The Use of Data Mining Methods for the Prediction of Dementia:
Evidence from the English Longitudinal Study of Aging

Hui Yang, Peter A. Bath

***Abstract*—Dementia in older age is a major health concern with the increase in the aging population. Preventive measures to prevent or delay dementia symptoms are of utmost importance. In this study, a large and wide variety of factors from multiple domains were investigated using a large nationally-representative sample of older people from the English Longitudinal Study of Ageing (ELSA). Seven machine learning algorithms were implemented to build predictive models for performance comparison. A simple model ensemble approach was used to combine the prediction results of individual base models to further improve predictive power. A series of important factors in each domain area were identified. The findings from this study provide new evidence on factors that are associated with the dementia in later life. This information will help our understanding of potential risk factors for dementia and identify warning signs of the early stages of dementia. Longitudinal research is required to establish which factors may be causative and which factors may be a consequence of dementia.**

***Index Terms*—Mental health, cognitive informatics, gerontechnology.**

I. Introduction

CURRENT EVIDENCE suggests substantial increases in the prevalence of dementia across the world [21]. It is estimated that approximately 47 million people worldwide were living in dementia in 2015: this number is projected to triple by 2050. As the incidence and prevalence of dementia increase in older people, individual suffering and the burden of dementia will have major implications for people with dementia, their families, and health/social care systems. Dementia in older age is thus a significant public health concern in the context of global population ageing. In the absence of a disease-modifying treatment or cure, timely diagnosis and interventions are key to providing the optimal care of older adults with dementia [25]. Therefore, a better understanding of the manifestation of dementia, and the associated factors, is important for public health.

Recent evidence [4, 26] suggests that several cardiovascular (CVD) and non-cardiovascular (non-CVD) risk factors (e.g., stroke, diabetes, and hypertension) may be associated with an increased risk of dementia. Evidence from several prospective cohort studies [5, 8, 24] highlights psychological factors, in particular depression, loneliness, and sleep disturbance, that are implicated in the risk of dementia. There is increasing evidence [9, 27] to support a link between dementia and sensory impairment on eyesight and hearing.

There is also promising evidence [18, 34] to suggest that social support and social engagement may have a positive effect on reducing the risk of developing dementia, and alleviating symptoms. Other evidence [7, 15, 20, 33] suggests that healthy lifestyles, such as physical activity, stopping smoking, reduced alcohol intake, or a Mediterranean diet, may contribute to maintaining cognitive health in later life.

Current studies on dementia in older adulthood have two main limitations. First, most existing research has examined the effect of a limited number of risk factors (e.g., 5-15 factors) for, and associated with, dementia. However, dementia is a complex and multidimensional condition, which is associated with a wide range of factors. Hence, a systematic and complete examination about the influence of factors from multiple domains could provide a more robust and pragmatic way to find the most significant risk or protective factors for dementia, and factors associated with the condition. Second, two systematic reviews [1-2] have synthesized findings from previous research on the use of data mining techniques (e.g., deep learning, SVM, Decision Tree J48, and Random Forest) in the clinical diagnosis of dementia. However, most of the datasets used for dementia detection have focused on genetic, clinical, or neuroimaging data using numerical values. In contrast, our data contain demographic, economic, social, lifestyle, and psychological information with a large number of categorical and binary values: analyzing these data could offer new insights in dementia research. Few, if any, studies have used advanced machine learning techniques for these types of psychosocial data. These differences in data characteristics require specific consideration when choosing appropriate ML techniques to build predictive models.

Most social science studies [5, 7-9, 15, 33-34] have used statistical methods to study dementia in elderly people: the overall contribution of this study is that it addresses the issue of dementia in older adults from a new perspective, i.e., using advanced data analytical methods for the identification of factors associates with dementia in older people. This will help overcome some of the limitations of previous studies on

Manuscript received on January 30, 2019. (*Corresponding author: Hui Yang.*)

H. Yang and P. A. Bath are in the Information School, the University of Sheffield, Sheffield, S1 4DP, UK (e-mail: hyang@sheffield.ac.uk).

dementia outlined above. Two research questions emerge from these limitations in knowledge. The first is methodological, i.e., to what extent can advanced data analytic methods enhance our understanding of factors associated with dementia? The second aims to inform our understanding of dementia, i.e., from the wide range of possible psychosocial factors associated with mental health problems in later life, which are the most important factors for predicting the presence of dementia?

With these research questions in mind, we first explored a wide variety of potential factors (i.e., over 400 variables) from multiple domains related to dementia using cross-sectional data from a large nationally representative sample, the English Longitudinal Study of Ageing (ELSA) [28]. We then built predictive models using seven different machine learning (ML) algorithms, followed by a model ensemble approach, to merge the results from individual base models to improve the prediction performance further. Finally, the factors that were important in each domain area were identified based on the ranking of feature importance scores in the predictive models. Our aim was to enhance understanding of the underlying psychosocial factors associated with dementia, which could help improve diagnosis and management of the condition.

II. Materials

A. ELSA Data

The English Longitudinal Study of Ageing (ELSA)[1] [28] is a nationally representative study of community dwelling adults aged 50 years and older from the general population of England, UK. The first ELSA survey was initiated in 2002-3 (wave-1), with follow-up waves every two years (waves 2-8). ELSA is an ongoing study that aims to explore relationships between health and disability, social participation, socioeconomic position, and quality of life in older age.

B. Measurement of Dementia (Outcomes)

Dementia was assessed in two ways: (1) a physician's diagnosis of dementia or Alzheimer's disease reported by the participants; (2) an adapted short-form IQCODE questionnaire [14], consisting of 16 items assessing the ability of the person to perform various functions (e.g., remembering things about family/friends) compared with 2 years ago, with ratings ranging from `1(much improved)` to `5 (much worse)`. Those with a score greater than 3.5 were classed as having dementia: this score has high specificity and good sensitivity [23].

