# Queen Mary University of London

EMS740U/P - Machine Learning and Artificial Intelligence

## Artificial Intelligence in Air Traffic Management (ATM)


| Group A9 | |
|---|---|
| Brandon Rutagamirwa | 210169058 |
| Zhen Wei Yap | 250045280 |
| Onur Kurt | 250914227 |
| Staines Rajith | 250790023 |
| Hardit Ravinder Saini | 220099811 |
| Osinachi Odemena | 251003948 |


### Introduction

With air traffic demand continuing to increase worldwide, the efficiency of airport airside operations has become a significant concern for the aviation industry. Specifically, predicting taxi time is necessary to optimise these functions. Taxi time, the duration an aircraft takes to travel between the gate and the runway, is subject to complex uncertainties related to airport congestion, aircraft speed, and operational geometry. It is necessary not only to develop robust flight timetables and identify choke points, but also to more accurately predict this variable, enabling government analysts to estimate airport capacity and compare the effects of regulations.

This case study examines Manchester International Airport (MAN), the second-busiest airport in the UK. The primary objective is to develop and compare predictive models that estimate taxi times using high-dimensional data comprising up to 25 operational features. To address the complexity of this data and reduce the risk of overfitting, we first applied Principal Component Analysis (PCA) to identify the most informative variables. Then we reduced the correlated variables to a lower-dimensional subspace.

Based on this processed data, we constructed and tested three different supervised learning structures: Linear Regression (LR) as a control structure to test a linear relationship; Neural Networks (NN), which is a Back-Propagation algorithmic framework, to test the complexity of non-linear hierarchies to establish the non-linear linkage; and Adaptive Neuro-Fuzzy Inference Systems (ANFIS) to take advantage of hybrid learning with interpretable fuzzy regulations. These models were trained and evaluated using the same training, validation, and test splits to enable a fair comparison. The performance of all models is assessed using the Root Mean Squared Error (RMSE), the Mean Absolute Error (MAE), the Coefficient of Determination (R2), and a statistical test of bias (Chicco, 2021).

**Data Preparation and Principal Component Analysis (PCA)**

The features.csv raw dataset contains 25 columns of correlated features, such as taxi distance and the number of aircraft currently on the runway, which directly influence taxi time. The primary issue of a high-dimensional dataset is the requirement of significant computational resources which comes at the cost of both time and financial burden for more potent hardware that can handle enhanced processing and analysis. Additionally, the abundance of features can lead to overfitting, a machine learning phenomenon in which models begin to capture noise and outliers rather than the underlying pattern due to excessive complexity (GeeksforGeeks, 2024). Regardless of overfitting, the model's interpretability becomes highly complex beyond 3 dimensions. Principle Component Analysis (PCA) is a dimensionality reduction technique which lessens high-dimensionality data sets while retaining important information by transforming correlated features into a smaller set of uncorrelated components (GeeksforGeeks, 2018) also known as principal components.

Prior to the application of PCA, the raw data was carefully treated to ensure that the analysis will accurately capture the underlying relationships within the dataset. Firstly, non-numerical features such as 'id' and 'rwy' were removed as they do not contribute to the numerical variance of the transformation of the data in any meaningful way. An outlier removal process was conducted using the interquartile range (IQR) method where any data outside of the 1.5 times IQR range was removed for continuous features such as 'distance' and 'angle'. This was to ensure that noise and clear outliers were extracted from the continuous features while the discrete columns such as 'QDepDep', which represents the current number of aircraft on the runway to depart, were omitted from this process as significant values often represented feasible situations. Furthermore, any rows with missing values were eliminated as PCA cannot handle null values. This method extracted clear anomalies while preserving meaningful data in extreme operational situations such as aircraft traffic on runways. Furthermore, 'angle error' was removed as it exhibited no variance and therefore would not influence the analysis. The 'type' feature had its string data converted to numerical by mapping 'arrival' to 0 and 'departure' to 1.

