Disease Outcome Prediction Using Graph Auto-encoders

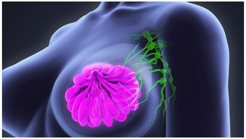
Juliette Valenchon

Department of Electrical and Computer Engineering, McGill University, Montreal
Under the supervision of Mark Coates

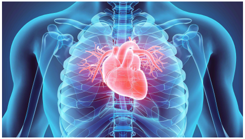
February 13, 2019

Motivation

Head and Neck Cancer[1]

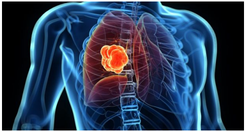


Breast Cancer[2]



Cardiovascular disease[3]

Colorectal Cancer[4]



Lung Cancer[5]

Alzheimer's Disease[6]

[1] Reproduced from "Head and Neck Cancer is not Just a Smoker's Disease Anymore", Mount Sinai News. [2] S. Roan, "Early Stage Breast Cancer: Do You Really Need Your Lymph Nodes Removed?", Everyday Health. [3] "Conquering Cardiovascular Disease", NIH. [4] "Colorectal Cancers", Dr. Fuhrman. [5] K. O'Sullican, "New drug approved for advanced lung cancer by HSE", The Irish Times. [6] M. Casalino, "Alzheimer's Association Offers Virtual Dementia Tour", Patch

Motivation

- Traditionally: risk calculator for possibility of disease development.
- Framingham study: prediction for hospitalization for long-term cardiovascular disease

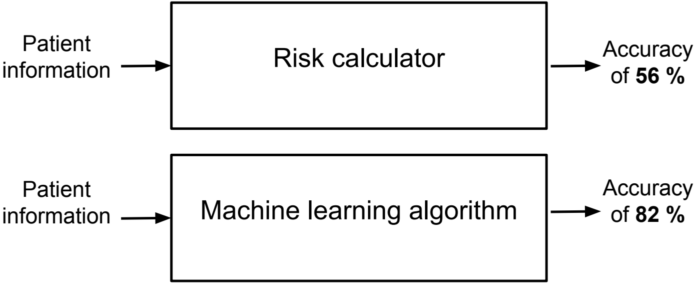
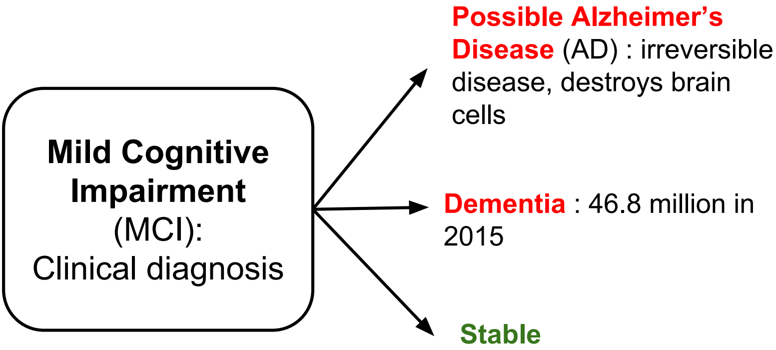


Figure 2: Comparison of a risk calculator and a machine learning algorithm[7]

[7] W. Dai et al., "Prediction of hospitalization due to heart diseases by supervised learning methods" Int. J. medical informatics, vol. 84, no. 3, pp. 189–197, 2015.

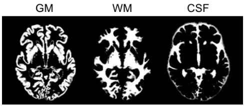
Motivation



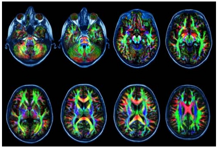
$\rightarrow$ **Early and accurate diagnosis** for an early treatment to improve the quality of life for some time

Goal

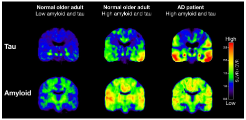
- Predict conversion from MCI to AD
- Multimodal data with missing values



(a) MRI[8]



(b) DTI[9]

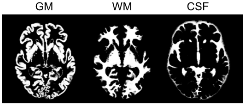


(c) PET[10]

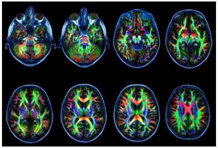
[8] Clinica developers, "Volume pre-processing - Clinica Documentation". [9] Rachel VanCott, "NOVA — scienceNOW — Diagnosing Damage image 3 — PBS". [10] University of California - Berkeley, "PET scans reveal key details of Alzheimer's protein growth in aging brains"

