Predicting risk of dementia with machine learning and survival models using routine primary care records

John Langham*, Daniel Stamate*,
Charlotte A. Wu, Fionn Murtagh
Data Science & Soft Computing Lab, and
Department of Computing
Goldsmiths, University of London
London, UK
j.langham@gold.ac.uk,
d.stamate@gold.ac.uk
*Joint-first authors and contact authors

Catharine Morgan, David Reeves,
Darren Ashcroft, Evan Kontopantelis,
Brian McMillan
NIHR School for Primary Care Research
Division of Population Health
Health Services Research & Primary Care
School of Health Sciences
The University of Manchester
Manchester, UK

***Abstract*—Worldwide, it is forecasted that 131.5 million people will suffer from dementia by 2050, and the annual cost of care will increase from 818 billion USD in 2016 to 2 trillion USD by 2030, with burgeoning social consequences. Given a timely prediction of a dementia outcome in patients, appropriate mitigating interventions can be applied to reduce risk. However such prediction facilities need to be made available to wider populations, and these facilities cannot rely on specialised, costly and invasive testing (such as neuroimaging, cerebrospinal fluid collection, etc which constitute important instruments used in diagnosis), for interventions to have a meaningful quantitative impact. Hence an emerging need exists for the wider application of prognostic measures which can be deployed using lower cost data sources such as longitudinal records routinely collected by general practices. This paper proposes an efficient prediction modelling approach to the risk of dementia, using CPRD data collected from GP practices in UK, and based on machine learning in particular the Gradient Boosting Machines model combined with a survival model such as the Cox Proportional Hazard, encapsulated in a semi-supervised learning and model calibration methodology.**

***Index Terms*—dementia risk, CPRD, primary care, prediction modelling, machine learning, classification, gradient boosting machines, Cox proportional hazards, model calibration**

I. Introduction

In 2016 there were 47 million estimated dementia sufferers worldwide, with a forecasted increase to 131.5 million by 2050 [18]. The cost of dementia estimated to 818 billion USD in 2016 is expected to increase to 2 trillion USD by 2030 [19]. For comparison, dementia has currently a health and social care cost higher than cancer, stroke and chronic heart disease, taken together.

Dementia diagnosis is problematic because there is currently no standardized dementia test. Moreover, the diagnosing procedure is a highly specific task based on the different sub-types of dementia [1]. Additionally, those who are diagnosed with mild cognitive impairment (MCI), while having a substantially higher risk of developing dementia [2], can either become cognitively stable or return to a healthy cognitive state [3]. Such complexities have contributed to a large proportion of people with dementia to go undiagnosed. The prevalence of undetected dementia is high globally, according to [17] which, based on reviewing 23 selected studies, concluded that the pooled rate of undetected dementia was 61.7% (95% CI 55.0% to 68.0%). Moreover, even with a successful dementia diagnosis, there is currently no cure [1]. As current thinking suggests that about a third of dementia cases could be prevented, the development of effective methods is crucial for early at-risk identification and proactive interventions [4], [18].

Machine learning implementation in the healthcare sector can provide an efficient means of using complex information to accurately predict diagnoses. The size and convolution of DNA sequences have been increasing in recent years; however, supervised machine learning methods, like a Bayesian Hidden Markov model, have been used to interpret DNA sequences for cancer prediction [5]. With the goal of moving away from a 'one-size fits all' approach to dementia prediction, the inclusion of machine learning methods enables the utilization of various data sources and predictive variables [9]. There are hundreds of possible predictors, but they can generally be categorized based on the following applicable models: neuropsychological based models, health-based models, multifactorial models and genetic risk scores [9]. Such models have been applied in various ways in relation to predicting dementia. The use of magnetic resonance imaging, in combination with multiplex neural networks, has been used to discriminate healthy from progressive mild cognitive impairment patients (pMCI), based on the structural atrophy of the brain because of Alzheimer's [6]. General practice patient records in UK have been used to develop a risk score model for estimating how at risk an individual may be of developing dementia [22]. Genetic markers have been used to create a polygenic hazard score test

whose results indicate how likely a person is to developing the Alzheimer's type of dementia over the course of the study [7]. Positron emission tomography scans and the regional analysis of the protein amyloid- β, have been used by a Random Forest classifier to identify patients with age-related stable mild MCI and pMCI [8]. Blood metabolites measurements data samples can be used to predict Alzheimer's dementia with powerful predictive models such as XGBoost, at least as well as with using the well-established but much more invasive to measure biomarkers based on the cerebrospinal fluid (CSF) [16].

