# Rethinking Recession Prediction: Comparing Classical and Deep Learning model using Yield Curve Spreads

Gordon Oboh and Yu-jiun Chen

## I. Abstract

This study investigates the effectiveness of classical machine learning models versus deep learning approaches in forecasting U.S. recessions using yield curve spreads. Specifically, we use the 10-year minus 3-month (10Y–3M) and 10-year minus 2-year (10Y–2Y) Treasury yield spreads as predictors, paired with recession labels derived from the National Bureau of Economic Research (NBER). We evaluate Logistic Regression, Easy Ensemble, and Long Short-Term Memory (LSTM) neural networks across multiple time frequencies. Our results show that traditional models (Logistic Regression and Easy Ensemble) consistently outperform LSTM models in terms of AUC-ROC scores, suggesting greater robustness and interpretability. However, LSTM's probability plots for monthly data reveal potential in capturing longer-term trends. These findings highlight the predictive power of simpler models in recession forecasting while leaving room for further exploration of deep learning architectures with alternative configurations.

**Keywords:** Recession prediction; Yield curve spread; Logistic Regression; Easy Ensemble; Long Short-Term Memory (LSTM); Time series forecasting; AUC-ROC

## II. Introduction

A recession is generally defined as a significant decline in economic activity lasting more than a few months, typically visible in key indicators such as real GDP, income, employment, industrial production, and wholesale-retail sales (National Bureau of Economic Research, n.d.). Translating this qualitative definition into a quantitative forecasting challenge has prompted economists and data scientists to investigate the predictive power of financial indicators

Bauer and Mertens (2018) underscore the predictive accuracy of the 10-year minus 1-year Treasury yield spread, noting that yield curve inversions have preceded every U.S. recession over the past 60 years, with only one false positive. Likewise, Aramonte and Xia (2019) emphasize the historical reliability of the 10-year minus 3-month spread, which inverted prior to each U.S. recession since 1973. These findings have reinforced the yield curve's status as one of the most consistent and early indicators of impending downturns (Aramonte & Xia, 2019; Bauer & Mertens, 2018).

Corresponding author: goboh@position.8shield.net
Corresponding author: joanne.chen97@gmail.com

Despite the rich literature on this topic, several gaps persist. Many studies focus narrowly on a single spread, such as the 10Y–3M, overlooking potentially informative alternatives like the 10Y–2Y (GS10–DGS2) or 10Y–3MO (GS10–DGS3MO). Additionally, the use of high-frequency data remains limited, with most research relying on monthly or quarterly averages. This may obscure important short-term dynamics critical for real-time forecasting. There is also insufficient integration of modern machine learning methods with macroeconomic data. While some work explores the use of classification algorithms, few studies address the class imbalance inherent in recession prediction tasks. Furthermore, side-by-side comparisons of classical statistical models and deep learning techniques remain rare, leaving unresolved questions about which modeling approaches are most robust across time scales and data regimes.

This study addresses these gaps by evaluating a range of machine learning models (Logistic Regression, Easy Ensemble, and LSTM networks) for recession prediction using daily and aggregated yield curve spread data. Theoretically grounded in the Expectations Hypothesis, which posits that long-term interest rates reflect average expected short-term rates, the research tests whether inversion patterns in the yield curve reliably signal recessions across different model classes and time frequencies (Estrella & Mishkin, 1998). The inclusion of ensemble and deep learning methods aims to assess whether newer, more complex techniques provide added predictive value over traditional models, especially in the context of imbalanced datasets.

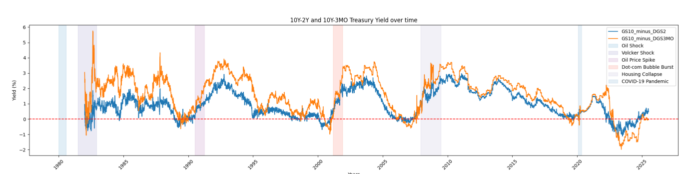


Figure 1: Time series of the 10Y–3M (GS10–DGS3MO) and 10Y–2Y (GS10–DGS2) yield spreads from January 1982 to July 2025. Shaded regions indicate NBER-designated recessions.