After preprocessing, it is imperative to compute the maximum possible variance in the dataset using principal components. First, each remaining feature was centre-shifted by computing the mean and subtracting it from all values in the column. This ensures that variance is captured relative to the mean of each feature, rather than by their absolute values. Following a similar methodology to Shizuya (2024), the covariance matrix was computed, from which the eigenvectors indicate the unit vector directions of the paths that capture the most variance in the data. This is paired with a corresponding eigenvalue, which quantifies the amount of variance captured along the eigenvector direction. The first principal component is the eigenvector associated with the largest eigenvalue; it is in the direction that captures the most variability and therefore the most information. The subsequent principal component, corresponding to the second-largest eigenvalue, is orthogonal to the first and captures the direction of the data with the second-largest variance. When selecting the appropriate number of principal components, it is critical to retain as much of the data's structure as possible while reducing dimensionality. To compute the variance and each principal component captures along with its approximation error, the equations below were used (Qmul.ac.uk, 2025):

$$Explained\ Variance = \frac{\sum_{i=1}^{K} \lambda_i}{\sum_{i=1}^{N} \lambda_i} \quad [1] \qquad \|\boldsymbol{x} - \hat{\boldsymbol{x}}\| = \frac{1}{2} \sum_{i=K+1}^{N} \lambda_i \quad [2]$$

Where $\lambda_i$ is the eigenvalues, K represents the principal components and N is the total number of principal components. Equation [2] is a theoretical measure of the reconstruction error. Information loss can also be computed by subtracting explained variance by one to determine discarded variance. The

table below details the cumulative effect of the principal components and their effectiveness in capturing the variance of the dataset:

*Table 1: Principal Component Selection Parameters*

| Principal Component | Cumulative Variance | Information Loss | Approximation Error |
|---|---|---|---|
| 1 | 0.7921 | 2.08E-01 | 1.91E+05 |
| 2 | 0.9887 | 1.12E-02 | 1.03E+04 |
| 3 | 0.9956 | 4.38E-03 | 4.03E+03 |
| 4 | 0.9997 | 2.90E-04 | 2.67E+02 |

Table 1 shows that selecting two principal components is appropriate as 98.87% of the variance is captured while significantly reducing dimensionality of the dataset. The approximation error is substantial however this can be attributed to the enormous dataset chosen for this analysis. Based off this decision, the PCA procedure was conducted:

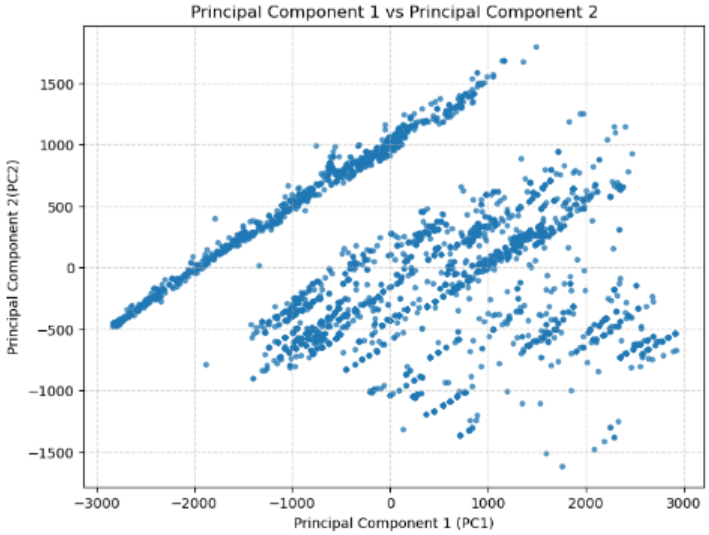


*Figure 1: Principal Component Analysis*

*Table 2: Top 5 most influential features for each Principal Component*

| Top 5 Feature Influence | |
|---|---|
| PC1 | PC2 |
| distance | distance_else |
| shortest_path | distance_long |
| distance_long | angle |
| distance_else | distance |
| angle | shortest_path |

PC1 is expected to account for the most significant proportion of variance in the dataset, as evidenced by its greater projection along its axis than along PC2. The distinct linear behaviour across the PCA plot can be attributed to similar top-feature influences on both principal components, as demonstrated in Table 2. There is a clear dominance of variability in the captured dataset among the continuous, geometric-based features. Table 1 indicates that the discarded principal components result in an information loss of 1.13% which is relatively insignificant and therefore provides confidence in the efficacy of the analysis.

