however, they are generally known to produce better results through the use of actual model evaluation [26].

BPSO works primarily on the same basis as the popular Particle Swarm Optimisation, which aims to optimise a problem by iteratively trying to improve a candidate solution with regard to a given measure of quality. It works by having a population of candidate solutions (dubbed particles), and moving these particles around in the search-space with each particle's movement influenced by its local best known position, but also guided toward the best known positions in the search-space, which are updated as better positions are found by other particles. In our case, each particle enables a certain feature subset out of original 66 selected features. The binary aspect of BPSO comes in the position updates by converting the computed velocities into probabilities using, for example, a sigmoid function. By comparing the feature probabilities to random numbers that are uniformly distributed between zero and one, the particle position is updated to include features where the random number is below the probability, and exclude / remove features otherwise. Evaluations are made for the new positions and the process continues until a threshold number of iterations has passed at which point the global best position, or feature set, is returned.

Unlike BPSO, RFE is not stochastic based, but systematic in that it recursively considers the relative importance of features at each iteration. The process starts by building and evaluating a model based on all available features which are then ranked according to a measure of importance such as, in the case of Support Vector Machines (SVM), coefficients learned by the model during training. From the ranked features, the least useful is removed from the set and a new model instance built and evaluated from the result. The process continues until a minimum threshold or a single feature remains. Based on the evaluation scores of the model for each feature subset, the set that produced the best values over all iterations is returned for use in the final model.

As the SOF classifer has no concept of coefficients or feature importance it is currently not possible to use RFE coupled with SOF directly for FS, as there is no metric to decide on what to remove. As such, SVM with a linear kernel was selected for the model as it has been successfully utilised in research [27]. The use of SVM was expanded for use in both RFE and BPSO to keep FS comparison fair between the two. With respect to FS model development, training of the SVM used stratified k-fold ($k = 5$) cross validation (CV) on the training set with the evaluation of the model in each case being based upon the accuracy of the predictions made on the held-out fold from the selection CV, averaged over all iterations.

D. Model Development

Model development is achieved using the SOF offline training mode described in previous sections. After preprocessing of the data and applying feature selection, the reduced dimensional data is input into the classifier. Upon starting training, the dataset is split into the three independent sets based on whether an observation is CU, MCI or AD according to the dependent variable. Each class is considered independently. For CU data, any duplicate observations (limited by the selected features) are removed leaving only the unique samples from the training set. For each unique CU observation the multi-modal density is calculated and these are added to a list in order of the least distance from the tail of the list, starting with the highest density seen in the subset. The CU observations producing a peak when considering the densities within the list are separated out to produce a set of centroids, to which all other listed observations are assigned based on which centroid they are closest to. This forms one or more CU data clouds from which a new cloud centroid is produced and a density computed. Based upon granularity parameter and the CU data points, neighbourhoods are determined that group nearby centroids and then reduce them to a single point by using the one with highest density. These CU centroids become the prototypes that are used to compare the distance to previously unseen data for classification describing the strength of the class. This process is repeated two more times: once for the MCI subset of data points, and once for AD resulting in three separate lists of prototypes groups.

IV. Experimentation

The methods discussed above were implemented in Python 3 using modules provided by Scikit Learn and PySwarms among others to provide functionality for preprocessing, oversampling, cross validation, FS and evaluation. The MATLAB engine for Python was used to integrate the SOF classifier built in MATLAB (developed by [21]) into the Python pipeline. The results of the experiments as described in the evaluation section above are shown in Table IV.

Multiple metrics are collated from model evaluations in order to determine the effectiveness of both the FS methods and the performance the SOF classifier in the given domain. The metrics considered include: the multi-class accuracy; balanced accuracy, which considers the imbalance of data classes by using the macro average of per-class recall; per-class precision and macro-precision; per-class recall and macro-recall; and per class F1 and macro-F1 (harmonic mean) scores. Macro in context meaning that the metrics are calculated per class and the mean taken. Each metric is averaged over all data folds with their respective standard deviations provided where appropriate. These are all common metrics used for evaluation of classification tasks and were selected for that reason [28]. The use of average F1 rather than F1 of averages was determined by considering recommendations made by [29]. Area under the ROC curve was also considered as a potential metric; however, due to the nature of the SOF classifier there are no thresholds available to compute probabilistic predictions, making the use of the metric unhelpful when only a single point is available.