Risk scores may be a very useful tool that can enable primary care facilitators to have an estimation of how at-risk an individual is of developing dementia. Once an individual is identified with a large enough probability of developing dementia, then proactive lifestyle interventions can be adopted to curb the development of this disease. Proven methods, such as those described in the Finnish Geriatric Intervention Study to Prevent Cognitive Impairment and Disability (FINGER), have illustrated how multidomain interventions (diet, exercise, vascular monitoring and cognitive training) could improve primary and secondary cognitive performance with age and reduce the risk of dementia [10]. This can be further extrapolated on with the advent of a mobile application that has users provide their information and yields the resulting risk score with intervention suggestions [11]. The focus on risk score methodology development has resulted in a widespread effort with more than 50 different dementia risk scores in 2010 [12]. However, a systematic review of these models has concluded that no single method could be recommended for a generalized screening procedure due to methodological weaknesses of the existing studies, concerning in particular population biases and lack of external validation [13]. This has spurred an emphasis on analysing existing research to inform future work's feature selection, for instance, the inclusion of predictors such as depression, anxiety, cognitive symptoms and others that are positively associated with dementia [14]. A significant part of such predictors can be retrieved from routine primary care data records.

In this paper we used primary care data records of patients in UK, available from the Clinical Practice Research Datalink (CPRD) [39], to propose an efficient prediction modelling approach to the risk of dementia based on machine learning, in particular the Gradient Boosting Machines model, combined with a survival model such as the Cox Proportional Hazard. In our framework these models are encapsulated in a semi-supervised learning procedure to make partial use also of a fraction of our dataset which is unlabeled with a diagnosis due mostly to censoring. We predict the risk of dementia occuring within the 5 years following the year an individual formally entered the study (their "index year"), and being between 60 and 79 years of age.

The main lines of developing our prediction modelling approach are:

- Working with noisy, highly dimensional data extracted from CPRD [39], where the proportion of dementia observations in uncensored data (patients who received a diagnosis) is 1.5%.
- Using Cox Proportional Hazards [24] (Cox PH) and Gradient Boosting Machines (GBM) [25], to obtain comparative results for the dementia risk prediction. Since Cox PH is a much lesser computationally expensive method when applied and compared to GBM on the large CPRD data sample used in this study, comprising 1,126,079 rows and 595 variables in its pre-processed form, we use the most important variables as indicated by the Cox PH cheaper model to select features and simplify the dataset before the application of the GBM method.
- Making use of (right) censored data (as patients who left the study prematurely and hence who have an undefined dementia outcome - comprising about 42.78% of the data) in non-linear models to develop a semi-supervised prediction modelling approach. We augment the training set, over a number of iterations, with an optimum subset of the censored observations which have been classified by GBM, in the semi-supervised approach.

In the remainder of this paper, Section II discusses the methodology that we propose, including the description of the CPRD dataset and its preprocessing for the analysis, and our prediction modelling approach based on Gradient Boosting Machines, Cox Proportional Hazards and semi-supervised learning. In Section III we present our prediction modelling approach's results, covering also the predictor importance, and the resulting model calibration. Section IV concludes the paper.

II. Data and Methodology

A. Clinical Practice Research Datalink (CPRD)

In this research we have used patient records from the Clinical Practice Research Datalink CPRD [39], which provides extracts of anonymised longitudinal patient data collected from GP practices in the UK, encompassing 60 million patients. Through CPRD, primary care information can be further enriched and extended through linkages with other patient data resources such as hospital records databases, but this latter resource was not available nor intended to be used in this study, as the latter focuses on medical data widely available on the general population to be used to inform the computation of the risk of dementia.

B. Description of the CPRD dataset

The dataset used in this study comprises 13,545,937 rows representing between 6 and 18 years of longitudinal history, plus the most recent predictor information known prior to the earliest year, for each of 1,126,079 patients. There are 595 variables of which 18 are either for identification purposes or are response variables, the remainder being predictors.

