Predicting different types of dementia using structured data

Miloš Pavlić*, Ognjen Francuski*, Igor Jovin*, Slivka Jelena*, Aleksandar Kovačević*
* Faculty of Technical Sciences/Computing and Control Department, Novi Sad, Serbia
pavlic.milosh@uns.ac.rs, ognjenfrancuski@uns.ac.rs, slivkaje@uns.ac.rs, igor.jovin@uns.ac.rs, kocha78@uns.ac.rs

***Abstract*—In this paper, we present a model for the automatic detection of dementia in patients based on their demographic data and the results of neurological and psychological tests. We classify patients according to four major dementia types, as defined by the domain experts. All patient features used in our model are obtained by noninvasive tests and can be easily collected by people who are not domain experts (neurologists or psychiatrists). In this way, our model could be applied to screen a much bigger population and would allow for more frequent screening of the individuals, compared to the traditional approach which requires manual diagnosis by the domain expert. In this way, we hope to alleviate the problem of early dementia diagnoses, which is critical for the patient's quality of life. We train and evaluate our model on the publicly available Open Access Series of Imaging Studies (OASIS) dataset and obtain the f-measure of 92% and an accuracy of 94%. To the best of our knowledge, this is the best performance achieved by using only the demographic, neurological, and psychological data. It should also be noted that similar solutions classify patients in two (patients with and without dementia) or three classes (patients with mild, severe and no dementia), while we perform multi-category classification according to four major types of dementia.**

I. INTRODUCTION

Alzheimer's disease is a significant public health concern. In 2010 alone, about 135 million around the world had dementia. The economic impact of this disease is estimated at $600 billion worldwide [2][3]. By 2050 it is estimated that dementia will affect 1.5 billion people [2].

The cause of Alzheimer's dementia is the reduction of neurotransmitter secretion with aging [4]. Although dementia is a progressive condition, and cannot be cured, its progress can be slowed down by certain treatments. Thus, a person who has dementia can prolong the time they are capable of doing certain activities, which would be significantly harder if dementia was to progress. This makes the early diagnosis of dementia of great importance for the patient.

As the number of old people increases much faster than the number of doctors with the expertise to diagnose dementia, the problem of early diagnosis becomes increasingly harder. In this paper, we present the model for automatic detection of dementia in patients based on their demographic data and the results of neurological and psychological tests. All patient features used in our model are obtained by noninvasive tests and can be easily collected by people who are not domain experts (neurologists or psychiatrists). The goal of our model is to hasten the diagnosis process by automatically identifying individuals with a high risk of dementia that needs additional tests performed by medical experts. By reducing the number of the patients required to be manually assessed by medical experts, our model could allow the screening of a much bigger population and more frequent screening of the individuals, thus leading to the early start of therapy which would significantly increase the quality of life of the patients who have dementia.

In this paper, we present the model for automatic detection of dementia using the following patient's data:

- demographic data (age, gender, etc.),
- neurological data (results of neurological tests: the ability to speak, the ability to control biological needs, etc.),
- psychological data (results of psychological tests: space and time orientation, ability to understand speech, ability to solve more complex tasks, etc.).

We consider the multi-category problem of classifying patients in four categories:

- Patients without dementia,
- Patients in starting stages of dementia,
- Patients with progressed Alzheimer's (AD) dementia,
- Patients with progressed non-AD dementia.

In section II we present previous papers on this subject. Methodologies and experiment approaches used in this research are described in section III. In section IV experiment results are presented. Section V shows future work. Paper is concluded in section VI.

II. RELATED WORK

To the best of our knowledge, all similar papers that address the same problem of predicting dementia from structured data, classify patients in two (patients with and without dementia) [1] or three classes (patients with mild, severe and no dementia) [5], even though there are differences between various types of dementia. Various types of dementia have entirely different therapies, so it is essential to know which patient is suffering from which type of dementia to know which kind of treatment should be applied.

The solution presented in [1] uses the most similar data to the one used in this paper. They have used machine

learning (ML) techniques to classify patients in two groups: those with mild dementia and those with no dementia. In total, they have used three datasets with 149 features. They have applied a classification model on each dataset with three different methods of preprocessing. The performed preprocessing methods differed only in feature selection/engineering. The methods used were:

- The whole feature set without feature selection/engineering,
- Features proposed by domain experts (28 features),
- Features engineered using the Principal Component Analysis method on the 28 features mentioned above.

The best accuracy achieved in [1] was 83.8% obtained by using the Naïve Bayes classifier in combination with the whole dataset without feature selection. It should be noted that the authors report accuracy, but do not specify whether the dataset was balanced or not. The datasets used in [1] contained the data about 583 patients total, while our dataset contains data about multiple examinations from 1098 patients, resulting in 6197 entries. In this paper, we strive to improve the performance of the model proposed in [1], as well as perform multi-category classification of patients according to the four major types of dementia.

