Artificial Intelligence In Medicine 117 (2021) 102097

Contents lists available at ScienceDirect

Artificial Intelligence In Medicine

journal homepage: www.elsevier.com/locate/artmed

Simultaneous imputation and classification using Multigraph Geometric Matrix Completion (MGMC): Application to neurodegenerative disease classification☆

Gerome Vivar [a,b], Anees Kazi [a], Hendrik Burwinkel [a], Andreas Zwergal [b], Nassir Navab [a], Seyed-Ahmad Ahmadi [a,b,*], for the Parkinson's Progression Markers and Alzheimer's Disease Neuroimaging Initiatives[1]

[a] Department of Computer Aided Medical Procedures (CAMP), Technical University of Munich (TUM), Boltzmannstr. 3, 85748 Garching, Germany
[b] German Center for Vertigo and Balance Disorders (DSGZ), Ludwig-Maximilians University (LMU), Fraunhoferstr. 20, 82152, Planegg, Germany

ARTICLE INFO

Keywords:
Computer-aided diagnosis
CADx
Deep learning
Multimodal medical data
Population-based studies

ABSTRACT

Large-scale population-based studies in medicine are a key resource towards better diagnosis, monitoring, and treatment of diseases. They also serve as enablers of clinical decision support systems, in particular computer-aided diagnosis (CADx) using machine learning (ML). Numerous ML approaches for CADx have been proposed in literature. However, these approaches assume feature-complete data, which is often not the case in clinical data. To account for missing data, incomplete data samples are either removed or imputed, which could lead to data bias and may negatively affect classification performance. As a solution, we propose an end-to-end learning of imputation and disease prediction of incomplete medical datasets via Multi-graph Geometric Matrix Completion (MGMC). MGMC uses multiple recurrent graph convolutional networks, where each graph represents an independent population model based on a key clinical meta-feature like age, sex, or cognitive function. Graph signal aggregation from local patient neighborhoods, combined with multi-graph signal fusion via self-attention, has a regularizing effect on both matrix reconstruction and classification performance. Our proposed approach is able to impute class relevant features as well as perform accurate and robust classification on two publicly available medical datasets. We empirically show the superiority of our proposed approach in terms of classification and imputation performance when compared with state-of-the-art approaches. MGMC enables disease prediction in multimodal and incomplete medical datasets. These findings could serve as baseline for future CADx approaches which utilize incomplete datasets.

1. Introduction

Large population-based studies in medicine, acquired at multiple institutions, are instrumental resources for a better clinical understanding of the diagnosis, progression and treatment of diseases. In medical health informatics, they serve as fundamental enablers for the design and analysis of novel clinical decision support systems (CDSS) and CADx [1]. Often, such datasets incorporate multimodal data, both imaging and non-imaging, in order to capture as many aspects of the disease as possible.

☆ This work was supported by the German Federal Ministry of Education and Health (BMBF) in connection with the foundation of the German Center for Vertigo and Balance Disorders (DSGZ) [grant 01 EO 0901], with partial support of "Freunde und Förderer der Augenklinik München", Germany.

* Corresponding author at: German Center for Vertigo and Balance Disorders (DSGZ), Ludwig-Maximilians-University Munich, Fraunhoferstr. 20, 82152, Planegg, Germany.
E-mail address: ahmadi@cs.tum.edu (S.-A. Ahmadi).

[1] Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf. Further data were obtained from the Parkinson's Progression Markers Initiative (PPMI) database www.ppmi-info.org/data. List of all PPMI funding partners can be found at www.ppmi-info.org/fundingpartners.

https://doi.org/10.1016/j.artmed.2021.102097
Received 16 November 2020; Received in revised form 4 May 2021; Accepted 5 May 2021
Available online 8 May 2021

Two prominent examples for such datasets in neurology and neuroscience were published by the Alzheimer's Disease (AD) Neuroimaging Initiative (ADNI) [2] and the Parkinson's disease (PD) Progressive Marker Initiative (PPMI) [3]. Together, AD and PD are the most common neurodegenerative diseases, with AD accounting for 60–80% of dementia cases, and PD affecting 1–2% of the global population over the age of 65. Neurodegenerative diseases result in a progressive decay and death of nerve cells [4]. Increasing rates of up to a million new AD cases per year [4], along with the prospect of novel models and care frameworks for dementia [5] as well as novel neuroprotective and disease-modifying therapeutics, in both AD and PD [6], motivate an early diagnosis of these diseases, ideally already at a pre-symptomatic stage.

Population-based datasets in medicine are often feature-incomplete, due to missing examinations of patients. Most ML-based CADx approaches require imputation before classification [7], and treat these steps sequentially and independently. Incomplete features are categorized into missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR), with MAR often lying at the basis of most modern imputation methods [8].

Related works: A recent review paper [9] on ML techniques for AD diagnosis has found that most recent methods treated multimodal feature modeling and classification separately, with a focus on the former. In addition, they suggested that more work is required in multimodal ML methods towards early AD diagnosis. In line with these findings, our proposed method addresses multimodal CADx for AD, with simultaneous feature imputation and classification.

Two commonly used methods to treat missing values in datasets are sample deletion or Mean-imputation, which either result in expensive loss of data or in biased and sub-optimal features. More advanced methods use multiple imputation or ML. Hedge et al. [7] compared Multiple Imputation Using Chained Equations (MICE) [10] with Probabilistic Principal Component Analysis (PPCA) on healthcare data, and found PPCA to be superior. A fundamentally different approach is matrix completion. Thung et al. [11,12] use Low-Rank Matrix Completion (LRMC) to predict conversion of the disease in patients with Mild Cognitive Impairment (MCI) to Alzheimer's Disease (AD). Zhou et al. [13] proposed to solve AD diagnosis using latent representation learning, by projecting both complete and incomplete modalities onto a common subspace. Both approaches by [12] and [13] assume a linear relationship between the input features and the target variable, and latent embeddings and linear classification are trained in two separate steps [13], which does not take advantage of end-to-end learning.

Recently, graph convolutional networks (GCN) have been introduced for CADx on multimodal medical datasets. Parisot et al. [14] introduced a novel concept for modeling patient populations as a graph: patient meta-data like demographics (e.g. sex, age, etc.) are used to compute similarities between patients, leading to an adjacency matrix with an associated graph Laplacian. Intuitively, the graph is akin to a "social network" of patients in the cohort. Several works since then have demonstrated that GCNs can significantly improve the accuracy of CADx in medicine [15–19]. Importantly, the graph definition crucially affects the CADx accuracy, and we have shown previously that parallel multi-graph models with attention, i.e. one graph for each meta-feature, can make GCNs more robust [16,17].

Importantly, like most other ML methods, GCNs assume feature-completeness and depend on imputation as a pre-processing step. Regarding incomplete datasets, Monti et al. [20] showed that geometric deep learning provides a principled framework for non-linear imputation, through geometric matrix completion (GMC). In our own previous work [21], we extended upon this work through multi-target training, which combined GMC with supervised classification, into a Recurrent Graph Convolutional Network (RGCN). Similar to [15], we constructed a patient graph from clinical meta-data (e.g. age and sex of patients). We concatenated the incomplete feature matrix and incomplete labels, and trained a GCN for signal diffusion, along with a Long-Short Term Memory (LSTM) network for iterative matrix reconstruction. Both GCN and LSTM were combined into a single-graph RGCN, which was trained end-to-end towards MCI to AD conversion prediction, with two weighted losses for simultaneous classification and imputation.

