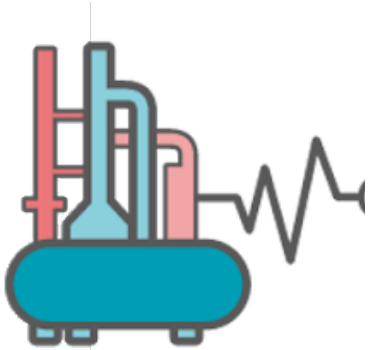
SHELL CHALLENGE

SHRI LEKKALA, SONJA N. TANG, THOMAS WONG, AND VALERIE JIA

MAIN QUESTION:

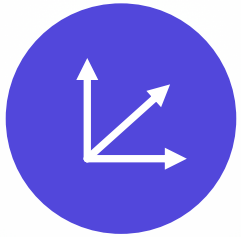
- Can we recognise compressor trips **before** they happen?



METHODS

DENOISING

CLUSTERING



PCA

RANDOM FOREST

CNN

INITIAL CLUSTERING ANALYSIS

`clustering_comparisons`

	# clusters	Spectral vs Hierarchical (ARS)	Spectral vs Target Spectral (ARS)	Hierarchical vs Target Spectral (ARS)	Spectral vs Hierarchical (AMIS)	Spectral vs Target Spectral (AMIS)	Hierarchical vs Target Spectral (AMIS)
0	k=5	0.277724	0.210947	0.122770	0.381221	0.325086	0.237709
1	k=6	0.407219	0.202407	0.053444	0.467338	0.318402	0.202923
2	k=7	0.096274	0.062718	0.097526	0.153558	0.125703	0.185334
3	k=8	0.094961	0.071076	0.117109	0.151608	0.133236	0.210010
4	k=9	0.412847	0.204177	0.059089	0.460193	0.332952	0.190610
5	k=10	0.409293	0.201726	0.051949	0.449166	0.332089	0.198953
6	k=11	0.400184	0.197048	0.071863	0.452202	0.329309	0.218881
7	k=12	0.106263	0.023188	0.073700	0.191049	0.141469	0.234088
8	k=13	0.152196	0.029327	0.084477	0.263598	0.152214	0.234232
9	k=14	0.150476	0.036467	0.092821	0.268156	0.162026	0.254231
10	k=15	0.284192	0.007361	0.082183	0.370819	0.166287	0.272286

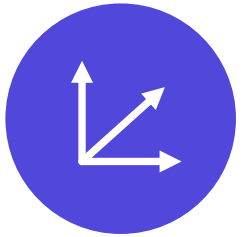
- **Spectral, Hierarchical and Target Spectral** clustering models were compared for different cluster numbers

ANALYSIS - CLUSTERING ON TIMESTAMP

1. Calculate the trailing variance on a window of 25
2. Calculate the percentage change on the trailing variance
3. Assign 1, 0, -1 to indicate drastic increase, relative stability and drastic decrease
4. Perform MiniBatchK-Means on the result dataframe with the number of clusters to be 10
5. Backfill the cluster result by 3 steps
6. Obtain a final dataframe with the latest cluster result added

- Majority of **timestamps fall into the 0 (stability) cluster**
- Majority of the 5 timestamps ahead of the given anomaly timestamp fall into one of the non-zero clusters

ANALYSIS - PCA



1. Calculate **simple moving average (SMA)** on the cleaned data on a window period of 25 to smoothen the noise
2. Employ PCA on the standardised SMA dataframe to cut down the dimensionality of the dataset
3. Purpose: increase efficiency in fitting machine learning models

