A Graph Machine Learning approach to Automatic Dementia Detection

Edoardo Stoppa[1,2], Guido Walter Di Donato[1], Isabella Poles[1], Eleonora D'Arnese[1], Natalie Parde[2], and Marco Domenico Santambrogio[1]

***Abstract*— Dementia is a term used to refer to a wide range of diseases that cause a decline in cognitive abilities. This decline is severe enough to impair daily life and it is extremely complex to diagnose in its early stages. In recent years multiple Natural Language Processing solutions have been proposed to automatically detect dementia. One of the main approaches to this problem is based on extracting manually engineered features from a set of patients' conversations and feeding them to traditional Machine Learning models. These features can be divided into very different groups, and we can define specific relations that connect one feature to the other. Thus, we introduce a new way to approach the problem by organizing all the extracted features in a graph structure and using Graph Machine Learning to detect dementia. We validate our method using a well-established score regression task and a newly proposed multi-class classification task. This new task is based on the mapping between the Mini-Mental State Examination score and multiple dementia severity levels. Compared to traditional Machine Learning, our Graph Machine learning technique achieves a relative increase in performance between 2.9% and 8% for the regression task, and between 4.4% and 7.9% for the classification task.**

I. INTRODUCTION

Thanks to improving health standards, the life expectancy of the world population is slowly increasing [1]. Perhaps unsurprisingly, this has been accompanied by a growing prevalence of health issues that primarily affect the older population. One example of this is an increase in the number of people with various forms of dementia [2]. Dementia is not a single disease, but rather a generic term that covers a wide range of specific medical conditions caused by abnormal brain changes. It triggers a decline in cognitive abilities affecting behavior, feelings, and relationship. This decline is severe enough to impair daily life and it is extremely complex to diagnose in its early stages because its symptoms resemble the normal process of aging. Yet it is precisely at the start of the disease that treatment has the most effect [3]. Language analysis has shown to be a valid tool to distinguish between patients with and without dementia [4], [5], especially during its first stages, offering a cheaper alternative to more complex medical examinations. For this reason, there is increasing interest in developing Machine Learning (ML) applications for automatic early Dementia Detection (DD) based on patients' conversations. This can be done in many different ways; however, our work is focused on one of the main approaches: the use of traditional ML models combined with various features extracted via Natural Language Processing (NLP), and categorized into different feature groups. We propose a novel mixed Graph-based and NLP technique for early DD. In particular, our contributions to this research domain are:

- We define an original graph ontology that explicitly encodes the relational information among the features extracted from the patients' conversations using NLP.
- We introduce a new Multiclass Classification task that was not previously studied in this context.
- We demonstrate that our Graph Machine Learning (GML) approach outperforms a traditional ML approach to automatic early DD.

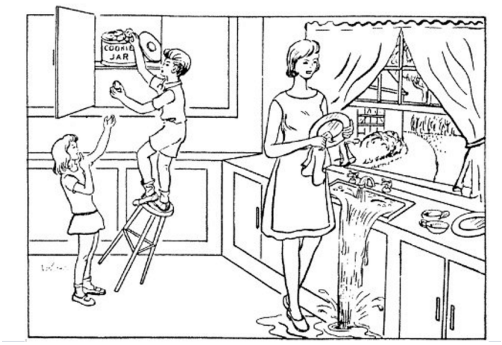

Fig. 1. Cookie Theft Picture, appears in "L. Cummings, Describing the cookie theft picture: Sources of breakdown in Alzheimer's dementia, Pragmatics & Society, 2019".

II. BACKGROUND AND RELATED WORKS

Dysphasia, also called aphasia, is a language disorder that affects how a person speaks and understands language, and it is a symptom that presents itself in most dementia patients. Physicians use a combination of different methods to reach a diagnosis, including Magnetic Resonance Imaging, Computerized Tomography, Neuroimaging, and specialized tests. Our work will rely on two of these specialized tests: the Cookie Theft Picture Test (CTPT), and the Mini-Mental Status Examination (MMSE) [6]. The CTPT, which is part of the Boston Diagnostic Aphasia Examination [7], consists of a simple descriptive task: the patient is asked to describe a special picture, called the Cookie Theft Picture (CTP) (Figure 1), while being as precise as possible. Our dataset (from DementiaBank[1]) is a collection of patients' conver-

[1]Affiliated with the Department of Electronics, Information Technology, and Bioengineering, Politecnico di Milano, edoardo.stoppa@mail.polimi.it, {guidowalter.didonato, eleonora.darnese, isabella.poles, marco.santambrogio}@polimi.it

[2]Affiliated with the Department of Computer Science, University of Illinois at Chicago, parde@uic.edu

[1]https://dementia.talkbank.org/

sations performing the CTPT, which is scored by marking the patients as "Dementia" or "No Dementia", and has associated an MMSE score (on a scale from 1 to 30, where 1 corresponds to "severe dementia", and 30 to "healthy").

