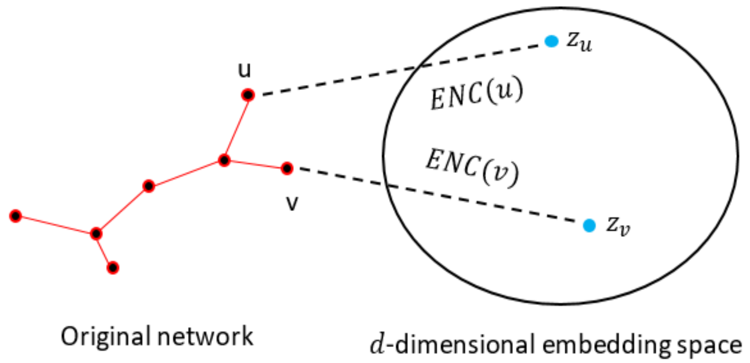
Horizon 2020
European Union Funding
for Research & Innovation

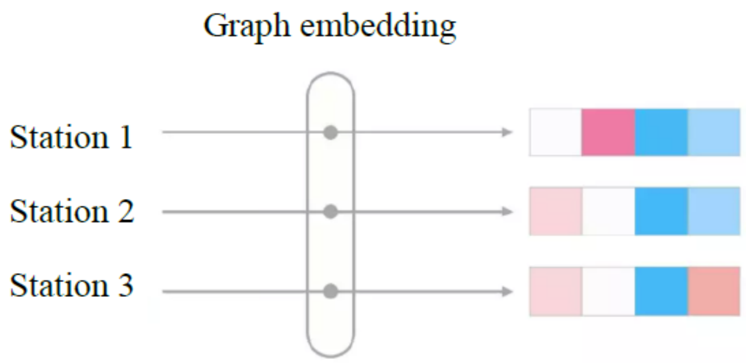
# Graph-Embedding based Primary Delay Prediction

Ruifan Tang, Ronghui Liu, Zhiyuan Lin
Institute for Transport Studies
University of Leeds

# Problem and Motivation

## --A better way to represent train routes in delay prediction




### Objective of the PoC

- Evaluate the effectiveness of route-embedding in retaining the characteristics of train route topology.
- Apply it to the prediction train primary delay

### Constraints / Requirements

- Route embedding vectors must be small and uniform in size.
- Representations can uniquely identified wrt the characteristics of the stations/routes

### Main Issues and Challenges

- Require large amount of data
- Lack of consideration of the influence of network structure on train delays
- Data pre-processing is complex

### Key Performance Indicators

- Model Stability and Reliability (with competitive methods)
- Computation Time
- Overall Prediction Accuracy

# Proof-of-Concept as a Benchmark

**AI Application**
Machine Learning

**AI Related Disciplines**
Structural Deep Network Embedding
Singular Value Decomposition

**AI Techniques**
Semi-supervised – SDNE module
Supervised – three ML-based Predictors (DT/RF/MLP)

**Inspiring Solutions**
Graph Embedding approaches
Singular Value Decomposition

**Datasets**
TransPennine Express timetables
Real-world network data
Synthetic Data generated by SMOTE

**Developments / Implementations**
SDNE for describing a Railway Network
Self-developed SVD for Route Embedding
Data feature engineering & Fed in for training
Training results for 5-delay-level classification

**Exploited Software and Framework**
Keras, Tensorflow, sklearn, Numpy, Pandas
DecisionTreeClassifier, RandomforestClassifier, MLPClassifier