C. Independent Variables

A variety of factors from different domains was directly selected or indirectly derived from the items in the ELSA questionnaire based on previous research (recall section I). We grouped these into six categories based on their characteristics.

Demographic & economic factors. Demographic variables included `age`, `sex`, `marital status`, `ethnicity`, `religion`, and `living arrangement`. Several indicators of socioeconomic status (SES) were used, which included `education`, `employment status`, `parental social class`, and `total net non-pension household wealth`. Socio-economic status was based on the eight-group version of the National Statistics Socio-Economic Classification (`NSSEC`).

Social engagement & social network factors. A wide range of social engagement variables referring to the nature of a person's social activities were adopted: 6-item `activities during last month`, 7-item `civic or culture activities`, 8-item `social organization or club`, 13-item `volunteer work`, 10-item `unpaid help`, 4-item `local amenities and services`, 7-item `public/private transport`, and 8-item `looking after` family members or friends. Each social activity type contained multiple questions: a corresponding score was created by summing the associated items.

Measures of social network were separated into three aspects: `social support`, `social contact`, and `social close`. `Social support` for each relationship tie (spouse/partner, children, family, friends) was calculated based on responses to a set of 6 questions. `Social contact` was assessed by frequency of contact with friends, relatives or children with whom they did not live. For each network type, a numerical score was calculated on the basis of a set of 4-item questions with ordinal multiple responses. `Social close` included the number of children, family, and friends with whom the participant had close relationships. A 9-item `neighborhood` variable, indicating the characteristics of the living area, was also included.

Physical health & disability factors. Several self-reported health indicators were included: `long-standing illness`, `limited work due to health`, `long-standing illness is limited`, `general health` and `health during childhood`. The number of co-morbidities, i.e. 7-item `CVD` and 9-item `non-CVD`, was assessed by self-reported doctor-diagnosed chronic diseases. Several `medications-taking` variables for the treatment or prevention of some common diseases (e.g., diabetes, hypertension) in older age were included. Variables known to be associated with sensory impairment, e.g., `eyesight` and `hearing impairment` and variables relevant to body `pain` (e.g., feet, hips, knees, and back) were also included.

Disabilities were assessed based on the participant's responses regarding perceived difficulties in 6 basic activities of daily living (`ADL`) and 7 instrumental ADLs (`IADL`). `Mobility impairment` was measured by 10 common leg and arm functions. Other disability variables included 7-item `mobility aids`, `gait walk`, history of `fall`, and `joint replacement`.

Psychological and mental health factors. `Depression` was measured using the shortened version of the 8-item Centre for Epidemiological Studies-Depression (CES-D) scale [29]. `Loneliness` was measured with the 3-item short form of the Revised UCLA loneliness scale [12]. Sleep quality was assessed by `sleep duration` and `disturbance`. In addition, 5-item sense of `control at home` and 12-item `work demand` were measured as psychosocial mediators. Intellectual activities like `internet` use conclude 6-item digit devices used

[1] http://www.elsa-project.ac.uk

to access the Internet and 13-item Internet activities. Psychological and social well-being was measured with 19 items from a quality of life (`CASP-19`) instrument [13].

Lifestyle factors. Several health behavior variables, e.g., `smoking` and `alcohol` drinking, were included. Variables recording the consumption of `fruit` or `vegetables` were included. Self-reported `physical activity` included questions about the frequency of participation in vigorous, moderate, mild and sedentary physical activities. Sedentary behaviors, such as `TV watching`, were also included.

Cognition factors. Memory was assessed by two cognition tests: 2-item `word recall` and 4-item `time orientation`. Executive function was measured by four cognition tests: `semantic verbal fluency`, 2-item `counting backwards test`, 5-item `number subtraction series`, and 5-item `naming questions`. Basic cognitive skills/abilities were estimated using two tests: 6-item `numeracy test` and 4-item `health literacy` test. Other cognition variables include 2-item `self-rated memory` and 2-item `self-rated mental ability`.

D. *ELSA Variables (Direct) vs. Derived Variables (Indirect)*

In this study, we created a feature set consisting of a total of over 400 variables to build the predictive models for dementia. The variables contained in the feature set were either *directly* extracted from the original ELSA surveys or *indirectly* derived based on ELSA item variables. Further details about the ELSA variables and how the derived variables were created can be found in the supplementary material (`Supplement_Variables.xlsx`).

The derived variables generally can be divided into two groups: one related to categorical variables (e.g., `age -> age group`) created using validated thresholds values of associated item variables in ELSA; the other referred to the composite variables that were calculated based on the summary or average of the scores from associated ELSA item variables, e.g., `8 CES-D scale items -> CES-D based depression*`. An underlying aim of these models was to investigate the usefulness of these derived variables in improving the predictive performance.

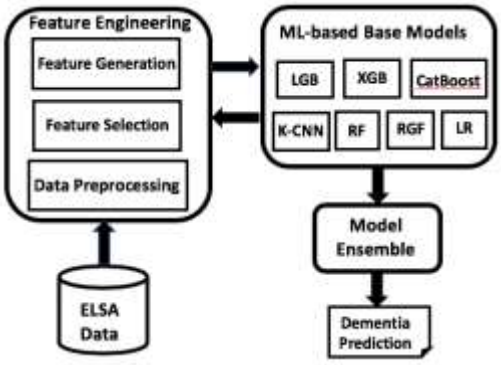

Fig. 1. System framework for the prediction of dementia.

