2020

Harnessing the Power of Machine Learning in Dementia Informatics Research: Issues, Opportunities, and Challenges

Tsang, Gavin

10.1109/rbme.2019.2904488
IEEE Reviews in Biomedical Engineering
Institute of Electrical and Electronics Engineers (IEEE)

Harnessing the Power of Machine Learning in Dementia Informatics Research: Issues, Opportunities and Challenges

Gavin Tsang, *Student Member, IEEE*, Xianghua Xie*, *Senior Member, IEEE*, and Shang-Ming Zhou*, *Member, IEEE*

***Abstract*—Dementia is a chronic and degenerative condition affecting millions globally. The care of patients with dementia presents an ever continuing challenge to healthcare systems in the 21st century. Medical and health sciences have generated unprecedented volumes of data related to health and wellbeing for patients with dementia due to advances in information technology, such as genetics, neuroimaging, cognitive assessment, free texts, routine electronic health records etc. Making the best use of these diverse and strategic resources will lead to high quality care of patients with dementia. As such, machine learning becomes a crucial factor in achieving this objective. The aim of this paper is to provide a state-of-the-art review of machine learning methods applied to health informatics for dementia care. We collate and review the existing scientific methodologies and identify the relevant issues and challenges when faced with big health data. Machine learning has demonstrated promising applications to neuroimaging data analysis for dementia care, while relatively less efforts have been made to make use of integrated heterogeneous data via advanced machine learning approaches. We further indicate the future potentials and research directions of applying advanced machine learning, such as deep learning, to dementia informatics.**

***Index Terms*—Alzheimer, Cognitive Assessment, Deep Learning, Dementia, Electronic Medical Records, Health Informatics, Machine Learning, NeuroImaging, NLP.**

I. Introduction

BEING a chronic and degenerative condition, dementia has a high prevalence within the elderly affecting 47.5 million people globally [1]. The most common cause of dementia being Alzheimer's Disease (AD) making up 60-80% of cases [2]. Within the UK, dementia affects 850,000 people with a forecast rate of prevalence of 1 million by 2025 and 2 million by 2051 [3]. The resulting cost of dementia in the UK totals £26.3 billion with two thirds being paid for by dementia patients and families in private social care as of 2018 [3].

The diagnosis and prognosis of dementia can be challenging. Dementia diagnosis generally requires a battery of clinical tests, such as cognitive assessments, patient histories and neuroimaging; whilst other potential causes must be ruled out for a more conclusive diagnosis [4]. A truly definitive diagnosis is only possible through a post-mortem autopsy [5]. This diagnostic procedure is very time-consuming and costly. On other other hand, due to advances in information technology, medical and health sciences have generated huge amounts of data related to dementia, such as genetics, neuroimaging, cognitive assessment, electronic health records etc. Data-driven techniques continually become increasingly important in making best use of these data resources. The complex interactions between patient physiology and clinical events, coupled with continually degenerative symptoms, lead up to diagnosis of dementia a difficult task via traditional statistical means. Data driven Machine Learning (ML) techniques have the capabilities of modelling such complex associations, as proven within other fields such as object recognition [6], natural language processing exampled by ubiquitous voice assistant applications: Apple Siri, Amazon Alexa and Microsoft Cortana [7]. Making the best use of these big health related data, ML techniques provide a way of delivering high quality personalised healthcare services in real time.

As a matter of fact, in medical informatics, ML has demonstrated promising applications to neuroimaging analysis. Orru *et al.* [8] surveyed the application of Support Vector Machines (SVMs), preceding 2011, in identifying imaging biomarkers of neurological and psychiatric diseases. Mosconi *et al.* [9] reviewed the existing scientific literatures involving the early detection of AD using neuroimaging. Mosconi *et al.* focused on the effectiveness of neuroimaging detection, possible risk factors and the progression from healthy to general cognitive impairment. Recent major reviews, such as by Ching *et al.* [10] provides overviews of ML within biology and medicine. These existing efforts of literature reviews are different from ours in this paper where we focus on ML applied to dementia diagnosis. We aim to evaluate the various ML approaches available to dementia diagnosis in addition to highlighting the ML models commonly used for differing types of data.

In what follows, a brief overview of the unique combination of challenges within dementia care is presented, followed by an in depth explanation and evaluation of common ML methodologies used within dementia informatics. We then summarise relevant literature and bring forward potential unexplored avenues of research.

II. Review Methodology

An electronic search was conducted for relevant academic papers using literature archives: PubMed and MEDLINE. The screening criteria used for our survey included any accessible academic paper from a 10 year period of 2008 to 2018 inclusive whose methodology involved the use of any ML applications in the screening, diagnosis or prognosis of dementia symptoms.

Preliminary research involved an online search of various combinations of the following keywords: "dementia, diagnosis, screening, machine learning, prediction, classification, detection, prognosis, deep learning".

Additional literature was gained through citations provided by papers found during preliminary research. On further analysis of preliminary papers, further keywords were compiled based upon the specific dementia sub-categories emphasised for diagnosis for each data-driven application category. For instance, great focus is placed upon AD diagnosis within neuroimaging whilst semantic dementia remains the sole focus of speech analysis applications. A full diagram of review steps taken is shown in Fig. 1.. The resulting, additional keywords compiled from said literature were included into our search criteria as follows: "Alzheimers, semantic dementia, patient records, patient history, electronic health records, cognitive impairment, clinical decision support system". The resulting literature search uncovered 923 publications fitting the aforementioned search criteria. The title, abstract and authors from the resulting literature

Submitted for review: 30-Jul-2018, accepted: 22-Feb-2019.

*Corresponding authors: X.Xie@swansea.ac.uk; S.Zhou@swansea.ac.uk

This study was undertaken with the support of the Health Data Research UK (NIWA1), the Farr Institute of Health Informatics Research (MR/K006525/1). This work was also supported by The Engineering and Physical Sciences Research Council (EP/N028139/1).

G. Tsang & X. Xie was with the Dept. of Computer Science, Swansea University, UK. Email: (see http://csvision.swan.ac.uk/)

S. Zhou was with the Inst. of Life Science, Swansea University, UK. Email: S.Zhou@swansea.ac.uk

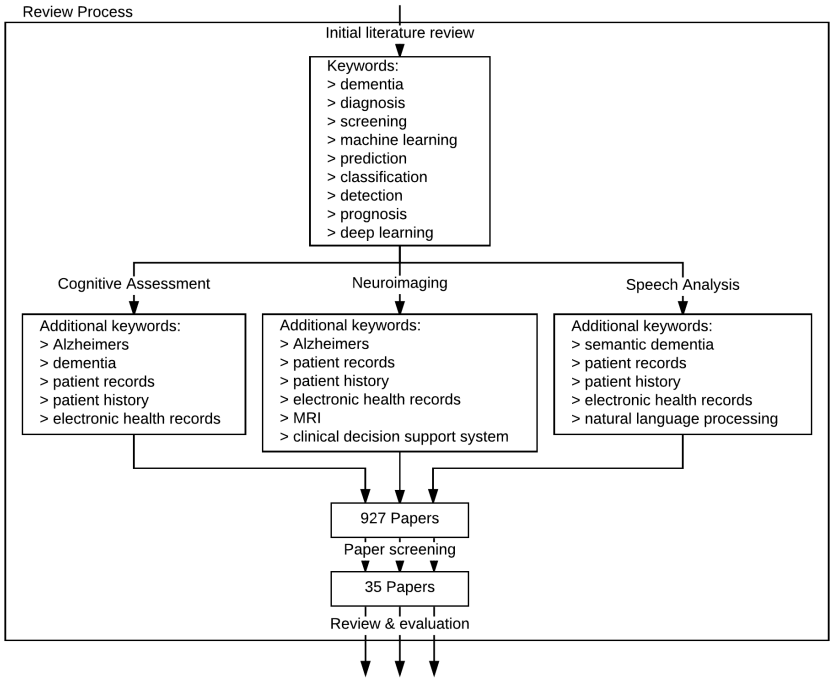


Fig. 1. Process diagram of steps taken during literature survey stage. After an initial preliminary survey, the review literature was categorized into three data centric categories. Each category was further surveyed in detail with additional keywords used before final screening, evaluation and review.

were then reviewed for duplicates and relevance to the topic at hand. Literature surveys were set aside as additional literature sources but not included as relevant literature. This resulted in a final selection of 35 articles which will be discussed further.

III. Dementia Care - Challenges

Dementia presents a unique assortment of varying challenges in regards to diagnosis originating from dementia being a chronic and degenerative set of conditions. The detection and diagnosis of the initial stages of dementia or Mild Cognitive Impairment (MCI) continues to be problematic [11] with a reliance on self reporting or reports by relatives. Compounded with symptoms being obscured by the regular effects of natural ageing, potential cases of dementia are generally reported 2 to 3 years after onset [4]. The delayed diagnosis results in continued unchecked decline which reduces the effectiveness of any care given upon discovery and diagnosis [12]–[14]. Such obscure symptoms also brings about the issue of physicians being unable to give a definitive diagnosis of dementia and as such apply for a full battery of cognitive assessments, expensive neuroimaging or invasive tests such as lumbar puncture [4].

The current diagnostic model begins with several screening procedures used to identify potential dementia patients for further evaluation leading to a definitive diagnosis, the most common being the Mini-Mental State Examination (MMSE) and the Consortium to Establish a Registry for Alzheimers Disease (CERAD) cognitive assessments [15]. However, current cognitive assessments remain problematic with the MMSE being called into question as an effective screening measure, Tombaugh *et al.* [16] provides a thorough review of literature evaluating screening effectiveness with a sensitivity varying from 21-100% and specificity of 46-100%.

Such a combination of factors results in a slow and ineffective system of diagnosis with an estimated two thirds of dementia cases remaining undetected [17], [18].

As mentioned previously, diagnosis generally involves the use of neuroimaging after screening to further assess the potential for dementia. The most common method being Magnetic Resonance Imaging (MRI) and Positron-emission Tomography (PET). Current evaluation of scan results requires the use of an expert radiologist or anatomist in order to correctly identify and perform manual measurements. Such manual tasks often result in excessive time consumption, variability between different medical professionals and are limited to only certain brain regions [19]. With current manual MRI evaluation [20] the resulting separation between normal elderly and probable dementia achieves accuracies ranging from 58-100%. Kloppel *et al.* indicates a significant difference in diagnostic ability between general radiologists and neuroradiologists in evaluation of clinical trial [21] which may clarify the apparent variation in accuracy. Kloppel also uses the results as an indication of the need for specialisation. All of which, speaks for the need of more straightforward and effective utilization of neuroimaging data.

Several shortcomings exist with the use of neuroimaging applications. The foremost challenge being cost at £169 and £844 per patient within the UK for MRI and PET scans respectively [5], [22]. Cheaper and more readily available alternative diagnostic procedures for dementia may serve to address the problematic issue of scarcity and cost of specialist neuroimaging services.

Interestingly, social stigma also factors into the challenge of needing a supplementary diagnostic procedure for diagnosis. Boustani *et al.* [2], [23] reports high refusal rates of 51% of potential Dementia screen patients for further diagnostic assessment after a positive screening result. Boustani's findings also suggested that patients believed dementia to be a devastating condition with no available treatment and would lead to issues such as depression, anxiety, social stigma, insurance coverage and loss of independence. Changes to diagnostic procedure which allows for immediate effective diagnosis

may alleviate the issues of patient refusal.

The prognosis of dementia is signified with continued degradation of mental ability, increased risk of co-morbidity [24] and increased risk of institutionalisation resulting in a significantly higher risk of death than non-demented patients [25]. The reduced health of dementia patients also produces a corresponding significant increase in healthcare costs [24]. The risks of acute conditions and events resulting in hospitalisation are also affected with an onset of Diagnosis. The fall rate of nursing home residents with dementia nearly doubles compared to residents without in a study by Van Doorn *et al.* [26]. Further research is required in effective prognosis and care post-diagnosis to alleviate the significant increase in co-morbidity risk posed by dementia.