III. METHODOLOGY

In this chapter we describe methodology applied in this paper. First the dataset is given, after that data preprocessing methods are presented. Lastly classification approaches used in this research are described.

A. Dataset

The dataset used in our paper can be obtained on request for scientific and research purposes on the Open Access Series of Imaging Studies (OASIS) organization website [6]. It consists of different types of patient data: demographical data, psychological and neurological tests results, and data acquired from Magnetic Resonance Imaging (MRI scans).

The total number of features we consider in our model is 146. We do not consider the features that contain over 50% missing values. There are a total of 1098 patients in the dataset, with a total of 6179 examinations by the doctors, classified in four classes:

- Patients without dementia (4493 instances) – Cognitively normal (CN),
- Patients in starting stages of dementia (506 instances) – Uncertain dementia (UD),
- Patients with progressed Alzheimer's dementia (1046 instances) (AD),
- Patients with progressed non-AD dementia (134 instances).

These four main categories were derived from 50 different types of diagnoses given by doctors with the help of domain experts.

Basic information about the dataset is given in .

TABLE I.
DATASET INFORMATION

Class	Age (years)	Entries	Entries per group (%)	Entries total (%)
Cognitively normal	<= 60	807	17.96	12.97
	> 60	3686	82.04	59.22
Uncertain dementia	<= 60	23	4.55	0.37
	> 60	483	96.45	7.76
Alzheimer's dementia	<= 60	44	4.21	0.71
	> 60	1002	96.79	16.10
Non AD dementia	<= 60	4	2.94	0.06
	> 60	130	97.06	2.09

B. Data processing

Dataset is not prepared for the direct application of machine learning algorithms, so some necessary processing steps had to be applied. Features with over 50% of missing values were removed. Dataset was split into training and test parts. The training set contained 80%, and the test set contained 20% of the entries in the whole dataset. After this, several processing steps were applied:

- Missing values imputation,
- Normalization,
- Feature selection,
- Oversampling the training set, and
- Outlier removal.

We have experimented with several different methods for imputing missing values: removing rows containing any missing values, mean imputation and ML model imputation. The best results were obtained by applying the K-Nearest Neighbours (KNN) algorithm to impute the missing values.

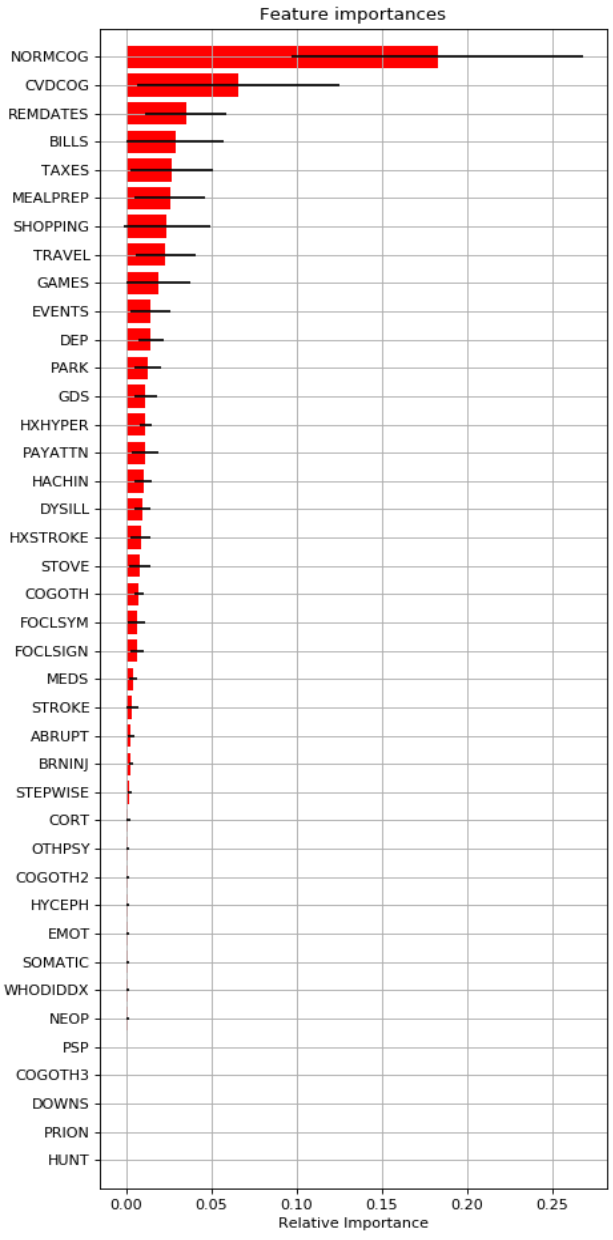
After imputation, feature values were normalized using the Z-Score method. In some of our experiments, normalization was followed by feature selection using Extra Trees classifier. Feature selection resulted in the final 40 features that were left in the dataset. We have also experimented with feature selection. We have tried two approaches: applying extra trees classifier in order to find the best feature set (40 features), and using only the features that have less than 50% missing values. In our experiments the best performance was obtained by using the whole set of features that have less than 50% missing values. Feature ranking by importance is presented in FIGURE I. The highest ranked features by importance are basic cognitive abilities described by NORMCOG and CVDCOG. Beside cognitive abilities, features representing memory like remembering important dates, paying bills and taxes (REMDATES, BILLS, TAXES) are of high importance.

