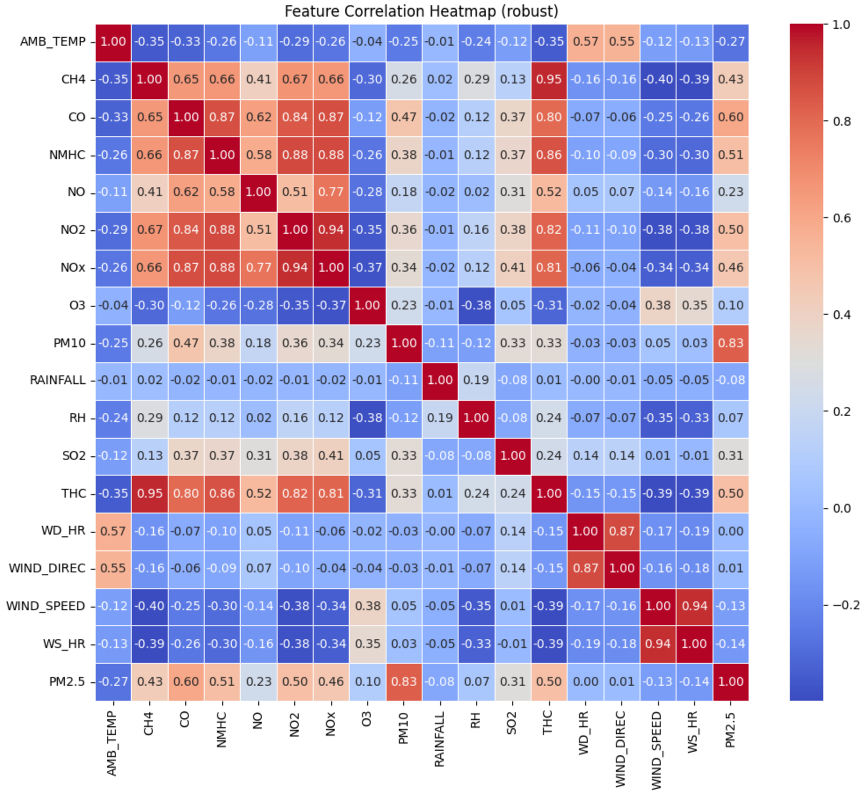
2025 Data Mining – HW 1

313553014 廖怡誠

How do you select features for your model input, and what preprocessing did you perform?

I selected features for my model input based on their correlation with the target variable, PM2.5. To do this, I computed the correlation matrix and identified features that had a strong relationship with PM2.5, using a threshold of 0.5 in absolute value. This helped ensure that only the most relevant variables were included while filtering out those with weaker associations. Additionally, I excluded non-predictive columns like dates and indices, keeping only meaningful environmental and meteorological factors. The heatmap visualization (Figure 1) further confirmed the relationships among variables, allowing me to refine the feature selection process and reduce redundancy.

Based on the heatmap results, I ultimately selected the following features for model input: ['CO', 'NMHC', 'NO2', 'PM10', 'THC', 'PM2.5']. These features showed relatively strong correlations with PM2.5 and were considered most relevant for capturing air quality patterns.