Contribution: We propose to solve disease classification in multimodal and incomplete datasets using Multi-graph Geometric Matrix Completion (MGMC). The contributions of this work are threefold: (1) we formulate the disease classification problem in multimodal and incomplete datasets using MGMC; (2) we propose a novel method which uses multiple non-autoregressive Recurrent Graph Convolutional Networks (RGCN) and a transformer-inspired self-attention mechanism for multi-graph fusion; (3) we validate the superiority of the proposed approach on two publicly available medical datasets and evaluate the effect of autoregressive LSTMs on MGMC architectures.

2. Materials and methods

We first introduce the notation used throughout the rest of the paper in Table 1, then elaborate on key background information in order to provide more context on our proposed approach.

2.1. Dataset and preprocessing

We used two publicly available datasets in this work: The Alzheimer's Disease Prediction Of Longitudinal Evolution (TADPOLE) [2] obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu) and the Parkinson's Progressive Marker Initiative (PPMI) dataset [3]. TADPOLE requires classification of subjects into three categories, normal control (NC), mild cognitive impairement (MCI), and Alzheimer's disease (AD). PPMI requires detection of Parkinson's disease (PD) vs. normal controls (NC).

In TADPOLE, we used 813 subjects coming from the ADNI protocol with 229 NC, 396 MCI and 188 AD diagnosed at baseline. This dataset contains pre-processed features [2] from cerebro-spinal fluid (CSF) markers, magnetic resonance imaging (MRI), positron emission tomography FDG (PET), diffusion tensor imaging (DTI), cognitive assessment scores, genetic information such as alipoprotein E4 (APOE4), and demographic information. Further pre-processing entailed a normalization of real-valued TADPOLE features to zero-mean and unit-variance. To match the classification task, we selected only features at baseline, and excluded features containing longitudinal information. We further

Table 1
Description of notations.

Notation	Dimension	Description
$\mathbf{X}$	$n \times m$	Observed feature matrix with n samples and m features
$\mathbf{Y}$	$n \times c$	Class label matrix with n samples and c number of class
$\mathbf{Z}$	$n \times (m+c)$	Concatenated X and Y matrices
$\widehat{\mathbf{X}}$	$n \times m$	Predicted feature matrix X
$\widehat{\mathbf{Z}}$	$n \times (m+c)$	Predicted matrix Z
$\overline{\mathbf{Z}}$	$n \times (m+c)$	Predicted matrix Z from a single RGCN
$\|\|.\|\|_{\mathrm{F}}^2$	–	Frobenius norm
$\|\|.\|\|_{\mathrm{D}}^2$	–	Dirichlet norm
$\mathcal{L}_{ce}(.)$	–	Cross-entropy loss
$\mathcal{L}_R(.)$	–	Reconstruction loss from GMC
$\boldsymbol{M}^{(i)}$	–	The ith meta-information
$\mathbf{M}$	–	Set containing $\{\boldsymbol{M}^{(1)}, \ldots, \boldsymbol{M}^{(I)}\}$
G_i	–	The ith graph constructed using meta-information $\boldsymbol{M}^{(i)}$
Ω_x, Ω_y	–	Denote whether input features and class lables, respectively, are known (1) or missing (0)
Θ, δ	–	Parameters from GCN and LSTM, respectively
$\gamma_{\{a,b,c\}}$	–	Hyper-parameters weighting loss terms
$\circ$	–	Hadamard product

removed features that were available for less than 10% of the available entries. In the end, the feature matrix had a dimensionality of 813 × 435, excluding label information.

In the PPMI dataset, we used all 75 healthy controls (HC) and 249 subjects with PD. PPMI data consists of brain MRI as well as non-imaging information such as Unified Parkinson's Disease Rating Scale (UPDRS), Montreal Cognitive Assessment (MoCA) scores, and demographic information (age and gender). The MRI information is used as input to the network while non-imaging information is used for the graph construction. As described in our previous GCN CADx approach [17], we pre-processed MRI volumes by co-registering each images to a normative space (SRI24 atlas [22]) to reduce variability in appearance, and further performed skull stripping using ROBEX [23]. Then we scaled each volume to an intensity range of [0,1]. Finally, to obtain a lower dimensional representation as input to the graph network, we used encoded raw image intensities coming from a 3D-autoencoder, which was pretrained towards anomaly detection. We refer the reader to [24] for a detailed discussion on the implementation of the pre-processing and 3D-autoencoder. The output at the bottleneck layer of the 3D-autoecoder was then used as the feature representation of the brain MRI volume.

Notably, our pre-processed PPMI dataset was 100% feature complete. In contrast, the TADPOLE dataset is inherently incomplete in native form, and was 83% feature-complete after our pre-processing pipeline. In the experimental section, we further removed known features artificially, to test classification and imputation robustness at various levels of data missingness. For better clarity throughout the rest of the paper, when denoting e.g. 50% data availability, we refer to the amount of data available at baseline (e.g. 50% for PPMI, and 41.5% for TADPOLE).

2.2. Graph construction

We use meta-information to construct separate graphs for each dataset. In the TADPOLE dataset, we use meta-information such as age, gender, and genetic risk factor (APOE4), all of which are known risk factors related to AD. For every given meta-information feature M, we calculate a separate graph using a pairwise similarity function. An edge between nodes i and j is defined using $W(i,j) = f(M(i), M(j))$ where

$$f(M(i), M(j)) = \begin{cases} 1 & \text{if } |(M(i) - M(j)| \leq \theta \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

$M(i)$ and $M(j)$ denote meta information of node i and j of a given meta-information M, and θ denotes a threshold value which is chosen empirically by the user, given domain expertise and depending on what can be regarded as a similar trait across patients [14,15].

To construct the graphs for the PPMI dataset, we use the same formulation in Eq. (1) and build graphs for every meta-information. Here we again use age and gender, along with two PD-related clinical scores of motor function (UPDRS) and of cognitive function (MoCA) to build the graph, following [17].

2.3. Geometric matrix completion

Consider an incomplete feature matrix $\mathrm{X} \in \mathbb{R}^{n \times m}$ where a certain proportion of values is missing at random. The goal is to recover the missing values in this matrix. One solution to this problem is by using rank minimization. However, as this is known to be computationally intractable, an alternative approximation is to constrain the predicted values to be smooth with respect to some geometric structure [25,26,20]. Here a graph structure is built based on the rows or columns of the matrix. Monti et al. [20] proposed to solve this using geometric deep learning on graphs, through a combination of GCN and LSTM networks. Compared to GMC recommender systems in [20], our CADx problem does not allow us to build a semantically meaningful column graph, especially since features stem from different modalities. Therefore, we modify the GMC approach to consider only a row graph derived from patient similarities to model the population. Nodes within a graph are the patient instances, their corresponding row vectors are the nodes' feature vectors, and the graph edges are based on patient similarities which are computed from meta-features, according to the metric in Eq. (1). Pair-wise similarities between nodes in the population graph connect patients that share the same risk-factor characteristics. The graph is then represented as $G = (V, E, \mathrm{W})$, with vertices $V = \{1, 2, \ldots, n\}$, and edges $E \subseteq V \times V$, which are weighted with non-negative weights. We represent the graph with a symmetric adjacency matrix $\mathrm{W} \in \mathbb{R}^{n \times n}$. The geometric matrix completion problem reduces to minimizing the loss:

$$\ell(\Theta, \delta) = ||\widehat{\mathrm{X}}_{\Theta,\delta}||_D^2 + \frac{\gamma}{2}||\mathbf{\Omega}_x \circ (\widehat{\mathrm{X}}_{\Theta,\delta} - \mathrm{X})||_{\mathrm{F}}^2 \tag{2}$$

where $\widehat{\mathrm{X}}_{\Theta,\delta}$ is the predicted matrix conditioned on the parameters of the GCN and LSTM, and $\circ$ denotes the Hadamard product. In Eq. (2), the first term on the right can be expressed as $\mathrm{tr}(\widehat{\mathrm{X}}^T \mathrm{L} \widehat{\mathrm{X}})$ [27] which contains a rescaled graph Laplacian ($\mathrm{L} \in \mathbb{R}^{n \times n}$) term such that its eigenvalues are in the interval $[-1, 1]$. This term keeps the prediction smooth with respect to the graph structure.