- # of components = **25**
- Ratio of variance explained overall = **86.3%**

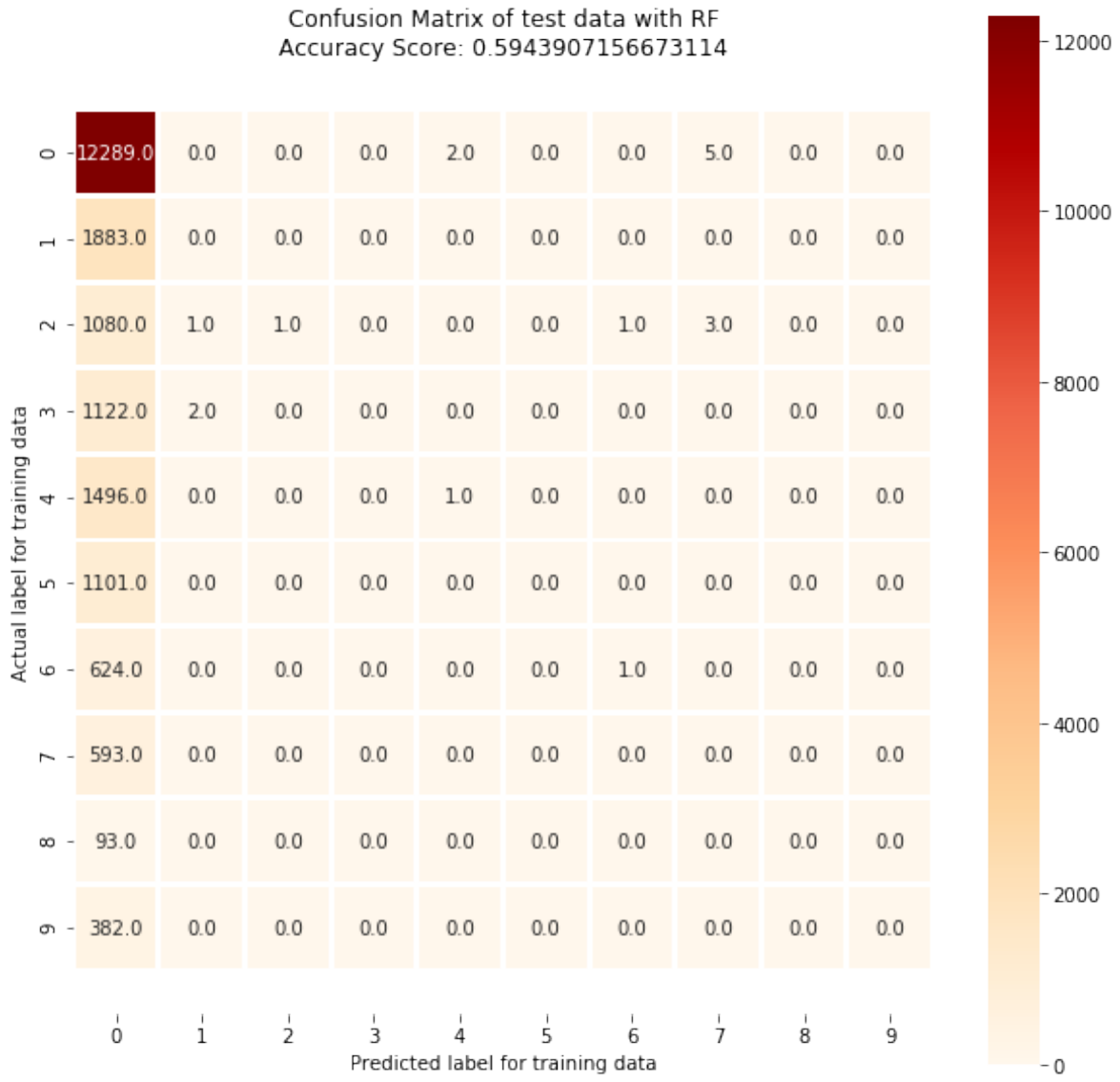
ANALYSIS - RANDOM FOREST 1

- Fit a random forest classifier on the **standardised PCA dataframe**
  - Max depth = 40, # of estimators = 120
- Training set (80%); Test set (20%)