The predictors can be grouped as:

- A set for each of 52 "events" which can be reported by a GP practice in a given year, and which relate either to a medical condition or a prescribed medicine (e.g. angina, aspirin). The main predictors for each event are concerned

with: (a) Is the event reported in this year? (b) Has the event been reported in any previous years? (c) What is the age of the patient when the event was first reported?
- Measurements of Body Mass Index (BMI), blood pressures and cholesterol level for the year.
- Information about smoking status and living arrangements for the year.

The response variables are demcase1, demcase3, demcase5, and demcase10, indicating whether or not a patient was diagnosed with dementia within 1, 3, 5, or 10 years, respectively. This work focuses on predicting dementia within 5 years, hence only demcase5 is kept as response variable in the study.

Patients are grouped in 418 GP practices.

C. Data transformation and enhancement

In this study the CPRD data was used to predict the risk of dementia occuring within 5 years. As the original data comprised multiple records per patient, capturing the longitudinal aspects of the data, for the purpose of this analysis we flattened the data, including new variables devised via transformations such as the following:

- Calculation of a new weighted sum predictor for event type; acute events (e.g. stroke) are weighted higher later in time, as opposed to chronic events (e.g. diabetes) which are weighted equally over time.
- Mapping of weighted sums on to new predictors for dementia comorbidities, and risk factors identified by Public Health England in [26].
- Mapping of weighted sums and other predictors to a new frailty predictor based on the Electronic Frailty Index [27].
- Calculation of a new smoking severity predictor based on weighted smoking history.
- Creation of 4 new predictors based on a longitudinal analysis of patients' living arrangements (latest value known, mode, latest year recorded, number of missing values).
- Devising new predictors counting missingness of values in variables per patient (over years).
- Imputing missing values per patient via interpolation for BMI, blood pressures and cholesterol level (i.e. horizonal imputation).
- Devising new predictors capturing the change over time for each of BMI, blood pressures and cholesterol level as per methodology proposed in [28], based on minimum, maximum, mean, variance, and velocity.
- Devising new predictors containing the means of number of consultations per year and polypharmacy count per year.
- Devising new predictors based on the mode (over time) of categorical interpretations of several numeric predictors.

This process led to building 2 datasets with 508 variables (12 of which were for identification or response purposes), namely:

- An uncensored dataset containing 644,306 patients from 406 GP practices, of which 9,656 were positive cases (i.e diagnosed dementia within the 5 years) and 634,650 controls.
- A censored dataset containing 481,773 patients from 406 GP practices, which left the study for which no diagnosis was available (this is unlabelled data, part of which to use in our semi-supervised approach).

D. Prediction models

As previously mentioned, the main two models that our predictive modelling approach relies on are Cox Proportional Hazards (Cox PH) [24] and Gradient Boosting Machines (GBM) [25].

Cox PH is a form of survival based models and is used to predict a probability of a hazard (e.g. dementia diagnosis) occuring at time t, based on the number of survivors at time t. In this regard it is able to use censored observations in the model training. Cox PH models have linear decision boundaries obtained through induced predictor coefficients, in a similar manner to certain other forms of linear models, and from which explicit risk factors can be obtained (hazard ratios). The "proportionality" in Cox PH is based on the assumption that predictor coefficients don't vary over time. The Cox PH hazard function $h(t)$ can be written as:

$$h(t) = h_0(t)e^{(\beta_0+\beta_1 x_1+\beta_2 x_2+...+\beta_n x_n)} \quad (1)$$

where $h_0(t)$ is the baseline hazard function, $h(t)$ is the hazard function, and $\{\beta_i\}$ are the intercept and predictor coefficients computed by the partial likelihood maximisation algorithm.

Stochastic Gradient Boosting Machines (GBM) [25] are models built from aggregations of a sequence of decision trees, each of which attempting to predict the residuals produced by the previous tree. The effect of each tree on the model as a whole is moderated by a learning rate.

$$\hat{f}(x) = \sum_{m=1}^{M} \lambda \hat{f}_m(x) \quad (2)$$

where λ is the learning rate, $\hat{f}_m$ is the prediction of the m^{th} tree, of the residual produced by the previous tree in the sequence, and $\hat{f}$ is the final prediction. For a classification problem, as we have here, a logistic function is used to transform final predictions into class probabilities.