GMC can also be extended to multi-target training on heterogeneous matrix entries. Consider a matrix $\mathrm{Z} \in \mathbb{R}^{n \times (m+c)}$, which contains a mixture of feature and label information, which is implemented by concatenation of the feature matrix $\mathrm{X} \in \mathbb{R}^{n \times m}$ and class label matrix $\mathrm{Y} \in \mathbb{R}^{n \times c}$, similarly to Goldberg et al. [28]. Following Eq. (2), we can add a classification loss term on the imputed class label matrix [21]. The combined loss for completion of matrix Z is then:

$$\begin{aligned} \ell(\Theta, \delta) = &\frac{\gamma_a}{2}||\widehat{\mathrm{Z}}_{\Theta,\delta}||_D^2 + \frac{\gamma_b}{2}||\mathbf{\Omega}_x \circ (\widehat{\mathrm{Z}}_{\Theta,\delta} - \mathrm{Z})||_{\mathrm{F}}^2 \\ &+ \gamma_c(\mathcal{L}_{\mathrm{ce}}(\widehat{\mathrm{Z}}_{\Theta,\delta} \circ \mathbf{\Omega}_y, \mathrm{Z} \circ \mathbf{\Omega}_y)) \end{aligned} \tag{3}$$

where $\widehat{\mathrm{Z}}_{\Theta,\delta}$ is the predicted matrix containing predictions for both $\widehat{\mathrm{X}}$ and $\widehat{\mathrm{Y}}$.

2.4. Multigraph Geometric Matrix Completion

MGMC[2] consists of multiple non-autoregressive RGCNs and Transformer-like self-attention. We first describe the motivation why we use multiple RGCNs then elaborate on the self-attention inspired aggregation scheme including the use of non-autoregressive RGCNs. First, as we described in our previous works [16,17], the rules for constructing a population graph from a medical dataset are crucial to the accuracy of a GCN's downstream task, e.g. diagnostic classification accuracy. Instead of collapsing all meta-features into a single patient similarity measure, we therefore construct multiple graphs, one for each meta-feature. We then propose to integrate multi-graph GCNs into matrix completion by training a dedicated GCN and LSTM for each graph in an end-to-end manner. We do this to learn better imputed feature representations for each graph which could be useful in the downstream classification task.

To aggregate separate signals from parallel RGCNs, we use a self-attention aggregation mechanism inspired by Transformer networks called Scaled Dot-Product Attention [29]. We do this by training separate RGCNs (which consists of GCN and LSTM) in an end-to-end manner as shown in Fig. 1, then aggregate every RGCN outputs using the weights learned from the self-attention layer. We calculate self-attention weights for every RGCN by first stacking the outputs of RGCN ($\widehat{\mathrm{Z}}_{\Theta,\delta}^{(i)}$) into a tensor of size (BxMxF) where B is the full-batch-size, M denotes number of RGCNs, and F the dimensionality of the RGCN output. Weights for every

[2] Code: https://github.com/pydsgz/MGMC.

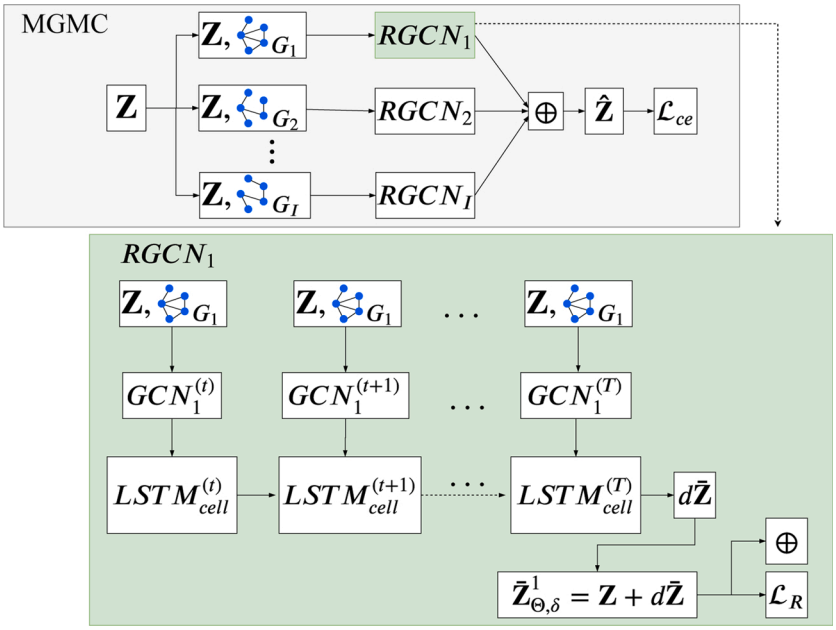

Fig. 1. Network architecture of MGMC which uses multiple Recurrent Graph Convolutional Network (RGCN) (top) including non-autoregressive RGCN layer (bottom). Information from a single RGCN branch will be aggregated ($\oplus$) together with the other outputs from other RGCN branches using a Scaled Dot-Product Attention mechanism. This output from a single RGCN is also used to calculate the reconstruction loss $\mathcal{L}_R$, which is the first term of the right-hand side of Eq. (5).

graph output are then calculated using Scaled-Dot-Product Attention [29]:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{4}$$

where Q, K and V are the linearly transformed outputs after stacking using learnable weight matrices (W_Q, W_K, W_V). In the end, this self-attention aggregation mechanism (denoted as $\oplus$ in Fig. 1) for outputs of every RGCN will yield an output $\widehat{\mathrm{Z}}$. Furthermore, we use muliple RGCNs, wherein each (unrolled) RGCN consists of a GCN and a non-autoregressive LSTM. Although multiple graphs and LSTMs have been used in previous methods ([20,19], one important difference of our proposed approach is the use of non-autoregressive LSTMs. As shown in Fig. 1, we only use the original input feature as input to the next timestep including the learned parameters from the previous LSTM cell-block. Such a non-autoregressive strategy is motivated in several ways. First, it limits the number of neighborhood hops and graph signal diffusion steps, as the input feature matrix to the GCN layer is the same at every time-step in the RGCN. Second, it allows the model to have better control on which graph-relevant information is useful for the imputation and downstream classification task. Third, by using the original input features as prior information at every optimization step, we reinforce the reconstruction of the input data, and prevent the model from diverging from the input data. As a result, this strategy prevents the model from suggesting non-realistic features as outputs. For the GCN layers, we use a Cheb-Net implementation [30,20]. This uses a Chebyshev polynomial basis ($\sum_{k=0}^{K} T_k(\widetilde{L})X\Theta_k$) to represent the spectral filters. For a more in-depth discussion regarding deep learning on graphs we refer the reader to [27]. The optimization loss for multi-graph GMC then boils down to minimizing the loss:

$$\ell(\Theta, \delta) = \sum_{i}^{M}\left(\frac{\gamma_a}{2}||\overline{\mathrm{Z}}^{(i)}_{\Theta,\delta}||^2_{D,r} + \frac{\gamma_b}{2}||\mathbf{\Omega}_x \circ (\overline{\mathrm{Z}}^{(i)}_{\Theta,\delta} - \mathrm{Z})||^2_{\mathrm{F}}\right) + \gamma_c(\mathcal{L}_{\mathrm{ce}}(\widehat{\mathrm{Z}}_{\Theta,\delta} \circ \mathbf{\Omega}_y, \mathrm{Z} \circ \mathbf{\Omega}_y)) \tag{5}$$

where $\overline{\mathrm{Z}}^{(i)}_{\Theta,\delta}$ is the ith predicted matrix from the ith graph (noting that this is conditioned on the parameters of the ith GCN and LSTM) and $\widehat{\mathrm{Z}}^{(i)}_{\Theta,\delta}$ is the aggregated predicted matrix coming from all GCNs and LSTMs.