**Hardware Requirements**
Google Colaboratory, Python 3
GPU(s) with CUDA cores and 16GB System RAM

# Dataset Description

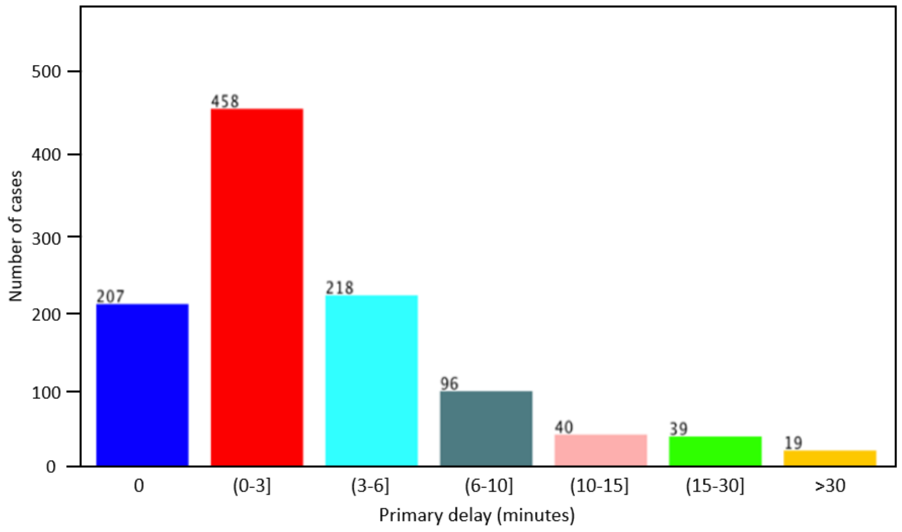
- Train operating data source: TransPennine Express (TPE)
- Time horizon: 28/05/2017-24/06/2017, and 27/05/2018-23/06/2018
- Size of the dataset: 1191 train delay instances, 177 stations and 192 edges/links

| Feature Categories | Name of Features |
|---|---|
| Temporal Features | Date of Service; Weekday/Holiday; Departure Time; Arrival Time |
| Numerical Features | Passenger Volumn; Total Margin; Speed Limit; Link Travel Time |
| Categorical Features | Rolling Stock Type; Train ID (headcode) |
| Network feature | Train’s route (origin, intermediate stops, destination) |
| Label Feature | Primary Delay |

# Network Illustration: TPE Network


# Approach: Feature engineering and label processing

| Date quality issue | Policy |
| --- | --- |
| Imbalanced dataset | SMOTE resampling strategy |
| Temporal features are not continuous | Cyclical variable projection – Polar coordinates systems |
| Outliners on numerical features | Z-score normalization (feature scaling) |

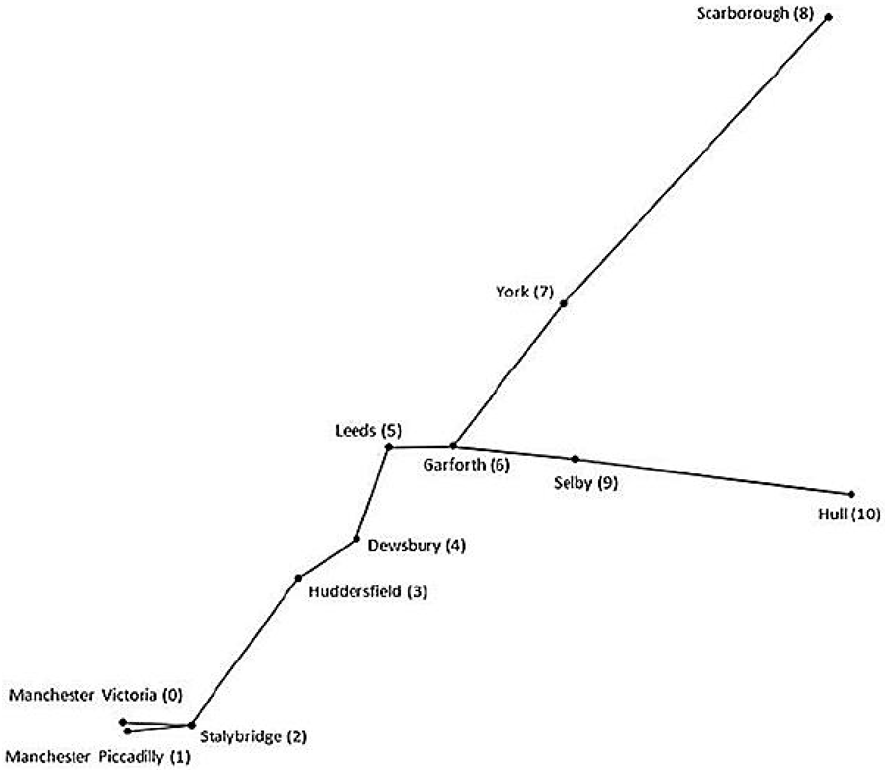



- Features pre-processing
  - Polices applied according to the data quality issue identified.
- Sci-kit learn library
  - Three different ML-based classifiers
- Dimension reduction on network features (strategies)
  - Principal component analysis (PCA) – competitor method
  - SDNE+SVD
- Label processing
  - Categorize train delay minutes into delay levels.