E. Feature selection and data pre-processing

The approach was constrained by the need to create training and test sets that were acceptable to the different technologies underpinning the Cox PH and GBM models we deployed. For example, GBM could handle missing values and categorical variables but Cox PH could not. Feature selection and model pre-processing were applied independently for each model building iteration, within each Monte Carlo sample [29] (i.e. each training and test set split).

We handled missing values in smoking and living arrangements by assigning a specific string value indicating missingness prior to dummification (one-hot encoding). All

predictors were scaled and centered by z-normalisation, and we removed those which were highly correlated with other predictors.

We used elastic net [30] regularized Cox PH to select features for GBM, using cross validation to find the optimum elastic net mixing parameter (α) and regularisation weight (λ), based on best Area Under the Curve (AUC) [37], when the model is applied to a large sample of the raw (i.e. before class balancing) training set.

F. Class balancing

The data had very adverse class balance with only 1.5% of uncensored patients having a dementia outcome. No benefit (in terms of AUC on the test set) accrued from the various over-sampling, under-sampling and mixed regimes (including Synthetic Minority Oversampling Technique [36]) that were tried; the most pragmatic approach was to use all dementia outcomes in the training set, and sample the control observations such that the dementia cases constituted 10% overall, broadly.

G. Model structure and content

Each Monte Carlo sample yielded an independent set of results from 1 training and test set sample (from uncensored data); hence model stability could be inferred from the variance in results metrics, and model performance from the means.

For each sample, up to 4 attempts were made to incrementally augment the training set with a selection from the set of censored observations, which had been classified by the GBM model built in each attempt. We used Positive Predicted Value (PPV) and Negative Predicted Value (NPV) metrics (as per Kuhn and Johnson [38]) to establish criteria for selection:

$$PPV = \frac{Se \times Pr}{Se \times Pr + (1 - Sp) \times (1 - Pr)} \tag{3}$$

$$NPV = \frac{Sp \times (1 - Pr)}{Pr \times (1 - Se) + Sp \times (1 - Pr)} \tag{4}$$

where Sp is Specificity, Pr is prevalence (of the positive class) in the population, Se is Sensitivity

The semi-supervised methodology followed the following steps:

1) A penalised Cox PH model was built from the (class balanced) training set
2) A GBM model was built using the same training set and the predictors selected by the Cox PH model
3) The GBM model was used to make predictions on a large sample of the current level of augmentation of the training set prior to class balancing (call this the "post procesing set"). The post processing set probabilities were used to search for optimum PPV and NPV thresholds, which were then used to select observations from the censored data set
4) From the post processing set, a "high PPV" probability threshold was found such that the number of observations with probabilities above that threshold was equal to the number of correct dementia predictions in that set, based on the Youden [32] threshold. The rationale was that a good proportion of those above the high PPV threshold would be correct
5) From the post processing set, a probability threshold corresponding to the maximum PPV was obtained
6) An "optimum PPV" probability threshold was sought, based on a scaled mid-point between the maximum PPV and the high PPV (the optimum scaling factor was 0.9)
7) The number of observations with probabilities above the optimum PPV threshold in the post processing set, was multiplied by the ratio of control to dementia observations in that set, to give an externally calibrated number of eligible control cases, n
8) An "optimum NPV" probability threshold was set from the n^{th} highest probabilty in the post processing set
9) Classified observations from the censored set with probabilities $\leq$ the optimum NPV threshold, and $>$ the optimum PPV threshold were moved from the the censored set to augment the training set and repeat the process; unless insufficient observations remained in the censored set.

We used Youden's method of determining a probability discrimination threshold since this yields the point on the ROC curve furthest from the main diagonal, which corresponds to a random guess model.

Platt scaling [33] and isotonic [34] calibration models were built on predicted post processing set probabilites from the best performing Cox PH and GBM models in each training set augmentation iteration. The calibration models redistribute predicted probabilities according to observed probabilities (in this case, dementia outcomes of 1 or 0), so as to to improve their (external) calibration. In a well calibrated model the probability distibution should reflect the population, e.g. a probability of around 0.7 should apply to around 70% of observations. For each best Cox PH and GBM result the calibration model with a slope nearest to 1 was chosen; this model was used to calibrate test set probabilities, from which the final performance metrics could then be obtained.