## III. Literature Review

A substantial body of literature affirms the predictive power of the yield curve in forecasting U.S. recessions. The term spread has long been considered a leading economic indicator. Estrella and Mishkin (1998) and Bauer and Mertens (2018) empirically demonstrate that inverted yield curves precede virtually every U.S. recession over the past six decades, with only one false positive. Aramonte and Xia (2019) confirm the reliability of the 10-year minus 3-month (10Y–3M) spread, emphasizing that each U.S. recession since 1973 was preceded by an inversion. These studies reinforce the notion that yield curve inversion reflects investors' expectations of declining economic activity and accommodative monetary policy.

Despite this consensus, recent discussions have raised questions about the robustness of the yield curve's signal, especially in environments with exceptionally low term premia or unconventional monetary policy. For instance, Aramonte and Xia (2019) caution that factors such as central bank asset purchases and price-insensitive investors may distort the informational content of the term spread. Nevertheless, statistical analyses from Bauer and Mertens (2018) and others suggest that the predictive power of the yield curve remains intact even after accounting for such structural changes.

Beyond traditional econometric analysis, a growing number of studies have applied machine learning techniques to yield curve data. Choi et al. (2023) explore recession prediction using various yield spreads and machine learning classifiers, including random forests and gradient boosting machines. Their findings suggest that yield spreads containing medium- and long-term maturities may outperform the widely used 10Y–3M spread. They also emphasize the importance of variable selection and model-specific tuning for improving forecast accuracy.

While much of the prior literature focuses on monthly or quarterly data, relatively few studies explore the utility of high-frequency (daily or weekly) yield spreads for near-term recession forecasting. This creates a gap in understanding whether classical or deep learning models perform better at such granular levels. Furthermore, issues such as class imbalance and temporal dependencies are often under-addressed in traditional forecasting frameworks. Most existing research also fails to compare a broad set of model architectures within the same empirical setup.

This study aims to bridge these gaps by evaluating the performance of Logistic Regression, Easy Ensemble, and Long Short-Term Memory (LSTM) networks in predicting U.S. recessions. Unlike many previous works, this research incorporates multiple yield curve spreads (10Y–3M and 10Y–2Y), explores varying time frequencies, and directly addresses class imbalance. By juxtaposing classical and deep learning methods under identical data conditions, this paper contributes to a growing literature on the application of machine learning in macroeconomic forecasting.

## IV. Methodology

### A. Data Preparation

This study uses macro-financial data from the U.S. Treasury yield curve, spanning the period from January 1, 1982 to July 18, 2025. The dependent variable is a binary recession indicator derived from National Bureau of Economic Research (n.d.), with a label of 1 indicating a period of recession and 0 otherwise. The independent variables include two yield spread features: the 10-year minus 3-month (10Y–3M) and the 10-year minus 2-year (10Y–2Y) Treasury spreads, corresponding to the GS10–DGS3MO and GS10–DGS2 series.

Missing values in the time series were handled using forward fill. The feature set was resampled to three distinct time frequencies; daily, weekly, and monthly. The aim is to evaluate model performance under different temporal granularities. Weekly values were computed as Friday averages, shifted back by two days to align with business days. Monthly values were computed using month-end averages and adjusted back by 14 days to account for mid-month representation.

### B. Modeling Approaches

We evaluated a range of classification models for recession prediction, including Logistic Regression, Easy Ensemble, and Long Short-Term Memory (LSTM) networks. Each model was trained on the same feature sets across different time frequencies.