Each of the above measures were computed without FS, with BPSO and with RFE for comparison. The use of static random seeds in the implementation allowed for reproducible results and for identical data sets to be produced for each experiment particularly with respect to the data splits created for repeated k-fold and SMOTE samples. This allows for direct comparison of selection techniques and an overall impression of the SOF classifier's performance.

With respect to FS, RFE-SVM proved to produce feature subsets that allowed the SOF classifier to attain the best results compared to BPSO and no FS by over 10% in both multi-class accuracy and F1 scores, while also improving on balanced accuracy. The results of BPSO proved surprising with an average feature set size of 50.6, compared to the much smaller average of 11 features selected by RFE. Examination of the BPSO particle positions over time appeared to show that particles got stuck exploring narrowly around a local optimum feature subset, suggesting that convergence of the particles happened much too early and resulted in limited exploration. This appears to be a known issues with BPSO as discussed by [26]. Multiple variations of BPSO parameters were attempted including increases of both the population of particles and maximum attempted iterations in addition to testing different values for cognitive, social and inertial weights. Results varied with the maximum accuracy seen around 0.8 for a single run of 10-fold CV, however, the computational time and power required for each iteration increased to the point where the process was no longer viable. The results shown in the table are from using the default 20 particles over 100 iterations with cognitive / social factors of 0.5 and an inertia weighting of 0.8 producing evaluations worse than using no FS at all.

Additionally, by using the feature sets derived by RFE, commonly seen features were tabulated (see Table I). The table shows how the optimal features span across the data modalities described previously. In particular, multiple neurocognitive assessment scores appear in every set (LDELTOTAL, FAQ, MMSE, ADAS13) suggesting that these particular tests are the most useful for classification while others (LIMMTOTAL, MoCA and RAVLT trials) are used in less than half showing that, perhaps, some aspects of brain function are unnecessary, or less helpful for AD classification. Age also appears often which, as discussed previously, is expected due to the higher chance of the condition later in life. Additionally, single features are picked from the MRI (Hippocampus), PET (Temporal_Left) and CSF (TAU_ABETA) modalities.

TABLE I
TOP 10 OCCURRENCES OF FEATURES SELECTED BY RFE

Feature	Modality	Occurances
LDELTOTAL	CM	50
FAQ	CM	50
MMSE	CM	50
ADAS13	CM	50
AGE	Patient	49
Hippocampus	MRI	45
Temporal_Left	PET	42
PTRACCAT_2	Patient	34
TAU_ABETA	CSF	34
LIMMTOTAL	CM	20

In terms of classification, when using the reduced features determined by RFE, the SOF classifier in offline training mode has been shown to produce good results, with an average accuracy and balanced accuracy of 0.81 and 0.82 respectively, with an F1-score of 0.81, all averaged over 5-repeated 10-fold CV. The accuracy values suggest that the use of SMOTE may have successfully reduced the effect of class imbalance. Specifically, these results were achieved using an SOF granularity of 1, the lowest available value, which results in a lower number of prototypes being generated for prediction purposes. Any value above 1 resulted in worse evaluation scores which, as discussed by [21] suggests that the lower values may allow the model to generalise better. By using more prototypes, it could be that overfitting was occurring due to the extra complexity caused by introducing more points (and radii) into the prototype space, effectively tightening the equivalent of the decision boundaries.

