# Forecast Difficulty Detection via Latent Geometry in N-BEATS

Lauryane Santos Siqueira[1(✉)], Leonardo Nogueira Matos[1], Paulo Novais[2], and Pedro José Costa de Oliveira[2]

[1] Department of Computer Science, Graduate Program in Computer Science (PROCC), Federal University of Sergipe (UFS), Sergipe, Brazil
{lauryanesiqueira,lnmatos}@academico.ufs.br
[2] Department of Informatics, Algoritmi Research Centre, University of Minho, Braga, Portugal
pjon@di.uminho.pt, pedro.jose.oliveira@algoritmi.uminho.pt

**Abstract.** Identifying difficult forecasts and assessing model reliability remain important challenges in time series forecasting. This work investigates whether geometric properties of the latent space associated with the interpretable coefficients of the N-BEATS model can provide informative signals about predictive difficulty. For this purpose, a metric based on the aggregated magnitude of latent activations is proposed and compared with probabilistic density models (*Gaussian Mixture Models* and *Kernel Density Estimation*) and baselines computed directly from the input signal, such as amplitude and variance. Experiments conducted on the M4 dataset showed that input-space baselines achieved the highest global AUC and Spearman correlation values, indicating a strong relationship between simple statistical properties of the original signal and forecasting difficulty. However, among the latent-space approaches, the proposed geometric metric consistently outperformed the probabilistic models, reaching AUC values up to 0.765 and Spearman correlation close to 0.59 in weekly series. In addition, t-SNE analyses indicated partial organization of samples according to difficulty levels in the latent space. The results suggest that, although part of the discriminative signal is associated with the original temporal signal itself, latent representations learned by N-BEATS also capture relevant structural information for characterizing forecasting difficulty.

**Keywords:** Time Series Forecasting · Interpretable Neural Networks · Latent Space Analysis · Uncertainty Estimation · N-BEATS · Predictive Difficulty

## 1 Introduction

Forecasting is widely used to support decisions related to resource planning and organizational strategies across different time horizons. In this context, time

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2027
O. Gervasi et al. (Eds.): ICCSA 2026 Workshops, LNCS 16759, pp. 163–180, 2027.
https://doi.org/10.1007/978-3-032-30518-3_11

series forecasting plays a fundamental role in several applied domains, including economics, energy, healthcare, and urban planning [9].

In recent years, deep learning architectures such as Recurrent Neural Networks (RNNs), Long Short-Term Memory networks (LSTMs), and Gated Recurrent Units (GRUs) have gained significant popularity in time series forecasting across various applied domains, demonstrating strong performance, particularly in scenarios characterized by large volumes of data. However, variability in the performance of these architectures, associated with factors such as random weight initialization, raises concerns regarding their consistency and reliability [28].

Most of these architectures, however, operate as "black-box" models, in which the relationship between input variables and predictions remains opaque, requiring the use of post hoc interpretability methods, such as SHAP and LIME, to assist in understanding the model's decision-making process [6].

In contrast to this trend, the N-BEATS model (Neural Basis Expansion Analysis for Time Series Forecasting) proposes a deep architecture capable of producing interpretable forecasts by decomposing the predicted signal into basis components, enabling the extraction of explainable factors analogous to traditional decomposition approaches into trend and seasonality [15].

Despite the structural interpretability introduced by models such as N-BEATS, understanding the reliability of forecasts in deep architectures remains a relevant challenge, as explainability methods often face limitations related to the trade-off between transparency, performance, and computational cost [12].

Predictive uncertainty estimation has been extensively investigated in the literature, since point forecasts are often insufficient to represent the variability and randomness inherent in modeled systems. Probabilistic approaches aim to address this limitation by providing additional information that supports more robust analysis and decision-making in complex scenarios [11].

However, many uncertainty quantification approaches in deep neural networks require modifications to model structures or inference strategies based on multiple executions, such as Bayesian methods, ensembles, and test-time augmentation techniques. These characteristics may imply increased computational and operational complexity, which can hinder their adoption in certain applied contexts [7].

In this sense, investigating whether internal representations learned by the model provide signals about forecast reliability emerges as a promising direction. In N-BEATS, forecasts are constructed through basis function expansions parameterized by interpretable coefficients, which can be viewed as a latent representation of temporal behavior. Deep learning studies suggest that such representations integrate correlated information and offer compact system descriptions, supporting both performance and interpretability [15]. Furthermore, latent representation learning has been associated with improved predictive capability across different settings [19].