➢ Test set accuracy = **59.4%**

ANALYSIS - RANDOM FOREST 1

- Confusion matrix
- Recall rate



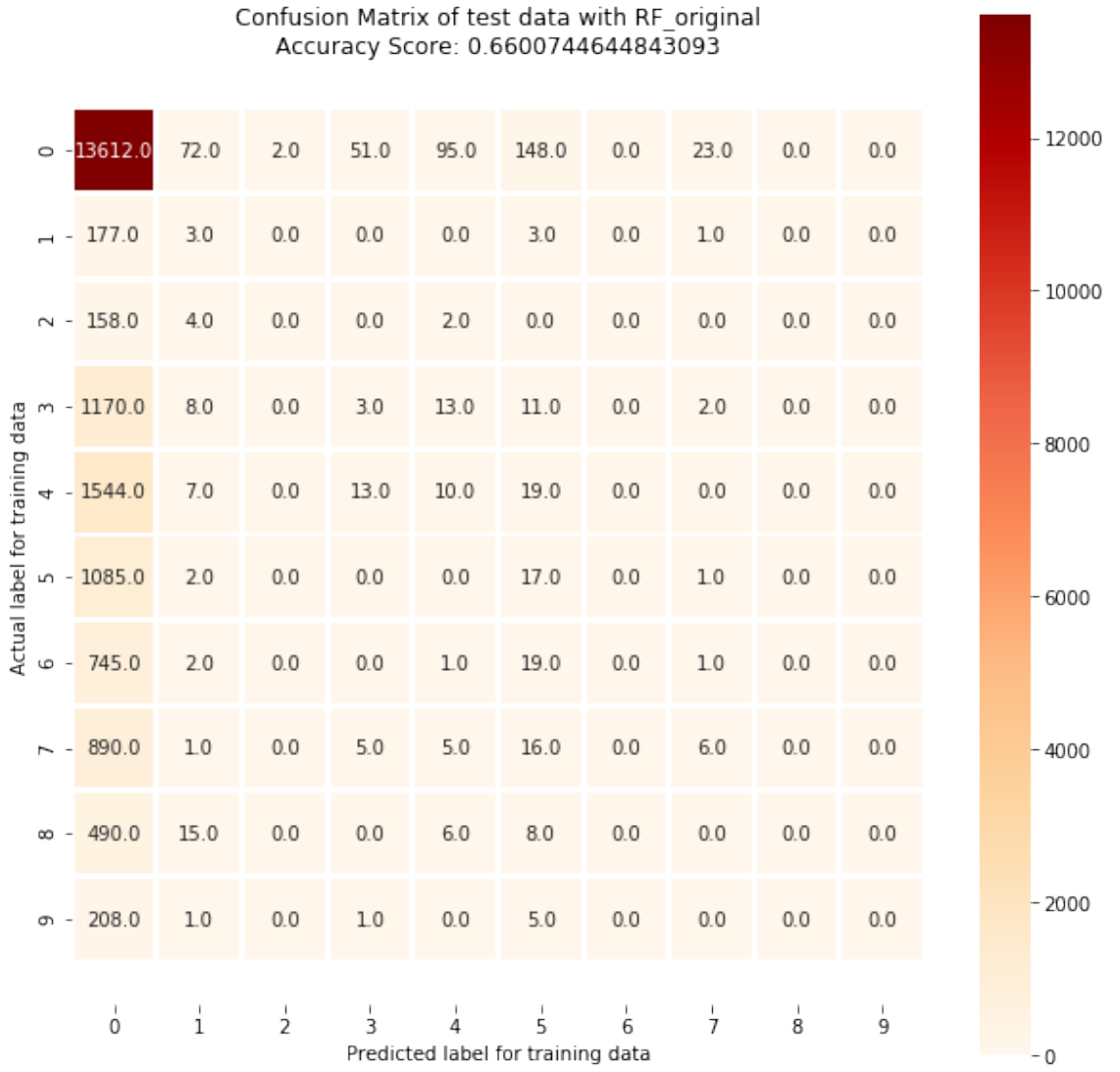
ANALYSIS - RANDOM FOREST 2

- Fit a random forest classifier on the **standardised cleaned dataframe**
  - Max depth = 40, # of estimators = 120
- Training set (40%); Test set (20%)