**Linear Regression:**

A linear regression model is used to establish a relationship between one or more independent variables and the dependent variable. This model assumes the data for the independent variables are available; therefore, it can learn the relationships among them by estimating regression coefficients proportional to the weights of the independent variables. For this model, taxi time is treated as the dependent variable, with the PCA components as independent variables, allowing the model to learn the required coefficients by minimising the sum of squared residuals between the actual and predicted taxi time values. Linear regression models assume that a linear relationship exists between the independent and dependent variables and that the residual errors are independent and normally distributed with constant variance. The $R^2$ scores and p-values are found by evaluating the fitted regression model against the

observed data, which measures the proportion of variance explained and the statistical significance of the regression coefficients. The model used to predict taxi time is a multiple regression model, which is shown as:

Y = a +$b_1X_1$ + $b_2X_2$ + u:

- Y = Dependent variable
- X = Independent variable
- a = y-intercept
- b = regression coefficients
- u = regression residual

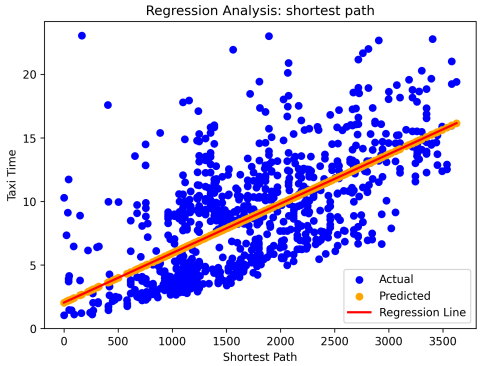


Figure 2

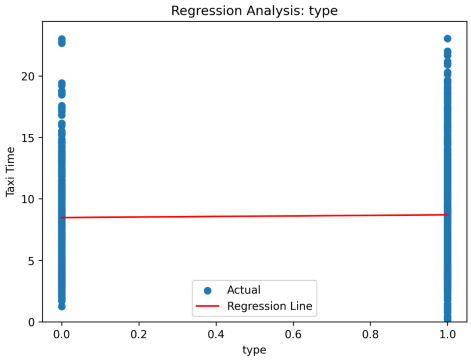


Figure 3

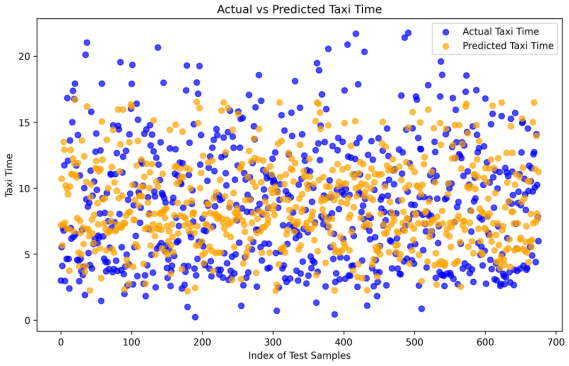


Figure 4

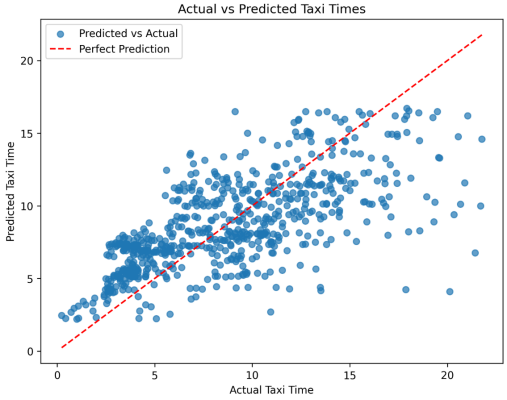


Figure 5

Based on the $R^2$ scores, the "shortest path" exhibited the strongest linear relationship with taxi time, accounting for approximately 43.5% of the variance when considered individually, as shown in Figure ... Through the "shortest path" feature's variance value, a strong claim can be made that it is the most influential predictor of taxi time among all variables analysed. However, the "type" feature achieved an $R^2$ of 0.000529, indicating that it explains only approximately 0.053% of the variance in taxi time and therefore provides minimal predictive value when used independently, as shown in Figure 3. As shown in Figure 4, the actual and predicted taxi times align closely; however, noticeable deviations occur at higher taxi times, indicating reduced predictive accuracy for extreme cases. Moreover, an MSE of 11.07 reflects the average squared difference between the actual and predicted taxi times. Due to squaring errors, larger deviations are penalised more heavily, making the MSE particularly sensitive to outliers. The RMSE is 3.33 minutes, which is typical for prediction error in the same units as taxi time.