IV. Overview of Commonly Used Machine Learning Technologies

Within the reviewed literature, a consistent set of ML technologies is applied within the diverse fields of dementia diagnosis. As shown in Table I, the distribution of underlying ML methodologies used indicates SVM as the popular methodology in use overall. Within the majority of common methodologies being applied to the the diverse data categories of neuroimaging, cognitive assessment data and speech analysis, this highlights the adaptability of such methods in dementia analysis in regards to data used.

Following on is an overview of some of the root ML technologies in use by reviewed literature in addition to state-of-the-art ML technologies with potential applications in dementia informatics.

A. Support Vector Machine

SVM is a supervised learning model which attempts to generate separating hyperplanes across groups of observations in accordance with class labels. Unlike Linear Discriminant Analysis (LDA), SVM makes no assumptions on data distribution, allowing for great flexibility in model generation.

By mapping observations from original feature space into a higher dimensional space through the use of linear or non-linear kernel functions, observations which were once non-linearly separable in feature space may be mapped into a higher dimensional space which supports separation by linear hyperplanes.

The separating hyperplanes within higher dimensional space are defined by

$$w^T \phi(x) + b = 0 \tag{1}$$

where w is the normalised normal vector to the hyperplane, b the normalised perpendicular distance of the hyperplane to the origin and $\phi(x)$ the linear or non-linear mapping function. The resulting classification function for any new observations is simply comparing the observation position in relation to constructed hyperplane (1). Since there exists an infinite set of hyperplanes which could potentially separate class boundaries, an optimal separating hyperplane must be generated based upon the structural risk minimisation principle. As such, the optimal separating hyperplane is arranged so that there is the greatest separating distance, or *margin*, between the borders of class distributions defined as parallel hyperplanes. The superficial observations which lie on the aforementioned parallel hyperplanes are called support vectors.

The optimal separating hyperplane can be found by maximising said margin distance using the Lagrangian dual of the optimisation function

$$\arg_a \min L_p(a) = \sum_{i=1}^{l} \sum_{j=1}^{l} a_i a_j y_i y_j K(x_i, x_j) \tag{2}$$

TABLE I
Table of underlying ML methodologies used within the reviewed literature.

ML model	Total Count	Literature
Cognitive Assessment		
SVM	3	[15], [27], [28]
NB	3	[15], [28], [29]
LR	3	[29], [30]
DT	2	[28], [29]
NN	2	[15], [28]
LDA	1	[30]
RF	1	[15]
Neuroimaging		
SVM	14	[8], [21], [31]–[42]
LR	5	[19], [31], [34], [43], [44]
LDA	4	[19], [34], [41], [45]
PCA	2	[39], [43]
NB	1	[34]
DT	1	[34]
NN	1	[34]
RF	1	[31]
CNN	1	[46]
DNN	1	[47]
Speech Assessment		
NN	2	[4], [48]
LR	1	[48]
DT	1	[48]
NB	1	[49]
PCA	1	[50]

CNN: Convolutional Neural Network, DNN: Deep Neural Network, DT: Decision Tree, LDA: Linear Discriminant Analysis, LR: Logistic Regression, NB: Naive Bayes, NN: Neural Network, PCA: Principal Component Analysis, RF: Random Forest, SVM: Support Vector Machine

subject to constraints:

$$\sum_{i=1}^{l} a_i y_i = 0 \tag{3}$$

$$a_i \geq 0 \text{ for } i = 1, \ldots, l \tag{4}$$

where a are optimal Lagrange multipliers found through quadratic optimisation and $K(x_i, x_j)$ is the kernel mapping matrix.

B. Random Forest

Random Forests (RFs) are an ensemble learning method involving the generation of multiple decision trees whose input dataset are a random sample of features with replacement (*feature bagging*) and also a random sample of observations with replacement (*tree bagging*). Typically in a classification problem of p features, a subset of $\lfloor \sqrt{p} \rfloor$ features are selected [51] for each tree. Final classification involves the aggregation, generally vote count, of prediction result from every tree.

The use of ensemble classification in RF reduces overfitting whilst also allowing for the evaluation of feature importance after training using the out-of-bag (OOB) error [51]. By permuting individual features across a dataset passed into a trained RF for evaluation, the resulting OOB error can be compared against the original training OOB error to determine feature importance with a greater difference indicating a greater importance for said feature and vice versa. The

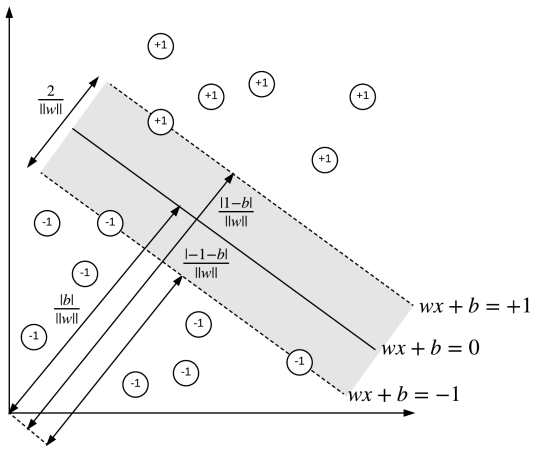


Fig. 2. Diagram of SVM separating hyperplane across observations of a binary class $\{-1, +1\}$. The optimal separating hyperplane produces the maximum margin distance between class boundaries defined as parallel hyperplanes which lie on superficially located observations called support vectors.

OOB error serves as a validation metric without the need for a entirely separate validation dataset by evaluating prediction accuracy of observations on decision trees who's training subset did not contain said observations.

Decision trees involve the generation of a directed acyclic graph in a tree like structure containing interior nodes corresponding to individual features containing edges whose conditional response is based on the set of possible values. Leaf nodes would correspond to the final classification or regression result. Feature selection of interior nodes is evaluated based on various possible metrics such as the Gini impurity.

C. Principal Component Analysis

Principal Component Analysis (PCA) is a methodology used to orthogonally transform a set of observations containing potentially correlated features into a set of linearly uncorrelated features called principal components. Principal components indicate the vector dictating the direction of greatest variance in a normally distributed dataset, *eigenvector*, whilst the *eigenvalue* corresponds to the variance along said vector. Subsequent principal components provide the next greatest variance along the orthogonal vector of all preceding eigenvectors.

Through the eigendecomposition of the covariance matrix, A, of a dataset x, given by $A = x^T x$, A can be decomposed into the matrix of eigenvectors V where $V_{i,:} = (v_1, \ldots, v_k)$ and eigenvalues Λ

$$A = V \Lambda V^T \tag{5}$$

where $\Lambda = \text{diag}(\lambda_1, \ldots, \lambda_k)$. For each eigenvalue λ_i, a specific eigenvalue equation exists which can be solved for to determine the set of eigenvectors associated to each eigenvalue. Whilst solving for such equations are trivial on small datasets, large feature and observation datasets are solved through the use of various iterative algorithms.

A common use of PCA involves its use as a precursor to dimensionality reduction. Lower order principal components of low variance or eigenvalue can be removed while higher order principal components are kept, reducing dimensionality whilst retaining as much variance in the dataset as possible. Subsequent observations can thus be transformed into eigenspace using the eigenvector matrix ,W, with lower order eigenvectors removed.

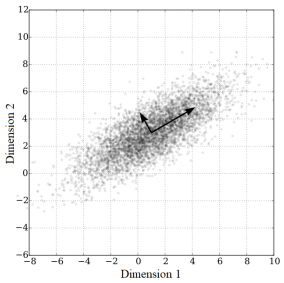


Fig. 3. PCA of a gaussian distribution showing the orthogonal eigenvectors. The arrow vectors shown indicate the first and second principal component eigenvectors. The first eigenvector pointing top-right lies on the direction of greatest variance as seen within the distribution whilst the second eigenvector lies orthogonal to the first indicating the direction of second greatest variance.

D. Linear Discriminant Analysis

LDA, one of the oldest classifiers still in use, is a supervised dimensionality reduction technique. Similar to PCA, LDA produces principal components or linear discriminants representing transformed axis vectors. However, whilst PCA, a non-supervised methodology, attempts to find eigenvectors which seek to maximise the variance within a distribution, the supervised LDA seeks to find linear discriminants which maximise the ratio of between-class distance $\tilde{\Sigma_B}$ to within-class class variance $\tilde{\Sigma_W}$

$$\arg_V \max J(V) = \frac{|\tilde{\Sigma_B}|}{|\tilde{\Sigma_W}|} = \frac{|V^T \Sigma_B V|}{|V^T \Sigma_W V|} \tag{6}$$

where V is the matrix of eigenvectors. As such, (6) can be rearranged into a classic eigenvalue problem using

$$S_W^{-1} S_B V = \lambda V \tag{7}$$

where V is the matrix of eigenvectors and λ the diagonal matrix of eigenvalues which can thus be solved through eigendecomposition to arrive at the linear discriminants.

Similar to PCA, dimensionality reduction can be performed by thresholding out the lowest ranked linear discriminants. Classification is also possible through a technique similar to Naive Bayes Gaussian where class probabilities along linear discriminants are calculated with new observations being classified based on the maximum *a posteriori* decision rule where the highest probability class is used as final classification decision.

E. Naive Bayes

Naive Bayes (NB) is a family of probabilistic classifiers based on Bayes' theorem. Unlike other machine learning methods, NB requires no iterative parameter estimation, minimising computational complexity; whilst also having a linear scaling in parameter count versus feature count, minimising model complexity. The simplicity of such a model is however reliant on the assumption of strong independence between all features, a rare occurrence.

Let x_i be a feature where $x_i \in x = (x_1, \ldots, x_n)$. The naive Bayes' probability model is formulated as

$$p(C_k|x) = \frac{p(C_k) \prod_{i=1}^{n} p(x_i|C_k)}{\sum_k p(C_k) p(x|C_k)} \tag{8}$$

where the prior probability of class C_k is $p(C_k)$, while $p(x|C_k)$ is the probability of feature vector x given class c_k. The denominator, $p(x)$ is subsequently a scaling factor indicating the probability of feature vector x across all classes. The Naive Bayes Gaussian method of calculation is used on continuous valued features with the assumption of a normal distribution. Whereas binary feature data can be measured using the Bernoulli naive Bayes model.

F. Neural Network

Neural Networks (NNs) are a diverse and robust ML application able to model a variety of input-output mappings through the use of various network architectures. NNs have been argued to be able to equate to any optimal statistical classifier [15]. As such, NNs have shown great promise and continued use in various disciplines and domains [52]–[57].

Nevertheless the capabilities of NNs come with the price of model complexity, with model parameters exponentially increasing with feature count and capacity. As a result, overly complex NNs suffer from long training times and issues with overfitting without a large enough dataset to match network capacity or the adoption of regularisation techniques [58]. Such high model complexity also suffers from the inability to validate trained models past that of empirical evidence from testing as opposed to theoretical validation. The black box nature of NNs limits the capability of understanding how a set of features and parameters are able to model a complex problem, only that it is able to.

NNs consist of sets of interconnected nodes arranged in layers called Multilayer Perceptrons (MLP). Each MLP maps multiple input signals a_j^{l-1} through an activation function

$$a_k^l = \sigma(w_{jk}^l a_j^{l-1} + b_k^l) \tag{9}$$

to form a singular output a_k^l. Each input signal is modulated through a weighting w_{jk}^l and bias b_k^l before being aggregated into an activation function σ. Various σ functions exist with various properties and uses, the most popular of which being the logistic sigmoid function.

Learning within a NN is based off the adjustment of weight and bias parameters in a feed-forward and back-propagation process. The forward pass consists of passing observations through a network consisting of a randomly initialised set of weights to generate an initial prediction. Through the use of a cost function, such as the mean squared error

$$C = \frac{1}{2n}\sum_i^n (y_i - \hat{y}_i)^2 \tag{10}$$

where n is the number of training observations, y the ground truth and $\hat{y}$ the model output, a loss can be formed on the distance error of prediction from ground truth. By minimising the cost and subsequent model parameters, a NN can be pushed towards modelling the problem space and subsequent classification. Many algorithms exist which enable cost minimisation. An example of one being the stochastic gradient descent algorithm providing the capability to back-propagate changes in cost back through the NN updating weights and bias.