Goal

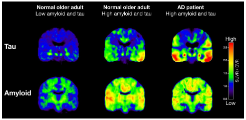
- Predict conversion from MCI to AD
- Multimodal data with missing values



(a) MRI[8]

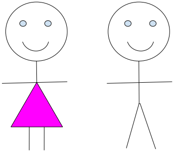


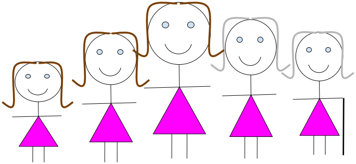
(b) DTI[9]



(c) PET[10]

- Use characteristics of subjects





[8] Clinica developers, "Volume pre-processing - Clinica Documentation". [9] Rachel VanCott, "NOVA — scienceNOW — Diagnosing Damage image 3 — PBS". [10] University of California - Berkeley, "PET scans reveal key details of Alzheimer's protein growth in aging brains"

Problem formulation

$\rightarrow$ Deal with missing values
$\rightarrow$ Perform classification

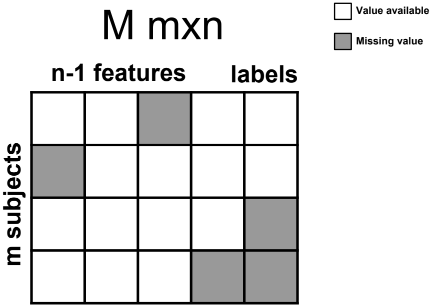
Problem formulation

$\rightarrow$ Deal with missing values	$\rightarrow$ Matrix completion
$\rightarrow$ Perform classification	$\rightarrow$ Label as feature

Matrix completion

- Recover missing values by solving optimization problem
- Loss function :

$$l = ||\Omega * (M - \tilde{M})|| + \gamma l_{\Omega_b}(M, \tilde{M}) + \beta \sum_{i=1}^{q} W_i \tag{1}$$



Graph methods for the prediction of MCI to AD conversion

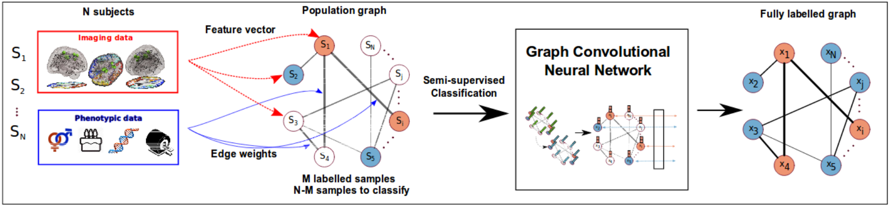


Figure 5: Overview of the pipeline used for classification of population graphs using Graph Convolutional Networks. Reproduced from Parisot et al. [1]

- No missing data
- One graph

Graph methods for the prediction of MCI to AD conversion

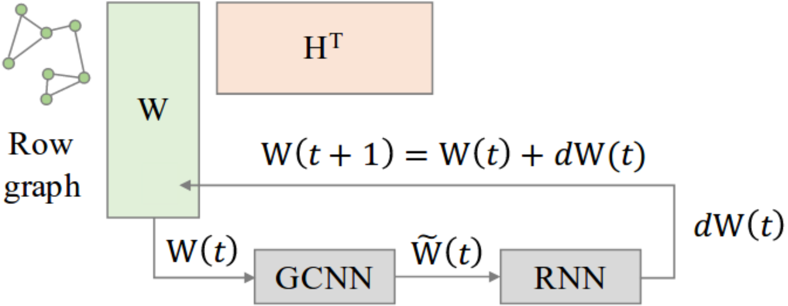


Figure 6: Vivar et al. [2]