Given the distribution of the classes in the dataset isn't balanced, the training set was oversampled using

ADASYN method [7]. This greatly improved the performance of the classification.

Finally, we experimented with two options for outlier removal. In the first approach, models were trained on the whole dataset without removing any outliers. In others, outliers were removed using several outlier removal algorithms. We considered Isolation Forest and Local Outlier Factor algorithms. The best results were achieved using the Isolation Forest algorithm [8].

FIGURE I.
FEATURES RANKED BY IMPORTANCE



C. Classification approaches

On the processed dataset, two classification approaches were applied. In the first approach, ML models were trained on the processed dataset to classify the patients in one of the previously named classes (cognitively normal, uncertain dementia, Alzheimer's dementia, non-AD dementia).

The second approach consists of two steps. In the first step, a classifier was applied to divide healthy patients from patients who are diagnosed with any type of dementia. In the second step, the second classifier diagnosed patients into one of the three classes that represent different types of dementia.

We have experimented with the following ML classifiers:

- Bagging,
- Random Forest,
- Extra Trees,
- Support Vector Machine (SVM), and
- Extreme Gradient Boosting (XGBoost).

Each of the ML models was trained on two different versions of the dataset:

- Basic – each model is trained on the processed dataset, where entries are classified in one of the four classes.
- Separated – each model is trained separately on two datasets. During the first run, models are trained to classify entries either as cognitively normal or diagnosed with any type of dementia. During the second run, models are classifying diagnosed patients into groups representing different types of dementia.

IV. EXPERIMENTAL RESULTS

To evaluate our solution, we divide the data by stratified random sampling into training (80%, total of 4943 examinations) and test (20%, total of 1236 examinations) set. The optimal parameter values for the considered ML models are chosen by performing 10-fold cross-validation on the training set. In each iteration, a different subset of training data is used as test data, while the remaining nine subsets are used as training data. As our dataset is unbalanced, precision, recall, and f1-measure are reported for each class.

The obtained results show the clear difference between the data collected by testing cognitively normal and patients diagnosed with any kind of dementia (as shown in TABLE III. and TABLE IV.). However, it is relatively hard to distinguish different types of dementia with current approaches given the results of these tests (as shown in TABLE III. and TABLE V.).

When diagnosing dementia or any other serious medical condition, it is essential to classify all the patients with the diagnosis correctly. Thus, the primary performance metric that should be considered is the recall of the classes representing dementia diagnosis.

Since similar papers use accuracy or f1-measure as the measure of performance [1][5], these measures of performance are reported to be able to compare with those

papers. Given the imbalance in the dataset, accuracy and average f1-measure do not genuinely show the performance of the models. Differences in results between the two approaches (basic and separated) used in this research are minimal, thus and accuracy and global f1-measure are reported only for the basic approach (TABLE II.).

TABLE II.
OVERALL RESULTS OF ML MODELS FOR BASIC APPROACH. THE BEST OVERALL RESULTS WERE ACHIEVED USING XGBOOST CLASSIFIER

Model	Accuracy	F1 score
SVM	0.9036	0.8940
Bagging	0.9265	0.9125
Random Forest	0.9301	0.9187
Extra Trees	0.9076	0.8977
XGBoost	**0.9355**	**0.9199**

TABLE III.
REPORT OF PRECISION, RECALL AND F1 MEASURES OF PERFORMANCE OF ML MODELS PER CLASS, WITH THE BASIC APPROACH. THE BEST RESULTS FOR CLASSES THAT REPRESENT DEMENTIA ARE EMPHASIZED