Note that a requirement was for models to be able to generalize across GP practices (which vary in record keeping procedure and coding, and in demographic and other factors), hence entire GP practices were allocated either to training or test sets.

H. Model tuning

Cox PH models were tuned using elastic net mixing parameter, α, values of 0, 0.25, 0.5, 0.75 and 1, where the regularization penalty for an estimated coefficient, $\hat{\beta}$, in the penalized partial maximum likelihood algorithm ("partial" because the baseline hazard function, $h_0(t)$, is excluded from consideration in Cox regression), is defined as:

$$\frac{1 - \alpha}{2} \lambda \hat{\beta}^2 + \alpha \lambda |\hat{\beta}| \tag{5}$$

Thus with $\alpha = 1$ regularization is entirely L1 (Lasso) and with $\alpha = 0$ regularization is entirely L2 (Ridge); L1 conferring the ability to shrink coefficients to 0 and thus reduce dimensionality.

For a given value of α, the inner cross validation process computed and selected from a range of regularization weights, choosing a value for λ that maximised partial likelihood .

The best Cox PH model was selected based on the highest AUC when applied to the post processing set.

GBM models were tuned by finding the smallest log loss in a random search of 25 models in a hyperparameter space of 348 models, using 3 fold cross validation. The variable hyperparameters were: learning rate (0.01, 0.005, 0.1, 0.3), learning rate annealing factor (0.99, 1), maximum tree depth 2, 5, 10), row sampling rate (0.1, 0.4, 0.632, 1.0), column sampling rate (0.1, 0.3,0.7, 1.0).

I. Hardware and software

Models were built and evaluated using parallel processing (wherever possible) on a data analytics cluster of 11 servers with Xeon processors and 832GB RAM. The software was predominantly R using packages: tidyverse, survival, glmnet, h2o, calibration, pROC.

III. Results

All results were obtained from a held-out test set which was not exposed to any model training processes. Cox PH and GBM results given are from the same Monte Carlo iteration sample and hence are directly comparable.

A. Predictor importance

Since it is not meaningful to estimate coefficient standard errors for penalized models, we have not used hypothesis testing to validate these results; however since all predictors were scaled and centered it is reasonable to use the (absolute) magnitude of an induced coefficient as an indicator of importance for Cox PH, as given in table I.

For GBM, predictor importance is determined by the impact to the model's squared error each time the predictor is used for a split; importance is shown as a fraction or percentage as per table II.

B. Discrimination and calibration

Tables III and IV give model performance metrics for Cox PH and GBM; we have included the Kappa statistic to give a measure of distance from expectation, and observe that this is low due to the unfavourable balance of dementia cases to controls. The final in both tables III and IV gives calibration slope and intercept; note that the column headings "LCI" and "RCI" denote left and right 95% confidence interval boundaries.

TABLE I
TOP 20 COX PH COEFFICIENTS

predictor	mean	sd
yb-age	0.9851	0.0217
add-age	0.1400	0.0285
YC-0-bmi-cat.normal	0.1226	0.0063
frailty	0.1068	0.0209
bmi-velocity-mode	-0.1057	0.0056
diabetes-age	0.1053	0.0272
dna-sum	0.0937	0.0166
YC-0-bpr-cat.normal	0.0771	0.0063
Resid-shelter-care-home	0.0724	0.0076
bmi-min	-0.0704	0.0036
antichol-sum	0.0680	0.0083
diabetes-cat-age.x-60-plus	-0.0625	0.0184
tca-age	-0.0614	0.0227
stroke-chronic	0.0589	0.0091
headinj-chronic	0.0530	0.0071
hyperten-sum	-0.0516	0.0184
diastolicmeandi-l20-mode.myNA	0.0487	0.0223
ae-sum	0.0473	0.0028
h2rec-sum	-0.0471	0.0123
bpr-cat-mode.normal	0.0465	0.0274