III. Methods

A. *System Framework*

Figure 1 shows the system framework for the prediction of the presence of dementia in older age. First, a set of features was directly extracted, or indirectly derived, from the ELSA dataset, and was used for building the predictive models. In addition, several data pre-processing methods were applied for data preparation, e.g., data resampling techniques targeted for imbalanced data. Second, several ML algorithms were first implemented to construct individual base models based on the generated full features. A subset of important features was then selected based on the ranking of feature importance scores in the predictive models, which was used to further refine the predictive models. Third, a model ensemble approach was proposed to combine the results from different ML-based base models to improve the system performance further. Finally, the final predicted outcomes were generated by the ensemble model. Figure 2 provides the pseudo codes for the description of system workflow on dementia prediction.

```
Input: Given ELSA data:
       Run Feature_generation module
       Run Data_preprocessing module
       Run Building_model module based on full features
       Run Feature_selection module on the predictive models
       Run Rebuilding_model module using selected features
       Run Model_ensemble module based on different ML models
Output: the merged prediction results
```

Fig. 2. Pseudo codes for system workflow on dementia prediction

B. *Machine Learning based Base Models*

In this paper, advanced ML algorithms were implemented for the prediction of the presence of dementia at the same time point. We selected these ML algorithms because they have previously achieved competitive results in binary classification tasks in data science competitions, e.g., the Kaggle challenges[2]. Further details describing the selected ML algorithms are in the supplementary material (`Supplement_material.doc, p.4-6`).

Gradient Boosting Machine (GBM: XGB & LGB & CatBoost). The GBM [10] builds an additive model in a forward stage-wise fashion; it allows for the optimization of arbitrary differentiable loss functions. At each stage a regression tree is fitted on the negative gradient of the given loss function. GBMs allow one to minimize more complicated loss function that cannot be minimized directly. Three GBM algorithms, eXtreme Gradient Boosting (XGB)[3] [6], Light Gradient Boosting Machine (LGB)[4] [17], and CatBoost[5] [22] were separately implemented to build the predictive models.

Keras-based Convolutional Neural Network (K-CNN). Convolutional neural networks (CNNs) [19] are a class of deep, feed-forward artificial neural networks. They use a variation of multilayer perceptrons designed to require minimal pre-processing. We implemented the CNN-based models using the Keras toolkit, a Python-based deep learning library[6]. The predictive models were built by a binary classifier with three fully connected (FC) layers and dropout,

[2] https://www.kaggle.com/competitions
[3] https://xgboost.readthedocs.io/en/latest/index.html#
[4] https://lightgbm.readthedocs.io/en/latest/
[5] https://tech.yandex.com/catboost/
[6] https://keras.io/

and trained using the ADAM algorithm and the advanced PReLU activation (see `Supplement_material.doc, p. 3` for more details).

Random Forest (RF). Random forest [11] is an ensemble method in which a classifier is constructed by combining several different Independent base classifiers. This technique is known as bagging, or bootstrap aggregation. In Random Forest, further randomness is introduced by identifying the best split feature from a random subset of available features.

Regularized Greedy Forests (RGF). RGF [16] is a tree ensemble machine learning method, which works directly with the underlying forest structure. RGF integrates two ideas: one is to include tree-structured regularization into the learning formulation; and the other is to employ the fully-corrective regularized greedy algorithm. Here the RGF-based model was carried out using the toolkit Scikit-learn Wrapper for Regularized Greedy Forest (`rgf_python 3.4.0`)[7].

Logistic Regression (LR). Logistic regression here was used as the baseline algorithm for performance comparison. In this study, both random forest and logistic regression algorithms were implemented using the scikit learning Python library[8].

Parameter tuning. In our predictive models, a hyperparameter optimization method, called the grid search technique, was adopted to determine the optimal parameters for individual ML algorithms. Further details about the optimal parameter setting of the different algorithms can be found in the supplement (`Supplement_material.doc, p. 2`).

C. Model Ensemble

Model ensemble is a process of running two or more related but different analytical models and then synthesizing the results into a single score to improve the accuracy of predictive analytics. Generally, the ensembled model (also called 2nd-level model) will outperform each of the individual models due to its smoothing nature and ability to highlight each base model where it performs best and discredit each base model where it performs poorly.

In this study, an ensemble model was implemented using a simple weighted linear model which was calculated as below:

$$y = \sum_{i=1}^{n} w_i x_i \; ; \;\; \sum_{i=1}^{n} w_i = 1$$

where x_i is the prediction results from the i base model, and the w_i is the weight assigned to the i base model. The sum of all the weights will be equal to 1. Base models with better performance will have a higher score. The ensemble model merged the results from different ML-based base models at the previous stage to improve the model performance further.

D. Dealing with Imbalanced Data

In this study, the ELSA data used for the prediction of dementia was an imbalanced data set in which the prevalence of dementia in the different ELSA waves ranged between 1-1.5%. Two commonly-used approaches were adopted to deal with these imbalanced data: (1) at the data level, two resampling techniques, SMOTE and ADASYN, were used to achieve a balanced distribution by under-sampling the majority class or over-sampling the minority class. In this study, different resampling techniques were implemented using a Python-based package[9] (`imbalanced-learning 0.4.3`). (2) At the algorithm level, built-in parameters specific to handling imbalanced data were set for their effectiveness in individual ML algorithms, for example, the parameter (`scale_pos_weight`) for the XGB, LGB, and CatBoost and the parameter (`class_weight`) for the K-CNN, RF, and LR.

IV. Experiments

A. Training and Test Data

The cross-sectional data used in this study consists of training and test datasets. 9,666 instances in ELSA wave-7 (2014-15) were used as the training data, which had 142 (1.46%) cases of dementia. 8,445 instances from ELSA wave-8 (2016-17) were used as the test data with 109 (1.29%) dementia cases. At the training stage, K-fold cross-validation was used for the training and validation of the models built for the prediction of dementia.

