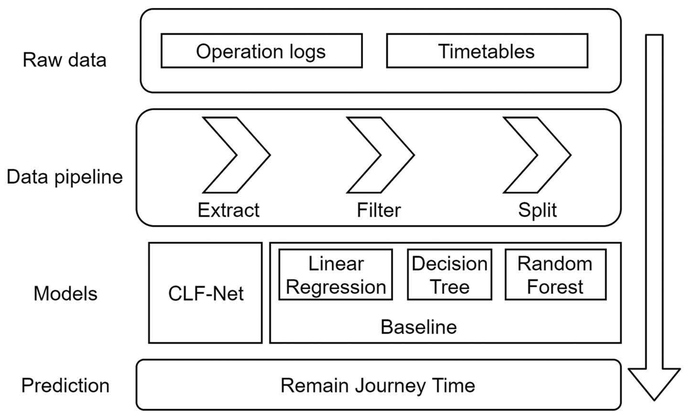
MTRC train data analysis
A machine learning approach

Objective & Scope

This project aims to **develop a predictive model**, driven by **neural network** learning techniques of machine learning, which is capable of f**orecasting the remaining journey time of a running train midway in the journey.** The predicted result will serve as an important piece of information for the reference of MTR train operators to improve the overall train punctuality, which in turn reduces the duration of the suspension and lessens the financial penalty to the MTR company.

The scope of the project is limited to analysing the train operation data of **Kwun Tong Line (KTL).** The end of the project include a neural network model for the prediction of remaining journey time.

Methodology



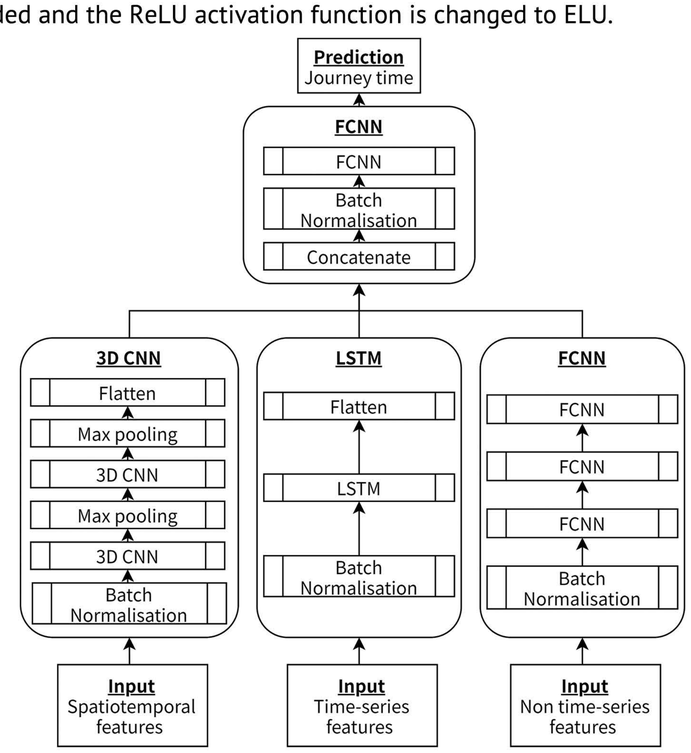
The methodology of this study is described as the figure above. We are using the actual & expected arrival/departure time from MTR data as input. After data cleaning in the pipeline, data is fed into a model for the final prediction, remaining journey time at the station midway of a journey.

CLF Net

In this study, we adopted a multi-network model, CLF-Net.

- 3D Convolution Network (3D CNN)
  => Spatiotemporal features
- Long short-term memory (LSTM) recurrent neural network
  => Time-series features
- Fully-connected neural network (FCNN)
  => Non-time-series features

Both input features, output, and model architecture have been changed. The model architecture is changed to enable the model to learn more complex logic and seek better performance. Specifically, the parameter size of the model is increased, batch normalization is added and the ReLU activation function is changed to ELU.



Methods	LR	DT	RF	CLF-Net
Mean Absolute Error **MAE**	32.02s	33.00s	28.31s	**24.67s**
Root-Mean-Square Error **RMSE**	48.27	51.33	47.10	**43.55**
% of samples w/ error < 1min	88%	86%	89%	**91%**

Table. Overall prediction result of different models

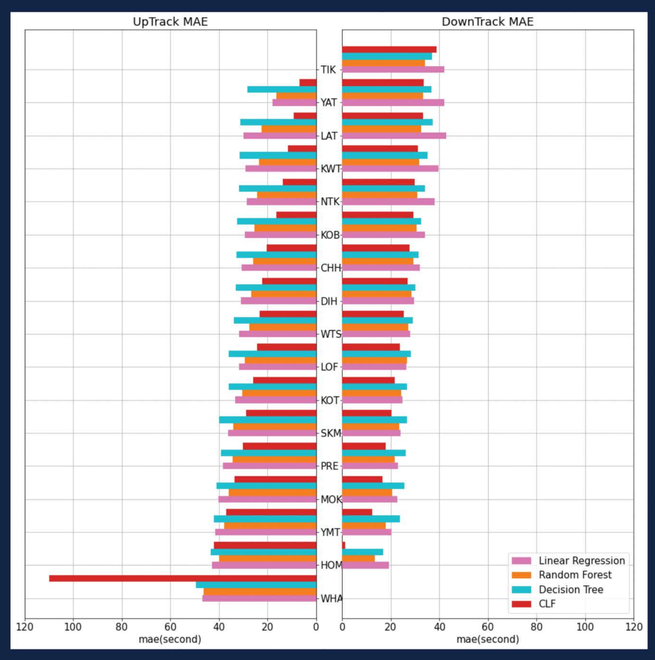


Figure. MAE of different models over stations

Overall, 91% of CLF-Net's prediction has error less than 1 minute

- The processed KTL train operation data in the preparation phase is used as the input data of the models and the performance of models is evaluated by MAE and RMSE
- The MAE of the CLF model is 24.67s, while the RMSE is 43.55s. Both of the errors are lower than the other three baseline models.
- The proportion of CLF prediction with MAE <= 1 minute = 91%, which is higher than that of baseline models.
- The CLF model outperforms the baseline models generally and the MAE <= 1 minutes, it meets the objective of this project.