- Logistic Regression was implemented with balanced class weights, 1,000 maximum iterations, and a fixed random state of 42.
- Easy Ensemble Classifier used 200 estimators with a fixed random state of 42.
- LSTM Models were implemented with TensorFlow and configured with 1–2 hidden layers, 4 or 8 units per layer, a dropout rate of 0.3, input sequence of length 32, and a sigmoid activation function in the output layer. The optimizer used was Adam. The LSTM loss function employed binary focal loss, defined as:

$$\mathrm{FL}(p_t) = -\alpha_t (1 - p_t)^{\gamma} \log(p_t) \tag{1}$$

where $\gamma = 2$, and $\alpha$ was dynamically computed from class weights as:

$$\alpha = \frac{\omega_1}{\omega_0 + \omega_1} \tag{2}$$

To handle class imbalance in Logistic Regression, a pipeline using Synthetic Minority Oversampling Technique (SMOTE) followed by Random UnderSampling was applied. The sampling strategies were 0.5 and 1.0, respectively. Other models, including Easy Ensemble and LSTM, inherently handled imbalance either through their architecture or custom loss functions.

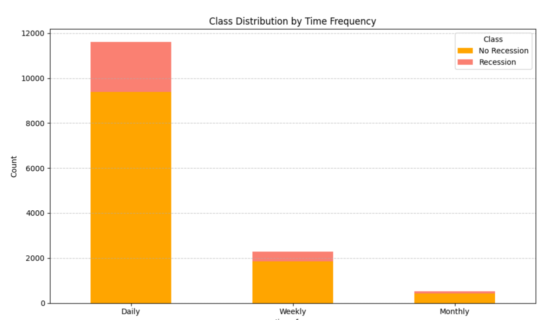


Figure 2: Stacked bar chart illustrating class imbalance across daily, weekly, and monthly datasets, emphasizing the disproportionate representation of non-recession versus recession periods.

### C. Train-Test Splitting and Evaluation

The dataset was chronologically split into training and test sets at January 1, 2015, resulting in a rough 68/32 training-test ratio across all time frequency. No explicit validation set was used; however, the LSTM models internally leveraged a validation split during training for early stopping and weight restoration. Model performance was primarily evaluated using the Area Under the Receiver Operating Characteristic Curve (AUC-ROC) on the held-out test set.

## V. Results

### A. Key Findings Overview

Model performance was evaluated using AUC-ROC$_{test}$ scores on the test set across daily, weekly, and monthly frequencies. Table I summarizes the results. Among traditional models, Logistic Regression (LR) achieved the most consistent performance across all frequencies, followed closely by the Easy Ensemble Classifier (EEC). Several LSTM variants performed competitively, with `LSTM_8` attaining the highest individual score (AUC-ROC$_{test}$ = 0.7604) on the weekly dataset. Models such as XGBoost, Balanced Random Forest, and traditional Random Forest were excluded from further analysis due to AUC-ROC$_{test}$ scores consistently below 0.5.

Table I: AUC-ROC$_{test}$ and AUC-ROC$_{train}$ Scores by Model and Time Frequency.

| Time Frequency | Model | AUC-ROC$_{test}$ | AUC-ROC$_{train}$ |
|---|---|---|---|
| Daily | LR | 0.7222 | 0.7166 |
| | EEC | 0.6918 | 0.7719 |
| | LSTM_4 | 0.7204 | 0.6915 |
| | LSTM_4_4 | 0.7329 | 0.7110 |
| | LSTM_8 | 0.7117 | 0.7232 |
| | LSTM_8_4 | 0.7286 | 0.7437 |
| | LSTM_8_8 | 0.7017 | 0.7712 |
| Weekly | LR | 0.7251 | 0.7158 |
| | EEC | 0.6853 | 0.7841 |
| | LSTM_4 | 0.7651 | 0.6837 |
| | LSTM_4_4 | 0.7357 | 0.7287 |
| | LSTM_8 | 0.6576 | 0.7682 |
| | LSTM_8_4 | 0.6912 | 0.7602 |
| | LSTM_8_8 | 0.7328 | 0.7931 |
| Monthly | LR | 0.7263 | 0.7134 |
| | EEC | 0.6675 | 0.8063 |
| | LSTM_4 | 0.5335 | 0.6461 |
| | LSTM_4_4 | 0.3880 | 0.2242 |
| | LSTM_8 | 0.3668 | 0.9691 |
| | LSTM_8_4 | 0.5935 | 0.2456 |
| | LSTM_8_8 | 0.5476 | 0.5509 |

### B. Performance of Classical Model

Figure 3 and Figure 4 presents the predicted recession probabilities from the Logistic Regression and Easy Ensemble models. Logistic Regression produced smoother, more gradual probability increases, typically peaking less than a year prior to recessions. Easy Ensemble, in contrast, yielded sharper spikes, approximately 52 weeks before recessions, indicating higher short-term sensitivity.