**Neural Networks:**

Advances in machine learning (ML) have relied heavily on deep learning methods, particularly neural networks (NNs), owing to their ability to learn hierarchical feature representations from large-scale data automatically. By stacking multiple nonlinear layers, NNs can model complex, high-dimensional relationships that were difficult to capture using traditional ML approaches. In this study, a Multiple-Layer Perceptron (MLP) architecture was implemented to predict aircraft taxi time using principal components derived from PCA.

A training set was used to update the weights of the NN, while the validation set was used to monitor generalisation performance and detect overfitting, and the test set was held out entirely for final performance evaluation.

To ensure efficient convergence during gradient descent, the input features were standardised to have a mean of 0 and a standard deviation of 1. Following data preprocessing, a Backpropagation Neural Network (BPNN) was implemented using NumPy.

| Layer | How many Neurons? Is there a Bias? | What was the Activation Function? |
|---|---|---|
| Input Layer | 2 (with bias) | n/a |
| Hidden Layer | 12 (with bias) | Sigmoid |
| Output Layer | 1 | Linear |

*Table 3: Architecture and Activation Functions of the MLP BPNN*

A single hidden layer was used for prediction, given the low-dimensional input space obtained via PCA. Furthermore, a single, compact hidden layer for nonlinear relationships while maintaining generalisation. 12 neurons were selected from the hidden layer to capture nonlinear patterns without increasing complexity. Balancing the model's efficiency and generalisation is necessary to achieve close alignment among training, validation, and test losses. (Gardner, 1998)

The training algorithm used batch gradient descent for 1000 epochs with a learning rate of 0.01. This training involved:

1. Forward Propagation: Computing the dot product of inputs and weights, applying the sigmoid activation for the hidden layer, and calculating the linear output.
2. Loss Calculation: The Mean Squared Error was used as the loss function to quantify the difference between predicted and actual taxi times.
3. Back Propagation (BP): Gradients of the weights were computed using the chain rule. BP was then used to adjust the weights to minimise loss.

Model performance was evaluated on the training, validation, and test datasets using three statistical metrics: Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and the coefficient of determination ($R^2$). These metrics were used to analyse and visualise the model's behaviour and error characteristics through loss curves, predicted-versus-actual plots, residual plots, and residual distribution histograms. The datasets used for this evaluation were train_reduced_gA.csv, validation_reduced_gA.csv, and test_reduced_gA.csv.