	AMB_TEMP	CH4	CO	NMHC	NO	NO2	NOx	O3	PM10	RAINFALL	RH	SO2	THC	WD_HR	WIND_DIREC	WIND_SPEED	WS_HR	PM2.5
AMB_TEMP	1.00	-0.35	-0.33	-0.26	-0.11	-0.29	-0.26	-0.04	-0.25	-0.01	-0.24	-0.12	-0.35	0.57	0.55	-0.12	-0.13	-0.27
CH4	-0.35	1.00	0.65	0.66	0.41	0.67	0.66	-0.30	0.26	0.02	0.29	0.13	0.95	-0.16	-0.16	-0.40	-0.39	0.43
CO	-0.33	0.65	1.00	0.87	0.62	0.84	0.87	-0.12	0.47	-0.02	0.12	0.37	0.80	-0.07	-0.06	-0.25	-0.26	0.60
NMHC	-0.26	0.66	0.87	1.00	0.58	0.88	0.88	-0.26	0.38	-0.01	0.12	0.37	0.86	-0.10	-0.09	-0.30	-0.30	0.51
NO	-0.11	0.41	0.62	0.58	1.00	0.51	0.77	-0.28	0.18	-0.02	0.02	0.31	0.52	0.05	0.07	-0.14	-0.16	0.23
NO2	-0.29	0.67	0.84	0.88	0.51	1.00	0.94	-0.35	0.36	-0.01	0.16	0.38	0.82	-0.11	-0.10	-0.38	-0.38	0.50
NOx	-0.26	0.66	0.87	0.88	0.77	0.94	1.00	-0.37	0.34	-0.02	0.12	0.41	0.81	-0.06	-0.04	-0.34	-0.34	0.46
O3	-0.04	-0.30	-0.12	-0.26	-0.28	-0.35	-0.37	1.00	0.23	-0.01	-0.38	0.05	-0.31	-0.02	-0.04	0.38	0.35	0.10
PM10	-0.25	0.26	0.47	0.38	0.18	0.36	0.34	0.23	1.00	-0.11	-0.12	0.33	0.33	-0.03	-0.03	0.05	0.03	0.83
RAINFALL	-0.01	0.02	-0.02	-0.01	-0.02	-0.01	-0.02	-0.01	-0.11	1.00	0.19	-0.08	0.01	-0.00	-0.01	-0.05	-0.05	-0.08
RH	-0.24	0.29	0.12	0.12	0.02	0.16	0.12	-0.38	-0.12	0.19	1.00	-0.08	0.24	-0.07	-0.07	-0.35	-0.33	0.07
SO2	-0.12	0.13	0.37	0.37	0.31	0.38	0.41	0.05	0.33	-0.08	-0.08	1.00	0.24	0.14	0.14	0.01	-0.01	0.31
THC	-0.35	0.95	0.80	0.86	0.52	0.82	0.81	-0.31	0.33	0.01	0.24	0.24	1.00	-0.15	-0.15	-0.39	-0.39	0.50
WD_HR	0.57	-0.16	-0.07	-0.10	0.05	-0.11	-0.06	-0.02	-0.03	-0.00	-0.07	0.14	-0.15	1.00	0.87	-0.17	-0.19	0.00
WIND_DIREC	0.55	-0.16	-0.06	-0.09	0.07	-0.10	-0.04	-0.04	-0.03	-0.01	-0.07	0.14	-0.15	0.87	1.00	-0.16	-0.18	0.01
WIND_SPEED	-0.12	-0.40	-0.25	-0.30	-0.14	-0.38	-0.34	0.38	0.05	-0.05	-0.35	0.01	-0.39	-0.17	-0.16	1.00	0.94	-0.13
WS_HR	-0.13	-0.39	-0.26	-0.30	-0.16	-0.38	-0.34	0.35	0.03	-0.05	-0.33	-0.01	-0.39	-0.19	-0.18	0.94	1.00	-0.14
PM2.5	-0.27	0.43	0.60	0.51	0.23	0.50	0.46	0.10	0.83	-0.08	0.07	0.31	0.50	0.00	0.01	-0.13	-0.14	1.00

Fig1. Heatmap visualization of correlation matrix

In this homework, the preprocessing steps were carefully designed to enhance the accuracy of PM2.5 prediction. The key aspects of preprocessing included handling invalid data, determining the optimal number of past hours to use as input features, and selecting an appropriate normalization method for both training and testing data.

To address missing or invalid values, two different imputation strategies were tested. The first approach involved replacing missing values with the mean of the corresponding column, while the second method filled in missing values using the nearest available data point, either from the previous or the next valid entry. The experimental results (figure 2.) showed that the second approach performed better. One possible explanation for this is that taking the mean still introduces some inaccuracies, whereas air quality measurements typically do not fluctuate drastically within a short period. By filling missing values with the nearest valid measurement, the introduced error remained relatively small.