3. Results

3.1. Implementation details

We used a 10-fold stratified cross-validation strategy to split the dataset into 10% test and 90% train (of which 10% as validation set) on all methods. For all deep learning based methods we use Adam optimization [31], with implementations in PyTorch [32], on a workstation with a single GPU (Nvidia GTX 1080 Ti). We automatically determine hyperparameters in Eq. (5) using hyperparameter optimization on the validation set with 120 iterations [33], with the following search spaces for the Chebyshev Polynomial parameters ($K \in \text{range}(1, 20)$), learning rate = $\text{uniform}([0.00001, 0.1])$, intermediate layers' hidden units $\in \text{range}(8, 512)$, and $\gamma_{(a,b,c)} = \text{uniform}([0.001, 1000])$.

We compared the proposed method with shallow learning methods in machine learning, gradient-based Matrix Completion (MC), and state-of-the-art (SOTA) graph-based methods which have shown to be highly effective for disease prediction. For shallow learning, we used Logistic Regression (LR) as the linear baseline, and Random Forest (RF) [34] as a competitive non-linear baseline. We also compared against MC which is a simple non-graph-based gradient based matrix completion approach. Previous graph-based methods included GCNs [14,15], GMC [21], and MG-RGCN [19]. As several algorithms (LR, RF and GCN) assume feature-completeness, we first need to impute the missing values in the feature matrix. We used five approaches to accomplish this: the commonly used Mean-imputation method, kNN imputation [35], MICE with linear regression (MICE_LR) [10], MICE with random regression forest (MICE_RF) [10], and PPCA [36]. For GCN, we use the empirically best-performing imputer. To test imputation performance, we artificially reduce the percentage of known data in the ADNI/PPMI feature matrices and perform imputation/classification at {100,75,50,25}% data availability (MAR assumption [8]). At each percentage level, we report the worst and the best performance for each imputer + classifier combination, to give an indication of the spread of possible outcomes. To report and compare classification outcomes, we visualize the three metrics Accuracy/F-measure/ROC-AUC in Fig. 2, and compare them

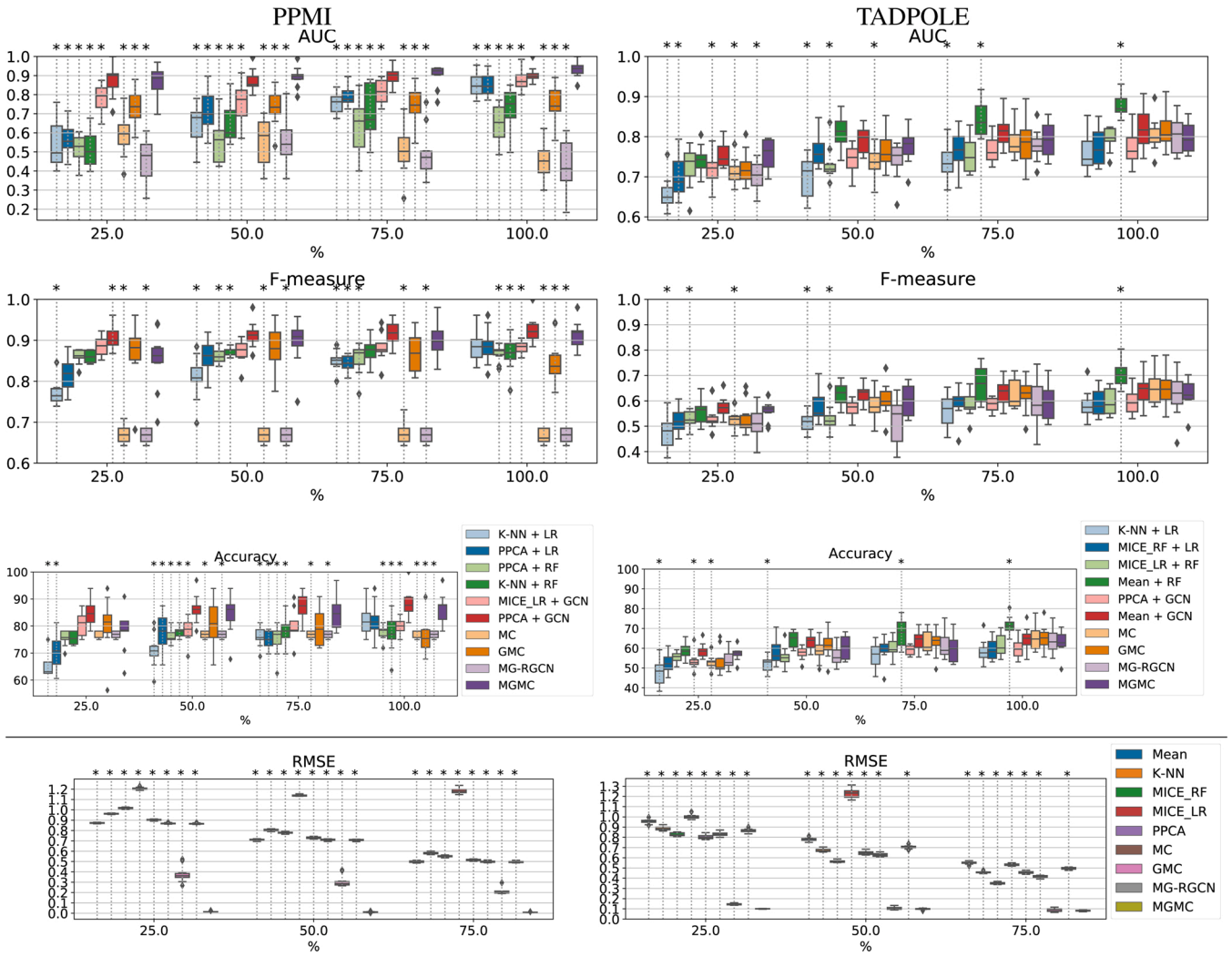

Fig. 2. PPMI (left panel) and TADPOLE (right panel) classification results (boxplots indicate the distribution of metrics over the 10 folds for each model): ROC-AUC (first-row), *F*-measure (second-row) and Accuracy (third-row). Imputation results (fourth-row) for PPMI and TADPOLE. Asterisk symbols (*) and dotted vertical lines denote that the tested model is statistically significantly different (two-tailed Wilcoxon rank-sum test, $p \leq 0.05$) to our proposed model (MGMC). *X*-axis values denote the percentage of available/known features prior to imputation and model training. (Best viewed in digital format).

quantitatively via a two-tailed Wilcoxon rank-sum hypothesis test, at an alpha-level of $p \leq 0.05$).