Model	Class	Precision	Recall	F1 score
SVM	CN	0.97	0.98	0.97
	AD	0.80	0.80	0.80
	Non AD	0.36	0.24	0.29
	UD	0.59	0.56	0.57
Bagging	CN	0.99	1.00	1.00
	AD	0.80	0.83	0.81
	Non AD	0.88	0.47	0.61
	UD	0.60	0.57	0.58
Random Forest	CN	0.98	0.99	0.99
	AD	0.89	0.63	0.74
	Non AD	0.42	0.67	0.51
	UD	**0.55**	**0.80**	**0.65**
Extra Trees	CN	0.98	1.00	0.99
	AD	0.93	0.44	0.60
	Non AD	**0.23**	**0.93**	**0.36**
	UD	0.57	0.78	0.66
XGBoost	CN	0.99	0.99	0.99
	AD	**0.77**	**0.88**	**0.82**
	Non AD	0.56	0.29	0.38
	UD	0.68	0.54	0.60

TABLE IV.
REPORT OF PRECISION, RECALL AND F1 MEASURES OF PERFORMANCE OF ML MODELS FOR COGNITIVELY NORMAL PATIENTS AND PATIENTS DIAGNOSED WITH DEMENTIA. THE BEST RESULTS FOR DIAGNOSED PATIENTS ARE EMPHASIZED

Model	Class	Precision	Recall	F1 score
SVM	CN	0.99	0.98	0.98
	Diagnosed	0.95	0.96	0.96
Bagging	CN	0.99	1.00	1.00
	Diagnosed	**1.00**	**0.99**	**0.99**
Random Forest	CN	0.99	1.00	1.00
	Diagnosed	**1.00**	**0.99**	**0.99**
Extra Trees	CN	0.99	1.00	1.00
	Diagnosed	1.00	0.98	0.99
XGBoost	CN	0.99	1.00	1.00
	Diagnosed	1.00	0.98	0.99

TABLE V.
REPORT OF PRECISION, RECALL AND F1 MEASURES OF PERFORMANCE OF ML MODELS FOR DIFFERENT TYPES OF DEMENTIA.

Model	Class	Precision	Recall	F1 score
SVM	AD	0.70	0.83	0.76
	Non AD	0.50	0.11	0.18
	UD	0.51	0.41	0.46
Bagging	AD	0.74	0.89	0.81
	Non AD	1.00	0.11	0.20
	UD	0.65	0.53	0.58
Random Forest	AD	0.76	0.76	0.76
	Non AD	0.79	0.62	0.67
	UD	0.76	0.76	0.75
Extra Trees	AD	0.75	0.81	0.78
	Non AD	0.35	0.50	0.41
	UD	0.76	0.56	0.64
XGBoost	AD	0.72	0.95	0.82
	Non AD	1.00	0.11	0.20
	UD	0.75	0.43	0.55

V. FUTURE WORK

The approach that uses only structured data given in the dataset used in this research can not differentiate the types of dementia with high performance. However, clinical signs of dementia can also be detected on MRI scans and in the future, the plan is to use computer vision techniques to build a model for automatic detection of dementia based on this type of information. In consultation with domain experts, it has been concluded that MRI scans can reliably show differences between the various types of dementia.

Additional experiments that include combining these models by using methods such as weighted voting and stacking will be conducted to improve the classification performance.

Finally, with the help of domain experts, additional feature engineering will be applied. Newly engineered features will be divided into groups to determine which features are the most significant for the task of dementia classification.

VI. Conclusion

The research has shown that ML models can differentiate cognitively normal patients from those diagnosed with dementia, using only structured data. However, it is challenging to classify patients diagnosed with dementia into the groups representing each type of dementia using only the structured data provided in the dataset. In consultation with domain experts, authors have concluded that using computer vision methods on MRI scans can lead to significantly better results when classifying patients depending on which type of dementia they have since the MRI scans differentiate types of dementia better than results obtained using neurological and psychological tests.

References

[1] Williams, J. A., Weakley, A., Cook, D. J., & Schmitter-Edgecombe, M. (2013, July). Machine learning techniques for diagnostic differentiation of mild cognitive impairment and dementia. In

[2] Prince, Martin, et al. "Recent global trends in the prevalence and incidence of dementia, and survival with dementia." Alzheimer's research & therapy 8.1 (2016): 23.

[3] Wimo, Anders, et al. "The worldwide economic impact of dementia 2010." Alzheimer's & Dementia 9.1 (2013): 1-11.

[4] Davies, Peter. "Neurotransmitter-related enzymes in senile dementia of the Alzheimer type." Brain research 171.2 (1979): 319-327.

[5] Shankle, William Rodman, et al. "Simple models for estimating dementia severity using machine learning." MedInfo 98 (1998).

[6] https://www.oasis-brains.org/#data

[7] He, Haibo, et al. "ADASYN: Adaptive synthetic sampling approach for imbalanced learning." 2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence). IEEE, 2008.

[8] Liu, Fei Tony, Kai Ming Ting, and Zhi-Hua Zhou. "Isolation forest." 2008 Eighth IEEE International Conference on Data Mining. IEEE, 200