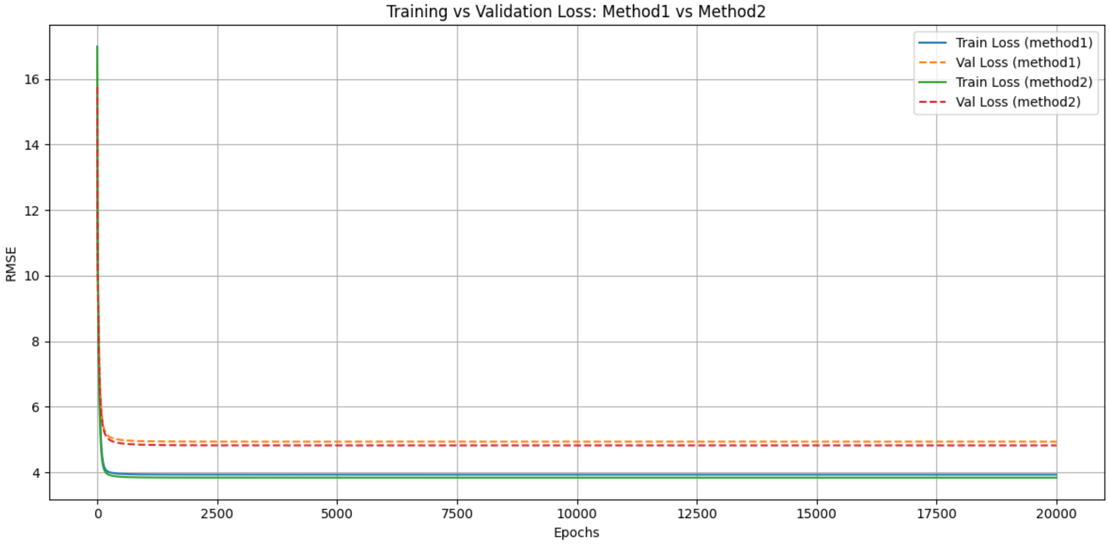


Fig2. Experimental results of different methods.

Another important preprocessing decision was determining how many past hours of data should be used as input features for the model. Intuitively, consecutive and more recent data points have higher relevance to the current PM2.5 concentration, so only continuous hourly data was considered. The test dataset constraints allowed a maximum input window of nine consecutive hours. A series of experiments were conducted using different input lengths, ranging from one to nine hours, to identify the optimal configuration. The results (Figure 3) indicated that although the performance of the models trained with eight and nine consecutive hours of data was quite similar during training, the actual testing scores revealed that the eight-hour configuration performed better. Therefore, the final model was trained using an eight-hour input window, as it provided more reliable results on unseen data.