Using this dataset researchers have explored different prediction tasks regarding automatic DD using various approaches. Despite the emergence of Deep Learning methods for DD, the main approach is still based on traditional ML that uses manually engineered features extracted using NLP techniques. We can observe various features and feature groups that can be extracted from these conversations. For example, Ahmed et al. [8] claimed that semantic, lexical content, and syntactic complexity features can diagnose dementia. In the work by Fraser et al. [9], 370 different features derived from Part of Speech (POS) tags, syntax analysis, grammatical constituents, psycholinguistic analysis, vocabulary richness, and acoustic attributes were used in order to detect Alzheimer's Disease. A more recent work [10] demonstrated the possibility of predicting MMSE scores using a mix of verbal and non-verbal features.

Our work defines a set of features and feature groups based on the current literature and organizes everything in a graph ontology defined specifically to represent the relational information that exists among these features. To demonstrate that our approach increases the final prediction performance with respect to traditional ML, we evaluate it using a GML algorithm called Graph2Vec [11]. This technique, given a series of N whole graphs as input, generates N vectors, called embeddings, that represent each original graph in a lower dimensional space. We specifically decided to use Graph2Vec because it is a well-established technique that can process both the graph's structure and the numerical features contained in each node. For instance, it has been used to create more informed word embeddings [12], and was adopted to detect phishing attempts on the Ethereum Blockchain [13].

III. DATASET AND METHODOLOGY

To validate our approach we used the most widely-used dataset in this research domain: the DementiaBank dataset. We used a specific subset of data contained in the Pitt Corpus [14], which contains transcripts and audio files of conversations related to the CTPT gathered by the Alzheimer and Related Dementias Study at the University of Pittsburgh School of Medicine. The dataset contains 510 conversations divided into two categories: conversations done by healthy patients (Control Group) and by patients with dementia (Dementia Group). Some metadata was recorded for each patient visit: age, sex, level of education, race, confirmed diagnosis, MMSE score, the audio recording, and the transcript (recorded using the CHAT format [15]) of the interview. Unfortunately, some of this information is missing for a small number of patients, so we decided to remove these conversations, bringing our dataset to a total of 459 interviews. The data were divided into 182 samples from the Control group (negative samples) and 277 positive samples from the dementia group (positive samples). Around 70% of

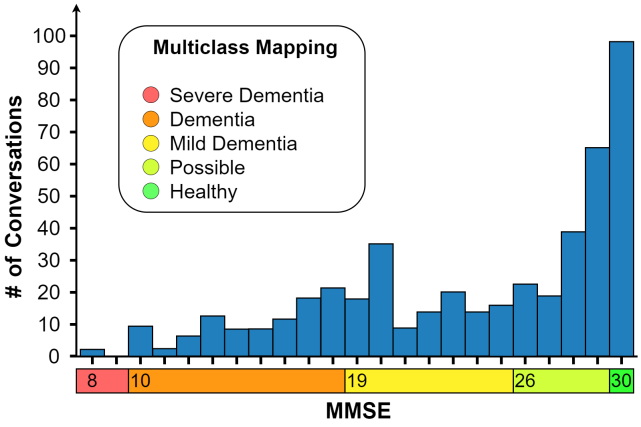

Fig. 2. Distribution of conversations over the MMSE score scale, coupled with the mapping between MMSE scores and dementia severity levels. The MMSE scale stops at 8 given that no conversation in the DementiaBank had a score less than that.

the conversations have an MMSE score between 26 to 30, and we did not have any samples with a score below 8 (see Figure 2). Since our work focused on the early detection of dementia, the infrequency of examples in the lower portion of the scale did not present a strong limitation for our study.