| Primary delay /mins | Delay level description | Indexed level |
| --- | --- | --- |
| 0 | None | 0 |
| (0,6] | Mild | 1 |
| (6,11] | Moderate | 2 |
| (11,16] | Serious | 3 |
| (16+) | Severe | 4 |

# Approach: Integrating station embedding vectors

- With the aim of predict delay level for each service, we extract the route information (ie, stations & sequences) from the original station matrix.




(a) an 11-node network

| | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **1** | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **2** | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **3** | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| **4** | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| **5** | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 |
| **6** | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 1 | 0 |
| **7** | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 0 |
| **8** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 |
| **9** | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 1 |
| **10** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |

(b) one-hot representation for the 11-node network

One-hot embedding

SDNE+SVD

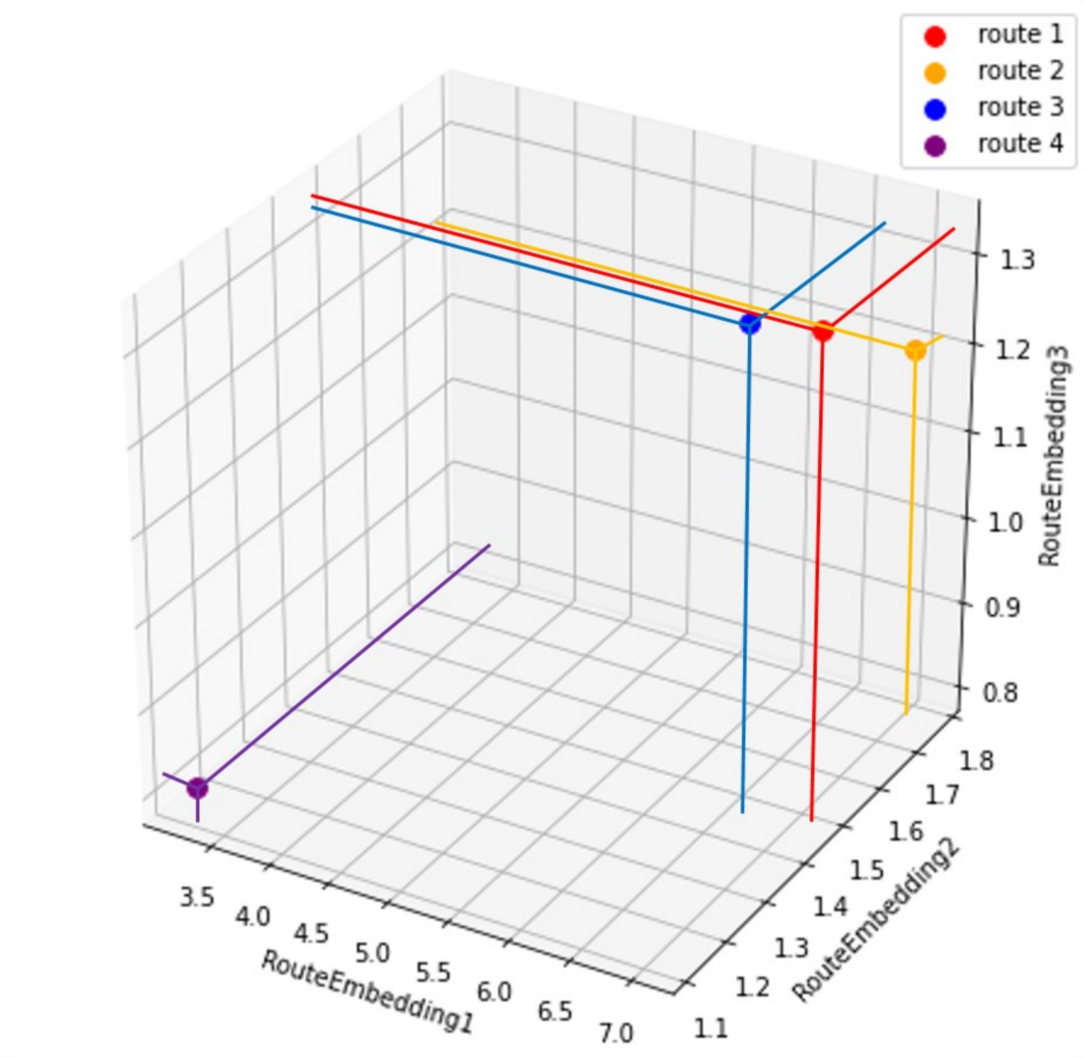
# Route effectiveness clustering

**Route1**:
Newcastle - Liverpool Lime Street: Newcastle, Chester-le-Street, Durham, Darlington, Northallerton, York, Leeds, Huddersfield, Manchester Victoria, Liverpool Lime Street

**Route2**:
Newcastle - Manchester Airport: Newcastle, Chester-le-Street, Durham, Darlington, Northallerton, York, Leeds, Huddersfield, Manchester Piccadilly, Manchester Airport

**Route3**:
Newcastle - Manchester Victoria: Newcastle, Durham, Darlington, Northallerton, York, Leeds, Huddersfield, Manchester Victoria

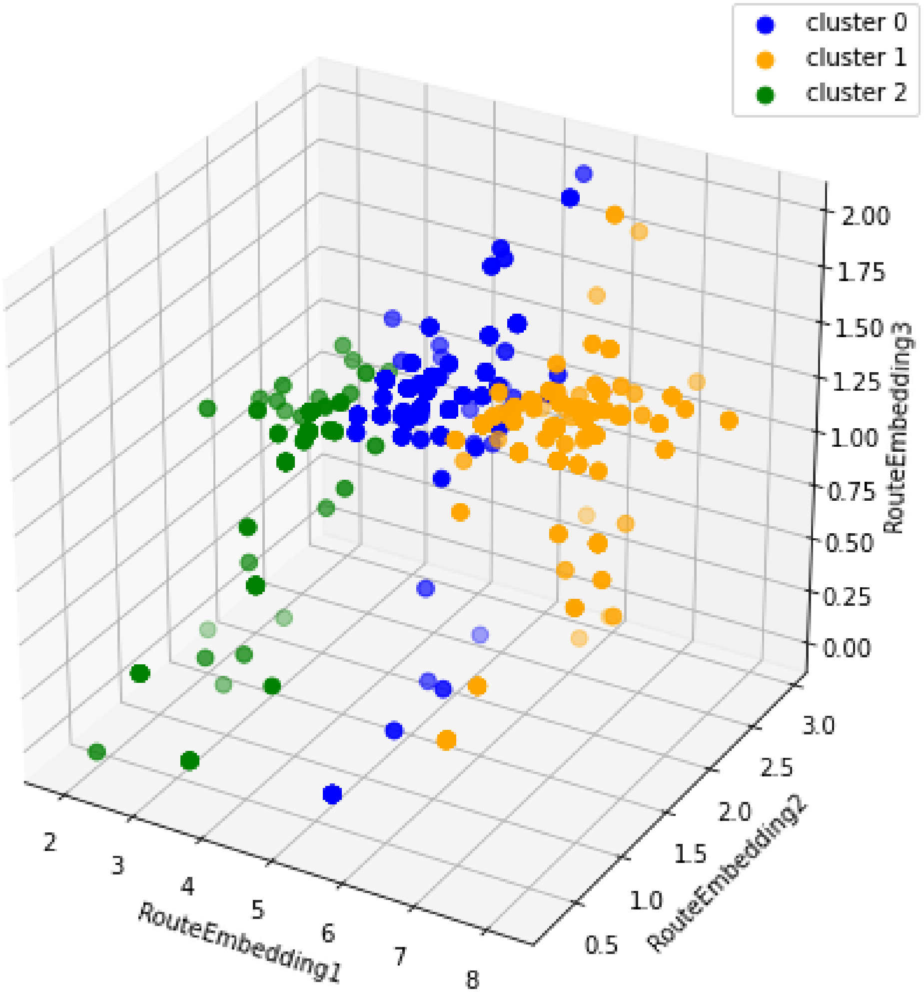
**Route4**:
Manchester Airport - Doncaster: Manchester Airport, Manchester Piccadilly, Stockport, Dore & Totley, sheffield, Meadowhall, Doncaster