TABLE II
TOP 20 GBM PREDICTOR IMPORTANCES

predictor	mean	sd
yb-age	0.3260	0.2104
add-age	0.0589	0.0129
antihyp-age	0.0456	0.0199
aspirin-age	0.0362	0.0108
ssri-age	0.0307	0.0272
bmi-min	0.0263	0.0039
nsaids-age	0.0253	0.0169
homev-sum	0.0246	0.0078
Resid-shelter-care-home	0.0222	0.0176
consultcount-mean	0.0180	0.0067
pulsepress-velocity-max-abs	0.0165	0.0083
frailty	0.0158	0.0078
hyperten-age	0.0149	0.0130
polypcount-mean	0.0145	0.0094
stroke-chronic	0.0135	0.0065
diastolic-min	0.0133	0.0105
dna-sum	0.0121	0.0027
tca-age	0.0112	0.0044
diabetes-age	0.0095	0.0019
antichol-sum	0.0093	0.0061

TABLE III
COX PH PERFORMANCE METRICS

metric	mean	sd	LCI	RCI
AUC	0.8242	0.0050	0.8145	0.8339
Youden	0.0143	0.0009	0.0126	0.0160
Sensitivity	0.7890	0.0186	0.7526	0.8254
Specificity	0.7193	0.0148	0.6903	0.7483
Accuracy	0.7204	0.0143	0.6923	0.7485
PPV	0.0417	0.0014	0.0388	0.0445
NPV	0.9955	0.0003	0.9949	0.9961
Kappa	0.0517	0.0024	0.0470	0.0565
C-intercept	0.0004	0.0005	-0.0006	0.0013
C-slope	1.0220	0.0481	0.9276	1.1163

TABLE IV
GBM PERFORMANCE METRICS

metric	mean	sd	LCI	RCI
AUC	0.8273	0.0050	0.8174	0.8371
Youden	0.0143	0.0007	0.0129	0.0158
Sensitivity	0.7412	0.0463	0.6504	0.8319
Specificity	0.7642	0.0324	0.7007	0.8277
Accuracy	0.7638	0.0313	0.7025	0.8251
PPV	0.0468	0.0041	0.0388	0.0548
NPV	0.9948	0.0006	0.9936	0.9960
Kappa	0.0611	0.0072	0.0469	0.0753
C-intercept	0.0038	0.0019	0.0001	0.0074
C-slope	0.7959	0.1292	0.5426	1.0492

IV. Discussion and conclusion

We proposed an efficient prediction modelling approach to the risk of dementia based on Gradient Boosting Machines (GBM) and Cox Proportional Hazard. While GBM algorithm offers its own method of ranking predictor variables according to their importance, it is very computationally intensive and less time-efficient when applied to our large routine primary care records dataset which, after pre-processing, had 1,126,079 rows and 595 variables. As such, we chose to combine the application of GBM with a less computationally intensive method such as Cox Proportional Hazards (Cox PH) with a two-fold purpose: (1) We use the most important variables selected by Cox PH, with the GBM method, and (2) We compare the performance of the predictions achieved by the GBM and Cox PH models.

As our whole dataset contained a large proportion of patients who didn't receive a diagnosis (42.78%), we used part of these censored patients to augment our training dataset by enhancing our models with a semi-supervised learning procedure in which GBM and Cox PH are encapsulated. Moreover, we calibrated our prediction models using Platt scaling and isotonic calibration. Our approach led to comparable AUC performances as described in [22].

As ongoing work, the approach we proposed here is currently extended with developing predictive models based on the XGBoost algorithm which is an optimised extension of GBM and hence suitable to be applied on large datasets such as CPRD that we used in this study. Due to the large volume of computation the analyses on CPRD require, XGBoost may be a better candidate algorithm which can make use also of the GPU hardware technologies. Another extension we currently focus on is making use of the large volume of data and developping deep learning models using also GPU faster hardware technologies, with (a) the use of autoencoders as an unsupervised learning tool to learn new features and reduce data dimensionality; (b) building optimised prediction models as standard deep feed-forward neural networks, and (c) using a state of the art algorithm GANOCC proposed in [15] and based on Generative Adversarial Networks - a modern concept in deep learning nowadays. GANOCC is a one-class classification method working well also on highly imbalanced datasets, as it is the case of CPRD used in this study.

Another direction of ongoing work is the investigation of the distribution of dementia outcomes amongst patients with similar medical histories, involving patient clustering, as a means of gaining insights into the contribution that certain groups of predictors may have on the prediction of risk of dementia.