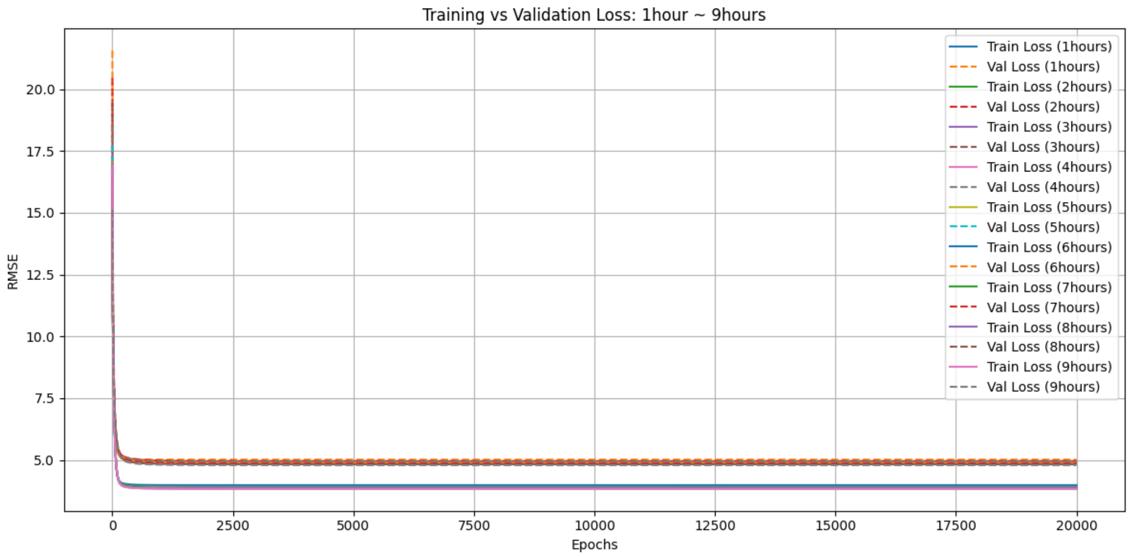


Fig3. Experimental results of different hours.

Finally, three normalization methods—Z-score, Robust scaling, and Min-max scaling—were evaluated to standardize the input data. As shown in Figure 4, robust scaling achieved the lowest training and validation RMSE, converging to around 4.0, while Z-score and min-max plateaued at approximately 5.0 and 6.0 respectively.

This superior performance is likely due to robust scaling's ability to handle outliers by using the median and interquartile range, which is especially important for air quality data that can contain sudden spikes. In contrast, Z-score is sensitive to outliers, and min-max scaling amplifies the influence of extreme values, leading to higher and less stable errors. Therefore, robust scaling was selected as the final normalization method.

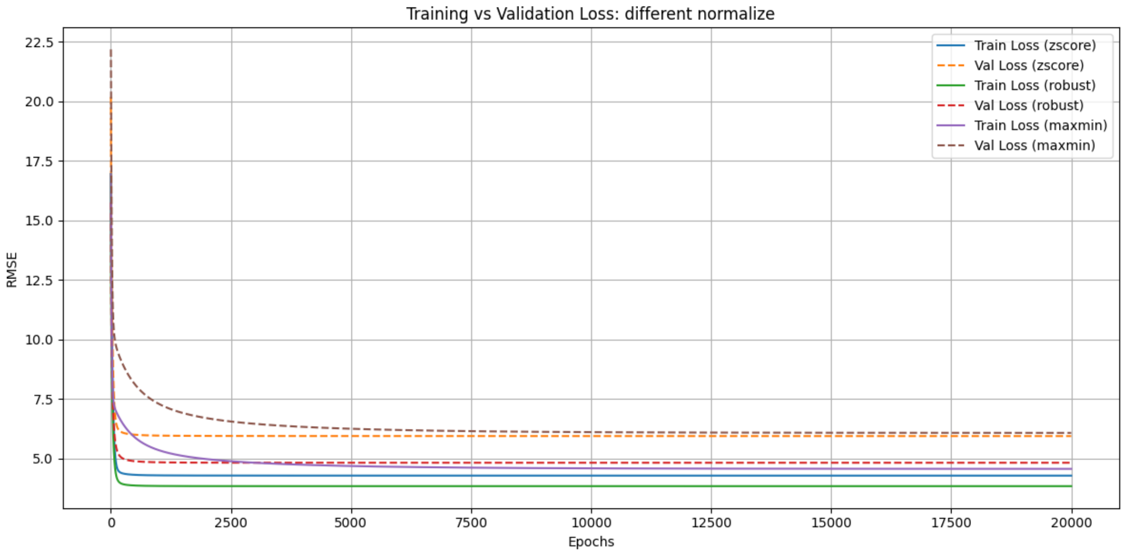


Fig4. Experimental results of different normalized methods.