Most of the researchers in this domain worked on two prediction tasks using this dataset: a binary classification task, which tries to predict if the patient has dementia or not, and a score regression task, which tries to predict the MMSE score associated with the patient. The binary classification task has been the subject of many more research studies than the other, so we decided to focus on the score regression task and propose an extension of the binary classification task in the form of a multiclass classification problem. This new task is based on a mapping between MMSE scores and dementia stages validated by the work of Perneczky et al. [16]. Instead of predicting a simple binary outcome, this task will provide much more information to the user, not only classifying the patient as healthy or with dementia, but also offering a more precise estimate of the severity of the disease. In Figure 2 can be observed both the distribution of patients across the MMSE score and the mapping between MMSE score and dementia severity.

Graph2Vec does not handle direct classification or regression; rather, the generated embeddings are used as input for ML models, which are trained to make the final prediction. This gives us the opportunity to work within a highly controlled environment, facilitating the systematic comparison of the traditional ML approach to our GML approach. In particular, we not only use the same models with the ML baseline and the Graph2Vec embeddings, but also the same method for fine-tuning hyperparameters, and the exact same testing pipeline (more details are provided in Section IV).

IV. FEATURES AND GRAPH ONTOLOGY

As discussed in Section II, a great variety of features and feature groups have been proposed for early DD. To define our set of features we mainly followed two criteria: all of

our final features must be widely used in the state of the art, and no one of those must be extracted using anything else other than the audio or plain transcript of the conversation (for example, we excluded the added information provided in the CHAT-based transcription of the DementiaBank dataset). This decision was taken in order to ensure a high-quality set of features, while at the same time making it possible to directly replicate our approach with any other CTPT interview-based dataset.

We ended up with 6 final sets of features: acoustic, anagraphic, discourse-based, lexicosyntactic, psycholinguistic, and spatial. All features can be extracted using the feature extraction framework detailed in our previous work [17]. The *Acoustic group* used the Mel-frequency Cepstral Coefficients (MFCCs) which are an audio feature generally used in speech processing designed to approximate the human auditory system's response to sound. The *Anagraphic group* contains part of the metadata discussed in Section III: age, sex, level of education, and race. The *Discourse-based group* describes the type distribution of Elementary Discourse Units based on the Rhetorical Structure Theory [18]. The *Lexicosyntactic group* has features focused on patterns associated with patients' vocabulary complexity and sentence construction. The *Psycholinguistic group* contains features grounded in linguistic theory, specifically regarding human language behavior like familiarity, concreteness, imageability, age of acquisition, and mean word frequency in the English language. The *Spatial group* is based on the work of Croisile et al. [19], which tries to detect the presence of spatial-visual impairment in a patient by checking the presence or absence of 20 Information Units (specific actions, objects, or people present in the picture).

These feature groups can be imagined as different perspectives of the same conversation, where each feature is related to other features in specific ways. Thus, to retain and emphasize these relationships, we organize them all into a graph structure. Each conversation is represented by a separate graph with different feature values corresponding to that specific conversation, but all graphs have the same structure. Every graph has a central node from which each feature group branch starts, and each node inside the graph is associated with 2 labels: a *generic* label and a *specific* one. The generic label describes the node type from among four categories:

- *Branch type*, which represents the first node in a feature group branch.
- *Category*, which represents the first high-level division into different feature sub-groups.
- *Sub-Category*, which represents all the non-feature nodes used to further diversify the feature sub-groups.
- *Feature*, which represents all the "leaf nodes" that contain a numerical feature.

The specific label on the other hand is a unique label used to identify each node. At the very end of each branch, we have *Feature* nodes. These are the only nodes that have an associated numerical value, representing the feature itself.

As a case example, we summarize the organization of the spatial branch here (Figure 3), where each color represents a node general label (as explained in the legend), and its name represents the specific label. To visualize the graph ontology in its entirety, we recommend looking at the ontology schema contained in our repository[2], which was used to generate these graphs.