Stochastic gradient descent involves iteratively stepping down a hyperplane formed by the cost function and model parameters towards zero error or, more generally, a local minima which closely approximates the correct output. By determining the direction of descent through the pre-defined partial derivatives of the cost function. The model error (11) and subsequent layer errors (12) can be determined

$$\delta^L = \frac{\partial C}{\partial \hat{y}} \odot \hat{y} \tag{11}$$

$$\delta^l = (W^{l+1})^T \delta^{l+1} \odot \vec{a}^l \tag{12}$$

where $\frac{\partial C}{\partial \hat{y}}$ is the cost function partial derivative, $\vec{a}^l$ is the activation vector of layer l and W^{l+1} is the weight matrix of layer l. The direction of travel for layer parameters weight and bias can thus be calculated by

$$\frac{\partial C}{\partial b_k^l} = \delta_k^l \tag{13}$$

$$\frac{\partial C}{\partial w_{jk}^l} = a_k^{l-1}\delta_j^l \tag{14}$$

where b_k^l is the bias parameter for MLP k in layer l, a_k^{l-1} is the activation signal for MLP k in layer l and w_{jk}^l the weight parameter for MLP k in layer l to MLP j in layer $l-1$.

Finally parameters are updated by iteratively stepping towards a minima based on (11) & (12) through the following equations

$$w_{jk}^l \rightarrow w_{jk}'^l = w_{jk}^l - \frac{\eta}{n}\sum_i^n \frac{\partial C_i}{\partial w_{jk}^l} \tag{15}$$

$$b_k^l \rightarrow b_k'^l = b_k^l - \frac{\eta}{n}\sum_i^n \frac{\partial C_i}{\partial b_k^l} \tag{16}$$

where η is the learning rate indicating the length of stride for each iteration, n the number of observations, b_k^l the bias parameter for MLP k in layer l, w_{jk}^l the weight parameter for MLP k in layer l to MLP j in layer $l-1$ and b' and w' the future bias and weight parameter.

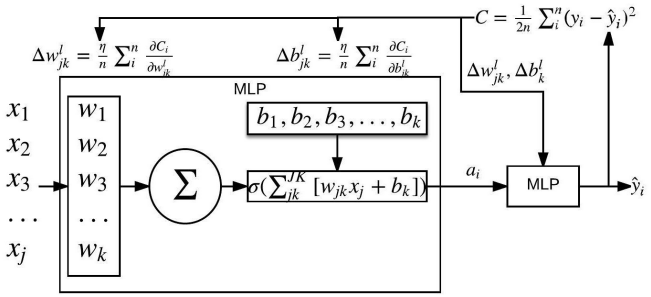


Fig. 4. Diagram of a NN and multilayer perceptron with input, hidden and output layer. Input is passed through the network as a forward pass for error calculation before being backpropagated through the network to update weights and bias.

G. Deep Learning

Whilst no strictly accepted definition of Deep Learning (DL) exists, it is generally accepted as a broad family of ML models based upon the use of large parameter spaces which attempt to learn latent representations of data and thus be used for or in support of training a more task orientated model. As a result, common representations of DL involve the use of multi-layer NNs. Such inclusion of multiple hidden layers allow for the construction of continually deeper feature representations on a layer by layer basis thus automatically engineering features suitable for complex tasks.

With the introduction of various advancements in optimisation algorithms [59], loss functions [60], training strategies [61] and regularisation techniques [62] in addition to powerful dedicated hardware and large datasets, DL has quickly found it's way to the top of many areas of research [6], [63]–[65]. Common examples of DL methods include deep NNs, autoencoders, Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs). Adoption of DL methods, however, have been slow within health informatics due to issues of model transparency, model validation and data availability. (See section VI-D for a more detailed evaluation of DL methods.)

In what follows will be a brief explanation of various DL models, namely: autoencoders, CNNs and RNNs as shown in Fig. 5..

1) Autoencoders: Autoencoders are a form of deep NN with a distinctive architecture. Specifically, autoencoders have a strictly mirrored architecture with input and corresponding hidden layers (called the encoding layers), reflecting output and corresponding hidden layers (called the decoding layers). Training objective is thus to output observations as close to what was input into the model. In order to stop encoding and decoding layers being trained into an identity mapping, layer capacity can be reduced to less than that of the input layer or through the use of regularisation techniques. Consequently, models are forced into learning a smaller and deeper representation of the data contained within the encoding layers. The encoding layers can thus be used as a feature encoder within a larger ML application much like PCA or LDA feature encoding.

2) Convolutional Neural Networks: CNNs are able to perceive spatial relationships between neighbouring features, such as images, unlike regular ML models in which flattened spatial data features lose all association with neighbouring features [6]. CNNs include the addition of convolution layers and pooling layers.

Convolution layers consist of convolutional filters which sweep through an input, applying a convolution operation to generate a feature vector. These convolution filters support update through back-propagation allowing for a number of features to be jointly represented by a small set of parameters reducing the size of, what would be, a large number of parameters in a regular NN. Said filters also represent neighbouring relationships by only convolving features within its receptive field. Due to sweeping convolutional filters generating potentially large feature vectors, pooling layers combine outputs of multiple clusters into a single output through various strategies. Some of which include max pooling or mean pooling, outputing the max or mean value of all inputs respectively.

Combinations of convolution and pooling layers can be stacked to generate a deep structure. A regular fully connected MLP NN is attached to perform final prediction based upon the learned encoded features of the above convolution and pooling layers.

3) Recurrent Neural Networks: RNNs are an adaptation of the MLP allowing for an internal memory state to be retained between observations [66], [67]. Consequently, RNNs are applicable to tasks involving time-series based data by "memorising" states from a previous time-step for use in a future prediction. RNNs include the addition of a weighted time-delayed recurrent connection which feeds a MLP output back into the MLP as an additional input. Said time-delayed recurrent connection enables a weighted output state to be stored and later included as a feature in future time-steps. Due to issues of vanishing and exploding gradients within the original implementation of RNNs limiting the availability of long term state memory, the Long Short-Term Memory (LSTM) unit was proposed using separate input,forget, update and output gates to form a single node.

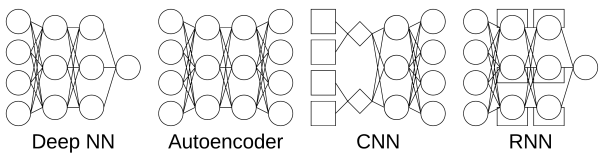


Fig. 5. Examples of DL architectures. Circles are MLP nodes, squares are convolutional filters and diamonds are pooling layers. Note how RNNs have a recurrent connection within the hidden layers.

V. Machine Learning Methodologies in Dementia Diagnosis

The field of ML in dementia diagnosis has been very active over recent years with applications making use of a variety of patient data and methodologies for diagnosis with the overall goal of discovering novel biomarkers for diagnosis or to improve upon diagnostic ability. Various papers have proposed the use of more novel patient data such as interview transcripts or Electronic Health Records (EHR) in an attempt to move away from the expensive use of neuroimaging [5], [22]. In regards to methodology, SVMs were the most popular ML model used within reviewed literature. Various other methodologies such as RF, LDA, Bayesian Network (BN), PCA and occasionally NN are used in support of or as diagnostic methodologies.

Following on will be a review on the applications proposed by various literature organised into categories of data used. As such, the various data categories can be evaluated for potential advantages and shortcomings in an informatics and clinical sense.

A. Cognitive Assessments

Initial screening of dementia patients is generally performed through the use of various neuropsychological assessments. Used to determine competency in various neurological abilities such as orientation, attention, recall and language; such assessments serve as a precursor to a more definitive and standardised diagnosis via MRI. With multiple assessment criteria available, definitions of categories of dementia vary amongst said assessments. While there is no official standardisation of assessments; the MMSE (Table II) and CERAD battery (Table III) are the common go to assessments [15]. However, modern evaluation on the screening performance of such assessments have shown highly varied results. A review of clinical trials evaluating the MMSE by Tombaugh *et al.* [16] produces results ranging from 21-100% in sensitivity and 46-100% in specificity. Consequently the need for improved, robust screening procedures remains an open area of research. The use of current screening procedures in conjunction with ML applications in improving diagnostic performance remains an active area.

In addition to the development of the assessment battery by CERAD, a central database of battery results along with demographic information was also introduced [68]. Containing both AD and corresponding control subjects with no confounding comorbidities, the dataset represents a "clean" group of AD subjects. Overall the dataset contains 1094 AD patients and 463 control subjects with the majority including follow up assessments annually for greater than 3 years.

Literature involving the use of neuropsychological assessments generally fall under categories of evaluation of ML model performance, neuropsychological assessment category selection or a combination of both. Literature considering diagnostic performance of ML methods only consist of So et.al [15] and Maroco *et al.* [30], both of which evaluate similar methodologies such as SVM, NN, RF and Logistic Regression (LR).

The dataset used by Maroco is based on a hand picked selection of 10 neuropsychological assessment categories spanning various cognitive attributes. Evaluation was based on classification between MCI and full dementia subjects with a 275 to 125 patient split respectively. Statistical evaluation of validation results by Maroco indates overall classification accuracy amongst all methods provide similar performance with median accuracies ranging from 63% to 76% with SVM providing statistically significant improvements in accuracy. Sensitivity and specificity, however, indicate a different picture. Methods such as SVM and NN show high sensitivities nearing 100% but fall short in sensitivity dropping below the baseline random

TABLE II
MMSE ASSESSMENT CRITERIA

Category	Description
Orientation to time	Indication of current date and time
Orientation to Location	Indication of location narrowing from country to hospital ward
Registration	Repeating name prompts
Attention and Calculation	Basic arithmetic
Recall	Recall names from Registration category
Language 1	Name 2 objects
Language 2	Repeat a verbal statement
Language 3	Obey a 3 stage verbal command
Language 4	Read and obey a written command
Language 5	Write a sentence
Copying	Copy an image of a pair of intersecting pentagons

TABLE III
CERAD BATTERY ASSESSMENT CRITERIA

Category	Description
Word fluency	Enumerate as many instances of animal as possible in 1 minute
Boston naming	Respond to the name of the picture shown
MMSE	See Table II
Memory function 1	Recall after knowing 10 noun words
Memory function 2	Recall previous 10 noun word list
Memory function 3	Repeat memory function category 1&2 on new noun word list
Visuospatial function 1	Transcribe four line drawings of increasing difficulty
Visuospatial function 2	Repeat previous category again
Mental flexibility 1	Draw a line linking sequential numbers
Mental flexibility 2	Link numbers and letters in an alternating sequence

selection of 50%. Such results stem from unbalanced datasets and an emphasis on predictions of the larger proportion class resulting in high specificity and low sensitivity. Such validation results emphasise the danger of evaluation based purely off of total accuracy and the need for further evaluation through metrics such as sensitivity, specificity and accuracy amongst others. Overall, RF and LDA were evaluated to have statistically significantly higher sensitivity and accuracy in relation to remaining methods.

So's approach to evaluation was a two stage evaluation process classifying control and MCI subjects using MMSE categorical scores, and classifying MCI and dementia subjects using CERAD battery categorical scores. Stage 1 consisted of 9799 control subjects and 4201 MCI patients representing the largest dataset used by several orders of magnitude in comparison to other studies within the same field. As such, results favour NNs with an accuracy of 97.2% (Sensitivity (Sens.): 97% Specificity (Spec.): 96%) The overall average accuracy across 7 classification models was 90.9%, considerably higher than any comparable methodology proposed by others, the use of significantly larger datasets is perhaps indicative to improved performance. However with confounding factors such as varied methodology and datasets in use, such a conclusion remains speculative. Stage 2 consists of a significantly smaller dataset of 663 MCI and 573 Demented subjects containing categorical scores from the CERAD battery. Highest performing model was evaluated as SVM with an accuracy of 74.03% (Sens.: 73.8% Spec.: 73.8%) whilst NN was worst performing. Results across all models were greatly reduced however, such difficulty in classifying MCI and dementia is apparent across all relevant studies which explore this classification task.