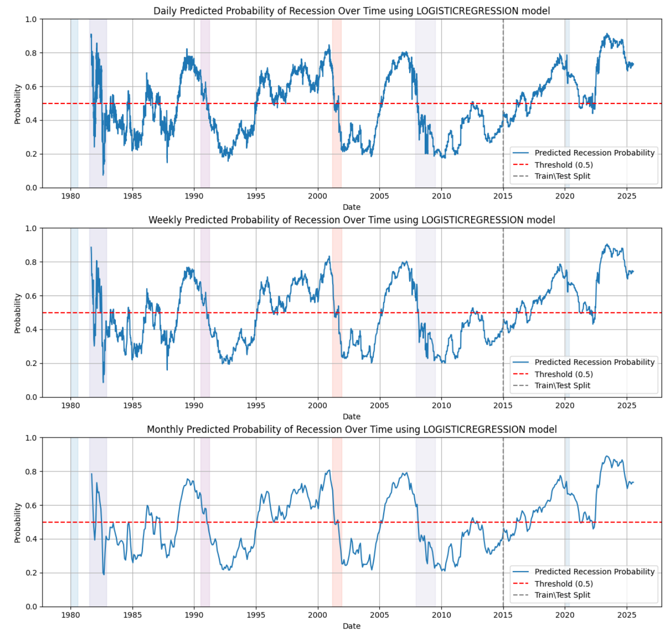


Figure 3: Predicted recession probabilities using Logistic Regression across time frequencies.

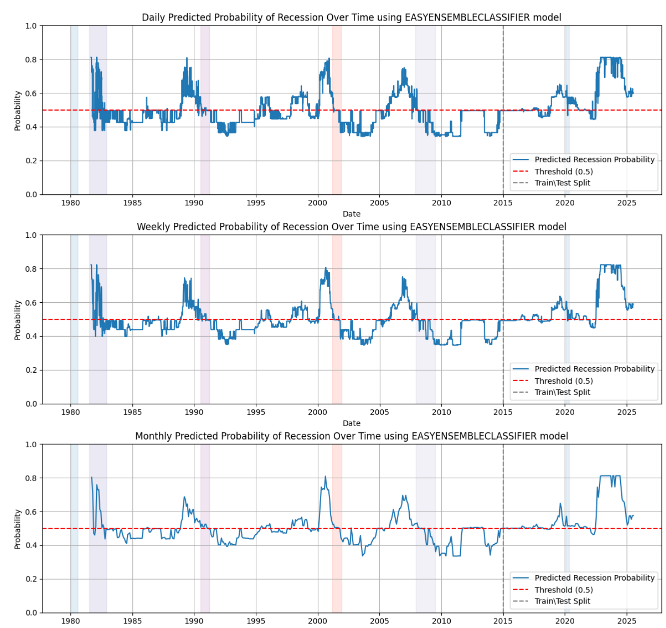


Figure 4: Predicted recession probabilities using Easy Ensemble across time frequencies.

### C. Performance of LSTM Variants

Figure 5 and Figure 6 shows the probability plots from the single-layer LSTM models. On daily data, `LSTM_4` exhibited moderate alignment with known recessions, with its probability crests appearing a few weeks prior to official start dates, while `LSTM_8` showed sharper peaks but it also over-predicted outside recession periods. On weekly dataset, `LSTM_4` achieved the highest AUC-ROC$_{test}$ score across all models (AUC-ROC$_{test}$ = 0.7651), but its probability plot did not show clear or timely signals around recession periods.

On monthly data, `LSTM_4` and `LSTM_8` models were less informative, with `LSTM_8` showing elevated probabilities after the official recession periods had ended.