We use Scikit-learn [37] implementations for cross-validation, pre-processing, imputation (PPCA [38]) and shallow classifier models (LR and RF). To make baseline algorithms as competitive as possible, we also perform hyperparameter optimization (also 120 max. iterations) for the standard machine learning models (i.e. LR and RF) [39]. We concatenate the meta-features (e.g. demographics) with the feature vectors for all baseline methods, to further ensure fairness, as our proposed graph-based method utilizes this information as well (i.e. for graph construction).

3.2. *PPMI and TADPOLE dataset results*

We plot classification (Fig. 2 rows 1–3) and imputation (Fig. 2 row 4) results on the PPMI dataset (Fig. 2 left panels) and on the TADPOLE dataset (Fig. 2 right panels).

In PPMI, the classification metrics show that our proposed MGMC method is consistently among the top-performing methods. In terms of ROC-AUC and Accuracy (cf. Fig. 2, rows 1 and 3), MGMC is often significantly better than other classifiers, at all levels of data availability. In the following, we will describe our results by focusing mainly on the aggregate metric *F*-measure, as it reflects the harmonic mean between precision and recall and is therefore better suitable to asses the classification of rare positives (as it is often desired in medicine). In Fig. 2 (middle-left panel), we can see that the average *F*-measure over the 10 folds for MGMC stays consistently high at 0.852/0.897/0.904/0.913 (25/50/75/100% data availability, respectively). The only other method that performs comparably high is another graph deep learning method, GCN with PPCA-imputation. The difference is significant at 25% data availability (0.905, $p < 0.05$), but not at the other levels (0.914/0.918/0.926, $p > 0.05$). It is important to note that all algorithms that require prior imputation have a noticeable difference of performance, given the same amount of available data. For example, LR combined with kNN performs on average lower than when combined with PPCA, especially at 25% of data (*F*-measure difference: 0.044) and 50% of data (*F*-measure difference: 0.053). The best vs. worst imputation combination of imputer + classifier is not consistent across models: for RF, PPCA is on average worst, kNN is best, while for GCN, MICE_LR is worst and PPCA is best. Compared to our previously proposed GMC method, MGMC performs significantly better at 100% data availability (0.913 vs. 0.850, $p < 0.05$), not significantly better on average (not significant, $p > 0.05$) at 50% (0.896 vs. 0.881) and 75% (0.904 vs. 0.872) data availability, and not significantly worse at 25% (0.852 vs. 0.870, $p > 0.05$). Another striking result in PPMI is that the matrix completion methods MC and MG-RGCN more or less failed to learn a

good classification, at all levels of data availability (*F*-measure <0.7). The implications of this low performance will be discussed in Section 4.

In TADPOLE, compared to PPMI, the classification accuracy does not benefit as clearly from the population graph or imputation in our method. Similar to PPMI, the metrics ROC-AUC and Accuracy show some cases where MGMC is significantly better than other methods, notably at 25% and 50% data availability, and compared to LR or algorithms that are matched with the worst imputation method. For a further analysis, as in PPMI, we focus on the *F*-measure. As Fig. 2 (middle-right panel) shows, most classifiers perform in a similar range if matched with a suitable imputation method. As with PPMI, the choice of imputation method can have a noticeable effect though. Again, this choice is not consistent across classifier models. For LR/RF/GCN, the worst/best classifiers are kNN/MICE_RF, MICE_LR/Mean, and PPCA/Mean, respectively. A noteworthy performance is achieved by the combination of RF classifier with Mean-imputation. This combination achieves a significantly higher *F*-measure than MGMC (and all other methods) at 100% data availability (and a significantly higher ROC-AUC/Accuracy also at 75% data availability). However, RF paired with the worst imputer MICE_LR leads to a significantly worse performance for 25% and 50% data availability.

In terms of imputation quality (RMSE), Fig. 2 (bottom row) shows that in both PPMI and TADPOLE, our method imputes better (i.e. lower RMSE) than all other methods, and highly significantly ($p < 0.001$) in all comparisons, except when comparing to our previously proposed method GMC at 50% and 75% data availability. Among the other methods, the best-performing imputers for PPMI were Mean-imputation and the two matrix completion methods MC and MG-RGCN, while for TADPOLE, the best imputer was PPCA at 25%, and MICE_RF at 50% and 75% data availability. MICE_LR was the worst-performing data imputer in both datasets. Furthermore, the trend is visible that all imputation methods impute with higher RMSE errors as fewer data is available in the feature matrix, whereas our proposed MGMC method provides fairly robust imputation results.

In our ablation experiments, we investigated how non-autoregressive LSTMs affect the imputation and classification performance. In Fig. 3 top, we observe that for the PPMI dataset, the non-autoregressive model yields significantly better results in terms of ROC-AUC, *F*-measure, and Accuracy at all levels of data missingness. For the TADPOLE dataset (Fig. 3 bottom), the proposed method classifies comparably well at 50%, 75% and 100% data availability, but significantly outperforms the autoregressive model at 25% data availability, demonstrating better classification robustness at lower levels of data availability.

4. Discussion

4.1. Classification performance when using all available data

In PPMI, we observed that our proposed approach achieved a consistently high classification performance in terms of ROC-AUC, *F*-measure, and Accuracy for PD prediction when compared with standard ML models (LR and RF), MC, MG-RGCN and GMC approaches, as shown in Fig. 2 (left panel row 1–3). The only method that was able to perform equally well (and significantly better at 25% data availability) was GCN, when optimally paired with PPCA imputation. In TADPOLE, we observed that our approach is mostly at par with baseline ML methods and SOTA approaches from literature, and could only be significantly outperformed by RF and at 75–100% data availability, and only if RF was optimally paired with Mean-imputation. As mentioned in the dataset descriptions, PPMI is 100% feature-complete at baseline, whereas TADPOLE is only 83% complete at baseline. It is noteworthy that at 100% data availability, MGMC already performs imputation in TADPOLE, but we cannot validate the imputed values due to a lack of groundtruth data for those missing features. Compared to previous studies, Zhou et al. [13] reported ∼60% classification accuracy and ∼0.6 ROC-AUC for the same AD classification problem posed in this paper for the TADPOLE dataset. Gray et al. [40] reported ∼60% classification accuracy and ∼0.7 ROC-AUC. In our study, we also achieve a classification accuracy on the order of ∼60%, however with higher ROC-AUC values on the order of ∼0.8. To interpret these results, we recall that the Accuracy metric represents the number of true positive and true negative cases among the total population, at a fixed threshold of the model's posterior. In comparison, the ROC-AUC gives an estimate of the likelihood that a classifier simultaneously achieves a high true positive rate and low false positive rate. This indicates that MGMC, compared to related works, and compared to baseline models at 25% data availability, achieves a more robust classification outcome, not only in terms of sensitivity, but also in form of a lower likelihood for type I errors. A likely reason for the ROC-AUC difference of ∼0.1 compared to [40] is that earlier (2013) versions of the ADNI dataset had a smaller sample size, which also makes comparisons to our work somewhat unfair. Compared to [13], the ROC-AUC difference of ∼0.2 can be likely attributed to the use of multi-graph convolutions in our work, which are trained end-to-end in a semi-supervised manner.