Although such approaches have shown great promise in diagnostic performance, within a clinical setting, the need to assess a patient for a potentially large amount of assessment categories for accurate diagnosis is infeasible. The proceeding literature attempts to demonstrate accurate diagnosis using a limited selection of assessment criteria across multiple assessments based on feature selection algorithms.

Williams *et al.* [28] uses a wrapper feature selection method [69] in which features are individually evaluated on predictive accuracy and greedily selected based on current highest performing feature. Subsequent iterations temporarily append features to be evaluated onto the set of previously selected features and evaluated based on accuracy. Greedy selection of best feature is permanently included into the new dataset until a set number of features are chosen, which in this case is 12 features. Several classification models (NB, Decision Tree (DT), SVM, NN) were also evaluated in combination with the full 149 feature dataset and the reduced 12 feature dataset. Classifying between control, MCI and dementia patients, final results show both increased and decreased accuracy when comparing the full and reduced feature dataset dependant on classification model with an average reduced dataset accuracy of 74.9% and an average full dataset accuracy of 73.3%. As such, wrapper feature selection was able to reduce feature size whilst not significantly altering accuracy statistically. In terms of model evaluation, NB provides the highest accuracy of 83.3% (Spec.: 90.0% Sens.: 81.9%) whilst the greatest accuracy for the full feature dataset was NN with 80.3% (Spec.: 89.0% Sens.: 78.4%).

Weakley *et al.* [29] also uses a wrapper feature selection method for classifying control, MCI and dementia subjects on a smaller 27 feature dataset. However, feature count threshold was determined as a $<2\%$ increase in accuracy across iterations. Performance on the full dataset was reported as accuracies between 66.2-79.7% (Sens.: 46.5-88.8% Spec.: 69.8-96.1%) whilst performance on the reduced dataset was consistently higher, accuracy of 80.6-87.6% (Sens.: 58.8-93.5% Spec.: 75.2-97.3%). Such indications of performance from the aforementioned studies confirm the capability of data-centric based feature selection.

Battista *et al.* [27] performs feature selection on a set of 131 neuropsychological assessment criteria based on the class discriminatory power of features via Fisher's Discriminant Ratio with a selection threshold of the top 5% of features. A baseline feature selection by 2 experienced neuropsychologists based on redundancy, overlap and relevance was also proposed. Selection methodologies were evaluated by classifications, using SVM, of subjects measured on the Clinical Dementia Ratings (CDR) scale, namely CDR $= \{0, 0.5, 1\}$ corresponding to control, MCI and demented respectively [15]. Comparisons were made against features selected by Fisher's Discriminant ratio and by neuropsychologists with no baseline full feature dataset result. Results show comparable performance (Accuracy (Acc.): 81%, Sens.: 78%, Spec.: 82%) compared with manual feature selection by neuropsychologists whom provide marginally improved performance (Acc.: 84%, Sens.: 82%, Spec.: 85%).

In retrospect feature selection methods, whilst still relatively unexplored, show promise in reducing the size of neuropsychological assessments whilst maintaining diagnostic performance comparable to that of current expert human opinion. As such, there is potential for the development of streamlined screening procedures incorporating specific criteria from currently available assessments in conjunction with machine learning techniques for improved or similar diagnostic

performance.

B. Neuroimaging

According to studies, the progression of neuropathology in AD can be observed years before clinical symptoms become apparent [9]. AD is commonly characterised by the formation of intracellular neurofibrillary tangles, extracellular β-amyloid plaques and significant brain atrophy [9]. The resulting neuropathology can thus be detectable using neuroimaging techniques with MRI being the most popular amongst the literature reviewed. Over the past few years, various biomarkers and ML approaches have been proposed in identifying patients with MCI and AD with varying results. With no definitive pre-mortem method of reliably diagnosing AD currently, the use of neuroimaging in diagnosis is still an open and challenging research application.

The popularity of MRI based applications stem from MRI being a non-invasive procedure whilst also able to provide high resolution and contrast imaging of the various brain soft tissues allowing for accurate segmentation and thus analysis of biomarkers within brain regions [31]. As such, the popularity and reliance on MRI diagnosis has allowed for the relatively high availability of datasets. Initiatives such as Alzheimer's Disease Neuroimaging Initiative (ADNI) [70] provide large, standardised neuroimaging datasets for research into AD biomarkers. With over 600 publications, ADNI has proven to be a popular dataset within neuroimaging research [71].

MRI scans however, suffer from a prohibitive cost [5], [22] and availability in regards to technology and training around the globe. As a result, the use of MRI is infeasible as an initial, definitive diagnosis with patients generally going through initial screening before MRI scans are used as mentioned previously. The need for an expert neuroradiologist for a consistently accurate diagnosis when compared to a general radiologist [21] further limits availability and effectiveness. With diagnosis determined through manual measurements of various brain regions for specific biomarkers, there is opportunity for automation through the use of ML.

Pre-processing

Normalisation and segmentation of scan data remains a open area for novel methodologies to be developed. Based in part due to variances in MRI hardware producing images of varying image, voxel and slice sizes in addition to variances in overall brain structure across patients. As such, comparisons across multiple individuals requires non-linear image transformations for normalisation whilst preserving features such as brain tissue volume and density. Various methodologies and applications to address the aforementioned issue have been proposed within literature.

With a heavy focus on White Matter (WM) regions of the brain specifically, in addition to increasingly larger MRI datasets becoming available such as from the aforementioned ADNI initiative, automated image segmentation applications allow for fast segmentation of relevant brain regions for further feature extraction. While manual labelling is still done [45], the use of automated segmentation allows for a push towards a fully automated clinical application while remaining comparatively accurate to human labelling across multiple segmentation methodologies within literature.

Costafreda *et al.* [33] uses an auto-context segmentation model for automated hippocampal segmentation based upon the use of only a few ground truth labelled brains [72] which has been reported to produce accurate segmentation based on ground truth labels by human experts [73].

A popular normalisation methodology after segmentation is the Regional Analysis of Volumes Examined in Normalized Space (RAVENS) voxel-based analysis method [74]. An adaptation of the Boxel method for measuring volumes of WM, Grey Matter (GM) and Cerebro Spinal Fluid (CSF) within different brain regions. The normalisation of the variances in individual's brain structures into a global atlas for analysis (Talairach space), results in deformation of original image volume. RAVENS preserves tissue volumes into intensity-based maps which can thus be overlaid onto a global atlas yielding a normalised intensity-based volume map of various brain regions. RAVENS has been shown to be as reliable and accurate as a trained operator in regards to volume computation of brain region volumes [74] and remains popular amongst current literature 20 years later [34], [35], [40].

In regards to specific brain region feature extraction, Shi *et al.* [75] provides a novel approach for hippocampal volume calculation by mapping an initial mesh representation of the hippocampus into a common mesh structure across all subjects with one-to-one vertex correspondence. This allows for a robust correspondence between patients invariant to orientation and position. Costafreda *et al.* [33] then proceeds to measure radial distances from center to each mapped vertex corner as features for use in conventional patient classification.

Regional Brain Feature Selection

Amongst literature, the use of pre-defined Region of Interest (ROI) such as the hippocampus have proven quite popular and effective at distinguishing between the various cognitive categories. However, several papers continue to propose novel ROI which also prove effective in diagnosis. Through the use of various feature selection methodologies, heat maps of MRI scans allow for visualisation of potential ROI.

Davatzikos *et al.* [35] combined the latter concepts by proposing the use of local voxel PCA on a region by region basis to determine brain regions of maximum "*effect*" between AD, MCI and healthy subjects using RAVENS maps. The resulting reduced feature set is passed to an SVM classification model for final diagnosis. As a result, feature selection was able to identify several regions of high effect for distinction between the various impairment categories with the capability to classify between MCI vs. control and MCI vs. AD with accuracies ranging from 84.3% to 100% respectively.

Whilst the aforementioned applications performing region selection separate to that of classification have provided effective diagnostic results, several papers have proposed an integrated region selection and classification methodology. The advantage of which is the capability to more tailor ROI towards final classification providing a more valid set of reduced final features, explicitly relevant to classification.

Examples of such integrated ROI selection include that of Teipel *et al.* [43] in which, the derived Jacobian determinant maps of the transformation matrix from a MRI scan to a data averaged brain atlas, were generated. The resulting Jacobian determinant maps were then segmented with GM and WM extracted and Gaussian smoothed. Finally, the use of PCA on the co-variance matrix of all scans allows for the determination of maximum variance between scans. The principal components or eigenimages can thus be attributed to AD based on correlation using Wilk's lambda [76]. As such, AD could be predicted based on specific regions of atrophy in brain matter (WM & GM) and regions of enlargement in CSF spaces.

Kloppel *et al.* [37] also proposes a ROI selection methodology, post-classification based upon the evaluation of SVM importance weightings. By summing the weight and label multiplied scans on a pixel by pixel basis, ROI which especially separate cognitive categories can be shown in heat map form.

Mesh mapping has also been attempted on whole brain scans as opposed to individual ROI. Lerch *et al.* [19] adapts a cerebral

TABLE IV
LITERATURE INVESTIGATING THE DIAGNOSTIC POTENTIAL FOR ALZHEIMERS' DISEASE OF VARIOUS ML MODELS WITHIN NEUROIMAGING.

ROI	Sample			Model	Avg. Accuracy(%)	Reference
	Healthy Control	MCI	Alzheimers' Disease			
Hippocampus	88	103	71	SVM	80	[33]
Hippocampus	21	14	15	LDA	89.7	[45]
Medial Temporal Lobe	75	-	75	LDA, SVM	92	[41]
Multiple Parahippocalampal Gyrus	17	-	19	LDA	94	[19]
All Brain Regions	52	99	51	AE, SVM	85.6	[47]
All Brain Regions	28	-	58	SVM	84	[36]
All Brain Regions	226	-	191	SVM	84.2	[32]
All Brain Regions	190	-	190	SVM	89.3	[38]
All Brain Regions	34	-	34	SVM	94	[21]
All Brain Regions	37	-	37	SVM	100	[35]
All Brain Regions	18	24	32	LR	83	[43]
All Brain Regions	232	411	200	CNN	89.47	[46]

AE: Autoencoder, CNN: Convolutional Neural Network, LDA: Linear Discriminant Analysis, LR: Logistic Regression, SVM: Support Vector Machine

cortex deformable model [77] to generate large vertex meshes (81,920 polygons) of WM and GM. The resulting meshes are then used to perform automated cortical thickness measurements of distances between linked vertices. The resulting methodology is a attempt at direct automation of the standard manual clinical procedure of regional dimension measurements for diagnosis which shows great promise with reported accuracies of up to 94% using an LDA model.

Deep learning models have also been proposed for brain region encoding. Stacked autoencoders were used to extract latent patterns from GM tissue volume of 93 ROI within MRI and PET scans before being passed into a SVM classification model to identify between AD, MCI and healthy subjects [47]. Stacked autoencoders outperformed direct input of unencoded features consistently by 1-5% depending on classification task.

Performance

On the high-end of diagnosis using ML with MRI scans, applications have proven to be extremely effective with diagnostic ability rivalling, if not beating that of expert neuroradiologists. Kloppel *et al.* [21] provides an example of such conclusion through a study involving the direct comparison of diagnostic performance between a SVM model and a selection of 6 radiologists with varying levels of experience. Over 3 datasets comprised of 52 confirmed AD cases with an equal, age matched control and Fronto-temporal Lobar Degeneration (FTLD) samples, the SVM model with an Acc. of 92.4% (Sens.: 92.8%, Spec.: 91.8%) was able to outperform human radiologists with an Acc. of 80.0% (Sens.: 77.2%, Spec.: 82.0%). The aforementioned dataset however, was limited to individuals showing no neurological co-morbidities resulting in reduced confounding variables or influences, poorly reflecting real world samples. In addition, AD patients were limited to definitively diagnosed cases and as such had limited mild AD cases. As a result, the overall dataset can be argued to not be a true representative sample of real world cases, providing a possible explanation to the model's high performance. Due to the human radiologists being evaluated on a the same data sample however, the conclusion that SVM is able to outperform their human equivalent is still very much applicable.