Embedding results

TU Delft

# Route effectiveness clustering

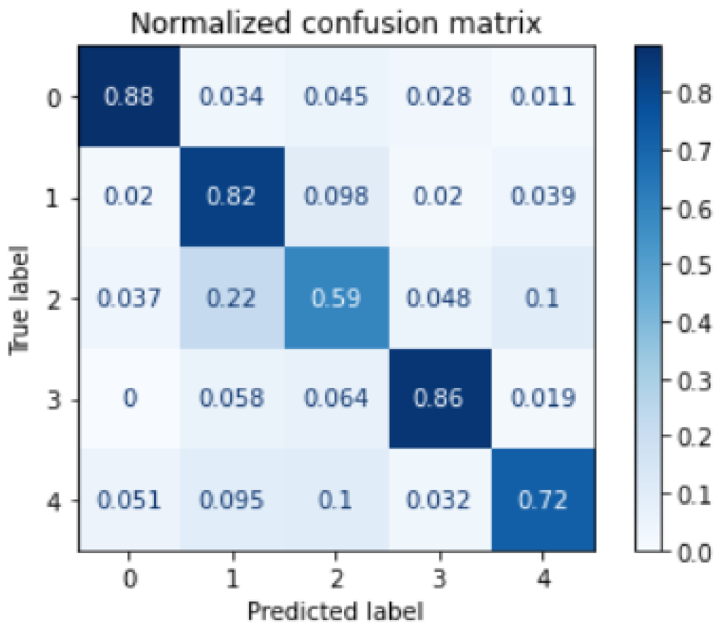

TU Delft

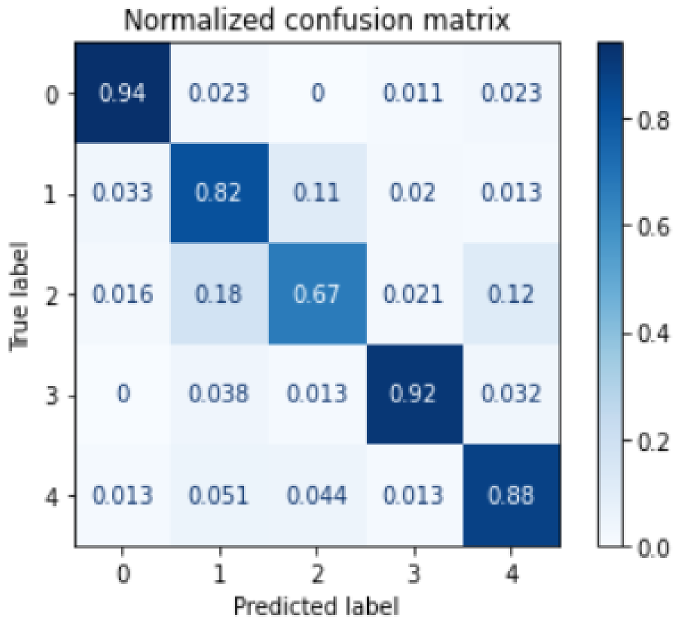
# Overall Discussion and Results

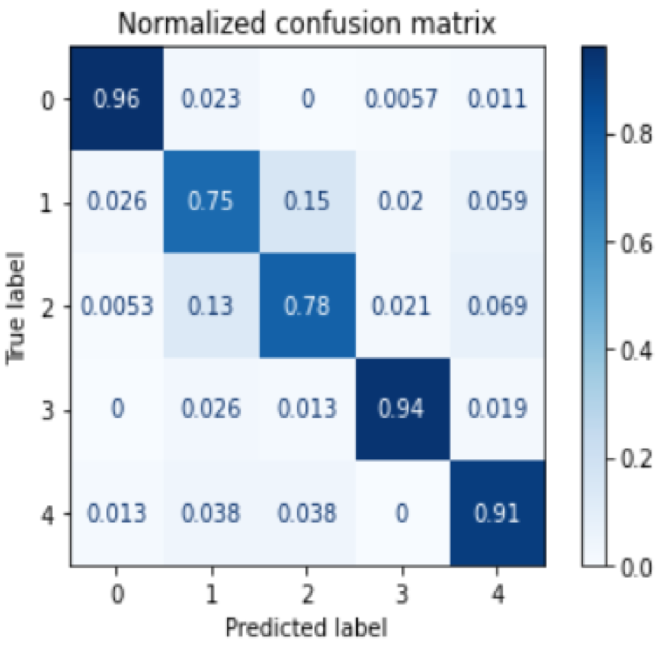
- 5-fold Cross Validation results

| Strategy | Algorithm | 1-fold | 2-fold | 3-fold | 4-fold | 5-fold | Average Score | Standard Deviation |
|---|---|---|---|---|---|---|---|---|
| PCA | DT | 0.7198 | 0.7564 | 0.7347 | 0.7113 | 0.7381 | 0.7321 | 0.0227 |
| | RF | 0.7773 | 0.8190 | 0.8084 | 0.7916 | 0.8239 | 0.8040 | 0.0174 |
| | MLP | 0.8138 | 0.8378 | 0.8393 | 0.8150 | 0.8511 | 0.8314 | **0.0146** |
| SDNE + SVD | DT | 0.7443 | 0.7537 | 0.7035 | 0.7249 | 0.7132 | 0.7279 | 0.0187 |
| | RF | 0.8362 | 0.8338 | 0.8138 | 0.8196 | 0.7941 | 0.8195 | 0.0152 |
| | MLP | 0.8436 | 0.8181 | 0.8421 | 0.8346 | 0.8313 | 0.8339 | **0.0091** |