- Matrix completion
- Missing data
- One graph

A novel graph-based method for the prediction of MCI to AD conversion

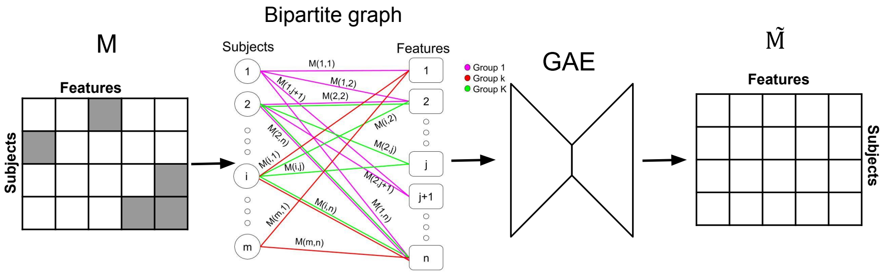
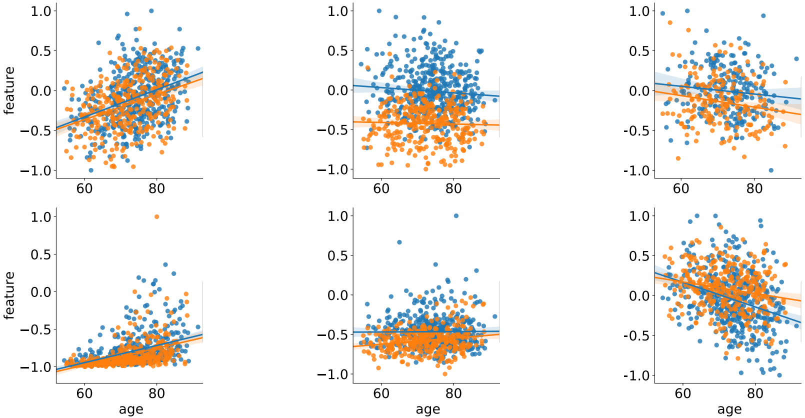


Figure 7: Proposed architecture

- Matrix completion : Van den Berg et al. [3]
- Missing data
- Multiple graphs

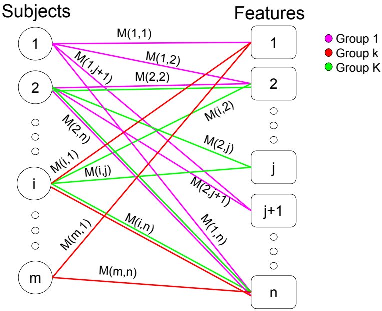
Defining the feature dependencies



Age-related features. **Sex**-related features. **Age & Sex** features.

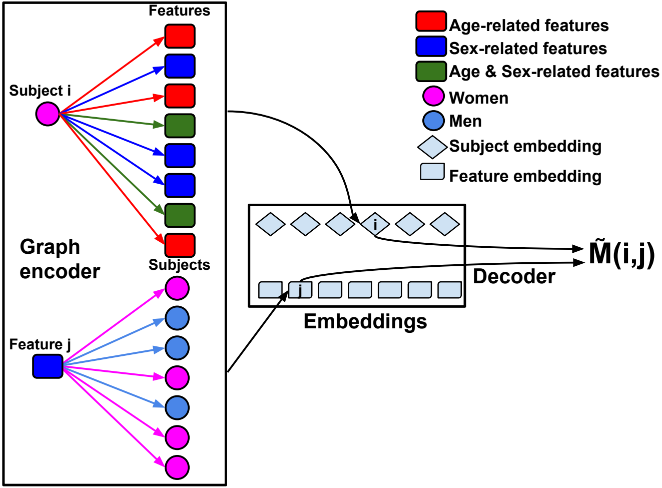
Figure 8: Relationships of age and sex (Men and Women) with six different features in the case of Alzheimer's disease.