In summary, preprocessing played a crucial role in improving the model's predictive accuracy. The selection of an appropriate imputation strategy, an optimal time window for input features, and a robust normalization technique significantly influenced the overall model performance. By carefully evaluating different options through experimentation, the most effective preprocessing techniques were identified, ensuring that the model was well-prepared to learn meaningful patterns from the data.

Compare the impact of different amounts of training data on the PM2.5 prediction accuracy. Visualize the results and explain them.

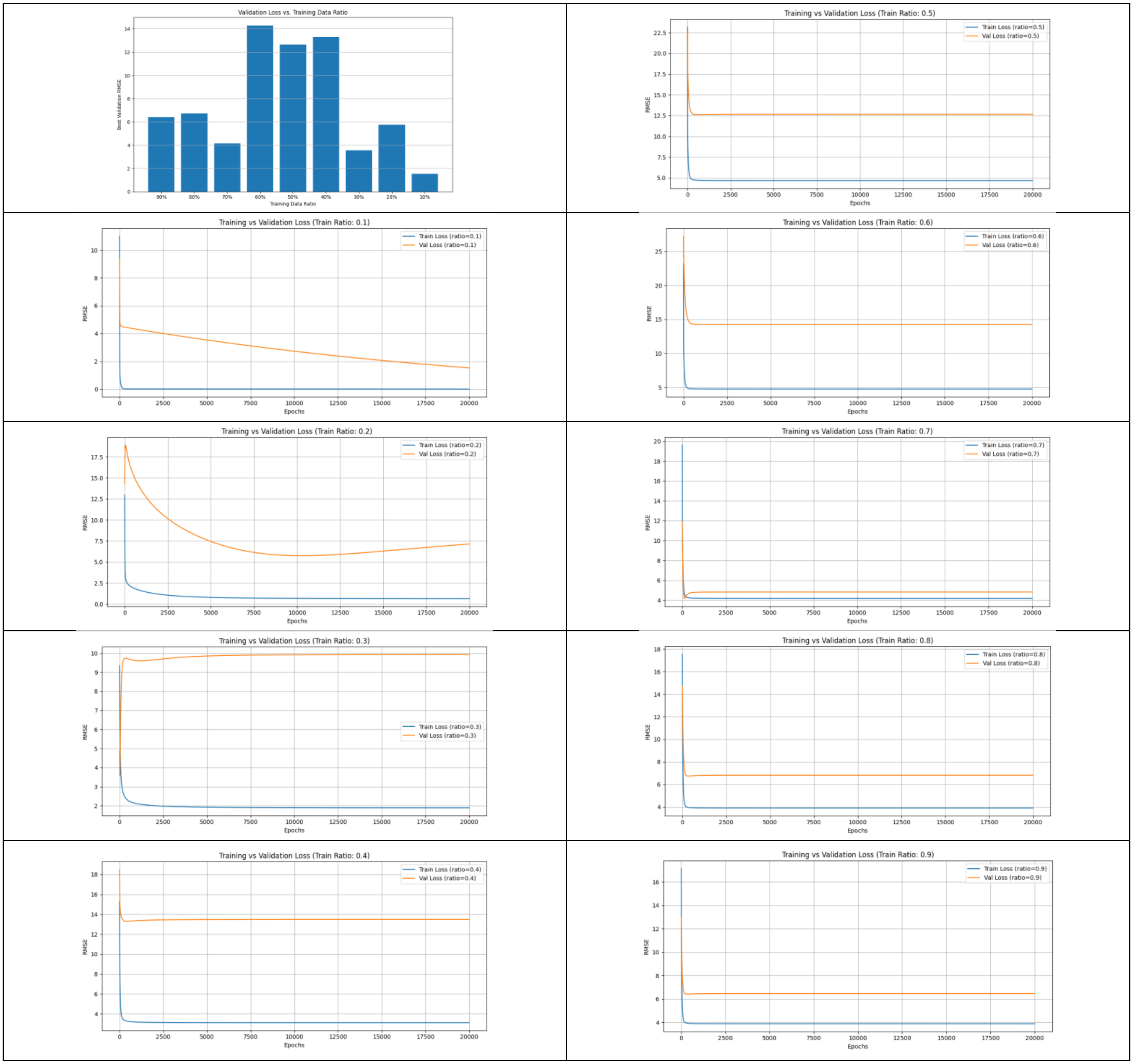
The impact of different amounts of training data on PM2.5 prediction accuracy is evident from the visualization (table 1.). The loss curves for various training ratios illustrate how the model learns and generalizes as the available training data changes.