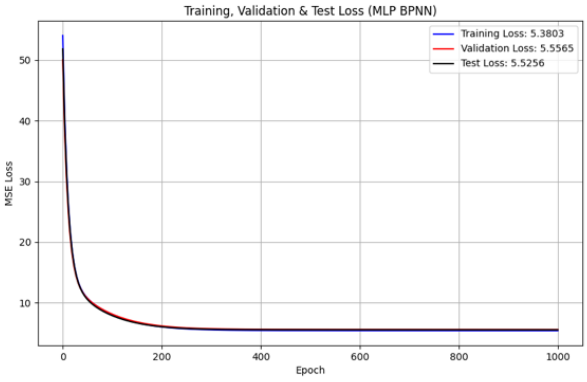


*Figure 6: Training, Validation, and Test Loss Curves*

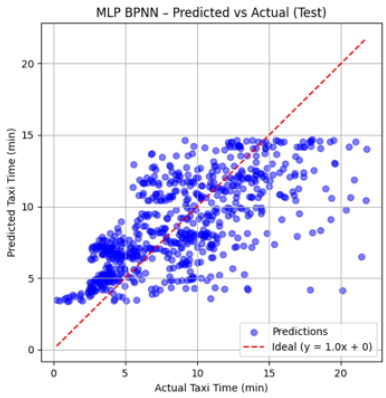


*Figure 7: Predicted vs Actual taxi times*

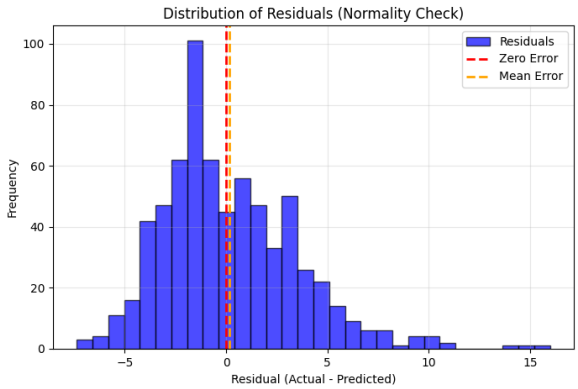


*Figure 8: Histogram of Prediction Residuals*

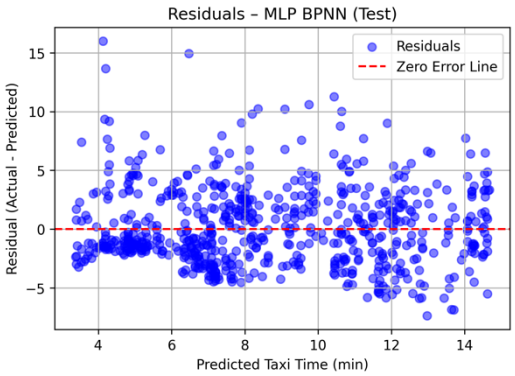


*Figure 9: Residuals vs Predicted taxi times*

The results shown in Figures 6–9 demonstrate that the MLP achieves stable learning and consistent generalisation across the training, validation, and test datasets. Training required approximately 2.5 minutes for 1000 epochs using a NumPy-based batch gradient descent implementation, reflecting the computational cost of a manual backpropagation approach. The loss curves plateau after approximately 300 epochs, indicating that the model has converged and that further training yields diminishing performance gains. Final mean squared error (MSE) values of 5.38, 5.56, and 5.53 were obtained for the training, validation, and test sets respectively, indicating minimal overfitting and good generalisation to unseen data.

As shown in Figure 7, the predicted versus actual taxi times exhibit a strong positive correlation, although greater dispersion is observed at higher values. This behaviour is further supported by the residual analysis presented in Figure 8, which shows errors centred around zero with an approximately bell-shaped distribution and increasing variance at higher predicted taxi times, indicating that the assumption of constant error variance does not hold in real-world traffic data. Overall, the model achieves a test RMSE of 3.32 minutes and an $R^2$ value of 0.46, indicating moderate predictive performance given the stochastic nature of aircraft taxi operations. While the model demonstrates strong generalisation capability, its transparency is limited by the difficulty of interpreting interactions among internal weights, reflecting the "black box" nature of neural networks. Consequently, the model's primary strength lies in its predictive performance and generalisation capability rather than interpretability.

**ANFIS**

The Adaptive Neuro-Fuzzy Inference System (ANFIS) is a hybrid computational system that combines the learning capabilities of artificial neural networks and the reasoning capabilities of fuzzy logic. ANFIS is effective at modelling nonlinear, complex relationships because it employs a hybrid learning process to adjust its parameters (Al-Hmouz, 2011). The method combines the flexibility of machine learning with the explainability afforded by the pertinent rules of fuzzy inference.

This comparison of results is presented in Tables 3 and 4 for the performance of ANFIS with cluster radii of 0.25, 0.30, and 0.501. The cluster radius of 0.30 yielded the best overall performance, with the lowest Root Mean Square Error (RMSE) of 3.0736 and the highest coefficient of determination (R^2) of 0.53752.

*Table 3*

| Cluster Radius | RMSE | MAE | R2 |
|---|---|---|---|
| 0.25 | 3.0958 | 2.3433 | 0.5308 |
| 0.30 | 3.0736 | 2.3598 | 0.5375 |
| 0.50 | 3.1978 | 2.4897 | 0.4994 |

*Table 4*

| Cluster Radius | Mean Error | 95% CI Lower | 95% CI Upper | Conclusion |
|---|---|---|---|---|
| 0.25 | -0.0919 | -0.3258 | 0.1420 | Unbiased |
| 0.30 | -0.0748 | -0.3071 | 0.1574 | Unbiased |
| 0.50 | -0.1006 | -0.3422 | 0.1410 | Unbiased |

For a cluster radius of 0.25, the model with radius=0.25, as shown in Figures 10 and 11, exhibits a smooth decrease in RMSE. Still, a slight difference between training and validation errors indicates slight overfitting. The regression plot affirms a moderate correlation (R^2 = 0.531).