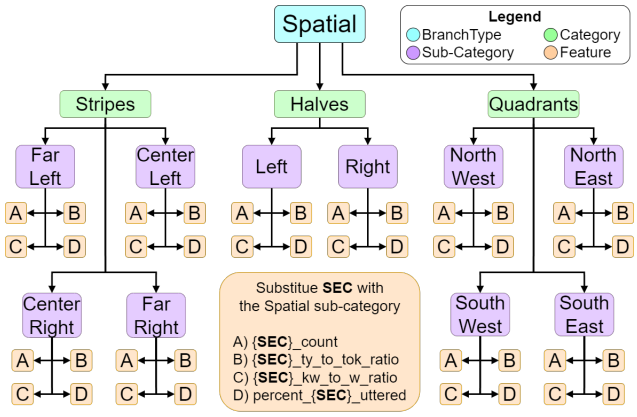

Fig. 3. Visualization of the Spatial branch full ontology.

V. EXPERIMENTAL EVALUATION

A. Experiments Design

To ensure a fair comparison between the ML approach and the GML one, we designed our evaluation workflow in a way that minimizes the number of differences between those two. The only difference between these two approaches is that after completing the feature extraction, the ML approach immediately uses the features to start training and testing, while the GML approach first generates the corresponding graph, then uses Graph2Vec to generate the embeddings, and in the end uses them for training and testing. In particular, we used for classification and regression the following models for both approaches: Support Vector Machine, Random Forest, Gradient Boosting, and Bagged Trees. In addition, we used Logistic Regression and Gaussian Naive Bayes only for the classification task, while we added Ridge Regression for the regression task. All hyperparameters for these models were automatically fine-tuned using an automatic grid search in order to ensure a fair comparison between ML and GML. This is because grid search guarantees that all models are tuned trying the same possible combinations of hyperparameters. When fine-tuning was complete, we used five-fold cross-validation in order to train and test our models, ensuring consistent results. The performance of each algorithm was evaluated using a set of standardized classification and regression metrics that are commonly used across literature pertaining to automated DD detection. Specifically, we measure accuracy, precision, recall, and F1 score for the classification task, and we use root mean squared error (RMSE), mean absolute error

[2] github.com/EdoStoppa/dementiaGraphGenerator

(MAE), and R^2 score for the regression task. The code for the entire evaluation pipeline can be found in our repository[3].

B. Experimental Results

The best model associated with the traditional ML approach for the Multiclass Classification task is Gradient Boosting, while the best model for the GML approach is Random Forest. We macro-averaged the precision, recall, and F1 scores across classes to facilitate a simple yet precise comparison. The GML approach shows a clear increase in performance in every collected metric. We observe a relative increase in accuracy of 4.4%, in precision of 5.7%, in recall of 7.9%, and in F1 score of 6.9% (as shown in Table I). Unfortunately, we can also see that the standard deviation of all metrics increased.

TABLE I

MULTICLASS CLASSIFICATION PERFORMANCE COMPARISON

	Machine Learning (Gradient Boosting)	Graph Machine Learning (Random Forest)
Accuracy	0.473 ± 0.038	**0.494** ± 0.053
Precision	0.489 ± 0.042	**0.517** ± 0.047
Recall	0.464 ± 0.033	**0.501** ± 0.061
F1 Score	0.476 ± 0.037	**0.509** ± 0.055

The best model associated with the traditional ML approach for the Score Regression task is Gradient Boosting, while the best model for the GML approach is Ridge Regression. The GML approach shows, once again, a definite improvement in performance across all metrics. We observe a relative decrease in MAE of 3.4%, in RMSE of 2.9%, and an increase in R^2 score of 8% (as shown in Table II), while also obtaining a lower standard deviation in all metrics due to the different nature of the task.

TABLE II

SCORE REGRESSION PERFORMANCE COMPARISON

	Machine Learning (Gradient Boosting)	Graph Machine Learning (Ridge Regression)
MAE	3.405 ± 0.269	**3.291** ± 0.201
RMSE	4.263 ± 0.277	**4.139** ± 0.249
R^2	0.436 ± 0.071	**0.471** ± 0.059

VI. CONCLUSIONS

In this work, we introduced a graph ontology that organizes a standard set of features extracted through NLP from CTPT-related conversations, explicitly encoding the relations among these different features. We generated embeddings for these graphs using Graph2Vec, in order to use them as input for multiple ML models. We introduced a new prediction multiclass classification task that extends the traditional binary classification task for DD, and also evaluated our graph-based approach on a score regression task. We ultimately found that our approach, based on adding semantic structure to the features by organizing them in a graph and combining these graphs with GML techniques, outperformed the ML approach for both the score regression and multiclass classification tasks across all recorded metrics.