For larger training data ratios, such as 90% and 80%, the training loss decreases steadily, and the validation loss follows a similar trend, although some overfitting may be present. The model benefits from having more data to learn from, leading to better generalization. However, as the training ratio decreases to 60% and 50%, the validation loss significantly increases, indicating that the model struggles to generalize. This suggests that with insufficient data, the model may not capture essential patterns, leading to poor predictive performance.

Interestingly, at lower training ratios, particularly 30% and 20%, the validation loss appears to improve again. This could indicate that while a smaller dataset limits the model's ability to capture complex relationships, it may also reduce exposure to noise, resulting in a more stable, albeit less complex, model. Additionally, the bar chart summarizing validation loss across different training ratios confirms that the worst performance occurs in the mid-range ratios, while the best validation loss is observed at either high or very low training ratios.

Overall, these results highlight that increasing training data generally improves accuracy, but there is a threshold where additional data provides diminishing returns or even introduces noise. The optimal training ratio depends on the dataset quality and the complexity of the underlying patterns.

However, due to the limited size of the training dataset—approximately 24 hours × 20 days × 12 months = 6,060 data points—I have decided to use the entire dataset (100%) for training. By utilizing all available data, I aim to maximize the model's ability to learn from as much information as possible. The downside of this approach is that without a separate validation set, I cannot assess whether the model is overfitting, which may impact its ability to generalize to unseen data.



Tab1. Loss curve of different ratio

Discuss the impact of regularization on PM2.5 prediction accuracy.

Regularization plays a crucial role in improving the generalization of a model by preventing overfitting. In the case of PM2.5 prediction, different regularization techniques have distinct effects on the model's accuracy, as shown in the visualized results.

From the validation RMSE comparison (table 2.), it is evident that the model without regularization achieves relatively low validation loss but does not necessarily outperform all regularized models. The training and validation loss curves for the normal model indicate that while the model learns quickly and stabilizes, there is still a risk of overfitting, as the validation loss does not decrease significantly after a certain number of epochs.