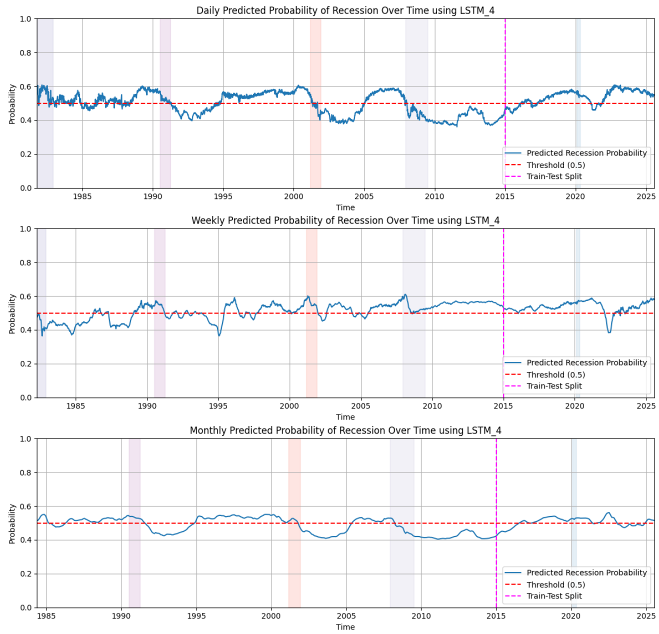


Figure 5: Recession probability forecasts using `LSTM_4` across time frequencies.

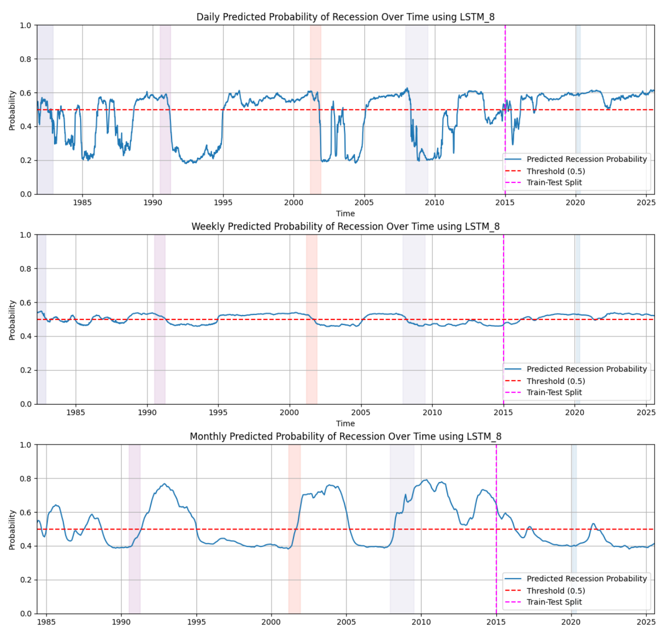


Figure 6: Recession probability forecasts using `LSTM_8` across time frequencies.

Figure 7, Figure 8 and Figure 9 shows the probability plot from the double-layer LSTM models with mixed results. For the daily datasets `LSTM_4_4` and `LSTM_8_4` 's crest showed modest alignment with known recessions (a few weeks before), while `LSTM_8_8` had more distinct crest a few weeks before recession periods. For the weekly datasets `LSTM_4_4` , `LSTM_8_4` and `LSTM_8_8` have distinct crest but tend to over-predict outside recession periods. For the monthly datasets, `LSTM_4_4` produced the most recession-aligned signals, despite having the lowest AUC-ROC$_{\text{test}}$ .