Classification model

Within reviewed literature, the popularity of SVM as the primary diagnostic model is highly prevalent [8] with methodology novelty originating from image preparation, feature selection and objectives. Alternative classification models have been proposed such as LDA [45], RF [78], NN [46] and LR [44]. With models exhibiting different strengths and weaknesses, comparisons against each other are difficult due to various confounding factors such as: dataset source, preparation and population demographics.

Stepping away from AD diagnosis, Chen *et al.* [34] attempts to evaluate multiple classification models (SVM, NB, LR, LDA, NN and DT) on general dementia diagnosis based upon a set of 113 dementia and control subjects using RAVENS GM maps as features. As shown in Table V: SVM, LR and LDA are amongst the highest performing classification models across all brain regions as well as specific ROI. However, with a total of 91 features across the entire RAVENS map and 12 on the specific ROI whilst having a limited training set of 33 dementia and 50 control samples, methodologies such as NN which require significantly larger datasets will remain limited in performance. As such, model evaluation can only feasibly be restricted to similar datasets as opposed to on a more widespread scale.

TABLE V
RESULTS FROM MULTIPLE CLASSIFICATION MODELS EVALUATED BY [34]

Method	Accuracy (%)	
	All brain regions	Medial temporal lobe
NB	69.9	77.1
DT	57.8	71.1
SVM	79.5	73.5
NN	68.7	69.9
LDA	75.9	80.7
LR	75.9	80.7

Classification was performed using RAVENS GM maps on whole brain scans and on specifically the medial temporal lobe.

Chen *et al.* also provides results on ROI feature selection based on model parameters taken from NB, DT, LDA and LR. In all of which, the right hippocampus features as the most influential ROI supporting the popularity of AD diagnosis based on the aforementioned ROI.

C. Speech Analysis

Semantic dementia (SD) is characterised by cognitive deterioration involving semantic memory, language and perceptual processes. All of which, exhibit heavily during periods of Spontaneous Speech (SS) resulting in reduced semantic fluency with an over-utilisation of generic vocabulary ('thing', 'stuff', that', etc.) [79]. As such, there is opportunity in the development of a non-invasive diagnosis technique without the need for specialised personnel or laboratory equipment via SS analysis.

With initial clinical studies in the past mainly focusing on individual words and concepts in an attempt to quantify level of deterioration [80], [81], only very recently have studies attempted continuous SS analysis. With relatively recent advancements in the field of ML and Natural Language Processing (NLP); the use of ML in continuous SS analysis has only recently been proposed.

Initial analysis of SS has shown distinguishable separation between discourse by SD subjects compared to their equivalent control. 42 transcripts by SD and control subjects of The Cookie Theft picture description task [82] based upon word frequency (Bag of Words) were analysed by Garrard *et al.* [50]. Transcripts from SD and control subjects were able to be clustered based on the first two principal components within PCA transformed space based upon transcript bag of words. With feature variance of said principal components emphasising grammatical function words, general uncertainty and generic concepts; said conclusions coincided with previous clinical studies providing indication of the potential of SS analysis in semantic dementia diagnosis.

Garrard *et al.* continues his work using continuous SS analysis for direct diagnosis of SD versus control whilst also attempting to predict temporal lobe asymmetry in either the left or right side in SD patients. Using bag of words to encode transcriptions from participants using the Western Aphasia Battery, a naive Bayes' classifier was used for classification resulting in an average accuracy across both models of 95%. Garrard also attempted feature selection of word categories based upon the Information Gain Algorithm by Mitchell [83], but was unable to discern a statistically significant improvement in accuracy based upon a reduced feature dataset.

Other methods of speech encoding have also been explored outside the bag of words approach. For instance Jarrold *et al.* [48] attempts a combination of acoustic-level and lexical feature extraction involving the measurement of word and letter durations, length of pause and other undisclosed features for acoustic-level feature extraction whilst lexical feature extraction measures frequency of word types such as nouns, verbs, etc. Using NN as the classification model across a set of 48 patients, reported accuracies of 88% in determining between AD and control show promising results.

Lopez *et al.* [4] expands on the concept of acoustic-level features through the evaluation of emotional response during SS through a novel emotional temperature feature proposed in earlier work [84]. Using a combination of emotional response and regular SS fluency features [85]. Using NN on a set of 20 AD patients and 50 control, the proposed method was able to reach validation accuracies of up to 97.7% outperforming established methods of diagnosis such as as MRI.

Speech analysis has shown great promise in diagnostic capability whilst being non-invasive and readily available. With a great focus on predictive accuracy and automation of diagnostic procedure, perhaps due in part to the complexities of human speech and emotion, without more work into the interpretability of features and models and validation on larger clinical trials; the use of speech analysis as a definitive diagnostic procedure is still far from ready. In regards to the methodology currently proposed, the potential for time-series based classification models such as RNN based off of NN is apparent. In fields of speech recognition, RNN has shown improvements over previous attempts [86].

VI. Opportunities & Future Research

Whilst the general field of Big Data analytics continues to mature, the current state of ML in dementia diagnosis remains behind current state of the art methodologies. Nonetheless many studies have proposed applications able to deliver promising dementia biomarkers or propose diagnostic procedures in collaboration with ML methods able to outperform current procedure. Several avenues of research still remain relatively unexplored in addition to advances in fields of ML opening previously unavailable avenues.

A. Potential unexplored machine learning methodologies

Within the fields of NLP, computer vision, and time-series based analysis, the incorporation of modern ML applications have enabled the potential for complex and novel modellings of various problem spaces.

1) Natural Language Processing: Current literature involving the application of NLP algorithms have shown promise in the diagnosis or screening of SD through analysis of semantic fluency over continuous periods of SS. Being non-intrusive and passive whilst maintaining reported performance metrics comparable to current established screening practice, such applications show great potential as complementary tools in conjunction to current procedures within social or primary care establishments.

NLP still remains in it's infancy however, with current state-of-the-art being described as lagging behind the pace of other growing and innovating technologies [87]. Cambria *et al.* [87] justifies this viewpoint through their envisioned evolution of NLP research indicating the three paradigms of current and future research direction. Of which, state-of-the-art NLP remains within the first category of syntactic analysis, the use of keyword, punctuation and co-occurrence frequency in analysing language. While the next major step in NLP technology is predicted to be the incorporation of word and phrase semantics and sentics into MLP applications.

The advancement of NLP technologies provide far reaching effects considering the continuing ubiquity of NLP applications such as voice assistants within smart phones and household appliances. Current dementia diagnosis applications using NLP, rely on bag of words or word2vec [88] methodologies for word encoding, following the first paradigm of NLP research as mentioned previously. Incorporation of word and phrase semantic encoding directly applies to applications analysing patient syntactic complexity and semantic fluency in spoken vocabulary. The reduced semantic fluency of patients suffering from SD can be directly attributed to semantic encoding complexity as compared to those not suffering from SD.

Further research directions include time-series based analysis of sequential words within sentences. With word order of appearance currently being largely ignored. Lopez lightly touches upon this concept with his use of sliding window emotional temperature analysis [4], however the flattening of said time-series features eliminates the potential for word order analysis. Such time-series based analysis techniques allow for the incorporation of the changing conditions shown by dementia subjects during the interview process. Several papers observe a noticeable drain on energy and patience in dementia subjects as interviews continue generally resulting in early termination of interviews. Analysis of such features coupled with Lopez's emotional temperature and time-series based modelling provide a possible novel approach to diagnosis.

2) Computer vision: A major challenge currently in neuroimaging based diagnosis consists of image normalisation prior to classification. The need for consistent features across various MRI hardware and brain compositions, current feature encoding focuses greatly on regional GM densities, ROI dimensions or flattening of images into non-spatial vectors. Applications such as CNNs which reference spatial relationships between neighbouring pixels enable the potential identification of ROIs simultaneously to classification training.

Common applications within the field of computer vision include the recently established methodology of transfer learning pre-trained large scale image classification models for use in various alternative ML tasks using limited observation datasets [89]. Fields of research within neuroimaging provide an excellent and novel area in which such methodologies can be leveraged for improved ROI identification and even Alzheimer's diagnosis. Additionally, CNN medical image segmentation continues to advance rapidly such as the recently proposed 3D volumetric medical image segmentation application by Milletari *et al.* [90] advancing segmentation of volumetric medical image data, such as MRI, from the previous 2D segmentation slice and merge methodologies to a simultaneous 3D segmentation whilst improving processing time.

Such applications however, are potentially limited by the black box nature of NN methodologies. The interpretability of such methodologies remain an ongoing research avenue within various fields [91], health informatics being one such potential field.

Mesh representation feature encoding of ROIs generate irregular domains such as in Electroencephalogram (EEG). As such, the use of regular domain ML methodologies which respect regional features such as CNNs produce inefficient sparse representations of said domain. Irregular domain ML applications such as graph based CNNs [92] would provide a better fit by respecting the irregular domain interactions between nodes.

With such a heavy reliance on edge measurements of various ROI, improved segmentation applications are also open for further research.

B. Health data linkage

A major prerequisite for any Big Data based complex modelling and applications is data availability. With the majority of research currently relying on small patient groups with observations in the hundreds to occasional thousands, various organisations have devoted immense effort into the creation of large-scale datasets appropriate for research. The ADNI dataset for MRI scans and patient demographics was mentioned previously along with the CERAD database for neuropsychological assessments. Several other databases exist, as shown in Table VI, for generic EHR which provide extremely large, full featured datasets of patient history. Opening a new avenue into dementia prognosis based on the chronic and degenerative nature of dementia providing continual data on individuals. Coupled with time-series modelling, EHR enable the exploration of dementia prognosis.

TABLE VI
SEVERAL LARGE SCALE EHR AND DATA LINKAGE DATABASES.

Name	Region
Secure Anonymised Information Linkage (SAIL) [93]	Wales, UK
Scottish Informatics Programme (SHIP) [94]	Scotland, UK
Data Linkage Western Australia [95]	Western AUS
Institute for Clinical Evaluative Sciences (ICES) [96]	Ontario, CAN
Manitoba Centre for Health Policy (MCHP) [97]	Manitoba, CAN

EHRs however, provide multiple challenges which limit potential applications. The predominant challenge being the wide-ranging and non-specific patient information recorded in such datasets. The resulting patient data presented are generally sparse and highly dimensional, compounded by a lack of prior knowledge in what constitutes as relevant data utilised in specific domains such as dementia diagnosis. The use of sparse, high dimensional EHR data within health informatics presents two major challenges: human interpretability, requiring the employment of sparse optimised feature selection, dimensionality reduction or representation learning for effective biomarker and risk factor identification; and adequate data coverage producing meaningless artefacts and bias termed *sparse data bias*, potential solutions of which exist [98].

In regards to EHR encoding, various avenues of research exist which address this challenge: representation learning technologies remain a constantly evolving field [99] with which to adapt into the field of EHR health informatics, whereas methodologies from alternative ML fields with similar data structures allow for potential adaptation into EHR encoding such as word representation methodologies within NLP, of which, methodology such as word2vec by Mikolov *et al.* remains highly popular [88]. With no single *de facto* methodology for EHR encoding, there remains great potential in the proposal of novel tailor-made encoding methodologies for EHRs.

Finally, relatively little research has focused on evaluation or diagnosis across simultaneously multiple data types. With health data linkage continuing to provide the possibility for full, structured and detailed records for individual care, the use of detailed assessments such as MRI, EEG, and cognitive assessments can be coupled with long term patient histories from EHRs allowing for the creation of fully fledged and thorough diagnostic support systems. While such work has been attempted using statistical methods [2], and through ML methods [5], little research has continued within such research avenue.