4.2. Classification performance with artificially removed data

To investigate the robustness of MGMC and baseline methods with respect to missing data, we randomly reduced the amount of available data in the feature matrix relative to the number of observed entries at

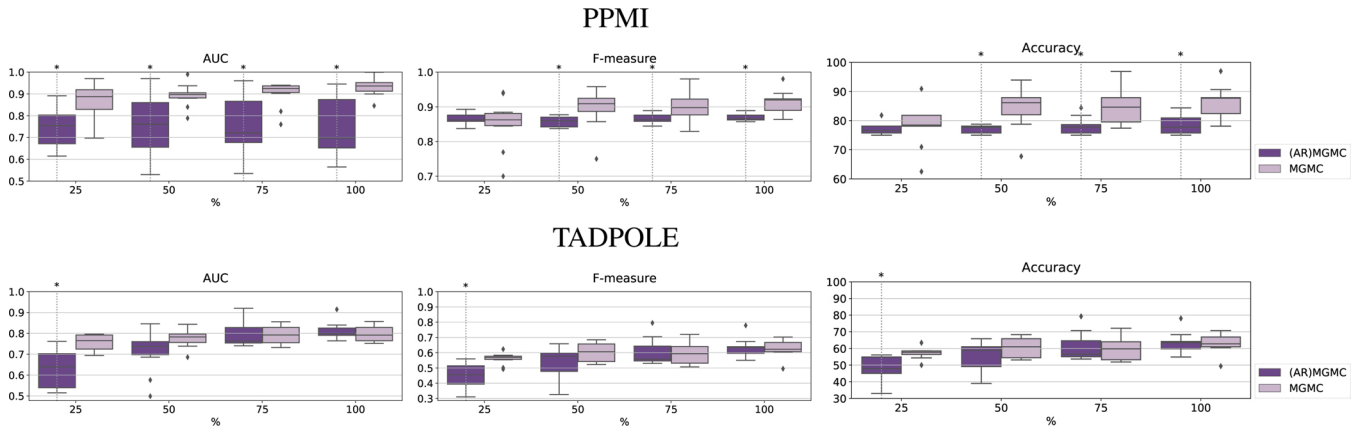

Fig. 3. PPMI (top) and TADPOLE (bottom) ablation results. ROC-AUC (left), *F*-measure (middle), and Accuracy (right) results on test dataset. Asterisk (*) and dotted vertical line denote model is statistically significantly different ($p \leq 0.05$) to proposed model. Values in *x*-axis denote relative percentage of features which are available to the network.

baseline, as shown in Fig. 2. We observed that the proposed approach has better and more stable classification and imputation results for PD prediction in PPMI when more information is missing. This effect is particularly visible in the ROC-AUC values, which may increase in standard deviation over the ten cross validation folds, but stay relatively stable in terms of median values above 0.9, even at low level of data availability around 25%. In comparison, LR, RF, MC, and MG-RGCN suffer from a noticeable drop in classification robustness. Interestingly, the single-graph GMC also yields relatively constant ROC-AUC values, but at a significantly lower level than MGMC. Furthermore, MC and MG-RGCN have an unstable and lower classification performance. This has two important implications. First, end-to-end learning of simultaneous imputation and classification, e.g. via geometric matrix completion, can improve the robustness of the CADx model towards the level of incompleteness in datasets up to a certain degree. Second, multiple RGCNs in parallel, e.g. fused by self-attention, improve both downstream tasks significantly, compared to using a single-graph or multiple graphs with a single RGCN. It is important to mention that we re-implemented MG-RGCN for comparison [19], as no reference implementation was available open-source. The on-par performance with many other algorithms on TADPOLE demonstrates a working re-implementation, however we have no clear explanation for the comparably low performance on PPMI. One factor that could partially contribute is that each graph in [19] utilizes a different feature set due to a graph-wise feature selection step as pre-processing. However, as none of the other algorithms in our comparison experiments used any sort of feature selection in the pre-processing stage, we also applied the full feature matrix to each branch of the RGCN, to make the comparison on same grounds. In MGMC, it is important to note that the two downstream tasks do not always benefit equally. In TADPOLE, for example, we observe a comparably stable classification performance at 75%, 50% and 25% data availability. However, a similar behaviour is observed for all other classifiers, and all classifiers in general classify similarly well. The only exception is the combination of Mean + RF where we observe a higher classification performance in TADPOLE. We hypothesize that one reason for this advantage could be due to the transductive imputation bias introduced in this model, since we performed imputation of the training set features together with the test set features. Another reason could also be the fact that we performed a hyperparameter tuning with nested cross-validation for all classifiers, including RF. For certain problems or datasets, apparently including TADPOLE, such hyperparameter optimization can achieve a noticeable performance boost, but not all translational studies of ML in medicine apply this step during their analyses. Only at 25–50% data availability, MGMC significantly outperforms other classifiers like GCN and LR, but only if these classifiers are matched with the worst-performing imputer (LR + kNN, and GCN + PPCA). As such, we consider this a negligible advantage for MGMC. Clearly, the main benefit of our proposed method on TADPOLE data lies not in an improved classification, but in a significantly more accurate imputation of missing values.

4.3. Joint classification and imputation performance

Most related literature in CADx naturally puts a focus on classification performance. Imputation is an often overlooked factor, even though it plays an important role in population-based and multimodal studies in medicine, as data missingness is a common problem here [41]. Considering the imputation performance in Fig. 2, our proposed approach is able to significantly outperform standard imputations (such as mean, kNN, and MICE, and PPCA) and other matrix completion approaches (MC and MG-RGCN) at all levels of missingness, on both datasets. This suggests that the proposed approach is able to take advantage of using known (semi-supervised) class label information in order to impute the features while simultaneously predicting the unknown class labels. It further suggests that the proposed method learns more class relevant feature representations compared to standard imputation approaches (mean, kNN, MICE, and PPCA) and other matrix completion methods (MC, MG-RGCN). We can also observe that population modeling and graph incorporation cannot always compensate for sub-optimal imputation, we would always have to find the right combination of imputer and classifier in order to achieve a comparable result with MGMC. Interestingly, even though MC and MG-RGCN also make use of the class label information, just like our GMC or MGMC approaches, their model performance did not significantly improve on both datasets. We hypothesize that this could be due to the feature representational capacity of MC and MG-RGCN. Additionally, MG-RGCN only makes use of a single RGCN which is autoregressive, just like GMC, and our experiments have shown that this could have a significant influence as can be seen in Fig. 3. One limitation to note is that we were not able to compare the imputation results to further matrix completion works in literature, e.g. [11–13], as those works do not report imputation fidelity, e.g. via RMSE. However, as a surrogate, we implemented an MC approach which is gradient-based and non-graph-based learning MC approach, and its results can serve as a stand-in for this family of methods. Furthermore, we can compare classification performance on TADPOLE data with [13], who used the same subjects (examinations at baseline) and classes (NC, MCI and AD) in TADPOLE as we did in our study. Here, authors explored classification performance of their proposed method, given 10% and 20% data missingness on either the MRI or SNP modality. As authors in [13] report, the results of our proposed approach are in line with their classification accuracy results at 20% data missingness (∼60% Accuracy) which corroborates our results on 75% data availability in Fig. 2 middle row. Finally, it is noteworthy that our proposed approach achieved a more accurate and stable classification performance for the PPMI prediction task than for the TADPOLE prediction task. A possible explanation is that distinguishing healthy controls from PD may be a simpler classification task than the three-class classification problem in TADPOLE (NC vs. MCI vs. AD). This notion is supported by clinical studies arguing that distinguishing NC, MCI, and AD based on clinical characteristics is a difficult problem at baseline [42].