In this context, this work investigates whether the geometry of the coefficient space generated by the N-BEATS model can be explored as a proxy measure of

forecasting difficulty. In particular, we analyze whether the magnitude of latent activations is associated with prediction error and can be used to identify difficult forecasts. Additionally, we compare this geometric approach with probabilistic latent-space density estimation methods, including *Gaussian Mixture Models* and *Kernel Density Estimation*, as well as with baselines computed directly from the input signal, such as amplitude and variance.

The main contribution of this work is the investigation of whether forecasting difficulty signals emerge both from simple statistical properties of the temporal signal and from the geometry of latent representations learned by N-BEATS. To this end, we propose a geometric metric based on the aggregated magnitude of latent activations and compare it with latent-space probabilistic models and input-space baselines.

This paper is organized as follows. Section 2 presents related work on time series forecasting, interpretability, and latent-space uncertainty. Section 3 describes the proposed methodology, including the N-BEATS architecture, the construction of latent representations, and the strategies used to model forecasting difficulty. Section 4 presents the experimental setup, including dataset, evaluated models, and metrics. Section 5 presents and discusses the empirical results. Finally, Sect. 6 concludes the paper and outlines directions for future research.

## 2 Related Work

Recent advances in time series forecasting have been driven by deep learning models, as well as by the growing interest in interpretability and uncertainty quantification. These studies indicate that internal representations learned by neural networks influence both predictive performance and the understanding of model behavior, motivating the investigation of structural and probabilistic properties of latent spaces.

### 2.1 Deep Learning for Time Series Forecasting

Deep learning models have become effective approaches for time series forecasting due to their ability to capture nonlinear relationships and automatically learn complex representations, as discussed by [4]. Different architectures, such as recurrent, convolutional, and attention-based networks, allow the exploration of temporal dependencies across multiple forecasting horizons.

Recent studies, such as [23], investigate more sophisticated architectures, including decomposition-based and multi-scale structures, reinforcing the role of deep learning as a dominant forecasting paradigm. However, as highlighted by [10], the increasing complexity of these models also amplifies challenges related to interpretability and the analysis of internal representations. Surveys such as [14] point out that latent representation learning plays a central role in capturing relevant temporal variation factors, although questions regarding generalization and transparency remain open.

### 2.2 Interpretability in Time Series Models

Interpretability in neural forecasting models has received increasing attention due to the opaque nature of these architectures. In this context, [8] demonstrate that interpretable recurrent models can learn variable-specific representations and identify their temporal relevance. Hybrid approaches combining neural networks and statistical decomposition also favor the interpretation of latent patterns, as discussed by [13], while intrinsically interpretable architectures aim to reduce the "black-box" character by incorporating structural mechanisms directly into the model [27].

The advancement of Explainable Artificial Intelligence (XAI) has expanded the repertoire of techniques aimed at interpreting forecasts, including global and local explanations and feature attribution methods, as summarized by [12]. From a methodological perspective, *ante-hoc* and *post-hoc* approaches can be distinguished, which in time series forecasting include attention mechanisms, saliency maps, and representative-pattern-based methods, still subject to challenges related to stability and long-range dependency interpretation [18].

### 2.3 Uncertainty Estimation in Forecasting

Uncertainty quantification in neural networks has been investigated under different paradigms. Surveys such as [7] organize these approaches into Bayesian methods, ensembles, test-time augmentation strategies, and deterministic solutions that estimate confidence from a single network.

Concerns regarding reliability are not recent: early studies already pointed out limitations related to noise sensitivity and out-of-sample generalization in neural networks applied to time series forecasting [5]. More recently, it has been argued that explanations produced by complex models should also be accompanied by explicit uncertainty measures in order to improve the practical usefulness of machine-learning-based decisions [21].

### 2.4 Latent-Space Uncertainty

Recent research lines investigate uncertainty estimation directly in the space of learned representations. The DUQ method proposed by [24] uses latent proximity measures to identify unreliable predictions or out-of-distribution samples. Complementarily, [16] show that the geometric organization of these representations reflects data coverage properties.

In generative models, latent vectors tend to organize complex variation factors into semantically coherent directions, favoring interpolation and probabilistic modeling [3]. Furthermore, analyses based on differential geometry indicate that latent spaces may assume the form of nonlinear manifolds associated with different levels of uncertainty and representational distortion [2].

## 3 Methodology

This section presents the proposed methodology to investigate the relationship between latent representations of the N-BEATS model and forecasting difficulty. First, the forecasting problem formulation and the adopted architecture are described. Next, the construction of the latent representation, the proposed geometric metric, and the comparison models are presented. Finally, the forecasting difficulty criteria and the evaluation strategy are defined. :contentReference[oaicite:0]index=0

### 3.1 Forecasting Problem Formulation

Time series forecasting consists of estimating future values of a variable from historical observations sequentially organized over time. It is assumed that patterns present in the historical data contain relevant information for forecasting, although they are subject to random variability and unobserved factors [9].