C. Prognosis

Within reviewed literature, the difficulty of classifying MCI versus dementia patients remains a continual observation. With reported evaluation accuracies indicating a consistent significant drop in comparison to control versus MCI or full dementia classification. The use of neuroimaging, cognitive assessment and discourse analysis have been unable to classify continued degeneration effectively whilst other approaches such as EHR remain unexplored.

In retrospect, little research has also gone into actual prediction of MCI and dementia conversion based on historical patient history. Such research applications would provide great potential into identification of risk factors and biomarkers indicating rates of cognitive decline. Several clinical studies have attempted equating cognitive decline to cognitive assessment scores statistically [100] however, the use of modern ML techniques may provide novel indications.

Continued cognitive decline provides a sequential time-line of discrete events indicating the gradual worsening of dementia symptoms, such time-series based data serves as a perfect example for modelling on time-series based methodologies. As mentioned in section IV-G, RNNs allow for short term memory of past events hence, ideal for applications modelling dementia conversion. Such applications can potentially provide improved predictions of cognitive decline allowing for personalized tailored medical care or identify future at-risk individuals for close monitoring. However, DL technologies such as RNN remain a concern in human interpretability and validation. Consequently, following on from the example of at-risk identification, indications to the reasoning behind predicting an individual as at-risk remain unknown. Such issues, remain an ongoing research challenge.

The degenerative nature of dementia results in an increase in comorbidities [24], institutionalisation [25] and fall rate [26]. Several studies have proven the statistical significance of dementia as a risk factor to hospitalisation [24], [101]–[103]. There remains untapped potential in prediction of institutionalisation risk and hospitalisation outcomes for dementia patients using ML applications.

D. Deep Learning

A highly advanced and influential field of ML, DL architectures such as deep NNs, RNNs and deep belief networks are at the heart of various high profile applications such as IBM Watson [63] and Google Translate [64] within the field of NLP or AlphaGo beating the then current human European Go champion [65]. In computer vision, deep CNNs revolutionised image classification during the 2012 ImageNet Large Scale Visual Recognition Challenge (ILSVRC) with SuperVision outperforming other entries by a margin of 11% image classification error [6]. Subsequent challenges have thus been dominated by various adaptations of deep CNNs [104]. DL has seen use within medical fields such as drug discovery [105], patient categorisation [106], imaging [107], biomedical text mining [108] and EHRs [109]. With DL proving capability in representing complex functions within other fields, the merger of such DL architectures with Dementia healthcare applications have potential for novel approaches to current research.

There is, however, criticism on various aspects of DL methodologies which limit widespread use within fields such as medicine. Criticisms which originate from the general NN model that architectures such as CNNs, RNNs and deep belief networks stem from. Issues of transparency or the "black box" nature of models limit validation of trained models to purely empirical evidence formed from unseen test data. As a result, the reliance on a model to correctly represent relevant aspects of the dataset are not always guaranteed as shown by Ribeiro *et al.* [110]. However, the limiting of such potential issues of dataset bias, indicated by Ribeiro, remain a primary concern within health informatics with great emphasis on minimising confounding bias in population demographics made on the majority of reviewed literature.

While attempts have been made to improve acceptance of ML among medical experts through the integration of existing medical knowledge into RF models for diagnosis [111], there remains potential for adapting DL architectures for improved acceptance within medicine. With acceptance, comes the possibility of adapting the proven capabilities of DL within other fields into the mainstream of health informatics.

VII. Conclusion

Health informatics in dementia research remains a open research field with a multitude of research avenues spanning into various other regions of ML and data analytics. Our findings suggest great and continued interest in adapting current diagnostic procedures with ML models in order to improve upon diagnostic performance or to identify novel, potential biomarkers as new avenues for diagnosis. Classical ML approaches such as SVM, PCA, RF and LR remain highly popular whilst modern DL methodologies are slowly being proposed with promising but mixed results.

This focus of this review on ML applications within dementia diagnosis, whilst thorough, focuses purely on ML based applications. As such, there is still further work available in direct comparison of ML against the more established statistical methodology of other clinical trials. Such an evaluation could provide greater illumination on the current state of ML as a whole within the space of health informatics. Outside of health informatics, there remains great interest in ML applications for individualised physical care such as in fall detection, robotics and support systems for dementia patients. Such applications fall outside of the scope of informatics, however, further review and evaluation of such systems remain open.

Current applications of ML have shown promising results; with continued advances in information technology enabling access to large scale patient records. The modern era of Big Data opens avenues into complex modelling and consequent potential for significant advances in dementia care. The chronic and degenerative nature of dementia provides a unique collection of challenges in the diagnosis, prognosis and care of patients which still remain to be fully addressed, there remains much to do in resolving the issue of dementia and in improving the care of current dementia sufferers.

References

[1] P. D. Sloane, S. Zimmerman, C. Suchindran, P. Reed, L. Wang, M. Boustani, and S. Sudha, "The Public Health Impact of Alzheimer's Disease, 20002050: Potential Implication of Treatment Advances," *Annual Review of Public Health*, vol. 23, no. 1, pp. 213–231, 2002.

[2] M. Boustani, C. M. Callahan, F. W. Unverzagt, M. G. Austrom, A. J. Perkins, B. A. Fultz, S. L. Hui, and H. C. Hendrie, "Implementing a screening and diagnosis program for dementia in primary care," *Journal of General Internal Medicine*, vol. 20, no. 7, pp. 572–577, 2005.

[3] Alzheimer's Society, "Dementia UK Report," Tech. Rep., 2018.

[4] K. López-de Ipiña, J.-B. Alonso, C. Travieso, J. Solé-Casals, H. Egiraun, M. Faundez-Zanuy, A. Ezeiza, N. Barroso, M. Ecay-Torres, P. Martinez-Lage, and U. Lizardui, "On the Selection of Non-Invasive Methods Based on Speech Analysis Oriented to Automatic Alzheimer Disease Diagnosis," *Sensors*, vol. 13, no. 12, pp. 6730–6745, 2013.

[5] J. Escudero, E. Ifeachor, J. P. Zajicek, C. Green, J. Shearer, and S. Pearson, "Machine learning-based method for personalized and cost-effective detection of alzheimer's disease," *IEEE Transactions on Biomedical Engineering*, vol. 60, no. 1, pp. 164–168, 2013.

[6] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet Large Scale Visual Recognition Challenge," *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211–252, 2015.

[7] G. López, L. Quesada, and L. A. Guerrero, "Alexa vs. Siri vs. Cortana vs. Google Assistant: a comparison of speech-based natural user interfaces," in *International Conference on Applied Human Factors and Ergonomics*. Springer, 2017, pp. 241–250.

[8] G. Orrù, W. Pettersson-Yeo, A. F. Marquand, G. Sartori, and A. Mechelli, "Using Support Vector Machine to identify imaging biomarkers of neurological and psychiatric disease: A critical review," *Neuroscience and Biobehavioral Reviews*, vol. 36, no. 4, pp. 1140–1152, 2012.

[9] L. Mosconi, M. Brys, L. Glodzik-Sobanska, S. De Santi, H. Rusinek, and M. J. de Leon, "Early detection of Alzheimer's disease using neuroimaging," *Experimental Gerontology*, vol. 42, no. 1-2, pp. 129–138, 2007.

[10] T. Ching, D. S. Himmelstein, B. K. Beaulieu-Jones, A. A. Kalinin, B. T. Do, G. P. Way, E. Ferrero, P.-M. Agapow, M. Zietz, M. M. Hoffman, W. Xie, G. L. Rosen, B. J. Lengerich, J. Israeli, J. Lanchantin, S. Woloszynek, A. E. Carpenter, A. Shrikumar, J. Xu, E. M. Cofer, C. A. Lavender, S. C. Turaga, A. M. Alexandari, Z. Lu, D. J. Harris, D. DeCaprio, Y. Qi, A. Kundaje, Y. Peng, L. K. Wiley, M. H. S. Segler, S. M. Boca, S. J. Swamidass, A. Huang, A. Gitter, and C. S. Greene, "Opportunities and obstacles for deep learning in biology and medicine," *Journal of The Royal Society Interface*, vol. 15, no. 141, p. 20170387, apr 2018.

[11] L. Glodzik, L. Mosconi, W. Tsui, S. de Santi, R. Zinkowski, E. Pirraglia, K. E. Rich, P. McHugh, Y. Li, S. Williams, F. Ali, H. Zetterberg, K. Blennow, P. Mehta, and M. J. de Leon, "Alzheimer's disease markers, hypertension, and gray matter damage in normal elderly," *Neurobiology of Aging*, vol. 33, no. 7, pp. 1215–1227, jul 2012.

[12] G. M. McKhann, D. S. Knopman, H. Chertkow, B. T. Hyman, C. R. Jack, C. H. Kawas, W. E. Klunk, W. J. Koroshetz, J. J. Manly, R. Mayeux, R. C. Mohs, J. C. Morris, M. N. Rossor, P. Scheltens, M. C. Carrillo, B. Thies, S. Weintraub, and C. H. Phelps, "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 and Dementia*, vol. 7, no. 3, pp. 263–269, 2011.

[13] American Psychiatric Association, *Diagnostic and statistical manual of mental disorders*, 5th ed. American Psychiatric Association, 2013.

[14] M. Lamar, S. M. Resnick, and A. B. Zonderman, "Longitudinal changes in verbal memory in older adults," *Neurology*, vol. 60, no. 1, pp. 82 – 86, jan 2003.

[15] A. So, D. Hooshyar, K. Park, and H. Lim, "Early Diagnosis of Dementia from Clinical Data by Machine Learning Techniques," *Applied Sciences*, vol. 7, no. 7, p. 651, 2017.

[16] T. N. Tombaugh and N. J. McIntyre, "The mini-mental state examination: a comprehensive review." *Journal of the American Geriatrics Society*, vol. 40, no. 9, pp. 922–935, 1992.

[17] V. G. Valcour, K. H. Masaki, J. D. Curb, and P. L. Blanchette, "The detection of dementia in the primary care setting." *Archives of internal medicine*, vol. 160, no. 19, pp. 2964–2968, 2000.

[18] C. M. Callahan, H. C. Hendrie, and W. M. Tierney, "Documentation and evaluation of cognitive impairment in elderly primary care patients." *Annals of internal medicine*, vol. 122, no. 6, pp. 422–9, mar 1995.

[19] J. P. Lerch, J. Pruessner, A. P. Zijdenbos, D. L. Collins, S. J. Teipel, H. Hampel, and A. C. Evans, "Automated cortical thickness measurements from MRI can accurately separate Alzheimer's patients from normal elderly controls," *Neurobiology of Aging*, vol. 29, no. 1, pp. 23–30, 2008.

[20] G. Chetelat and J.-C. Baron, "Early diagnosis of alzheimer's disease: contribution of structural neuroimaging," *NeuroImage*, vol. 18, no. 2, pp. 525–541, feb 2003.

[21] S. Klöppel, C. M. Stonnington, J. Barnes, F. Chen, C. Chu, C. D. Good, I. Mader, L. A. Mitchell, A. C. Patel, C. C. Roberts, N. C. Fox, C. R. Jack, J. Ashburner, and R. S. J. Frackowiak, "Accuracy of dementia diagnosis - A direct comparison between radiologists and a computerized method," *Brain*, vol. 131, no. 11, pp. 2969–2974, 2008.

[22] Department of Health and Social Care, "NHS reference costs 2015 to 2016," Tech. Rep., 2016.

[23] M. Boustani, L. Watson, B. Fultz, A. J. Perkins, and R. Druckenbrod, "Acceptance of dementia screening in continuous care retirement communities: A mailed survey," *International Journal of Geriatric Psychiatry*, vol. 18, no. 9, pp. 780–786, 2003.

[24] J. P. W. Bynum, P. V. Rabins, W. Weller, M. Niefeld, G. F. Anderson, and A. W. Wu, "The relationship between a dementia diagnosis, chronic illness, medicare expenditures, and hospital use," *Research and Practice in Alzheimer's Disease*, vol. 10, pp. 160–164, 2005.