K-fold cross-validation. In K-fold cross-validation, the original sample is randomly partitioned into k equal sized subsamples. Of the k subsamples, a single subsample is retained as the validation data for testing the model, and the remaining *k-1* subsamples are used as training data. The cross-validation process is then repeated k times (the folds), with each of the k subsamples used exactly once as the validation data. Here, stratified 10-fold cross-validation was applied to ensure that each fold had the same proportion of dementia observations.

B. Experiment Metric

The prediction of dementia can be treated as a binary classification task in which the output is labelled as 1 (dementia), or 0 (no dementia). Given a new instance, the classifier will assign a probability of having dementia to the instance rather than simply yielding the most likely class label. The predicted probability will fall in the range of [0, 1]. A higher predicted probability means that a participant is more likely to experience the dementia symptoms. In this study, a commonly-used evaluation metric, the Normalised Gini Coefficient, was applied to estimate model performance.

Normalised Gini Coefficient (Gini). During scoring, observations are sorted from the largest to the smallest predictions. Predictions are only used for ordering observations; therefore, the relative magnitude of the predictions are not used during scoring. The scoring algorithm then compares the cumulative proportion of positive class observations to a theoretical uniform proportion. The Gini that is used to normalize the AUC (Area Under Curve) is calculated as 2*AUC-1 so that a *random* classifier scores 0, and a *perfect* classifier scores 1. The *higher* the Gini score, the *better* the performance.

[7] https://pypi.org/project/rgf_python/
[8] http://scikit-learn.org/stable/
[9] https://pypi.org/project/imbalanced-learn/

V. Results

A. Performance Comparison in Different Feature Sets

To compare the usefulness of different types of features on predicting the presence of dementia, three feature sets were separately constructed:

- *Feature-set-I*: full feature set without derived variables
- *Feature-set-II*: full feature set without cognition variables
- *Feature-set-III*: full feature set

The usefulness of the derived variables. As discussed above, some features used for dementia prediction were generated for these analyses, and they were derived from individual ELSA item variables. By comparing the results between *feature-set-I* and *feature-set-III* in Table I, the derived variables showed some improvement in the prediction of dementia with a small increase in Gini scores, ranging from 0.3-1.4% using the different ML algorithms on the test data.

TABLE I
System Performance (Gini Scores) of Different Feature Sets

	Feature set I (validate/test)	Feature set II (validate/test)	Feature set III (validate/test)
XGB	0.865/0.899	0.719/0.858	0.909/0.913
LGB	0.897/0.897	0.879/0.860	0.888/0.904
CatBoost	0.904/0.888	0.834/0.858	0.872/0.891
K-CNN	0.899/0.896	0.853/0.874	0.919/0.907
RF	0.937/0.912	0.908/0.872	0.946/0.918
RGF	0.861/0.904	0.769/0.853	0.887/0.911
LR	0.926/0.862	0.863/0.831	0.917/0.868

The usefulness of cognition variables. A set of cognition variables created by a series of cognition tests was utilized when building the predictive models. To investigate the importance of these cognition variables for the recognition of dementia, another set of experiments (*feature-set-II*) was conducted by removing the cognitive features from the full features (*feature-set-III*). As shown in Table I, the performance in each algorithm generally deteriorated without the presence of the cognition features, and the corresponding Gini score generally dropped between 3-6%.

B. Top-ranking Factors for Dementia Prediction

Initially, over 400 variables were used during the training of the predictive models. However, not all variables are useful for predicting dementia, therefore, it was vital to identify those features that were of the greatest importance. Several selected ML algorithms have a built-in function that could assign a feature importance score to each independent variable. Due to space limitations, Table II provides only the top-40 ranked features from the XGB, LGB, and CatBoost based models.

The feature selection step based on the top-ranking features from the different algorithms was carried out to further refine the predictive models. This assumes that each ML algorithm captures different characteristics of dementia symptoms, thereby generating different top-ranking feature lists as indicated in Table II. Several subsets of top-ranking features were created below, each of which combined the top-N features in the three ML algorithms, i.e., XGB, LGB, and CatBoost.

- *Topset-1*: top-10 ranking features (XGB+LGB+CatBoost)
- *Topset-2*: top-20 ranking features (XGB+LGB+CatBoost)
- *Topset-3*: top-30 ranking features (XGB+LGB+CatBoost)
- *Topset-4*: top-40 ranking features (XGB+LGB+CatBoost)
- *Topset-5*: top-50 ranking features (XGB+LGB+CatBoost)

Table III gives model performance of different top-ranking feature sets on the test data. For each ML algorithm, the refined predictive models that were retrained by different top-ranking feature sets generally achieved better performance than the models with all features (recall Table I, *feature-set-III*).