TABLE II
SUMMED RFE-SVM BASED CONFUSION MATRIX OVER ALL ITERATIONS

		Predicted			
		CU	MCI	AD	Total
Actual	CU	816	148	6	970
	MCI	193	812	85	1090
	AD	1	42	337	380
	Total	1010	1002	428	2440

Table II shows the confusion matrix generated across all folds when using RFE. In general, the classifier did particularly well when separating CU and AD patients with only seven misclassifications made between the two, only one of which resulted in a prediction of being unimpaired when the patient was positive AD. On the other hand, most of the errors occurred when attempting to distinguish between MCI and AD / NC. This may be in part due to MCI being an intermediate stage with overlapping features that make definitive separations difficult between the two main diagnosis classes.

TABLE III
PRECISION AND RECALL RESULTS OF EXPERIMENTS USING FS VARIATIONS AND THE SOF CLASSIFIER

Feature Selection	Avg. Features	Precision				Recall			
		NC	MCI	AD	Avg	NC	MCI	AD	Avg
None	$*$	0.72 ± 0.07	0.72 ± 0.11	0.68 ± 0.13	0.71 ± 0.07	0.79 ± 0.11	0.57 ± 0.09	0.85 ± 0.13	0.74 ± 0.07
RFE	11	0.81 ± 0.08	0.82 ± 0.09	0.81 ± 0.14	$\mathbf{0.82 \pm 0.07}$	0.84 ± 0.09	0.74 ± 0.13	0.89 ± 0.13	$\mathbf{0.82 \pm 0.07}$
BPSO	50.6	0.73 ± 0.08	0.71 ± 0.09	0.65 ± 0.12	0.70 ± 0.06	0.79 ± 0.10	0.56 ± 0.13	0.86 ± 0.10	0.73 ± 0.06

TABLE IV
ACCURACY AND F1 RESULTS OF EXPERIMENTS USING FS VARIATIONS AND THE SOF CLASSIFIER

Feature Selection	Avg. Features	Avg. Accuracy / Bal. Accuracy	F1			
			NC	MCI	AD	Avg
None	$*$	0.70 ± 0.06 / 0.74 ± 0.07	0.75 ± 0.07	0.63 ± 0.08	0.75 ± 0.10	0.71 ± 0.06
RFE	11	$\mathbf{0.81 \pm 0.07}$ / $\mathbf{0.82 \pm 0.07}$	0.82 ± 0.06	0.77 ± 0.09	0.84 ± 0.10	$\mathbf{0.81 \pm 0.07}$
BPSO	50.6	0.69 ± 0.06 / 0.73 ± 0.06	0.75 ± 0.07	0.61 ± 0.10	0.73 ± 0.09	0.70 ± 0.06

V. CONCLUSION AND FUTURE WORK

This report considers the use of fuzzy systems, specifically the recent SOF classifier for use in predicting whether baseline observations for patients seen in the ADNI data indicate that they suffer from AD, MCI or are CU while using FS methods. SOF is shown to be appropriate and useful in the domain of AD prediction, though the results did not reach as high in accuracy or F1 as other works including [12]. While the results fell short, the methods used to achieve them are much more robust and close enough to warrant further consideration. In particular, given the low inaccuracy between the main CU and AD classes, the experiment could potentially be used independently or, to reduce the likelihood of false positives of CU / false negatives for AD, as part of an ensemble AD screening tool using multiple ML techniques.

The limitations of this experiment revolve primarily around the data. First, the experiment was limited by the features and patients selected, better results may be possible by making fuller use of the ADNI data by considering more of the features and patients available. Second, the class imbalance within the sourced data, particularly between CU and AD, can cause issues with classifications even though attempts were made to reduce its impact by oversampling and considering balanced accuracy. As such considering the use of data imputation such as advanced interpolation techniques [30] in future work rather than removal of incomplete samples may produce better results by retaining a better class balance, or keep enough data to consider undersampling. Third, the the removal of the CDRSB variable is likely to have had a negative impact on the results as it is an important factor for diagnosis. Therefore, repeating the experiments with the CDRSB variable available as an independent feature may improve on the presented results in future works.