[25] C. Jagger, K. Andersen, M. M. B. Breteler, J. Copeland, C. Helmer, M. Baldereschi, L. Fratiglioni, A. Lobo, H. Soininen, A. Hofman, and L. J Launer, "Prognosis with dementia in Europe: A collaborative study of population-based cohorts. Neurologic Diseases in the Elderly Research Group," *Neurology*, vol. 54, pp. S16–20, 2000.

[26] C. Van Doorn, A. L. Gruber-Baldini, S. Zimmerman, J. Richard Hebel, C. L. Port, M. Baumgarten, C. C. Quinn, G. Taler, C. May, and J. Magaziner, "Dementia as a Risk Factor for Falls and Fall Injuries Among Nursing Home Residents," *Journal of the American Geriatrics Society*, vol. 51, no. 9, pp. 1213–1218, sep 2003.

[27] P. Battista, C. Salvatore, and I. Castiglioni, "Optimizing neuropsychological assessments for cognitive, behavioral, and functional impairment classification: A machine learning study," *Behavioural Neurology*, vol. 2017, 2017.

[28] J. Williams and A. Weakley, "Machine Learning Techniques for Diagnostic Differentiation of Mild Cognitive Impairment and Dementia," *Workshops at the ...*, pp. 71–76, 2013.

[29] A. Weakley, J. A. Williams, M. Schmitter-Edgecombe, and D. J. Cook, "Neuropsychological test selection for cognitive impairment classification: A machine learning approach," *Journal of Clinical and Experimental Neuropsychology*, vol. 37, no. 9, pp. 899–916, 2015.

[30] J. Maroco, D. Silva, A. Rodrigues, M. Guerreiro, I. Santana, and A. De Mendonça, "Data mining methods in the prediction of Dementia: A real-data comparison of the accuracy, sensitivity and specificity of linear discriminant analysis, logistic regression, neural networks, support vector machines, classification trees and random forests," *BMC Research Notes*, vol. 4, no. 1, p. 299, 2011.

[31] E. Moradi, A. Pepe, C. Gaser, H. Huttunen, and J. Tohka, "Machine learning framework for early MRI-based Alzheimer's conversion prediction in MCI subjects," *NeuroImage*, vol. 104, pp. 398–412, 2015.

[32] A. Abdulkadir, B. Mortamet, P. Vemuri, C. R. Jack, G. Krueger, and S. Klöppel, "Effects of hardware heterogeneity on the performance of SVM Alzheimer's disease classifier," *NeuroImage*, vol. 58, no. 3, pp. 785–792, 2011.

[33] S. G. Costafreda, I. D. Dinov, Z. Tu, Y. Shi, C. Y. Liu, I. Kloszewska, P. Mecocci, H. Soininen, M. Tsolaki, B. Vellas, L. O. Wahlund, C. Spenger, A. W. Toga, S. Lovestone, and A. Simmons, "Automated hippocampal shape analysis predicts the onset of dementia in mild cognitive impairment," *NeuroImage*, vol. 56, no. 1, pp. 212–219, 2011.

[34] R. Chen and E. H. Herskovits, "Machine-learning techniques for building a diagnostic model for very mild dementia," *NeuroImage*, vol. 52, no. 1, pp. 234–244, 2010.

[35] C. Davatzikos, S. Resnick, X. Wu, P. Parmpi, and C. Clark, "Individual patient diagnosis of AD and FTD via high-dimensional pattern classification of MRI," *NeuroImage*, vol. 41, no. 4, pp. 1220–1227, jul 2008.

[36] L. Ferrarini, W. M. Palm, H. Olofsen, R. Van Der Landen, M. A. Van Buchem, J. H. C. Reiber, and F. Admiraal-Behloul, "Ventricular shape biomarkers for Alzheimer's disease in clinical MR images," *Magnetic Resonance in Medicine*, vol. 59, no. 2, pp. 260–267, 2008.

[37] S. Klöppel, C. M. Stonnington, C. Chu, B. Draganski, R. I. Scahill, J. D. Rohrer, N. C. Fox, C. R. Jack, J. Ashburner, and R. S. J. Frackowiak, "Automatic classification of MR scans in Alzheimer's disease," *Brain*, vol. 131, no. 3, pp. 681–689, 2008.

[38] P. Vemuri, J. L. Gunter, M. L. Senjem, J. L. Whitwell, K. Kantarci, D. S. Knopman, B. F. Boeve, R. C. Petersen, and C. R. Jack, "Alzheimer's disease diagnosis in individual subjects using structural MR images: Validation studies," *NeuroImage*, vol. 39, no. 3, pp. 1186–1197, 2008.

[39] C. Davatzikos, Y. Fan, X. Wu, D. Shen, and S. M. Resnick, "Detection of prodromal Alzheimer's disease via pattern classification of MRI," *Neurobiology of Aging*, vol. 29, no. 4, pp. 514–523, 2009.

[40] Y. Fan, N. Batmanghelich, C. M. Clark, and C. Davatzikos, "Spatial patterns of brain atrophy in MCI patients, identified via high-dimensional pattern classification, predict subsequent cognitive decline," *NeuroImage*, vol. 39, no. 4, pp. 1731–1743, feb 2008.

[41] S. Duchesne, A. Caroli, C. Geroldi, C. Barillot, G. B. Frisoni, and D. L. Collins, "MRI-based automated computer classification of probable AD versus normal controls," *IEEE Transactions on Medical Imaging*, vol. 27, no. 4, pp. 509–520, 2008.

[42] Z. Lao, D. Shen, Z. Xue, B. Karacali, S. M. Resnick, and C. Davatzikos, "Morphological classification of brains via high-dimensional shape transformations and machine learning methods," *NeuroImage*, vol. 21, no. 1, pp. 46–57, 2004.

[43] S. J. Teipel, C. Born, M. Ewers, A. L. Bokde, M. F. Reiser, H. J. Möller, and H. Hampel, "Multivariate deformation-based analysis of brain atrophy to predict Alzheimer's disease in mild cognitive impairment," *NeuroImage*, vol. 38, no. 1, pp. 13–24, 2007.

[44] H. Wolf, A. Hensel, F. Kruggel, S. G. Riedel-Heller, T. Arendt, L. O. Wahlund, and H. J. Gertz, "Structural correlates of mild cognitive impairment," *Neurobiology of Aging*, vol. 25, no. 7, pp. 913–924, 2004.

[45] A. M. Jauhiainen, M. Pihlajamäki, S. Tervo, E. Niskanen, H. Tanila, T. Hänninen, R. L. Vanninen, and H. Soininen, "Discriminating accuracy of medial temporal lobe volumetry and fMRI in mild cognitive impairment," *Hippocampus*, vol. 19, no. 2, pp. 166–175, 2009.

[46] A. Payan and G. Montana, "Predicting Alzheimer's disease: a neuroimaging study with 3D convolutional neural networks," *Journal of Artificial Intelligence Research*, vol. 4, pp. 237–285, feb 2015.

[47] H.-I. Suk and D. Shen, "Deep Learning-Based Feature Representation for AD/MCI Classification," in *16th International Conference on Medical Image Computing and Computer Assisted Intervention*, vol. 8150 LNCS, 2013, pp. 583–590.

[48] W. Jarrold, B. Peintner, D. Wilkins, D. Vergryi, C. Richey, M. L. Gorno-Tempini, and J. Ogar, "Aided diagnosis of dementia type through computer-based analysis of spontaneous speech," *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pp. 27–37, 2014.

[49] P. Garrard, V. Rentoumi, B. Gesierich, B. Miller, and M. L. Gorno-Tempini, "Machine learning approaches to diagnosis and laterality effects in semantic dementia discourse," *Cortex*, vol. 55, no. 1, pp. 122–129, 2014.

[50] P. Garrard and R. Forsyth, "Abnormal discourse in semantic dementia: A data-driven approach," *Neurocase*, vol. 16, no. 6, pp. 520–528, 2010.

[51] J. Friedman, T. Hastie, and R. Tibshirani, *The elements of statistical learning*. Springer series in statistics New York, NY, USA:, 2001, vol. 1, no. 10.

[52] S.-M. Zhou, H.-X. Li, and L. Xu, "A variational approach to intensity approximation for remote sensing images using dynamic neural network," *Expert Systems*, vol. 20, no. 4, pp. 163–170, 2003.

[53] T. Ching, X. Zhu, and L. X. Garmire, "Cox-nnet: An artificial neural network method for prognosis prediction of high-throughput omics data," *PLOS Computational Biology*, vol. 14, no. 4, pp. 1–18, 04 2018.

[54] S. M. Zhou and L. D. Xu, "A new type of recurrent fuzzy neural network for modeling dynamic systems," *Knowledge-Based Systems*, vol. 14, no. 5, pp. 243 – 251, 2001.

[55] D. A. Elizondo, S.-M. Zhou, and C. Chrysostomou, "Light source detection for digital images in noisy scenes: A neural network approach," *Neural Comput. Appl.*, vol. 28, no. 5, pp. 899–909, May 2017.

[56] S. M. Zhou, "Combining dynamic neural networks and image sequences in a dynamic model for complex industrial production processes," *Expert Systems with Applications*, vol. 16, no. 1, pp. 13 – 19, 1999.

[57] D. R. Kelley, Y. Reshef, M. Bileschi, D. Belanger, C. Y. McLean, and J. Snoek, "Sequential regulatory activity prediction across chromosomes with convolutional neural networks." *Genome Research*, 2018.

[58] S. Mc Loone and G. Irwin, "Improving neural network training solutions using regularisation," *Neurocomputing*, vol. 37, no. 1-4, pp. 71–90, 2001.

[59] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," pp. 1–15, 2014.

[60] D.-A. Clevert, T. Unterthiner, and S. Hochreiter, "Fast and Accurate Deep Network Learning by Exponential Linear Units (ELUs)," *International Conference on Learning Representations*, pp. 1–14, nov 2016.

[61] Y. Bengio, P. Lamblin, D. Popovici, and H. Larochelle, "Greedy Layer-Wise Training of Deep Networks," *Advances in Neural Information Processing Systems*, vol. 19, no. 1, p. 153, 2007.

[62] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[63] R. High, "The Era of Cognitive Systems: An Inside Look at IBM Watson and How it Works," *International Business Machines Corporation*, vol. 1, no. 1, pp. 1–14, 2012.

[64] M. Johnson, M. Schuster, Q. V. Le, M. Krikun, Y. Wu, Z. Chen, N. Thorat, F. Viégas, M. Wattenberg, G. Corrado, M. Hughes, and J. Dean, "Google's Multilingual Neural Machine Translation System: Enabling Zero-Shot Translation," *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 11–19, nov 2016.

[65] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. van den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, S. Dieleman, D. Grewe, J. Nham, N. Kalchbrenner, I. Sutskever, T. Lillicrap, M. Leach, K. Kavukcuoglu, T. Graepel, and D. Hassabis, "Mastering the game of Go with deep neural networks and tree search," *Nature*, vol. 529, no. 7587, pp. 484–489, jan 2016.

[66] S. M. Zhou, "Combining dynamic neural networks and image sequences in a dynamic model for complex industrial production processes," *Expert Systems with Applications*, vol. 16, no. 1, pp. 13–19, 1999.

[67] R. Pascanu, C. Gulcehre, K. Cho, and Y. Bengio, "How to Construct Deep Recurrent Neural Networks," pp. 1–13, dec 2013.

[68] G. G. Fillenbaum, G. van Belle, J. C. Morris, R. C. Mohs, S. S. Mirra, P. C. Davis, P. N. Tariot, J. M. Silverman, C. M. Clark, K. A. Welsh-Bohmer, and A. Heyman, "Consortium to Establish a Registry for Alzheimer's Disease (CERAD): The first twenty years," *Alzheimer's & Dementia*, vol. 4, no. 2, pp. 96–109, mar 2008.

[69] M. Gutlein, E. Frank, M. Hall, and A. Karwath, "Large-scale attribute selection using wrappers," in *2009 IEEE Symposium on Computational Intelligence and Data Mining*. IEEE, mar 2009, pp. 332–339.