Consider a univariate time series $y_t{}_{t=1}^{T}$. Given an input window of length $L$, the objective is to learn a forecasting function

$$f : \mathbb{R}^L \to \mathbb{R}^H, \tag{1}$$

where $H$ represents the forecasting horizon. Thus, the model produces point estimates $\hat{\mathbf{y}}_{t+1:t+H}$ from past observations $\mathbf{y}_{t-L+1:t}$.

### 3.2 N-BEATS Architecture

The N-BEATS (*Neural Basis Expansion Analysis for Time Series Forecasting*) model consists of a fully connected deep architecture designed for univariate time series forecasting tasks. The network is organized into stacks composed of multiple residual blocks, which operate sequentially over the input signal, allowing the progressive modeling of different temporal components [15].

Consider an input vector $x \in \mathbb{R}^L$ corresponding to the historical observation window used as model input, and a forecasting horizon $H$. Each block $b$ receives as input a residual vector $x_b \in \mathbb{R}^L$ and produces two sets of expansion coefficients: $\theta_b^b$, associated with the reconstruction component (*backcast*), and $\theta_b^f$, associated with the forecasting component (*forecast*). These coefficients are obtained through a stack of fully connected layers with nonlinear activations.

From these coefficients, the block generates its two outputs through specific basis functions, given by

$$\hat{x}_b = g_b^b(\theta_b^b), \qquad \hat{y}_b = g_b^f(\theta_b^f), \tag{2}$$

where $\hat{x}_b \in \mathbb{R}^L$ represents the block backcast and $\hat{y}_b \in \mathbb{R}^H$ represents its partial contribution to the future forecast.

The residual flow across blocks is iteratively updated as

$$x_{b+1} = x_b - \hat{x}_b, \tag{3}$$

so that each subsequent block operates on the portion of the historical signal not yet explained by previous blocks. In parallel, the forecast contributions are aggregated to compose the final model prediction:

$$\hat{y} = \sum_{b=1}^{B} \hat{y}_b, \tag{4}$$

where $B$ denotes the total number of blocks in the architecture.

Figure 1 illustrates this organization at different levels: the basic block, the residual structure within each stack, and the hierarchical aggregation of partial forecasts until the final model output is obtained.

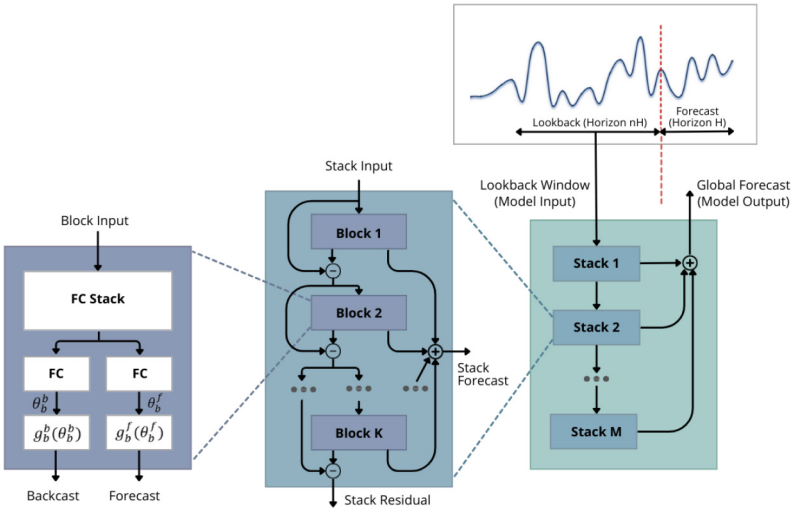


**Fig. 1.** General architecture of the N-BEATS model. Each block uses a stack of fully connected layers to produce expansion coefficients $\theta^b$ and $\theta^f$, which are subsequently transformed by the basis functions $g^b(\cdot)$ and $g^f(\cdot)$ into backcast and forecast components. The final prediction is obtained through the aggregation of contributions across residual stacks. Adapted from [15].

### 3.3 Latent Representation in the Coefficient Space

A central characteristic of N-BEATS is the construction of forecasts through an expansion in basis functions parameterized by coefficients learned by the network. In each block, the forecast and backcast outputs are obtained from linear projections of latent coefficient vectors.