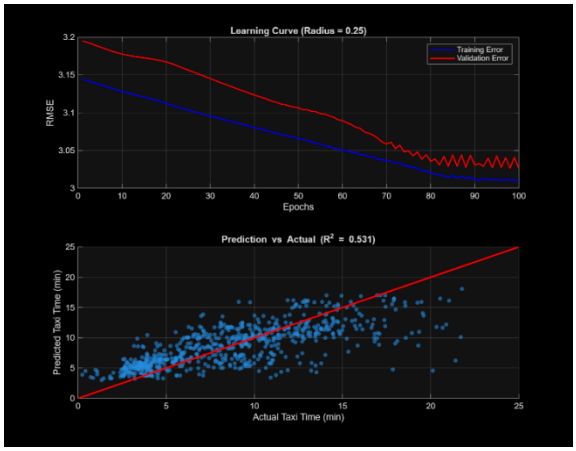


*Figure 10*

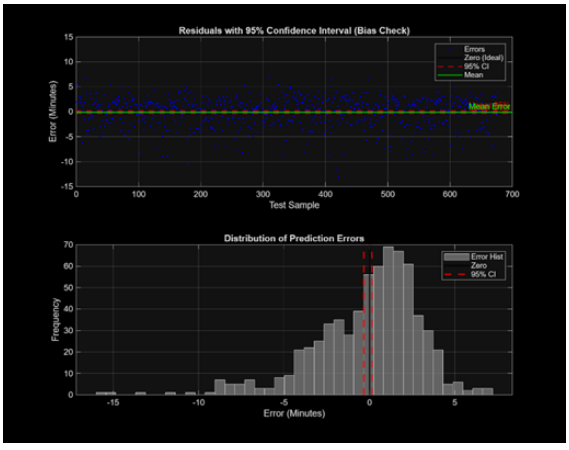


*Figure 11*

Statistical checks indicate that errors are randomly distributed around zero. The histogram indicates a normal distribution with a mean near zero, which supports the conclusion that the model is unbiased.

For a cluster radius of 0.30 (optimal model), as shown in Figures 12 and 13, this radius made the best predictions. The learning curve is smoother than in the 0.25 case, indicating better generalisation. In addition, the data points cluster closely along the regression line, yielding the highest R^2 of 0.53711.

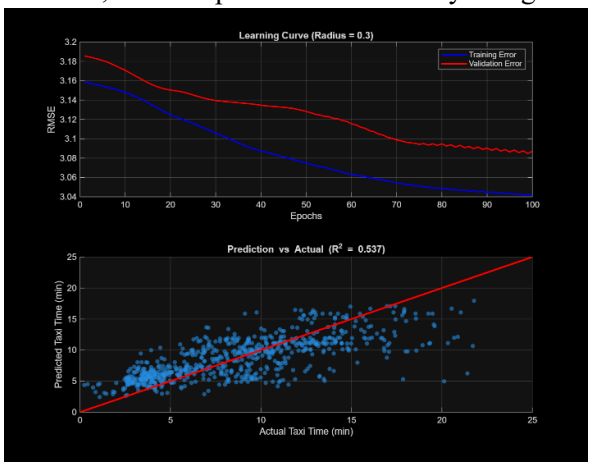


Figure 12

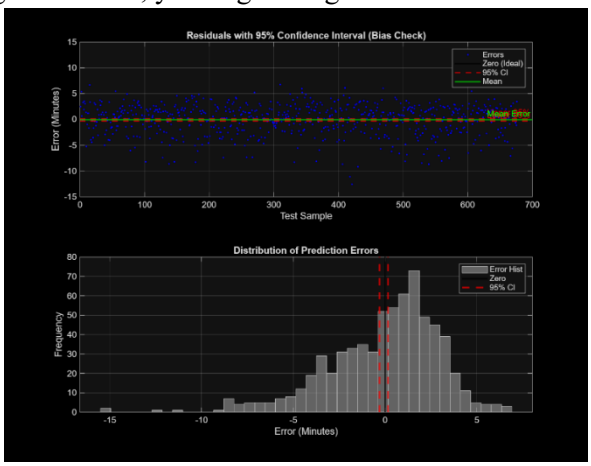


Figure 13

This model is robust, as shown by the residual analysis. The error distribution is symmetric, and the mean error is -0.0748, the nearest to zero across all situations.