[70] "Alzheimer's Disease Neuroimaging Initiative." [Online]. Available: http://adni.loni.usc.edu/

[71] M. W. Weiner, D. P. Veitch, P. S. Aisen, L. A. Beckett, N. J. Cairns, J. Cedarbaum, M. C. Donohue, R. C. Green, D. Harvey, C. R. Jack, W. Jagust, J. C. Morris, R. C. Petersen, A. J. Saykin, L. Shaw, P. M. Thompson, A. W. Toga, and J. Q. Trojanowski, "Impact of the Alzheimer's Disease Neuroimaging Initiative, 2004 to 2014," *Alzheimer's and Dementia*, vol. 11, no. 7, pp. 865–884, 2015.

[72] Zhuowen Tu and Xiang Bai, "Auto-Context and Its Application to High-Level Vision Tasks and 3D Brain Image Segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 10, pp. 1744–1757, oct 2010.

[73] J. H. Morra, Z. Tu, L. G. Apostolova, A. E. Green, C. Avedissian, S. K. Madsen, N. Parikshak, X. Hua, A. W. Toga, C. R. Jack, M. W. Weiner, and P. M. Thompson, "Validation of a fully automated 3D hippocampal segmentation method using subjects with Alzheimer's disease mild cognitive impairment, and elderly controls," *NeuroImage*, vol. 43, no. 1, pp. 59–68, oct 2008.

[74] A. F. Goldszal, C. Davatzikos, D. L. Pham, M. X. H. Yan, R. N. Bryan, and S. M. Resnick, "An Image-Processing System for Qualitative and Quantitative Volumetric Analysis of Brain Images," *Journal of Computer Assisted Tomography*, vol. 22, no. 5, pp. 827–837, sep 1998.

[75] Y. Shi, P. M. Thompson, G. I. de Zubicaray, S. E. Rose, Z. Tu, I. Dinov, and A. W. Toga, "Direct mapping of hippocampal surfaces with intrinsic shape context," *NeuroImage*, vol. 37, no. 3, pp. 792–807, sep 2007.

[76] C. Chatfield, *Introduction to multivariate analysis*. Routledge, 2018.

[77] D. MacDonald, N. Kabani, D. Avis, and A. C. Evans, "Automated 3-D Extraction of Inner and Outer Surfaces of Cerebral Cortex from MRI," *NeuroImage*, vol. 12, no. 3, pp. 340–356, sep 2000.

[78] S. Mathotaarachchi, T. A. Pascoal, M. Shin, A. L. Benedet, M. S. Kang, T. Beaudry, V. S. Fonov, S. Gauthier, and P. Rosa-Neto, "Identifying incipient dementia individuals using machine learning and amyloid imaging," *Neurobiology of Aging*, vol. 59, pp. 80–90, 2017.

[79] J. R. Hodges and K. Patterson, "Semantic dementia: a unique clinicopathological syndrome," *The Lancet Neurology*, vol. 6, no. 11, pp. 1004–1014, nov 2007.

[80] P. Garrard and E. Carroll, "Lost in semantic space: a multi-modal, non-verbal assessment of feature knowledge in semantic dementia," *Brain*, vol. 129, no. 5, pp. 1152–1163, may 2006.

[81] S. V. Pakhomov, L. S. Hemmy, and K. O. Lim, "Automated semantic indices related to cognitive function and rate of cognitive decline," *Neuropsychologia*, vol. 50, no. 9, pp. 2165–2175, jul 2012.

[82] H. Goodglass, *Boston diagnostic aphasia examination: Short form record booklet*. Lippincott Williams & Wilkins, 2000.

[83] T. M. Mitchell, "Machine learning. 1997," *Burr Ridge, IL: McGraw Hill*, vol. 45, no. 37, pp. 870–877, 1997.

[84] K. López-de Ipiña, J. B. Alonso, N. Barroso, M. Faundez-Zanuy, M. Ecay, J. Solé-Casals, C. M. Travieso, A. Estanga, and A. Ezeiza, "New Approaches for Alzheimer's Disease Diagnosis Based on Automatic Spontaneous Speech Analysis and Emotional Temperature," in *Proceedings of the 4th international conference on Ambient Assisted Living and Home Care*, ser. Lecture Notes in Computer Science, J. Bravo, R. Hervás, and M. Rodríguez, Eds., vol. 7657, no. February. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 407–414.

[85] "Praat: Doing Phonetics by Computer." [Online]. Available: www.fon.hum.uva.nl/praat

[86] S. Fernández, A. Graves, and J. Schmidhuber, "An Application of Recurrent Neural Networks to Discriminative Keyword Spotting," in *The 17th international conference on Artificial neural networks*, 2007, pp. 220–229.

[87] E. Cambria and B. White, "Jumping NLP Curves: A Review of Natural Language Processing Research [Review Article]," *IEEE Computational Intelligence Magazine*, vol. 9, no. 2, pp. 48–57, may 2014.

[88] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.

[89] M. Oquab, L. Bottou, I. Laptev, and J. Sivic, "Learning and Transferring Mid-Level Image Representations using Convolutional Neural Networks," *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1717–1724, 2014.

[90] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation," jun 2016.

[91] J. Yosinski, J. Clune, T. Fuchs, A. Nguyen, T. Fuchs, and H. Lipson, "Understanding neural networks through deep visualization," *arXiv preprint arXiv:1506.06579*, 2015.

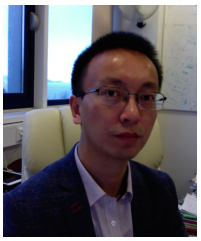
[92] M. Edwards and X. Xie, "Graph Convolutional Neural Network," in *Procedings of the British Machine Vision Conference 2016*. British Machine Vision Association, 2016, pp. 114.1–114.11.

[93] R. a. Lyons, K. H. Jones, G. John, C. J. Brooks, J.-P. Verplancke, D. V. Ford, G. Brown, and K. Leake, "The SAIL databank: linking multiple health and social care datasets," *BMC medical informatics and decision making*, vol. 9, no. 1, p. 3, dec 2009.

[94] Scottish Health Informatics Programme, "SHIP: A Blueprint for Health Records Research in Scotland," Scottish Health Informatics Programme, Tech. Rep., 2012.

[95] C. D. J. Holman, A. J. Bass, I. L. Rouse, and M. S. T. Hobbs, "Populationbased linkage of health records in Western Australia: development of a health services research linked database," *Australian and New Zealand journal of public health*, vol. 23, no. 5, pp. 453–459, 1999.
[96] "Institute for Clinical Evaluative Sciences." [Online]. Available: https://www.ices.on.ca/
[97] L. L. Roos, M. Brownell, L. Lix, N. P. Roos, R. Walld, and L. MacWilliam, "From health research to social research: privacy, methods, approaches," *Social science & medicine*, vol. 66, no. 1, pp. 117–129, 2008.
[98] S. Greenland, M. A. Mansournia, and D. G. Altman, "Sparse data bias: a problem hiding in plain sight," *BMJ*, vol. 353, p. i1981, apr 2016.
[99] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 8, pp. 1798–1828, 2013.
[100] A. Kluger, S. H. Ferris, J. Golomb, M. S. Mittelman, and B. Reisberg, "Neuropsychological prediction of decline to dementia in nondemented elderly," *Journal of Geriatric Psychiatry and Neurology*, vol. 12, no. 4, pp. 168–179, 1999.
[101] E. A. Phelan, S. Borson, L. Grothaus, S. Balch, and E. B. Larson, "Association of Incident Dementia With Hospitalizations," *JAMA*, vol. 307, no. 2, p. 165, jan 2012.
[102] S. Toot, M. Devine, A. Akporobaro, and M. Orrell, "Causes of Hospital Admission for People With Dementia: A Systematic Review and Meta-Analysis," *Journal of the American Medical Directors Association*, vol. 14, no. 7, pp. 463–470, jul 2013.
[103] M. E. Soto, S. Andrieu, S. Gillette-Guyonnet, C. Cantet, F. Nourhashemi, and B. Vellas, "Risk factors for functional decline and institutionalisation among community-dwelling older adults with mild to severe Alzheimer's disease: one year of follow-up," *Age and Ageing*, vol. 35, no. 3, pp. 308–310, may 2006.
[104] Stanford Vision Lab, "ImageNet." [Online]. Available: http://image-net.org/
[105] G. E. Dahl, N. Jaitly, and R. Salakhutdinov, "Multi-task neural networks for QSAR predictions," *arXiv preprint arXiv:1406.1231*, 2014.
[106] J. Tan, M. Ung, C. Cheng, and C. S. Greene, "Unsupervised feature construction and knowledge extraction from genome-wide assays of breast cancer with denoising autoencoders," in *Pacific Symposium on Biocomputing Co-Chairs*. World Scientific, 2014, pp. 132–143.
[107] N. C. F. Codella, Q.-B. Nguyen, S. Pankanti, D. A. Gutman, B. Helba, A. C. Halpern, and J. R. Smith, "Deep learning ensembles for melanoma recognition in dermoscopy images," *IBM Journal of Research and Development*, vol. 61, no. 4, pp. 1–5, 2017.
[108] S. Liu, B. Tang, Q. Chen, and X. Wang, "Effects of semantic features on machine learning-based drug name recognition systems: word embeddings vs. manually constructed dictionaries," *Information*, vol. 6, no. 4, pp. 848–865, 2015.
[109] R. Chalapathy, E. Z. Borzeshi, and M. Piccardi, "Bidirectional LSTM-CRF for clinical concept extraction," *arXiv preprint arXiv:1611.08373*, 2016.
[110] M. T. Ribeiro, S. Singh, and C. Guestrin, ""Why Should I Trust You?"," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '16*. New York, New York, USA: ACM Press, 2016, pp. 1135–1144.
[111] M. J. Pazzani, S. Mani, and W. R. Shankle, "Acceptance of rules generated by machine learning among medical experts," *Methods of Information in Medicine*, vol. 40, no. 5, pp. 380–385, 2001.

Gavin Tsang Gavin Tsang is a third year Ph.D. candidate within the Department of Computer Science, Swansea University, Swansea, U.K. He received his M.Eng. (Hons) degree in computing from Swansea University in addition to the award for "best performance by a student" in 2016. His research interests include machine learning and pattern recognition within the fields of healthcare and finance. He is a student member of the IEEE and a member of the BMVA.



Xianghua Xie Xianghua Xie received the M.Sc. (with commendation) and Ph.D. degrees in computer science from the University of Bristol, Bristol, U.K., in 2002 and 2006, respectively. He is an Associate Professor with the Department of Computer Science, Swansea University, Swansea, U.K. He was a recipient of an RCUK academic fellowship, and he is currently leading the Computer Vision and Machine Leaning Lab at Swansea University. His research interests include various aspects of pattern recognition and machine intelligence and their applications to real-world problems. He has published more than 140 refereed conference and journal publications and (co-)edited several conference proceedings. He is an associate editor of a number of journals, including IET Computer Vision. He is a senior member of IEEE and a member of BMVA.

Shang-Ming Zhou Shang-Ming Zhou is with the Health Data Research UK at Swansea University. His research interests include machine learning and AI in healthcare and medicine, big data analytics, biomedical and health informatics; computational intelligence (fuzzy logic, deep learning/neural networks), information aggregation and integration (type-1 OWA operators, type-2 OWA operators), data-driven methods for epidemiology and public health. Dr Zhou is the Member of IEEE, and the Member of International Federation for Information Processing (IFIP) Technical Committee on Information Systems Working Group on Enterprise Information Systems (WG-8.9), and the Member of IEEE Systems, Man and Cybernetics Society Technical Committee on Enterprise Information Systems. He was the recipient of "Outstanding Reviewer Award" from IEEE Transactions on Cybernetics, Applied Soft Computing, Fuzzy Sets and Systems, Journal of Biomedical Informatics, Knowledge Based Systems, Expert Systems with Applications, Journal of Science and Medicine in Sport, and the "Outstanding Academic Service Award" from the IFIP- WG-8.9 respectively.