Bipartite graph



- Relationship between a group of subjects and a group of features

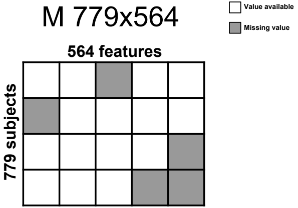
Graph Auto-encoder



Datasets

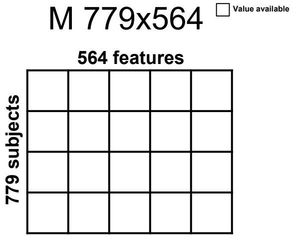
TADPOLE dataset

- 779 subjects
- 564 features
- 21 % missing data

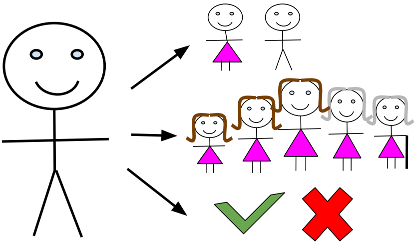


Creation of a synthetic dataset

- 779 subjects
- 564 features
- No missing data

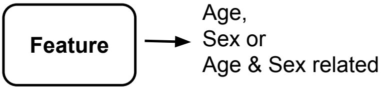


Creation of the synthetic dataset



$$M(i,j) = m_j f_{ij} + i_j + \epsilon_{ij} + v_j * y_i \quad (2)$$

$$\begin{aligned}
f_{ij} &= x_i \text{ if age} && (3)\\
&= s_i \text{ if sex} && (4)\\
&= s_i x_i \text{ if age \& sex} && (5)\\
m_j &\sim \mathcal{U}[-m, m] && (6)\\
i_j &\sim \mathcal{U}[a, b] && (7)\\
\epsilon_{ij} &\sim \mathcal{N}(0, \sigma) && (8)\\
v_j &\sim \mathcal{U}[c, d] && (9)
\end{aligned}$$



Evaluation measure for performance

- Chosen Metrics: Integral of ROC : AUC (Area Under the Curve)
- ROC measures the true positive rate, relative to the number of false positives
- Integral of ROC ranges from 0 to 1, with 1 being the best

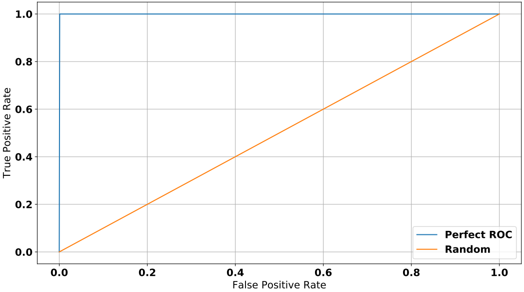
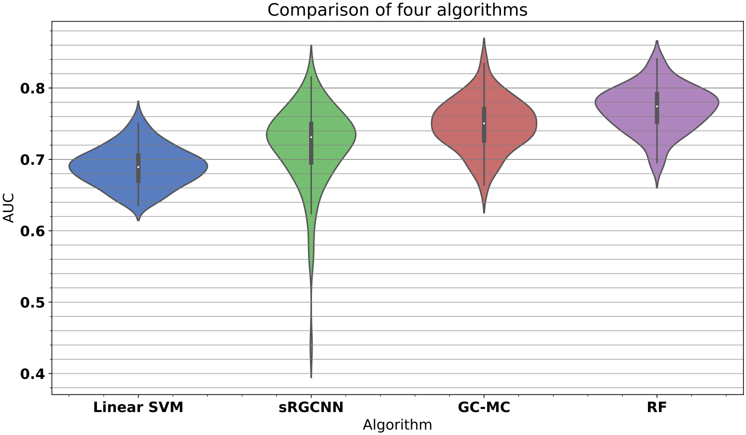
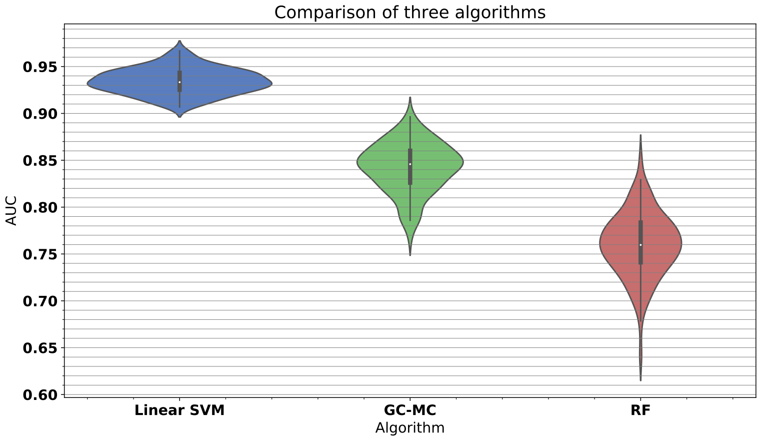


Figure 9: Perfect and random ROC curve

Results on the real dataset



Results on the synthetic dataset



Conclusion

- Better than baseline methods linear SVM and MLP
- Better performance than sRGNN by 2.9 %
- Random Forest performs better

Future work:

- Remove missing values in the dataset