Although our GML approach clearly demonstrated its strength in our evaluation, many opportunities remain for future work. For example, we could design a more complex GML architecture instead of relying only on a single algorithm to generate the embeddings. Furthermore, we also plan to explore different and more complex ontologies in order to increase the final performance of our models.

[3] github.com/EdoStoppa/multitaskBaselinePerformance

REFERENCES

[1] J. Ortman, V. Velkoff, and H. Hogan, An aging nation: The older population in the united states, Current population reports 2014. US Census Bureau, Jan 2014.

[2] C. C. Brück, F. J. Wolters, M. A. Ikram, and I. M.C.M. de Kok, Projected prevalence and incidence of dementia accounting for secular trends and birth cohort effects: a population-based microsimulation study, Eur J Epidemiol, vol. 37, pages 807-814, Aug 2022.

[3] S. G. Gauthier, Alzheimer's disease: the benefits of early treatment, European Journal of Neurology, vol 12, pages 11–16, 2005.

[4] B. Mirheidari, D. Blackburn, T. Walker, M. Reuber, and H. Christensen, Dementia detection using automatic analysis of conversations, Computer Speech & Language, vol. 53, pages 65–79, 2019.

[5] K. P. Ng, H. J. Chiew, L. Lim, P. Rosa-Neto, N. Kandiah, and S. Gauthier, The influence of language and culture on cognitive assessment tools in the diagnosis of early cognitive impairment and dementia, Expert Review of Neurotherapeutics, vol. 18, pages 859–869, 2018.

[6] M. F. Folstein, L. N. Robins, and J. E. Helzer, The mini-mental state examination, Archives of General Psychiatry, vol. 40, Jul 1983.

[7] The Boston Diagnostic Aphasia Examination, Philadelphia, Pa, Lippincott Williams & Wilkins, 3rd edition, 2001.

[8] S. Ahmed, A.-M. Haigh, C. Jager, and P. Garrard, Connected speech as a marker of disease progression in autopsy-proven Alzheimer's disease, Brain: a journal of neurology, vol. 136, Oct 2013.

[9] K. Fraser, J. Meltzer, and F. Rudzicz, Linguistic features identify Alzheimer's disease in narrative speech, Journal of Alzheimer's disease: JAD, vol. 49, Oct 2015.

[10] S. Farzana and N. Parde, Exploring MMSE score prediction using verbal and non-verbal cues, In INTERSPEECH, 2020.

[11] A. Narayanan, M. Chandramohan, R. Venkatesan, L. Chen, Y. Liu, and S. Jaiswal, Graph2Vec: Learning distributed representations of graphs, 2017.

[12] W. Li, J. Xue, X. Zhang, H. Chen, Z. Chen, and Y. Cai, Word-Graph2vec: An efficient word embedding approach on word co-occurrence graph using random walk sampling, 2023.

[13] Z. Yuan, Q. Yuan, and J. Wu, Phishing Detection on Ethereum via Learning Representation of Transaction Subgraphs, Blockchain and Trustworthy Systems (BlockSys), 2020.

[14] J. T. Becker, F. Boller, O. L. Lopez, J. Saxton, and K. L. McGonigle, The natural history of Alzheimer's disease: description of study cohort and accuracy of diagnosis, Archives of Neurology, 1994.

[15] B. MacWhinney, The CHILDES Project: Tools for Analyzing Talk, Child Language Teaching and Therapy, Jan 2000.

[16] R. Perneczky, S. Wagenpfeil, K. Komossa, T. Grimmer, J. Diehl, and A. Kurz, Mapping scores onto stages: Mini-mental state examination and clinical dementia rating, The American Journal of Geriatric Psychiatry, vol. 14, pages 139–144, 2006.

[17] E. Stoppa,G. W. Di Donato, N. Parde, and M. D. Santambrogio, Computer-aided dementia detection: How informative are your features?, 2022 IEEE 7th Forum on Research and Technologies for Society and Industry Innovation (RTSI), pages 55–61, 2022.

[18] W. Mann and S. Thompson, Rhetorical structure theory: A theory of text organization, Jan 1987.

[19] B. Croisile, B. Ska, M.-J. Brabant, A. Duchene, Y. Lepage, G. Aimard, and M. Trillet, Comparative study of oral and written picture description in patients with Alzheimer's disease, Brain and Language, vol. 53, pages 1–19, 1996.