4.4. Ablation experiments

In Section 2.4, we described our proposed improvements for usage of multiple RGCNs, specifically the usage of non-autoregressive LSTMs over autoregressive ones. Autoregressive RGCNs always use the output from the previous timestep and information from the previous LSTM cell-block as input. In contrast, non-autoregressive RGCNs always use the original input features as input at every timestep. Our motivation for using non-autoregressive LSTMs in MGMC is that the current output is always conditioned on the original input features. Intuitively, this should help the reconstructed output to avoid diverging from the input data, which is a desirable behaviour in matrix completion. Here, we perform and discuss an ablation experiment, where we compare the effect of both, as shown for PPMI and TADPOLE in Fig. 3. We observe that by using non-autoregressive LSTMs, we obtain a significantly better classification performance for all levels of data availability in PPMI. In TADPOLE, this tendency is not as clear, and a significant improvement is only achieved at 25% relative available data. At 50%, 75% and 100% available data, non-autoregressive LSTMs do not improve classification, but neither do they worsen the performance. This result suggests that it is indeed preferable to use non-autoregressive LSTMs in each parallel graph branch in MGMC. We attribute this to the intuitive notion explained above: by conditioning the reconstructed output on the original input data at every optimization timestep, we stabilize the reconstruction and achieve a better classification performance.

4.5. Overall implications

The main differences of our proposed approach to recent works that use RGCNs for matrix completion [21,19,20] are three-fold, namely (i)

the use of multiple LSTMs which are non-autoregressive, (ii) the use of self-attention weighting to aggregate information from (iii) multiple graphs representing different neighborhood relationships between patients in the population. Previous RGCN/GMC methods [21,19] use a single LSTM, while in our approach, we utilize one separate LSTM for every graph, which results to multiple recurrent graph convolutional networks. Notably, Monti et al. [20] also use a multi-graph formulation, but their approach differs from our method, since they consider both the rows and columns of the feature matrix as two separate graph structures. Instead, in this work, we consider multiple meta-information as separate graphs that contain rows of a feature matrix as the node features, similarly to [17].

A general take-away from our experiments is that the best choice of the imputation method is apparently not really dependent on the data, but mostly depends on the classification algorithm following imputation instead. Almost every imputation method that we tested in this work (Mean, kNN, MICE_RF and PPCA) appeared either as the best or worst imputation method, depending on which data it was applied and in combination with which classification algorithm. Only MICE_LR was consistently a bad match, for any classifier, and the RMSE analyses revealed that it was probably due to a consistently bad imputation performance. Overall, the data under observation, the chosen imputation and the classification models together form a complex interplay, which makes a careful examination and benchmarking necessary. In translational ML works on medical data, e.g. for CADx, such exhaustive analyses are rarely made. This is probably due to the fact that an exhaustive testing of all possible combinations of classifiers and imputation methods can quickly lead to very large numbers of experiments. When adding hyper-parameter optimization for every possible combination (as we did in our experiments), the required computational effort for nested cross-validation and the evaluation of all model setups may become a challenge. It is precisely this variability that highlights the attractiveness of our proposed MGMC approach. Imputation and classification are learned end-to-end, in a single model. Although it is not guaranteed that MGMC always achieves the best classification performance, our experiments provide evidence that the imputation is significantly better in all settings, and the classification is top-ranking compared to a wide range of classification methods, both shallow and deep, both transductive and inductive, and using matrix completion or not.

Finally, our work has certain limitations, which may suggest interesting avenues for future contributions. Following [14,15], our graph construction heuristic assumes a simple static graph. Recently, it has been shown that is possible to learn a clinical population graph end-to-end, along with the classification downstream task [43]. The resulting graph is optimally suited for e.g. classification. Consequently, an alternative approach would be to use the meta-information and the feature matrix information in parallel to build or learn the graph adjacency. The advantages could be potentially several-fold: (i) the classification accuracy might benefit from a better graph, (ii) the robustness might increase even further, compared to our applied heuristics for graph construction, (iii) no domain expertise would be necessary to manually define the optimal thresholds θ (cf. Eq. (1)) that determine patient similarity and connectedness in the graph, and (iv) the learned graph might be an end in itself, and serve as a form of knowledge discovery in medicine (e.g. discovery of previously unknown, yet connected sub-populations) [43]. Both approaches could potentially lead to better performances of the downstream tasks (classification and imputation). Another limitation is that we benchmarked our proposed MGMC method to several baseline methods (LR and RF) which are all inductive learning approaches. In contrast, our approach is inherently transductive, as we rely on spectral graph convolutions in the parallel graph-convolutional layers. We believe that it should be possible to incorporate imputation losses into the objective functions of GraphSAGE [44] or GAT [45] to obtain an inductive form of MGMC, and it is worth investigating whether the same benefits can be observed as in our experiments. Furthermore, future works could compare against other non-deep learning based techniques that tackle missing data such as [46] and [47] and address non-MAR scenarios of missingness.

5. Conclusion

In conclusion, we propose a novel automatic disease classification method which can handle multimodal data with missing information, a common setup in medical population based studies and datasets. We accomplish this by using Multi-graph Geometric Matrix Completion (MGMC). We train our architecture through Multiple Recurrent Graph Convolutional Networks, which are optimized in an end-to-end manner. Experimental results suggest the effectiveness of our proposed approach on two well-known and challenging population based studies of neurodegenerative Parkinson's and Alzheimer's diseases. Furthermore, ablation experiments highlight the importance of using non-autoregressive LSTM including the effect of self-attention weighting. These results could serve as a baseline for future works on disease classification in incomplete datasets. In addition, this could be useful in other domains where incomplete, multimodal, and high-dimensional data is an issue.

Conflicts of interest

None declared.

Acknowledgment

Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Therapeutic Research Institute at the University of Southern California. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

References

[1] Dash S, Shakyawar SK, Sharma M, Kaushik S. Big data in healthcare: management, analysis and future prospects. J Big Data 2019 Dec;6:54.
[2] Marinescu RV, Oxtoby NP, Young AL, Bron EE, Toga AW, Weiner MW, et al. Tadpole challenge: prediction of longitudinal evolution in alzheimer's disease. 2018 (arXiv preprint), arXiv:1805.03909.
[3] Marek K, Jennings D, Lasch S, Siderowf A, Tanner C, Simuni T, et al. The parkinson progression marker initiative (ppmi). Prog Neurobiol 2011;95(4):629–35.
[4] Alzheimer's Association. 2019 Alzheimer's disease facts and figures. Alzheimer's Dement 2019;15(3):321–87.
[5] Koumakis L, Chatzaki C, Kazantzaki E, Maniadi E, Tsiknakis M. Dementia care frameworks and assistive technologies for their implementation: a review. IEEE Rev Biomed Eng 2019;12:4–18.