The tabulated FS results show how useful features are split over multiple modalities. However, as shown by the number of observations remaining in the provided dataset, limited numbers of patients have data available for all considered modalities with, for example, more having undertaken neurocognitive testing and less for MRI / PET / CSF. Future work may consider an ensemble system where models are developed per modality with the final decision being made based on the results of each model for which data is available. The use of such an ensemble could fill a gap in screening that would allow an overall prediction per patient for all modality data available per patient with confidences determined by the support available for a particular class for stratification.

VI. ACKNOWLEDGEMENTS

Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Therapeutic Research Institute at the University of Southern California.

ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

REFERENCES

[1] Alzheimer's Research UK, "Global prevalence," https://www.dementiastatistics.org/statistics/global-prevalence/, 2018.

[2] World Health Organisation, "Dementia," https://www.who.int/news-room/fact-sheets/detail/dementia, 2019.

[3] NHS, "New type of dementia identified - NHS," 2019. [Online]. Available: https://www.nhs.uk/news/neurology/new-type-dementia-identified/

[4] ——, "Alzheimer's disease - Symptoms - NHS," 2018. [Online]. Available: https://www.nhs.uk/conditions/alzheimers-disease/symptoms/

[5] Alzheimer's Research UK, "Prevalence by age in the UK," https://www.dementiastatistics.org/statistics/prevalence-by-age-in-the-uk/, 2018.

[6] Alzheimer's Society, "Drug treatments for Alzheimer's disease — Alzheimer's Society," 2019. [Online]. Available: https://www.alzheimers.org.uk/about-dementia/treatments/drugs/drug-treatments-alzheimers-disease

[7] H. Kour, J. Manhas, and V. Sharma, "Evaluation of Adaptive Neuro-Fuzzy Inference System with Artificial Neural Network and Fuzzy Logic in Diagnosis of Alzheimer Disease," in *2019 6th International Conference on Computing for Sustainable Global Development (INDIACom)*. IEEE, 2019, pp. 1041–1046. [Online]. Available: https://ieeexplore.ieee.org/document/8991423

[8] F. Zhu, X. Li, D. McGonigle, H. Tang, Z. He, C. Zhang, G. U. Hung, P. Y. Chiu, and W. Zhou, "Analyze Informant-Based Questionnaire for the Early Diagnosis of Senile Dementia Using Deep Learning," *IEEE Journal of Translational Engineering in Health and Medicine*, vol. 8, 2020.

[9] S. Cahill, M. Clark, H. O'Connell, B. Lawlor, R. F. Coen, and C. Walsh, "The attitudes and practices of general practitioners regarding dementia diagnosis in Ireland," *International Journal of Geriatric Psychiatry*, vol. 23, no. 7, pp. 663–669, jul 2008. [Online]. Available: http://doi.wiley.com/10.1002/gps.1956

[10] A. Bradford, M. E. Kunik, P. Schulz, S. P. Williams, and H. Singh, "Missed and delayed diagnosis of dementia in primary care: prevalence and contributing factors," *Alzheimer disease and associated disorders*, vol. 23, no. 4, p. 306, 2009.

[11] NHS, "Guide to NHS waiting times in England - NHS," 2019. [Online]. Available: https://www.nhs.uk/using-the-nhs/nhs-services/hospitals/guide-to-nhs-waiting-times-in-england/

[12] M. Bucholc, X. Ding, H. Wang, D. H. Glass, H. Wang, G. Prasad, L. P. Maguire, A. J. Bjourson, P. L. McClean, S. Todd, D. P. Finn, and K. F. Wong-Lin, "A practical computerized decision support system for predicting the severity of Alzheimer's disease of an individual," *Expert Systems with Applications*, vol. 130, pp. 157–171, sep 2019.

[13] G. G. C. Lee, P. W. Huang, Y. R. Xie, and M. C. Pai, "Classification of Alzheimer's Disease, Mild Cognitive Impairment, and Cognitively Normal Based on Neuropsychological Data via Supervised Learning," in *IEEE Region 10 Annual International Conference, Proceedings/TENCON*, vol. 2019-October. Institute of Electrical and Electronics Engineers Inc., oct 2019, pp. 1808–1812.