TABLE II.
Top-40 Ranking Factors in Terms of Different Machine Learning Algorithms

rnk	XGB	LGB	CatBoost
1	**Age**	**Social close: friends**	**Activity last month: paid work**
	Cognition: word recall*	**Cognition: word recall***	**Cognition: word recall***
	Disability: IADL*	**Self-rated memory**	**Disability: IADL***
	Self-rated memory	**Cognition: verbal fluency**	**Cognition: time orientation (day)**
	No. of people living with	**Disability: IADL***	**Cognition: word recall (delay)**
	Cognition: verbal fluency	**Self-rated mental ability**	**Disability: ADL & IADL***
	Cognition: time orientation	**Disability: ADL & IADL***	**Cognition: naming questions***
	Self-rated mental ability	Food: fruit portions	**Cognition: time orientation**
	Disability & Mobility impair*	**Food: veg portions**	Cognition: naming (president)
10	Hearing: problem in bkgrd. noise	**Age**	**Limited work due to health**
	Social support: family*	**No. of people living with**	Working demand: physical demand
	Social close: family*	Food: fruit & veg*	Working demand: new skill
	Social contact: all*	**Social support: all***	Working demand*
	CES-D based depression*	Disability: Mobility impair*	Cognition: time orientat. (month)
	Pain: walking on foot	**Social contact: family***	working demand: work fast
	Disability: ADL & IADL*	**Social contact: all***	Cognition: word recall (immedia.)
	Depression: feel lonely	Social contact: family(phone)	Working demand: salary adequate
	Cognition: time orientat. (day)	**Disability & Mobility impair***	Working demand: approval by work
	Gait walk	CASP-19: look fwd. to each day	Working demand: job security
20	**Cognition: word recall (delay)**	**Social support: children***	**Disability & Mobility impair***
	Activity last month: paid work	**Self-reported hearing**	IADL: manage money
	Transport: transport tools	**Cognition: time orientat. (day)**	**Cognition: number subtraction***
	Self-reported eyesight	**Cognition: time orientation**	**Self-rated mental ability**
	Self-rated long-standing ill.	Transport: car driving (past)	Working demand: support received
	Social support: child (let down)	Social contact: family(message)	Working demand: freedom to decide
	Marital status	Social support: partner (nerve)	**Self-rated memory**
	Self-reported hearing	**CASP-19 quality of life***	IADL: read map
	Food: veg portions	**Cognition: word recall (delay)**	Cognition: time orientat. (year)

	Limited work due to health	**Internet: use frequency**	**Activities during last month***
30	Pain: other body part	CASP-19: do the things they want	IADL: work in the house/garden
	Cognition: naming questions*	**Cognition: number subtraction***	Cognition: time orien. (day of M)
	Social contact: family*	Social support: partner*	Cognition: naming (desert plant)
	Smoking: no. of cigar weekday	**Activities during last month***	Unpaid work: decorate home
	IADL: difficult in taking drug	Internet: used application*	CASP-19: out of control
	Social contact: friend by meet	Self-rated general illness	Cognition: naming (prime minster)
	Transport: use of car as driver	**Limited work due to health**	**Internet: use frequency**
	CASP-19 quality of life*	**Gait walk**	Volunteer work: committee member
	Social support: children*	Disease: non-CVD disease*	volunteer work: befriend people
	Social close: friends	Father (died of diabetes)	**Social support: child (let down)**
40	**Social support: all***	Retired with pension	Internet: device (smartphone)

NOTE: THE DERIVED VARIABLES ARE MARKED WITH THE ASTERISK (*); VARIABLES IN **BOLD** ARE PRESENT IN TWO OR MORE OF THE TOP-40 RANKED FACTORS.

Table III gives model performance of different top-ranking feature sets on the test data. It is noted that for each ML algorithm, the refined predictive models that were retrained by different top-ranking feature sets generally achieved better performance than the models with full features (recall Table I, *feature-set-III*). This suggests that feature selection is a useful step for performance improvement.

It can be seen that the best performance for individual ML algorithms was obtained on different top-ranking feature sets as shown in Table III. In summary, the optimal importance features ranged between 20-64 top-ranking features. After that, the performance of the predictive models deteriorated to some extent when further top-ranking features of less importance were added to the system.

TABLE III
SYSTEM PERFORMANCE (GINI SCORES) OF DIFFERENT TOP SETS ON THE TEST DATA

	Topset-1 (20 feat.)	Topset-2 (43 feat.)	Topset-3 (64 feat.)	Topset-4 (83 feat.)	Topset-5 (99 feat.)
XGB	0.913	0.918	**0.923**	0.919	0.911
LGB	**0.915**	0.914	0.913	0.910	0.905
CatBoost	0.910	0.917	**0.921**	0.914	0.909
K-CNN	0.926	**0.930**	0.924	0.922	0.912
RF	0.921	**0.929**	0.928	0.923	0.925
RGF	0.923	**0.927**	0.911	0.914	0.914
LR	**0.906**	0.893	0.883	0.879	0.879

Note: the best performance in each algorithm is highlighted in **bold**

C. Model Ensemble

In this paper, for each ML algorithm, we used the best performance achieved on the test data at the feature selection stage (recall Table III) as the prediction result before conducting model ensemble. Table IV provides the Gini scores on the test data in terms of individual base models, and Figure 3 indicates the correlation matrix between different ML predictions using Pearson's test. It is noted that the XGB and RF are highly correlated (0.9574) whereas the RGF and LGB have the lowest correlation (0.4677).

TABLE IV
THE BEST GINI SCORES OF INDIVIDUAL ML ALGORITHMS ON THE TEST DATA

XGB	LGB	CATBOOST	K-CNN	RF	RGF	LR
0.9234	0.9153	0.9218	0.9307	0.9295	0.9276	0.9069

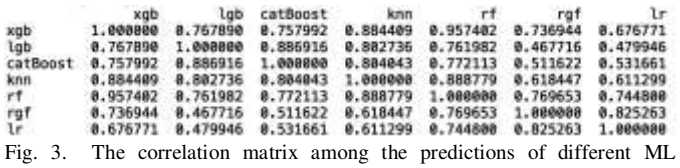

Fig. 3. The correlation matrix among the predictions of different ML algorithms in the word recall test

Here, a simple weighted linear model was applied to combine the results from different base models, and several ensemble strategies were applied as below:

- *Ensemble-1* (**E1**): K-CNN + RF + RGF
- *Ensemble-2* (**E2**): K-CNN + RF + XGB
- *Ensemble-3* (**E3**): K-CNN + RGF + XGB
- *Ensemble-4* (**E4**): K-CNN + RGF + CatBoost
- *Ensemble-5* (**E5**): K-CNN + RF + RGF + CatBoost
- *Ensemble-6* (**E6**): K-CNN + RF + RGF + XGB

TABLE V
SYSTEM PERFORMANCE ON THE TEST DATA USING DIFFERENT ENSEMBLE STRATEGIES

E1	E2	E3	E4	E5	E6
0.9332	0.9331	0.9325	0.9322	0.9332	0.9333

Table V presents the system performance on the test data using different ensemble strategies. It is clear that the results of the ensemble models were generally better than those of individual base models (recall Table IV) as the Gini score increased slightly by a range of 0.2-0.3%. The best result, with a Gini score of 0.9333, was achieved when merging the results from four different models, the K-CNN, RF, RGF and XGB.

VI. DISCUSSION

This study sought to answer two separate, but related, research questions. The first question was to what extent can advanced data analytic methods enhance our understanding of factors associated with dementia? An ensemble model that combined the results from several ML algorithms was implemented for predicting the presence of dementia in older adults and helped to improve the performance of the individual predictive models. The second research question was which are the most important factors for predicting the presence of dementia? The study utilized the ELSA data in which a large number of factors from six domains were considered and significant factors that were associated with dementia were identified. The findings are discussed below.

Important factors from the different domains. A wide variety of factors (over 400 variables) from multiple domains were initially explored for predicting dementia. A number of key factors in individual domains relevant to dementia symptoms were identified and analyzed in detail (see the file: `Supplement_material.doc, p.7-10`). The evidence from our study indicates that the top-ranking factor lists identified by the different ML algorithms had a relatively high level of overlap, but they differed in the ranked order (recall Table II).

When the larger number of factors were analysed together, some of the factors, such as social factors (e.g., `social support`, `contact` and `close`), disability factors (e.g., `ADL`, `IADL` and `mobility impair`), lifestyle factors (e.g, `Food: fruit & veg` and `smoking`), and several cognition tests (e.g., `Self-rated memory`, `word recall`, `verbal fluency` and `time`

orientation), were identified as being significantly associated with the presence of dementia: this accords with findings of other studies [20, 33, 34]. Nevertheless, some factors (e.g., cardiovascular diseases, alcohol and physical activity) became less important or even lost their importance, in contrast with other studies [15, 26]. In addition, some new factors (e.g., transport tools, CASP-19 quality of life and Working demand) were identified for the first time in our study as being closely associated with dementia.

Data collection and clinical assessment of older people for dementia are both costly and time-consuming. The findings regarding factors closely associated with dementia presented here could help improve the effectiveness and efficiency of collecting information relevant to dementia in older adults, both in future research and in diagnosing dementia, as discussed below. Such information will also be useful for health practitioners to develop more effective intervention strategies for managing dementia.

Dealing with Imbalanced Data. In this paper, two approaches were applied to mitigate the disadvantages of imbalanced data used for predicting the presence of dementia. The experimental results from our study showed that the resampling approach at the data level increased the Gini score by a range of 0.5-0.8% in the four algorithms, the XGB, CatBoost, RGF, and RF. In contrast, parameter setting at the algorithm level worked well only with the K-CNN, with an elevated Gini score of about 2%.

However, this methodological issue may also help tackle clinical problems in identifying people with early-stage dementia. The problem of imbalanced data, i.e., there being a low proportion of cases with dementia, reflects the real world, where a General Practitioner, family doctor or physician may only have a small proportion of their patients with a condition. Identifying these patients among all of the others is challenging, and testing all patients is not practical or cost-efficient, especially if genetic test or neuroimaging are required. Our findings identified psychosocial and other factors associated with the presence of dementia, e.g., self-rated memory (loss) (recall Table II): this information may be useful for clinicians in identifying individual patients who are in the early stages of dementia, and may need further testing, without the need for screening large numbers of patients. Furthermore, identifying patients with dementia at an earlier stage in the disease may help to develop interventions to manage the condition, and support the patient and their family.

Feature Selection. Here, feature selection based on top-ranking features was shown to be an important step in refining the predictive models for improving performance. The selection of the top-ranking features was the combination of important factors identified by several ML algorithms, each of which might capture different characteristics associated with dementia symptoms. In this study, ML algorithms generally performed best within the range of the 20-64 top-ranking features.

Model Ensemble. The model ensemble method did not contribute greatly to enhancing prediction capability: the Gini scores only had a slight improvement of 0.2-0.3 percent using a simply linear-weighted model. Our experimental results also showed that the model stacking technique implemented by a random forest model even deteriorated the overall system performance to some extent, when merging the prediction results from multiple base models.

ML algorithms and techniques used in dementia. Our study differs from other dementia studies using ML techniques and their application in two ways: (1) the ELSA data studied here focus on the information related to demographic, economic, psychosocial, health, and cognitive measures in older age, and the majority of influencing factors are categorical or binary. However, most previous ML applications for dementia recognition [1-2] were built based on biological, clinical or neuroimaging data using numerical data. Such differences in the data types were taken into account in our system design. (2) new ML algorithms, such as GMB-based algorithms (e.g., XGB, LGB, and CatBoost) and RGF, were first used in our prediction system, and performed competitively (recall Table IV) compared with algorithms (e.g., CNN-based deep learning and RF) used in previous research [1-2]. These new ML algorithms were important in improving performance in our ensemble models (recall Table V).