[6] Kim K-S. Toward neuroprotective treatments of Parkinson's disease. Proc Natl Acad Sci USA 2017 Apr;114:3795–7.
[7] Hegde H, Shimpi N, Panny A, Glurich I, Christie P, Acharya A. Mice vs ppca: missing data imputation in healthcare. Inform Med Unlocked 2019;17:100275.
[8] Van Buuren S. Flexible imputation of missing data. CRC Press; 2018.
[9] Tanveer M, Richhariya B, Khan R, Rashid A, Khanna P, Prasad M, et al. Machine learning techniques for the diagnosis of alzheimer's disease: a review. ACM Trans Multimed Comput Commun Appl (TOMM) 2020;16(1s):1–35.
[10] Buuren Sv, Groothuis-Oudshoorn K. MICE: multivariate imputation by chained equations in R. J Stat Softw 2010:1–68.
[11] Thung K-H, Adeli E, Yap P-T, Shen D. Stability-weighted matrix completion of incomplete multi-modal data for disease diagnosis. Intl. conf. on medical image computing and computer-assisted intervention 2016:88–96.
[12] Thung KH, Yap PT, Adeli E, Lee SW, Shen D. Conversion and time-to-conversion predictions of mild cognitive impairment using low-rank affinity pursuit denoising and matrix completion. Med Image Anal 2018;45:68–82.
[13] Zhou T, Liu M, Thung KH, Shen D. Latent representation learning for alzheimer's disease diagnosis with incomplete multi-modality neuroimaging and genetic data. IEEE Trans Med Imaging 2019;38(10):2411–22.
[14] Parisot S, Ktena SI, Ferrante E, Lee M, Moreno RG, Glocker B, et al. Spectral graph convolutions for population-based disease prediction. Intl. conf. on medical image computing and computer-assisted intervention 2017:177–85.
[15] Parisot S, Ktena SI, Ferrante E, Lee M, Guerrero R, Glocker B, et al. Disease prediction using graph convolutional networks: application to autism spectrum disorder and alzheimer's disease. Med Image Anal 2018;48:117–30.
[16] Kazi A, krishna S, Shekarforoush S, Kortuem K, Albarqouni S, Navab N. Self-attention equipped graph convolutions for disease prediction. 2019 IEEE 16th intl. symposium on biomedical imaging (ISBI 2019) 2019:1896–9.
[17] Kazi A, Shekarforoush S, Arvind Krishna S, Burwinkel H, Vivar G, Wiestler B, et al. Graph convolution based attention model for personalized disease prediction. In: Shen D, Liu T, Peters TM, Staib LH, Essert C, Zhou S, Yap P-T, Khan A, editors. Medical image computing and computer assisted intervention – MICCAI 2019. Springer Intl. Publishing; 2019. p. 122–30 (Cham).
[18] Kazi A, Shekarforoush S, Arvind Krishna S, Burwinkel H, Vivar G, Kortüm K, et al. InceptionGCN: receptive field aware graph convolutional network for disease prediction. Information processing in medical imaging, vol. 11492. Springer Intl. Publishing; 2019. p. 73–85.
[19] Valenchon J, Coates M. Multiple-graph recurrent graph convolutional neural network architectures for predicting disease outcomes. ICASSP 2019 – 2019 IEEE intl. conf. on acoustics, speech and signal processing (ICASSP) 2019:3157–61.
[20] Monti F, Bronstein MM, Bresson X. Geometric matrix completion with recurrent multi-graph neural networks. Proc. intl. conf. neural information processing systems (NeurIPS) 2017:3700–10.
[21] Vivar G, Zwergal A, Navab N, Ahmadi S-A. Multi-modal disease classification in incomplete datasets using geometric matrix completion. Graphs in biomedical image analysis (GRAIL), vol. 11044; 2018. p. 24–31.
[22] Rohlfing T, Zahr NM, Sullivan EV, Pfefferbaum A. The sri24 multichannel atlas of normal adult human brain structure. Hum Brain Mapp 2010;31(5):798–819.
[23] Iglesias JE, Liu C-Y, Thompson PM, Tu Z. Robust brain extraction across datasets and comparison with publicly available methods. IEEE Trans Med Imaging 2011;30 (9):1617–34.
[24] Baur C, Wiestler B, Albarqouni S, Navab N. Deep autoencoding models for unsupervised anomaly segmentation in brain mr images. Intl. MICCAI brainlesion workshop 2018:161–9.
[25] Rao N, Yu H-F, Ravikumar P, Dhillon IS. Collaborative filtering with graph information: consistency and scalable methods. Neural Inf Process Syst (NIPS) 2015:1–9.
[26] Kalofolias V, Bresson X, Bronstein M, Vandergheynst P. Matrix completion on graphs. 2014. arXiv:1408.1717.
[27] Bronstein MM, Bruna J, Lecun Y, Szlam A, Vandergheynst P. Geometric Deep Learning: going beyond Euclidean data. IEEE Signal Process Mag 2017;34(4): 18–42.
[28] Goldberg A, Recht B, Xu J, Nowak R, Zhu X. Transduction with matrix completion: three birds with one stone. Advances in neural information processing systems (NIPS). 2010. p. 757–65.
[29] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. Advances in neural information processing systems. 2017.
[30] Defferrard M, Bresson X, Vandergheynst P. Convolutional neural networks on graphs with fast localized spectral filtering. Advances in neural information processing systems (NIPS). 2016. p. 3844–52.
[31] Kingma DP, Ba J. Adam: a method for stochastic optimization. 2014 (arXiv preprint), arXiv:1412.6980.
[32] Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, et al. Pytorch: an imperative style, high-performance deep learning library. Advances in neural information processing systems. 2019. p. 8024–35.
[33] Bergstra J, Komer B, Eliasmith C, Yamins D, Cox DD. Hyperopt: a python library for model selection and hyperparameter optimization. Comput Sci Discov 2015;8(1): 014008.
[34] Breiman L. Random forests. Mach Learn 2001;45(1):5–32.
[35] Troyanskaya O, Cantor M, Sherlock G, Brown P, Hastie T, Tibshirani R, et al. Missing value estimation methods for dna microarrays. Bioinformatics 2001;17(6): 520–5.
[36] Tipping ME, Bishop CM. Probabilistic principal component analysis. J R Stat Soc: Ser B (Stat Methodol) 1999;61(3):611–22.
[37] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: machine learning in python. J Mach Learn Res 2011;12(Oct):2825–30.
[38] Stacklies W, Redestig H, Scholz M, Walther D, Selbig J. pcamethods – a bioconductor package providing pca methods for incomplete data. Bioinformatics 2007;23(9):1164–7.
[39] Komer B, Bergstra J, Eliasmith C. Hyperopt-sklearn: automatic hyperparameter configuration for scikit-learn. In: ICML workshop on AutoML, vol. 9; 2014.
[40] Gray KR, Aljabar P, Heckemann RA, Hammers A, Rueckert D, Initiative ADN, et al. Random forest-based similarity measures for multi-modal classification of alzheimer's disease. NeuroImage 2013;65:167–75.
[41] Little RJ, D'Agostino R, Cohen ML, Dickersin K, Emerson SS, Farrar JT, et al. The prevention and treatment of missing data in clinical trials. N Engl J Med 2012;367 (October):1355–60.
[42] Langa KM, Levine DA. The diagnosis and management of mild cognitive impairment: a clinical review. JAMA 2014;312(23):2551.
[43] Cosmo L, Kazi A, Ahmadi S-A, Navab N, Bronstein M. Latent-graph learning for disease prediction. Medical image computing and computer assisted intervention – MICCAI 2020, vol. 12262. Springer International Publishing; 2020. p. 643–53. Series Title: Lecture Notes in Computer Science.
[44] Hamilton W, Ying Z, Leskovec J. Inductive representation learning on large graphs. Advances in neural information processing systems. 2017. p. 1024–34.
[45] Veličković P, Cucurull G, Casanova A, Romero A, Liò P, Bengio Y. Graph attention networks. Intl. conf. on learning representations 2018.
[46] Wang G, Deng Z, Choi K-S. Tackling missing data in community health studies using additive ls-svm classifier. IEEE J Biomed Health Inform 2016;22(2):579–87.
[47] Venugopalan J, Chanani N, Maher K, Wang MD. Novel data imputation for multiple types of missing data in intensive care units. IEEE J Biomed Health Inform 2019;23(3):1243–50.