At most stations, CLF-Net performs better than the baseline models

- MAE decreases when the train travels closer to the terminal. It is reasonable to have a higher prediction error as many unforeseeable factors could affect the journey time of the train.
- CLF model suffers a large prediction error in station WHA at up track. It is because the train starts from station WHA and HOM alternatively while the CLF model requires consecutive trains at consecutive k stations for a complete record. It results in the number of training data at WHA is significantly fewer than other stations. As a result, the model performs worse in station WHA.

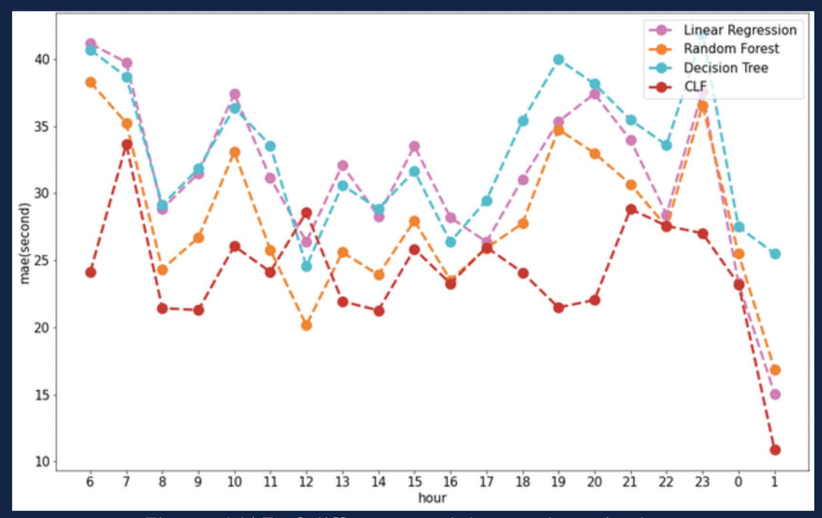
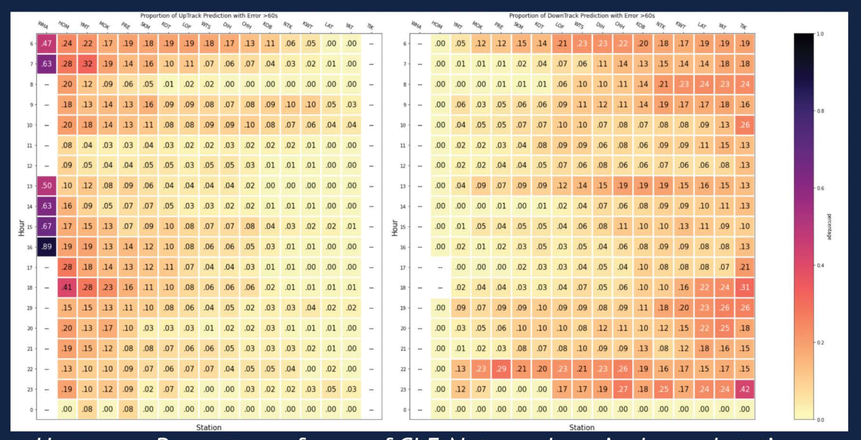


Figure. MAE of different models over hour in day

In peak hours, CLF-Net performs better than the baseline models

- The MAE of prediction of the baseline models follow the same trend over time meanwhile CLF-Net only follows the same in the early morning and late night.
- In general, CLF-Net performs better than the baseline model, particularly in morning and evening peak hours.



Heat-map. Percentage of error of CLF-Net over hour in day and station
Darker colour means more prediction have >1 min error

CLF-Net is stable and performs well in most situation

- The prediction of the model in the early stage of the journey is worse than the late stage. In many cases, the percentage of MAE falls outside 1-minute error is less than 15% which is acceptable.
- The model performs worse in down-track, late-night (around 23:00), TIK station, the percentage of prediction falls outside 1-minute error equals to 42%.
- After investigating the related cases, we found that it is possibly due to the abnormally short journeys.
- Overall, the large error predictions are low in ratio

Limitation & Suggestion

- Subject to the input feature shape of CLF-Net, information available for training at early stages of journeys is relatively less in the current dataset. To improve the model's performance at early stages of journeys, several techniques of data augmentation e.g. extrapolation from existing data may help. Future studies can trial different techniques and investigate their effects on model performance.
- To understand the robustness of CLF-Net against train incidents across the system, further analyses should be carried out when sufficient and reliable incident-labelling has been integrated into the existing dataset.
- More extensive and comprehensive historical train data, preferably covering an entire year, should be used to further investigate the robustness of CLF-Net model against changing train running patterns during different seasons, festivals and public events across the year.
- When historical train data on other MTR lines are available, similar experiments can be conducted to test the applicability of CLF-Net on other lines.

COMP4801 2020/21
FYP20033
Supervisor: Prof. Reynold Cheng
Students: Chung Hok Kan, Lam Kai I, Leung Chun Yin, Kwok Olivier Yuk-ting, Wong Chun Ming