References

[1] E. Barrett and A. Burns, "Dementia Revealed. What Primary Care Needs to Know," Department of Health, 2014.

[2] R. Petersen, R. Roberts, D. Knopman, B. Boeve, Y. Geda, R. Ivnik, et al. "Mild cognitive impairment: ten years later," Arch Neurol., 2009.

[3] A.J. Mitchell and M. Shiri-Feshki,"Rate of progression of mild cognitive impairment to dementia—meta-analysis of 41 robust inception cohort studies," Acta Psychiatr Scand, 2009.

[4] M. Prince, E. Albanese, M. Guerchet, and Prina, M., "Dementia and risk reduction: An analysis of protective and modifiable factors," World Alzheimer Report, 2014.

[5] G. Manogaran, V. Vijayakumar, R. Varatharajan, P.M. Kumar, R. Sundarasekar, and C.H. Hsu, "Machine Learning Based Big Data Processing Framework for Cancer Diagnosis Using Hidden Markov Model and GM Clustering," Wireless Pers Commun, 2017.

[6] N. Amoroso, M. L. Rocca1, S. Bruno, T. Maggipinto, A. Monaco, R. Bellotti, et al., "Brain structural connectivity atrophy in Alzheimer's disease," arXiv:1709.02369 [physics.med-ph], 2017.

[7] R.S. Desikan, C.C. Fan, Y. Wang, A.J. Schork, H. Cabral, L.A. Cupples, et al., "Genetic assessment of age-associated Alzheimer disease risk: Development and validation of a polygenic hazard score," PLOS Medicine, 2017.

[8] D. Brauser, "'Machine Learning' Tool May Predict Dementia Development Up to 10 Years Later.," Medscape, 2016.

[9] L. Robinson and M. Trenell, "Dementia: risk reduction," Newcastle University Institurte of Ageing, 2016.

[10] M. Kivipelto, A. Solomon, S. Ahtiluoto, T. Ngandu, J. Lehtisalo, R. Antikainen, et al., "The Finnish Geriatric Intervention Study to Prevent Cognitive Impairment and Disability (FINGER): Study design and progress," Alzheimer's Association, vol. 9, issue 6, pp. 657-665, January 2013.

[11] S. Sindi, E. Calov, J. Fokkens, T. Ngandu, H. Soininen, J. Tuomilehto, and M. Kivipelto, "The CAIDE Dementia Risk Score App: The development of an evidence-based mobile application to predict the risk of dementia," Alzheimers & Dementia: Diagnosis, Assessment & Disease Monitoring, vol. 1, issue 3, pp. 328-333, 2015.

[12] S. C.M. Blossom, C. Brayne, C. Dufouil, T. Kurth and F. Matthews, "Dementia risk prediction in the population: are screening models accurate?" Nat. Rev. Neurol, 2010.

[13] S. C.M. Blossom, E. Tang, G. Muniz-Terrera, "Composit risk scores for predicing dementia", Curr Opin Psychiatry, vol. 29, no. 2, 2016.

[14] E. Ford, N. Greenslade, P. Paudyal, S. Bremner, H. E. Smith, S. Banerjee, et al., "Predicting dementia from primary care records: A systematic review and meta-analysis," PLoS ONE, March 2018.

[15] M. Ermaliuc, D. Stamate, G. Magoulas, I. Pu, "Creating Ensembles of Generative Adversarial Network Discriminators for One-class Classificatio", Proceedings of 22nd Intl Conference of Engineering Applications of Neural Networks, EANN 2021, Springer, 2021.

[16] D. Stamate, M. Kim, P. Proitsi, S. Lovestone, C. Legido-Quigley, et al., "A metabolite-based machine learning approach to diagnose Alzheimer's-type dementia in blood: Results from the European Medical Information Framework for Alzheimer's Disease biomarker discovery cohort", Alzheimer's & Dementia: Translational Research & Clinical Interventions, Vol. 5, 2019, pp 933-938, Elsevier, 2019.

[17] L. Lang, A. Clifford, L. Wei, et al., "Prevalence and determinants of undetected dementia in the community: a systematic literature review and a meta-analysis", BMJ Open.;7(2):e011146, 2017.