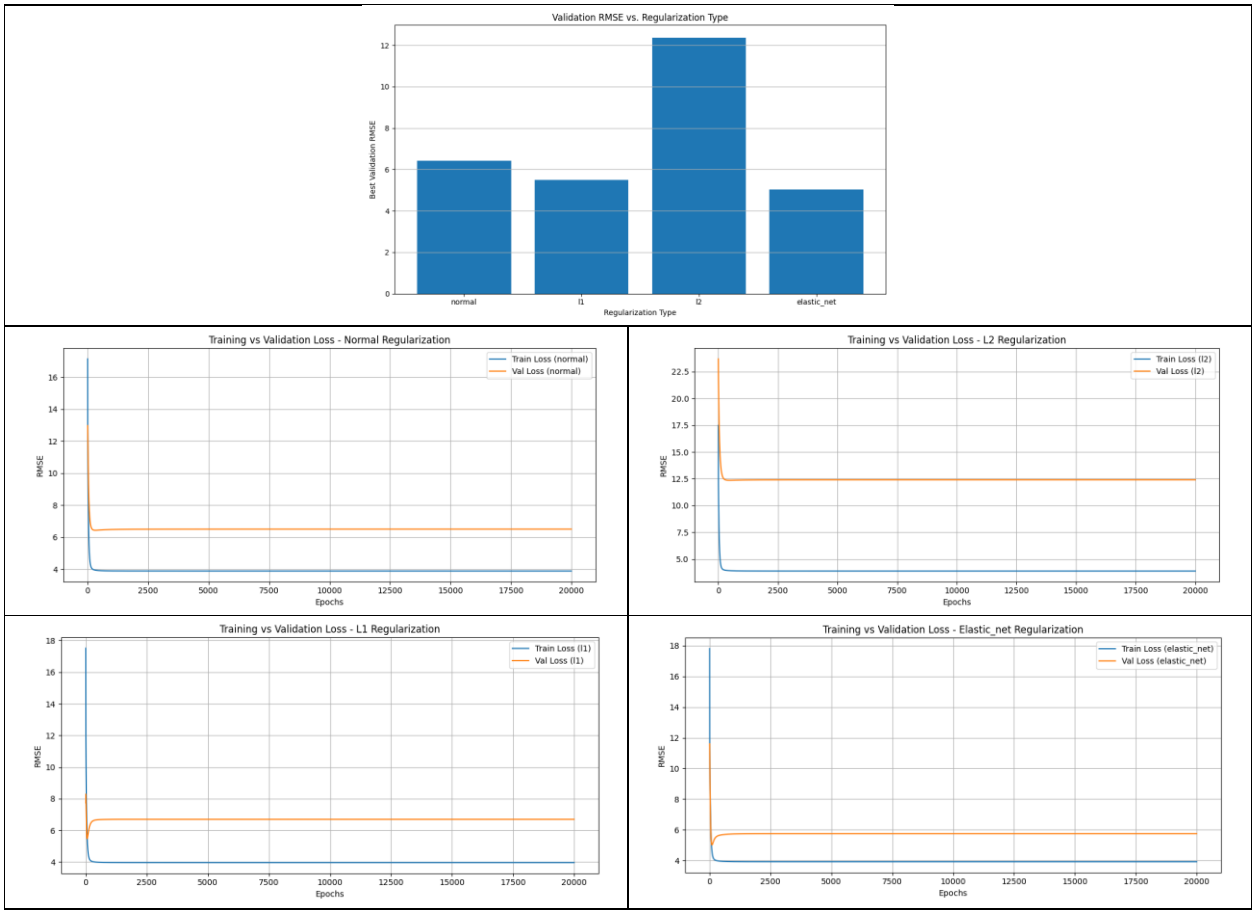
L1 regularization (Lasso) encourages sparsity by forcing some weights to become zero. This can be beneficial when dealing with high-dimensional data with redundant features. The results show that L1 regularization maintains a validation RMSE comparable to the normal model, suggesting that it may have helped the model focus on more relevant features while discarding noise.

L2 regularization (Ridge), on the other hand, penalizes large weights, leading to a smoother and more stable model. However, the validation loss for L2 regularization is noticeably higher, which suggests that while it prevents overfitting, it may also limit the model's flexibility, resulting in underfitting. This is further confirmed by its loss curve, where validation loss remains high throughout training.

Elastic Net regularization, which combines both L1 and L2 penalties, seems to provide the best performance in terms of validation RMSE. By balancing feature selection and weight shrinkage, it allows the model to generalize better than either L1 or L2 alone. The validation loss for Elastic Net remains lower, indicating that it effectively reduces overfitting while still maintaining the model's ability to learn complex patterns.

Overall, these results demonstrate that while regularization helps control overfitting, choosing the right type is essential. L1 and Elastic Net seem to be more effective in maintaining lower validation RMSE, whereas L2 may overly constrain the model, leading to higher prediction errors. For PM2.5 prediction, where noise and variability are inherent, Elastic Net appears to strike the best balance between bias and variance, making it a favorable choice for improving prediction accuracy.

Interestingly, when I submitted the predictions from all three regularization methods to Kaggle for evaluation, the model trained with L2 regularization achieved the best public leaderboard score. Although this might suggest that the model is overfitting to the training data, I ultimately chose to use the L2 method for the final submission due to its superior performance on the actual test set.



Tab2. The validation RMSE comparison