➢ Test set accuracy = **66%**

ANALYSIS - RANDOM FOREST 2

- Confusion matrix
- Recall rate



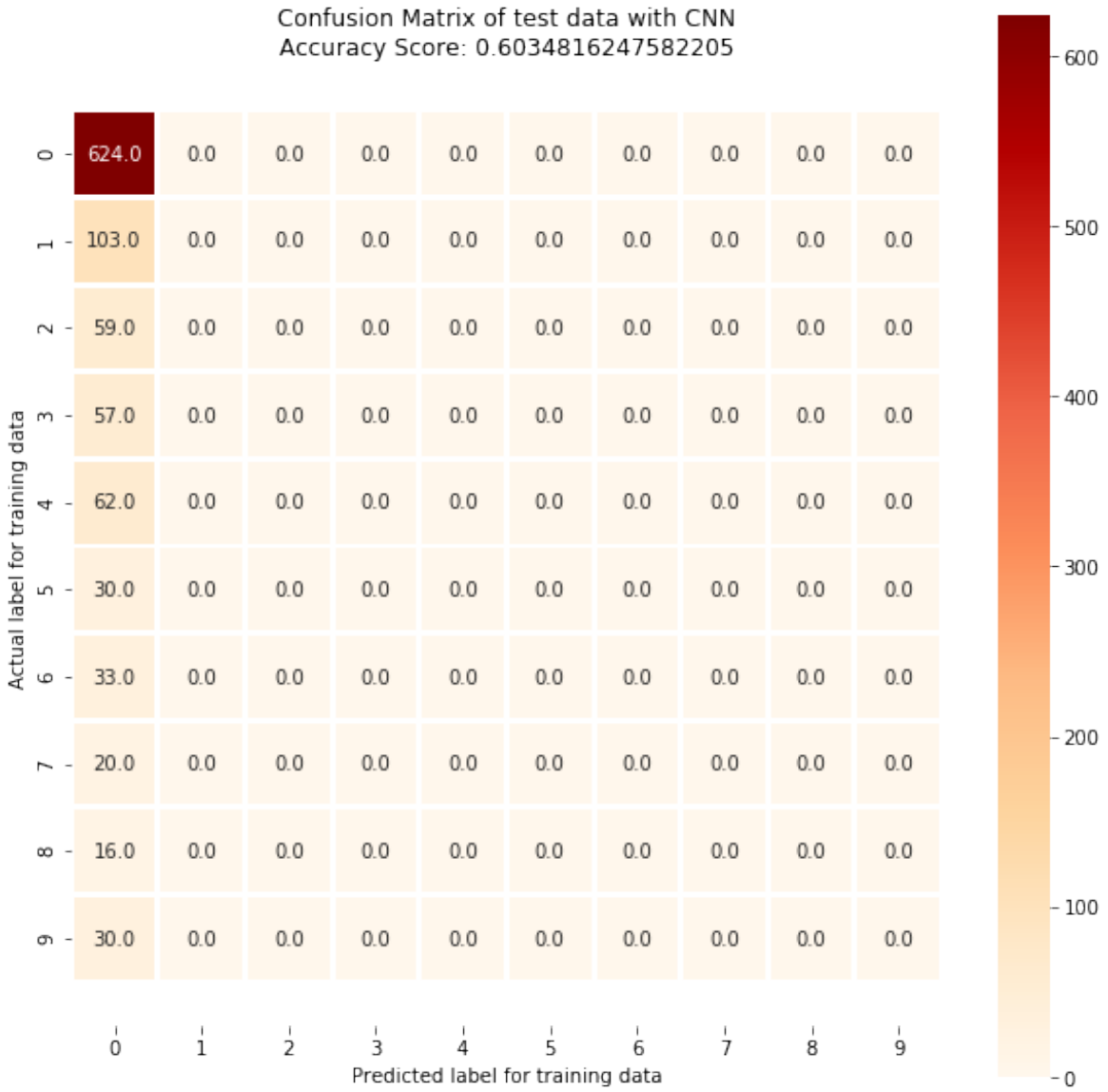
ANALYSIS - CONVOLUTIONAL NEURAL NETWORK

- Perform CNN on the stardardised SMA dataframe
  - # of classes = 10, batch size = 128, # of epochs = 15, batch size = 100, learning rate = 0.005
  - Two convolutional and subsampling layers
  - One fully connected layer
  - Activation function = ReLU
  - Optimisation method = Stochastic gradient descent
- Training set (n = 3619); Test set (n = 1034)
- Input: 250*362 matrix as one image
- Output: the cluster result in the following row