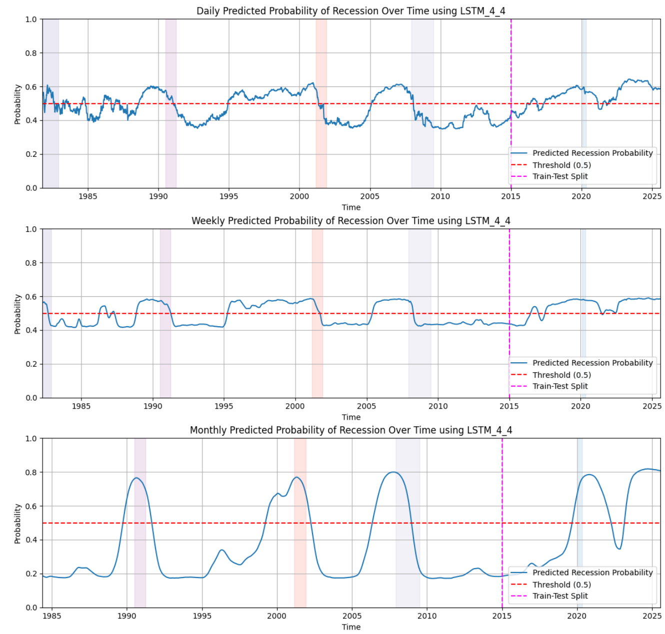


Figure 7: Predicted recession probabilities from `LSTM_4_4` across time frequencies.

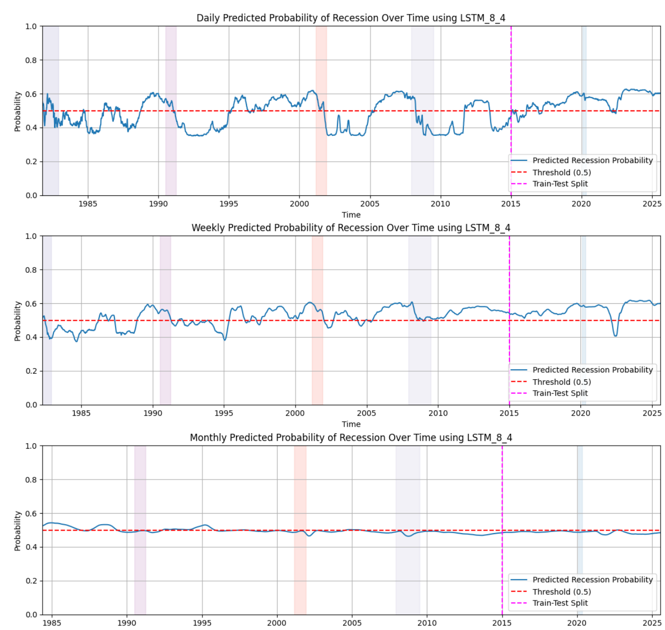


Figure 8: Predicted recession probabilities from `LSTM_8_4` across time frequencies.

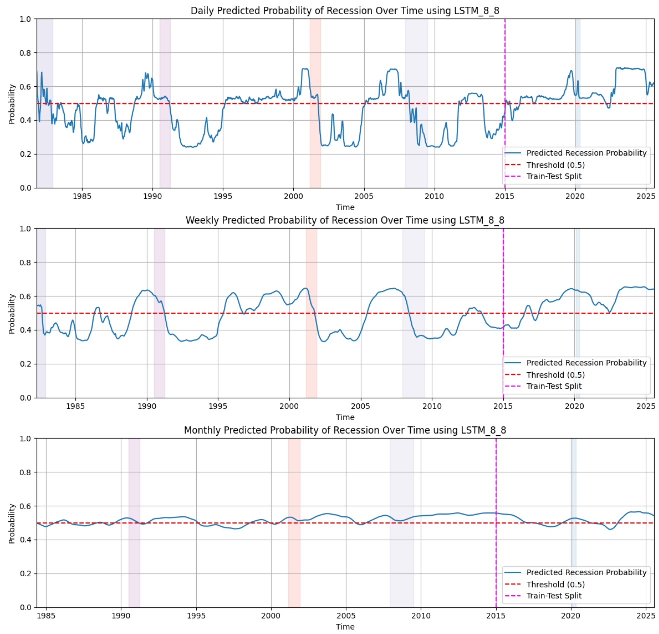


Figure 9: Predicted recession probabilities from `LSTM_8_8` across time frequencies.