For a cluster radius of 0.50, as shown in Figures 14 and 15, performance decreased as the radius increased to 0.50. The error curve associated with validation levels off earlier, resulting in a larger final RMSE. The regression analysis shows greater scatter and a lower R^2 of 0.49916.

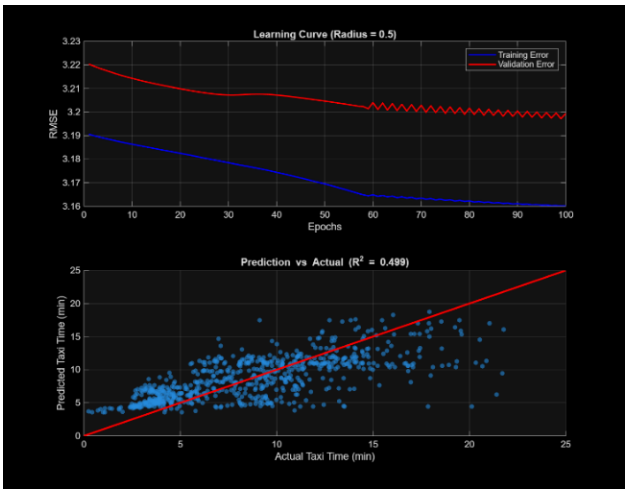


Figure 14

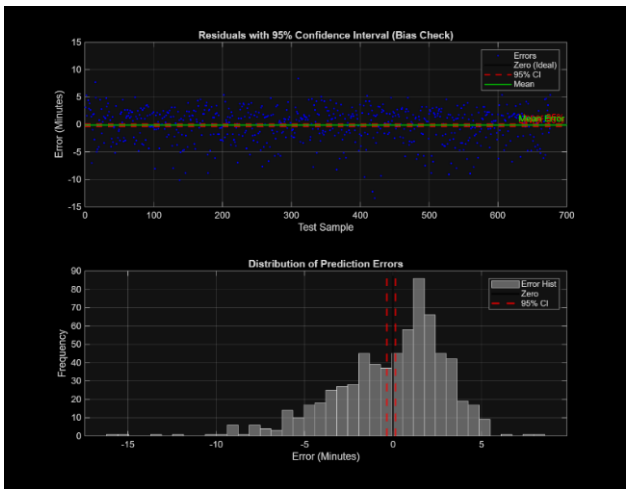


Figure 15

The model is statistically valid, although the mean error is imprecise and falls within the confidence interval. Nevertheless, the histogram shows a broader distribution of mistakes than that of the optimal model. The optimal ANFIS model generated eight logical rules via subtractive clustering. These regulations specify the mapping between the input variables and the predicted output.

## Comparison

To identify the most effective method for predicting the time required for aircraft taxiing at Manchester International Airport, we conducted a comparative study of three models: linear regression (LR), a

neural network (NN), and an adaptive neuro-fuzzy inference system (ANFIS). This analysis will focus on three key dimensions: predictive accuracy, generalisation capacity, and model transparency.

With respect to raw predictive ability, the ANFIS architecture was the strongest, with the lowest root-mean-square error (RMSE) of 3.07 minutes and the highest coefficient of determination (R^2) of 0.54, accounting for more than half of the variation in taxi times. In comparison, the LR and NN models exhibited high error rates. The LR variant, which was primarily motivated by the shortest-path feature, achieved an RMSE of 3.33 minutes (R^2=0.44), whereas the NN configuration achieved a similar RMSE of 3.32 minutes (R^2=0.46).

More importantly, the statistical tests of bias support the ANFIS model's accuracy. A t-test on the residuals indicated a statistically neutral effect, with a mean error of -0.07 minutes and a 95 per cent interval of [-0.31, 0.16] that includes zero. The residuals of the NN and LR models are also centred around zero. Still, the narrower ANFIS confidence interval indicates a reduced tendency toward systematic drift under more complex operating conditions.

Generalisation refers to a model's ability to perform well on unseen data. The NN exhibited stable generalisation, as indicated by the convergence of the training, validation, and test loss curves after 300 epochs. The fact that it is a similar performer to the base LR model, however, raises the possibility of challenges in capturing deep, nonlinear patterns in the sparser PCA-reduced features. LR (which is a linear model by definition) had a problem of underfitting; it could generalise reasonably well only insofar as it was able to ignore the stochastic complexity of ground movements. The ANFIS model, using subtractive clustering to partition the data into logically coherent rules, achieves an optimal balance and generalises well to the test dataset without overfitting, as indicated by the close match between its training and validation errors.