# Overall Discussion and Results

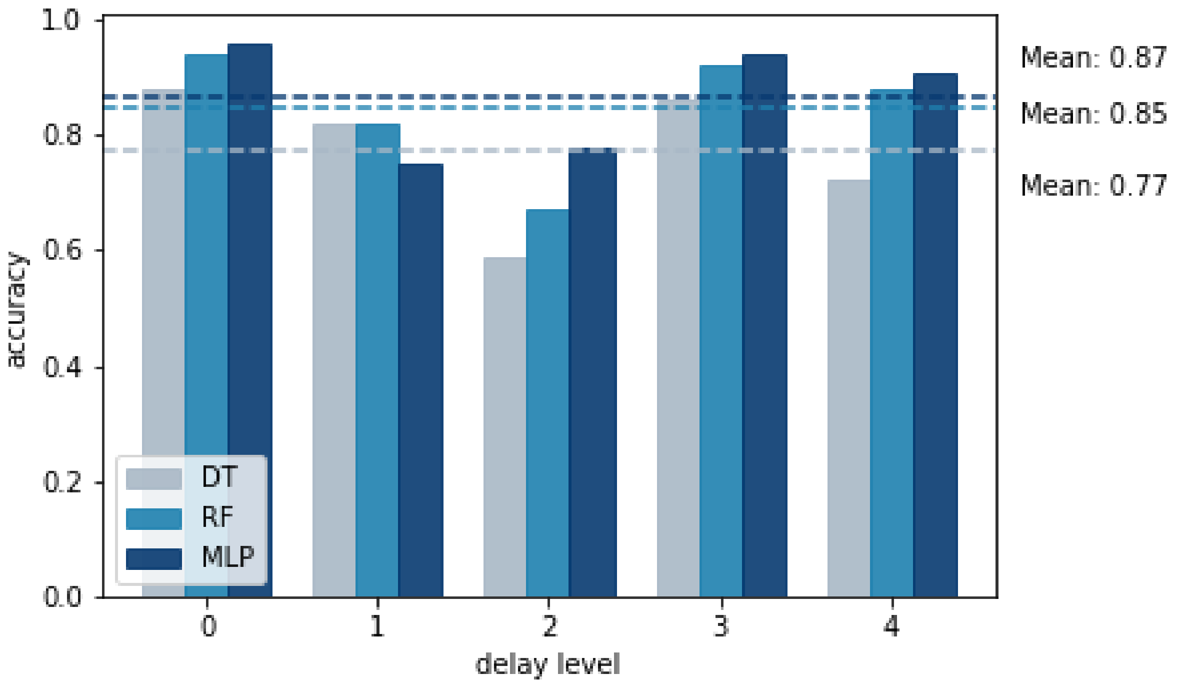


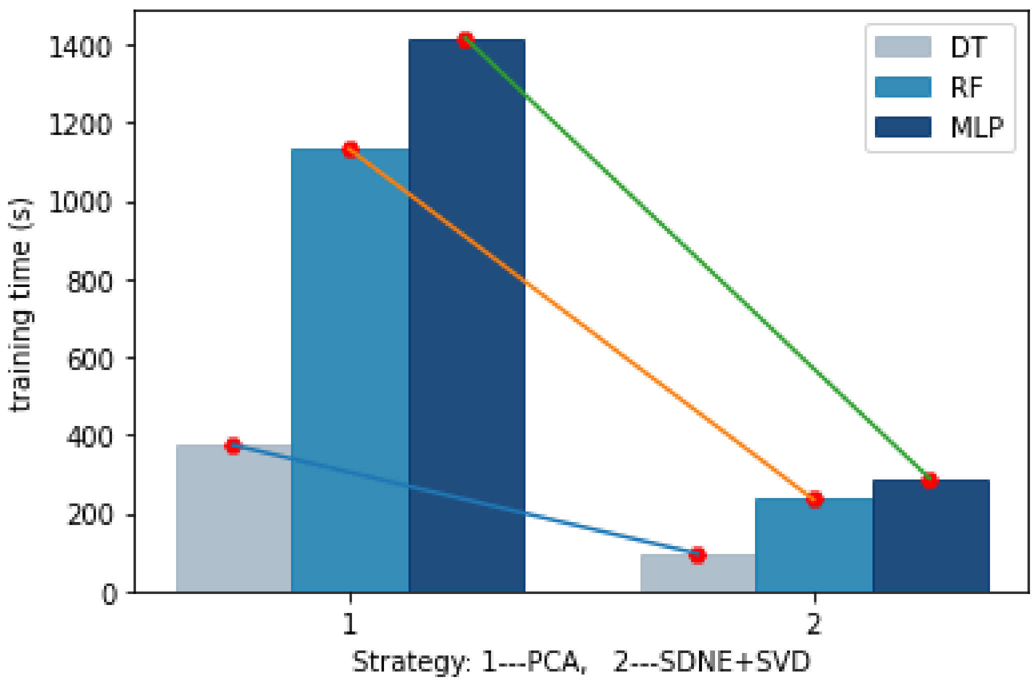




| **Strategy** | PCA | | | SDNE + SVD | | |
|---|---|---|---|---|---|---|
| **Algorithm** | DT | RF | MLP | DT | RF | MLP |
| **Overall Accuracy** | 76.68% | 82.09% | 83.89% | 77.02% | 84.24% | **86.64%** |
| **Overall Training time (s)** | 374 | 1133 | 1417 | 96 | 235 | **289** |

# Overall Discussion and Results



# SWOT Analysis of the Investigated Approach

Captures network topology
Non-linear mapping
Scalability and Efficiency
Robustness to noise/missing data

Data requirements
Interpretability
Primary delay only


Enhanced prediction accuracy
Incorporation of additional features

Generalization to new environments
Overreliance on network structure

# Discovery and Conclusion

- The TPE data fits better on more complex models, for example, MLP and RF.
- A better prediction accuracy on non-delay cases and serious/severe delay instances can be achieved on SDNE+SVD method.
- The overall accuracy increases thanks to the SDNE+SVD strategy, regardless of the baseline or specific delay level.
- The necessary training time for model to converge has been reduced significantly – this improvement is particularly meaningful and helpful to further implementing short-term or even real-time delay prediction paradigm.

# Thank you for your attention!

RAILS Roadmaps for AI Integration in raiL Sector

- ✓ *Deliverable D4.1: WP4 Report on case studies and analysis of transferability from other sectors (Railway planning and management)*
- ✓ *Deliverable D4.2: WP4 Report on AI approaches and models*
- *Deliverable D4.3: WP4 Report on experimentation, analysis, and discussion of results*
- *Deliverable D4.4: WP4 Report on identification of future innovation needs and recommendations for improvements*

*Available at: https://rails-project.eu/downloads/deliverables/*