[14] F. Er, P. Iscen, S. Sahin, N. Çinar, S. Karsidag, and D. Goularas, "Distinguishing age-related cognitive decline from dementias: A study based on machine learning algorithms," *Journal of Clinical Neuroscience*, 2017.

[15] T. Chen, G. Antoniou, M. Adamou, I. Tachmazidis, and P. Su, "Automatic diagnosis of attention deficit hyperactivity disorder using machine learning," *Applied Artificial Intelligence*, 2019.

[16] T. Chen, P. Su, C. Shang, and Q. Shen, "Weighted fuzzy rules optimised by particle swarm for network intrusion detection," in *Fuzzy Systems, 2018 IEEE International Conference on*. IEEE, 2018, pp. 1–7.

[17] P. Su, C. Shang, T. Chen, and Q. Shen, "Exploiting data reliability and fuzzy clustering for journal ranking," *IEEE Transactions on Fuzzy Systems*, vol. 25, no. 5, pp. 1306–1319, 2017.

[18] T. Chen, P. Su, C. Shang, and Q. Shen, "Reliability-guided fuzzy classifier ensemble," in *Fuzzy Systems, 2017 IEEE International Conference on*. IEEE, 2017, pp. 1–6.

[19] P. Su, T. Chen, J. Xie, B. Ma, H. Qi, J. Liu, and Y. Zhao, "A density and reliability guided aggregation for the assessment of vessels and nerve fibres tortuosity," *IEEE Access*, 2020.

[20] T. Chen, P. Su, C. Shang, R. Hill, H. Zhang, and Q. Shen, "Sentiment classification of drug reviews using fuzzy-rough feature selection," in *2019 IEEE International Conference on Fuzzy Systems*. IEEE, 2019, pp. 1–6.

[21] X. Gu and P. P. Angelov, "Self-organising fuzzy logic classifier," *Information Sciences*, vol. 447, pp. 36–51, jun 2018.

[22] J. Han, J. Pei, and M. Kamber, *Data mining: concepts and techniques*. Elsevier, 2011.

[23] R. Blagus and L. Lusa, "SMOTE for high-dimensional class-imbalanced data," *BMC Bioinformatics*, vol. 14, no. 1, pp. 1–16, mar 2013.

[24] L. Carlos Molina, L. Belanche, and À. Nebot, "Feature selection algorithms: A survey and experimental evaluation," in *Proceedings - IEEE International Conference on Data Mining, ICDM*, 2002, pp. 306–313.

[25] J. Kennedy and R. C. Eberhart, "Discrete binary version of the particle swarm algorithm," in *Proceedings of the IEEE International Conference on Systems, Man and Cybernetics*, vol. 5. IEEE, 1997, pp. 4104–4108.

[26] J. Too, A. R. Abdullah, and N. Mohd Saad, "A New Co-Evolution Binary Particle Swarm Optimization with Multiple Inertia Weight Strategy for Feature Selection," *Informatics*, vol. 6, no. 2, p. 21, may 2019.

[27] S. Maldonado, R. Weber, and F. Famili, "Feature selection for high-dimensional class-imbalanced data sets using Support Vector Machines," *Information Sciences*, vol. 286, pp. 228–246, dec 2014.

[28] M. Hossin and M. Sulaiman, "A review on evaluation metrics for data classification evaluations," *International Journal of Data Mining & Knowledge Management Process*, vol. 5, no. 2, p. 1, 2015.

[29] J. Opitz and S. Burst, "Macro F1 and Macro F1," nov 2019. [Online]. Available: http://arxiv.org/abs/1911.03347

[30] T. Chen, C. Shang, J. Yang, F. Li, and Q. Shen, "A new approach for transformation-based fuzzy rule interpolation," *Fuzzy Systems, IEEE Transactions on*, 2019. [Online]. Available: https://doi.org/10.1109/TFUZZ.2019.2949767