There is a clear trade-off between transparency and accuracy. The LR model offers the highest level of transparency: the coefficients are directly interpretable, and analysts can accurately quantify the effects of specific variables (e.g., distance or angle) on taxi time. ANFIS is a grey-box system that offers a trade-off: the decision logic can still be provided in fuzzy form (e.g., matching cluster radii to the level of congestion). On the contrary, the NN is a black box. Though it models non-linearities, its internal decision-making, distributed to the twelve hidden neurons with sigmoid activation, is opaque.

Although the LR model is the easiest to interpret and the NN can scale to larger datasets, ANFIS is the best of the three models in this application. It skilfully combines the interpretability of fuzzy logic with the learning capabilities of neural networks, thereby achieving high accuracy and statistical reliability.

## Conclusions

In this study, a comparative analysis of Linear Regression (LR), Neural Networks (NN), and Adaptive Neuro-Fuzzy Inference Systems (ANFIS) was conducted to predict aircraft taxi time at Manchester International Airport. The empirical results identify ANFIS as the best architecture, achieving the highest accuracy (R^2 = 0.54, RMSE = 3.07 min) while maintaining the necessary grey-box interpretability. NN, on the other hand, exhibited black-box opacity, and LR proved insufficient for the complex nonlinear interactions inherent in ground operations.

The transformation of these models into operational Air Traffic Management (ATM) systems requires consideration of several critical implementation issues. The explainable artificial intelligence is first needed to comply with the safety roadmap of the European Union Aviation Safety Agency (EASA); this is why hybrid models like ANFIS are more appropriate to be certified than opaque NNs. Second, the systems will need to be developed to be based not only on historical stationary data but also on real-

time surveillance inputs (e.g., Advanced Surface Movement Guidance and Control System, A-SMGCS) and meteorological information, thereby making them more resistant to data drift (Group, 2020). Lastly, given the approximately 46% unexplained variance, a human-in-the-loop architecture is necessary. The tools must be used solely as decision-support tools and therefore allow controllers to intervene in unusual occurrences (Paul, 2023).

Finally, although ANFIS is a more advanced predictive model, its implementation in critical infrastructure requires strict verification of safety and transparency, as well as effective real-time data integration (Kabashkin, 2023).

**References**


Aishwarya. (2025, November 13). *GeeksforGeeks*. Retrieved from Principal Component Analysis (PCA): https://www.geeksforgeeks.org/data-analysis/principal-component-analysis-pca/

Al-Hmouz, A. a.-H. (2011). Modeling and simulation of an adaptive neuro-fuzzy inference system (ANFIS) for mobile learning. *IEEE transactions on learning technologies*, 226--237.

Chen, J. (2022). *QMPlus*. Retrieved from EMS740U/P - 2025/26. Week 3.3 - Dimensionality Reduction & Principal Component Analysis: https://qmplus.qmul.ac.uk/pluginfile.php/4778850/mod_resource/content/3/EMS740UP%20W3.3%20Slide.pdf

Chicco, D. a. (2021). he coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation. *Peerj computer science*, e623.

Gardner, M. (1998 ). Artificial neural networks (the multilayer perceptron)—a review of applications in the atmospheric sciences. *Atmospheric Environment* , 32 .

Group, E. A. (2020). The FLY AI Report Demystifying and Accelerating AI in Aviation/ATM. EUROCONTROL Brussels, Belgium .

Kabashkin, I. a. (2023). Artificial intelligence in aviation: New professionals for new technologies. *Applied Sciences*, 11660.

Mehta, L. (2025, July 23). *GeeksforGeeks*. Retrieved from Managing High-Dimensional Data in Machine Learning: https://www.geeksforgeeks.org/machine-learning/managing-high-dimensional-data-in-machine-learning/

Paul, S. a. (2023). Assurance of Machine Learning-Based Aerospace Systems: Towards an Overarching Properties-Driven Approach.

Shizuya, Y. (2024). *Mathematical understanding of Principal Component Analysis*. Retrieved from Medium: https://medium.com/intuition/mathematical-understanding-of-principal-component-analysis-6c761004c2f8