## Discussion

The objective of the numerical evaluation was to compare predictive capabilities of the various models employed on binary classification tasks for FX time-series forecasting, particularly on the JPY/USD exchange rate (1-day, 5-day, and 10-day forward binary returns). The models were assessed based on their ability to generalise across unseen data and to distinguish directional market movements. In our study, two key evaluation metrics were utilised:

**Training and Validation Accuracy:** Measure how well each model fit the training data and how well it generalises to the validation data.

**Test AUC:** Measure a threshold-independent performance measure, focusing on the model’s discriminative power.

An important pattern observed across all models shows that predictive performance tended to be better for the 10-day forward return classification task compared to 1-day and 5-day. To be precise, all models demonstrated higher validation accuracies and test AUCs when tasked with predicting the 10-day forward return.

A potential reason for this trend could be that short-term returns (1-day and 5-day) tend to be much noisier, heavily influenced by microstructure effects, random price perturbation, and short-term market sentiments shifting. On this note, it could be extremely difficult for machine learning models to predict reliably. On the other hand, longer-term returns (10-day) may better reflect underlying macroeconomic forces, broader market trends, and technical indicators, which are considerably more stable and thus, learnable over time by sequential models deployed. To add on, longer forecast horizons also smooth out the random walk characteristics of price changes, making directional movements easier to classify.

## Results

Based on all models generalising better on the 10-day forward return, shown below are the numerical results of each model on 10-day forward return.

|  |  |  |  |
| --- | --- | --- | --- |
| Model | Train Accuracy | Validation Accuracy | Test AUC |
| LSTM | 52.5% | 36.8% | 57.2% |
| Transformer | 61% | 59% | 60% |
| GRU | 55.2% | 54.1% |  |
| CNN-Transformer | 75.2% | 67.6% | 64% |

Each model showed distinct characteristics in terms of training stability and predictive performance. Shown below are the pros and cons of each based on the results.

|  |  |  |
| --- | --- | --- |
| Model | Pros | Cons |
| CNN-Transformer | * Achieved the highest training and validation accuracy, and the best test AUC * Effectively captured both local patterns (via the CNN layer) and long-range dependences (via self-attention) * Regularisation terms like dropout and L2 penalty helped to mitigate overfitting despite its complexity | * There is a noticeable gap between training and validation accuracy, suggesting mild overfitting |
| Transformer | * Good generalisation with close training and validation accuracy values * Demonstrated that a light-weight Transformer (only three encoder blocks) can effectively model sequential relationships found in FX time-series | * Lacked the convolutional feature extraction power as seen in the CNN-Transformer model, slightly limiting performance * Slightly lower AUC compared to the CNN-Transformer, indicating room for improvement |
| GRU | * Simpler architecture and efficient training, supported by effective regularization strategies (dropout, batch normalization, L2 regularization). | * May underperform in highly volatile or deeply non-linear regimes compared to models with more flexible attention mechanisms. |
| LSTM | * Theoretically well-suited for long-sequence modeling due to its memory gates and attention integration. | * Poor validation performance and low test AUC, indicating generalisation issues * Susceptible to overfitting or getting trapped in local minima when trained on noisy data * More parameter-heavy than GRU, making it less stable |

## Interpretations

The potential success of the CNN-Transformer model can be attributed to its ability to integrate local feature extraction (via convolutional layers) and global temporal dependencies (via multi-head self-attention), thus, being able to model both short-term and long-term patterns.

The Classic Transformer demonstrated that even with a reduced number of encoder layers, self-attention mechanisms are effective in capturing sequence-wide relationships without relying on complex structures.

The GRU model’s strong performance reinforces the notion that simpler recurrent architectures, when modified with careful regularisation techniques, can still be effective.

Lastly, the LSTM model, despite its theoretical advantages for long-sequence memory retention, can struggle in practice, possibly due to model complexity and sensitivity to noisy financial data.

## Conclusion

The results suggest that deep learning models, particularly those that combine convolutional and attention mechanisms like the CNN-Transformer, show strong potential for financial time-series forecasting. Despite that, the challenges remain in modelling extremely short-term returns due to inherent noise, as evident by the poor model accuracy on 1-day and 5-day forward returns. Future work could explore hybrid modelling approaches, data augmentation strategies, and enhanced regularisation to prevent overfitting to improve robustness of model and enhance predictive accuracy.