### *D. Interpretation and Unexpected Outcomes*

Although LSTM models achieved AUC-ROC$_{\text{test}}$ scores comparable to those of Logistic Regression and Easy Ensemble on daily and weekly datasets, their probability plots often told a different story. In many cases, LSTM forecasts were either unstable or mistimed, failing to offer clear early signals before recessions. For instance, `LSTM_4` produced the highest AUC-ROC$_{\text{test}}$ of 0.7651, yet its weekly forecast lacked distinct crests near actual recession periods.

This underperformance, particularly given LSTM's theoretical advantages in capturing temporal dependencies, was unexpected. Several variants generated high false positives or delayed responses, reducing their practical utility as early warning systems.

Interestingly, while LSTM models sometimes produced smoother or more interpretable probability plots than their AUC-ROC$_{\text{test}}$ scores suggested, this highlights a key limitation of relying on AUC-ROC$_{\text{test}}$ as the sole evaluation metric in time-series settings. These findings challenge the assumption that increased model complexity directly translates to better real-world performance in macroeconomic prediction.

*1) Limitations and Future Work:* Several limitations should be acknowledged. First, the dataset used, while extensive in temporal coverage, it remains limited in dimensionality, relying primarily on two yield spread features. Finally, LSTM models were evaluated using a fixed architecture and limited hyperparameter tuning. It is possible that more refined configurations or alternative architectures (e.g., Bidirectional LSTMs or Transformers) may yield improved results.

## VI. Conclusion

This study evaluated the performance of traditional machine learning and deep learning models in forecasting U.S. recessions using Treasury yield spreads as predictive features. We compared Logistic Regression, Easy Ensemble, and multiple LSTM configurations across daily, weekly, and monthly datasets, using AUC-ROC$_{\text{test}}$ scores and probability plots to assess predictive effectiveness.

Our findings indicate that simpler models offered consistent and interpretable performance, often matching or exceeding the accuracy of more complex LSTM architectures. While some LSTM variants achieved high AUC-ROC$_{\text{test}}$ scores, their probability forecasts were often unstable, mistimed, or overly reactive, limiting their reliability for early warning applications. These results challenge the assumption that increased model complexity necessarily yields superior performance in macroeconomic time-series forecasting.

Future work may benefit from exploring alternative deep learning architectures, incorporating additional macroeconomic features, and adopting evaluation metrics better suited to the temporal nature of recession forecasting.

## Data Availability Statement

The data used in this study are publicly available from the Federal Reserve Bank of St. Louis FRED database. Yield spread series can be accessed via https://fred.stlouisfed.org/. All code and processed datasets used for model training and evaluation are available at: https://github.com/GordonOboh/RecessionPredictionML.

## Funding Statement

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

## AI Use Disclosure

Portions of this manuscript were prepared with the assistance of AI tools, including ChatGPT-4o (OpenAI) and Grammarly. These tools were used primarily for editing, formatting, language refinement, LaTeX debugging and basic Plagiarism check. All research design, data analysis, and interpretation of results were conducted and reviewed by the authors.

## Conflict of Interest Statement

The authors report no conflicts of interest.

## References

Aramonte, S., & Xia, D. (2019). Yield curve inversion and recession risk. *BIS Quarterly Review*. https://www.bis.org/publ/qtrpdf/r_qt1909.htm

Bauer, M. D., & Mertens, T. M. (2018). Economic forecasts with the yield curve. *FRBSF Economic Letter*, (2018-07). https://www.frbsf.org/economic-research/publications/economic-letter/2018/march/economic-forecasts-with-yield-curve/

Choi, J., Ge, D., Kang, K., & Sohn, S. (2023). Yield spread selection in predicting recession probabilities. *Journal of Forecasting*, *42*, 1772–1785. https://doi.org/10.1002/for.2980

Estrella, A., & Mishkin, F. S. (1998). Predicting us recessions: Financial variables as leading indicators. *Review of Economics and Statistics*, *80*(1), 45–61.

National Bureau of Economic Research. (n.d.). US Business Cycle Expansions and Contractions [Accessed: 2025-07-07]. https://www.nber.org/research/data/us-business-cycle-expansions-and-contractions