[18] M.J. Prince, A. Comas-Herrera, M. Knapp, M. Guerchet and M. Karagiannidou, "Improving healthcare for people living with dementia: Coverage, quality and costs now and in the future," World Alzheimer Report, 2016.

[19] World Health Organization (2017), "Global action plan on the public health response to dementia 2017–2025," WHO Geneva: Licence: CC BY-NC-SA 3.0 IGO, 2017.

[20] D. Stamate, W. Alghamdi, J. Ogg, R. Hoile and F. Murtagh, "A machine learning framework for predicting dementia and mild cognitive impairment", 17th IEEE International Conference on Machine Learning and Applications (ICMLA), Orlando, FL, pp. 671-678, 2018.
[21] D. Stamate, R. Smith, R. Tsygancov, R. Vorobev, J. Langham, D. Stahl, D. Reeves, "Applying Deep Learning to Predicting Dementia and Mild Cognitive Impairment," Proc. of 16th Intl. Conference of Artificial Intelligence Applications and Innovations (AIAI). IFIP Advances in Information and Communication Technology, vol 584. Springer, 2020.
[22] K. Walters, S. Hardoon, I. Petersen, et al., "Predicting dementia risk in primary care: development and validation of the Dementia Risk Score using routinely collected data," BMC Med 14, 6, 2016.
[23] "The Health Improvement Network. THIN (R)," https://www.the-health-improvement-network.com/
[24] D.R. Cox, D. Oakes, "Analysis of Survival Data," New York: Chapman & Hall. ISBN 978-0412244902, 1984.
[25] J.H. Friedman, "Stochastic gradient boosting," Comput. Stat. Data Anal. 38, 4, 367–378. 2002.
[26] D. Leese D, M. Jackson, M. Szczepaniak, J. Verne, S. Foster, S. Sandhu - Public Health England (2019), "Dementia: comorbidities in patients - data briefing," https://www.gov.uk/government/publications/dementia-comorbidities-in-patients/dementia-comorbidities-in-patients-data-briefing
[27] A. Clegg, C. Bates, J. Young, R. Ryan, L. Nichols, E.A. Teale, M.A. Mohammed, J. Parry, T. Marshall, "Development and validation of an electronic frailty index using routine primary care electronic health record data," Age and Ageing, Volume 45, Issue 3, pp. 353–360, 2016
[28] D. Stamate A. Katrinecz, D. Stahl, S.J.W. Verhagen, P.A.E.G. Delespaul, J. van Os, S. Guloksuz, "Identifying psychosis spectrum disorder from experience sampling data using machine learning approaches," Journal of Schizophrenia Research, Elsevier, 2019
[29] D.P. Kroese, T. Brereton, T. Taimre, Z.I. Botev, "Why the Monte Carlo method is so important today," WIREs Comput Stat. 6 (6): 386–392, 2014.
[30] H. Zou, T. Hastie, "Regularization and Variable Selection via the Elastic Net," Journal of the Royal Statistical Society, Series B. 67 (2): 301–320, 2005.
[31] N.V. Chawla, K.W. Bowyer, L.O. Hall, W.P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," Journal of Artificial Intelligence Research, Vol 16, 2002.
[32] W.J. Youden, "Index for rating diagnostic tests," Cancer. 3: 32–35, 1950.
[33] J. Platt, "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods," Advances in Large Margin Classifiers. 10 (3): 61–74, 1999.
[34] R.E. Barlow, D.J. Bartholomew, J.M. Bremner, H.D. Brunk, "Statistical inference under order restrictions; the theory and application of isotonic regression," New York: Wiley, 1972.
[35] N.E. Breslow, "Discussion of the paper by D.R. Cox," J R Statist Soc B 34:216–217, 1972.
[36] N.V. Chawla, K. Bowyer, L.O. Hall, W.P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," Journal of Artificial Intelligence Research, 16(1):321-357, 2002.
[37] J.A. Hanley, B.J. McNeil, "The Meaning and Use of the Area Under a Receiver Operating Characteristic (ROC) Curve," Radiology 143(1):29-36, 1982.
[38] M. Kuhn, K. Johnson, "Applied predictive modeling," Springer, 2018.
[39] Clinical Practice Research Datalink (CPRD) https://www.cprd.com/