It is worth remarking that the K-CNN performed best among all the ML algorithms at the feature selection stage (recall Table IV). In deep-learning based image recognition, data augmentation methods that help enrich the existing data by adding new instances with translational invariance have been shown to be useful in improving prediction accuracy. In this paper, the up-sampling techniques (i.e., SMOTE, ADASYN), which produce more new minority class rows using the K nearest neighbor technique can be considered as one of the data augmentation methods. However, this method worked well only on decision-based algorithms, such as the XGB, CatBoost, and RGF, but not for the K-CNN. In future work, we plan to explore other data augmentation techniques in the K-CNN models.

In this study, only a simple 3-layer CNN architecture was utilized in the predictive model (see Supplement_material.doc, p. 3 for more details). We also tried other deep learning networks, e.g., CNN with more hidden layers and various activation functions (e.g., sigmoid, tanh, linear, relu, and PReLU) and optimizers (e.g., SGD, RMSprop, Adagrad, and Adam), and Recurrent Neural Network (RNN) with multiple embedding layers. Unfortunately, they did not work satisfactorily on this dataset. Future research could investigate more complicated deep learning networks for dementia detection. For instance, how to adapt successful deep learning networks on image recognition [30-32] utilizing the characteristics of the ELSA data, and then integrate them into our current system.

Error Analysis. For the test data, which included 109 dementia cases, the '*perfect*' prediction on dementia should be the scenario in which the top-109 instances with high dementia probability score are the dementia instances in the ground truth. However, in the actual prediction, 34 false negatives (dementia instances that were not in the top-109 ranking) and 34 false positives (non-dementia instances within the top-109 ranking) were examined manually, and the possible causes for misprediction can be summarized as: (1) the overall IQCODE score (>3.5) was used as one of the criteria for dementia judgment. It was found that a number of the false negatives and positives had an IQCODE score either

side of this value, i.e., that ranged between 3.2-3.8, which resulted in judgement ambiguity for the predictive models; (2) some mis-predicted cases had inconsistent probability scores from different ML algorithms, which increased the prediction errors in the model ensemble. (3) a number of dementia instances lacked distinguishing characteristics associated with disease symptoms, and were thus incorrectly labelled with low probability scores by the predictive models.

Study limitations and future research. A number of limitations in our study need to be considered. First, dementia is inherently highly complex, and may be affected by a diverse range of biological, psychosocial, clinical, and neurological factors. The ELSA data used in our study examined demographic, economic, social, psychological/cognitive health/disability and lifestyle variables, in relation to dementia. While this constitutes a wide range of domains, to more comprehensively understand the etiology of dementia and further improve predictive accuracy, future research could also include predictive biomarkers from other sources such as genetic, clinical assessments, and brain imaging data [1-2, 30-32].

Secondly, the potential for confounding, i.e., the presence of hidden variables that affect both dependent and independent variables, within data from cohort studies such as ELSA, requires consideration. In our study, we included demographic variables, e.g., age, gender, education, socioeconomic class, etc., which will have reduced the potential for confounding, and indeed age was identified as an important variable in the top-40 ranked factors (recall Table II).

Thirdly, the analyses of the ELSA data reported here were cross-sectional, and restricted to each of two single waves of the survey. Therefore, the factors identified in the predictive models were associated with the presence of dementia at a given time point, rather than risk factors for developing dementia over time. Future longitudinal research could develop models to utilize information from multiple consecutive waves for prognosis (e.g., gradual deterioration of dementia over time) and diagnosis (e.g., mild cognitive impairment developing into dementia) of dementia.

Finally, while ELSA is a nationally representative sample, and the findings are generalizable to older people in England, the factors associated with dementia might vary to some extent in individual countries due to differences in geographic location, environment, culture and behaviors. Future research could compare risk factors across countries, using harmonized data for cross-national comparisons [3].

VII. Conclusion

Dementia in later life has emerged as a significant public health challenge in recent years with continuing increases in life expectancy. Strategies for preventing or alleviating dementia symptoms in older people are much needed. In this study, a wide variety of factors (over 400 variables) from multiple domains, from a large nationally representative sample of older people (ELSA), were explored to identify characteristics associated with dementia in older adults. Seven machine-learning algorithms were used to build the predictive models for performance comparison. A simple model ensemble model was developed to merge the results from individual base models to further improve prediction performance. A range of factors associated with dementia was identified in the models. These provide important insights into possible risk factors for dementia and symptoms of dementia. This information may help clinicians and public health detect early stage dementia and develop interventions to support people with dementia.

References

[1] M. R. Ahmed, Y. Zhang, Z. Feng, B. Lo, O. T. Inan, and H. Liao, "Neuroimaging and Machine Learning for Dementia Diagnosis: Recent Advancements and Future Prospects," *IEEE Rev. Biomed. Eng*., vol.12, pp. 19-33, 2019.

[2] S. G. Alonso, *et al*., "Data Mining Algorithms and Techniques in Mental Health: A Systematic Review," *J. Med. Syst*., vol. 42, no. 9, pp. 161, 2018.

[3] P. A. Bath, D. Deeg and J. Poppelaars, "The harmonisation of longitudinal data: a case study using data from cohort studies in The Netherlands and the United Kingdom," *Ageing and Society*, vol. 30, no. 8, pp.1419-1437, 2010.

[4] F. Bunn, *et al*., "Comorbidity and dementia: a scoping review of the literature," *BMC Med*., vol. 12, pp. 192, 2014.

[5] A. L. Byers, and K. Yaffe, "Depression and risk of developing dementia," *Nat. Rev. Neurol*., vol. 7, no. 6, pp. 323–331, 2011.

[6] T. Chen and T. He, "XGBoost: a scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.*, 2016, pp.785-794.

[7] C. Cheng, *et al*., "Alcohol-related dementia: a systemic review of epidemiological studies," *Psychosomatics*, vol. 58, no. 4, pp. 331-342, 2017.