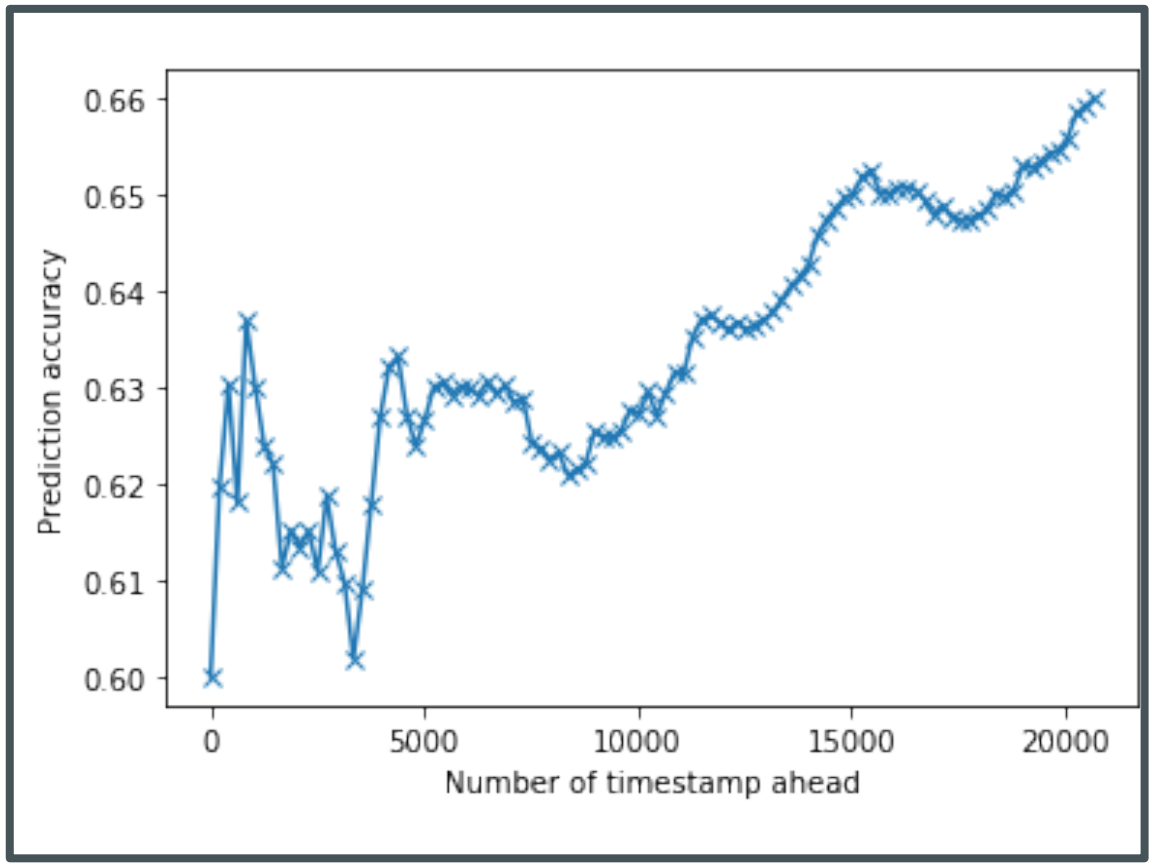
➢ Test accuracy = **60.4%**

ANALYSIS - CNN

- Confusion matrix
- Recall rate



MODEL COMPARISONS



- The **random forest classifier** with the standardised cleaned dataframe produced the **highest prediction accuracy**
- The random forest classifier is **efficient** in terms of running time
- Based on the figure, the prediction accuracy produced by the classifier fluctuates when we try to predict a smaller amount of timestamp ahead and eventually increases as the amount of timestamp ahead increases

RESULTS

Can we recognise compressor trips **before** they happen?

- ✓ We could **cluster the timestamp** into categories, which could potentially refer to **different types of warning messages** before **a trip**.
- ✓ Using the random forest classifier, we could **predict the category each timestamp could belong to** based on the values collected from the compressor.

THANK YOU FOR LISTENING!