[8] G. Cipriani, *et al*., "Sleep disturbances and dementia," *Psychogeriatrics*, vol. 15, no. 1, pp. 65-74, 2015.

[9] H. R. Davies, *et al*., "Hearing impairment and incident dementia: findings from the English Longitudinal Study of Ageing," *J. Am. Geriatr. Soc*., vol. 65, no. 9, pp. 2074-2081, 2017.

[10] J. Friedman, "Greedy function approximation: a gradient boosting machine," *Ann. Stat.*, vol. 29, no. 5, pp. 1189-1232, 2001.

[11] T. K. Ho, "Random decision forests," In *proc. 3rd Int. Conf. Doc. Anal. Recog.*, 1995, pp. 278–282.

[12] M. E. Hughes, *et al*., "A short scale for measuring loneliness in large surveys: Results from two population-based studies," *Res. Aging*, vol. 26, no. 6, pp. 655-672, 2004.

[13] M. Hyde, *et al*., "A measure of quality of life in early old age: the theory, development and properties of a need satisfaction model (CASP-19)," *Aging Ment. Health*, vol. 7, pp. 186-194, 2003.

[14] A. Jorm, "A short form of the informant questionnaire on cognitive decline in the elderly (IQCODE): development and cross-validation," *Psychol. Med.*, vol. 24, pp. 145-153, 1994.

[15] C. B. Guure, *et al*., "Impact of physical activity on cognitive decline, dementia, and its subtypes: meta-analysis of prospective studies," *Biomed. Res. Int.,* vol. 2017, pp. 9016924, 2017.

[16] R. Johnson and T. Zhang, "Learning Nonlinear Functions Using Regularized Greedy Forest,", *IEEE Trans. Pattern Anal. Mach. Intell.* vol. 36, no. 5, pp. 942-954, 2014.

[17] G. Ke, *et al*., "Lightgbm: a highly efficient gradient Q4 boosting decision tree," *Adv. Neural Inf. Process. Syst*., vol. 30, pp.3149–3157, 2017.

[18] M. Khondoker, *et al*., "Positive and negative experiences of social support and risk of dementia in later life: an investigation using the English Longitudinal Study of Ageing,", *J. Alzheimers Dis*., vol. 58, no. 1, pp. 99-108, 2017.

[19] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436-444, 2015.

[20] S. D. Petersson and E. Philippou, "Mediterranean diet, cognitive function, and dementia: a systematic review of the evidence," *Adv Nutr.*, vol. 7, no. 5, pp. 889–904, Sep. 2016.

[21] M. Prince, *et al*., "World Alzheimer Report 2016: Improving healthcare for people living with dementia," In *Alzheimer's Dis. Int*., 2016.

[22] L. Prokhorenkova, *et al*., "CatBoost: unbiased boosting with categorical features," *Adv. Neural Inf. Process. Syst*., vol. 31, pp. 6637-6647, 2018.

[23] T. J. Quinn, *et al*., "Informant Questionnaire on Cognitive Decline in the Elderly (IQCODE) for the diagnosis of dementia within community dwelling populations," *Cochrane Database Syst. Rev*., vol. 10, pp. CD010079, 2014.

[24] S. B. Rafnsson, *et al*., "Loneliness, social integration, and incident dementia over 6 years: prospective findings from the English Longitudinal Study of Ageing," *J. Gerontol. B. Psychol. Sci. Soc. Sci*., Jun. 2017. doi: 10.1093/geronb/gbx087.

[25] L. Robinson, E. Tang, and J. P. Taylor, "Dementia: timely diagnosis and early intervention," *BMJ*., vol. 16, no. 350, pp. h3029, Jun. 2015.

[26] M. B. Snowden, *et al*., "Dementia and co-occurring chronic conditions: a systematic literature review to identify what is known and where are the gaps in the evidence?" *Int. J. Geriatr. Psychiatry*., vol. 32, no. 4, pp. 357–371, Apr. 2017.

[27] O. Spierer, *et al*., "Correlation between vision and cognitive function in the elderly - a cross-sectional study," *Medicine (Baltimore)*, vol. 95, no. 3, pp. e2423, Jan. 2016.

[28] A. Steptoe, *et al*., "Cohort profile: the English Longitudinal Study of Ageing," *Int. J. Epidemiol*, vol. 42, pp. 1640-1648, 2013.

[29] S. van de Velde, K. Levecque, and P. Bracke, "Measurement equivalence of the CES-D 8 in the general population in Belgium: a gender perspective," *Arch. Public Health*, vol. 67, pp. 15–29, 2009.

[30] W. Wang, J. Shen, and L. Shao, "Video Salient Object Detection via Fully Convolutional Networks," *IEEE Trans. Image Process*., vol. 27, no. 1, pp. 38-49, 2018.

[31] W. Wang and J. Shen, "Deep Visual Attention Prediction," *IEEE Trans. Image Process*., vol. 27, no. 5, pp. 2368-2378, 2018.

[32] W. Wang, J. Shen, and H. Ling, "A Deep Network Solution for Attention and Aesthetics Aware Photo Cropping," *IEEE Trans. Pattern Anal. Mach. Intell*., 2018. doi: 10.1109/TPAMI.2018.2840724.

[33] G. Zhong, *et al*., "Smoking is associated with an increased risk of dementia: a meta-analysis of prospective cohort studies with investigation of potential effect modifiers," *PLoS One*, vol. 10, no. 3, pp. e0118333, 2015.

[34] Z. Zhou, P. Wang, and Y. Fang, "Social Engagement and Its Change are Associated with Dementia Risk among Chinese older adults: a longitudinal study," *Sci Rep*., vol.8, pp. 1551, 2018.