Formally, let $\theta_b^{(f)} \in \mathbb{R}^{d_f}$ and $\theta_b^{(b)} \in \mathbb{R}^{d_b}$ be the coefficient vectors associated with the forecast and reconstruction components of block $b$, respectively. The block outputs can be expressed as

$$f_b = G_f \theta_b^{(f)}, \quad b_b = G_b \theta_b^{(b)}, \tag{5}$$

where $G_f \in \mathbb{R}^{H \times d_f}$ and $G_b \in \mathbb{R}^{L \times d_b}$ represent basis-function matrices.

The coefficient vectors produced by the blocks can be interpreted as a structured latent representation of the temporal behavior modeled by the network. Considering the forecast coefficients of all blocks, the aggregated representation is defined as

$$\Theta = \left[\theta_1^{(f)}, \theta_2^{(f)}, \ldots, \theta_B^{(f)}\right], \tag{6}$$

and its vectorized form as

$$z = \mathrm{vec}(\Theta), \tag{7}$$

where $z \in \mathbb{R}^d$ encodes the temporal patterns used to generate the forecast. This latent representation forms the basis for constructing the forecasting difficulty metrics proposed in this work.

### 3.4 Geometric Metric

After model training, the forecast coefficients of each interpretable block are extracted. To handle differences in vector length and the presence of padding, the vectors are truncated at the last valid coefficient, ensuring alignment with the evaluation variables.

For each block $b$, the Euclidean norm of the coefficients is computed as

$$s_b = \|\theta_b^f\|_2, \tag{8}$$

defining the latent activation vector

$$\mathbf{s} = [s_1, s_2, \ldots, s_B]. \tag{9}$$

From this vector, the global scalar metric is defined as

$$S = \|\mathbf{s}\|_2, \tag{10}$$

denoted as *radial_norm*, interpreted as an indicator of the internal activation intensity mobilized by the model to generate each forecast. This metric is evaluated as a proxy for forecasting difficulty under the hypothesis that greater latent activation intensity is associated with higher representational effort by the model.

Figure 2 presents the adopted methodological pipeline.

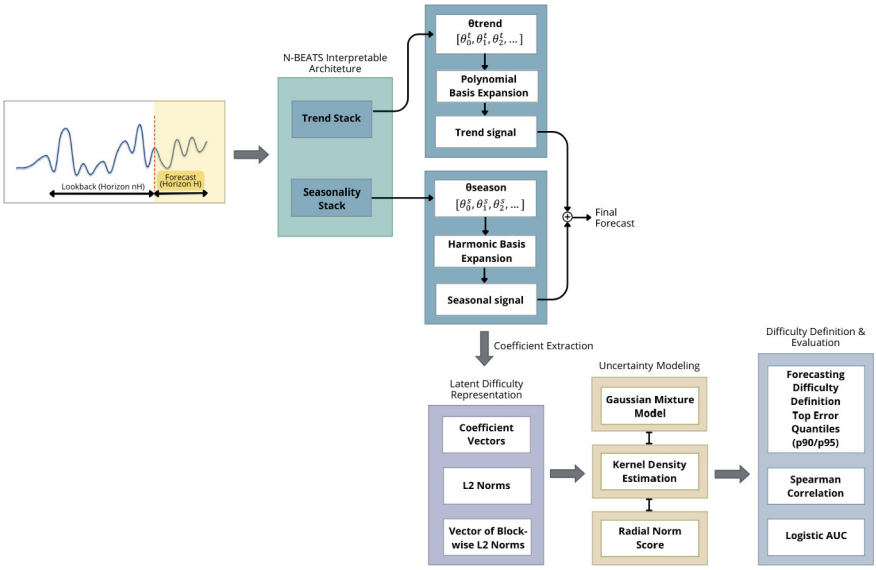


**Fig. 2.** Overview of the proposed methodological pipeline. The interpretable N-BEATS architecture generates latent coefficients associated with trend and seasonality blocks, which are transformed into geometric representations used for uncertainty modeling and forecasting difficulty definition.

### 3.5 Latent-Space Uncertainty Modeling

As a comparison to the proposed geometric metric, classical probabilistic models applied to the latent space were considered.

Gaussian mixture models are widely used in signal and time series modeling due to their ability to represent complex and multimodal distributions through combinations of simple components. Approaches using GMM to model trajectories in reconstructed phase spaces have demonstrated good performance in temporal classification [17]. Discriminative extensions based on large margins have also been proposed to improve class separability when modeling latent distributions [20]. In addition, applications in different domains indicate that GMM provides a robust strategy for representing structural variations and supporting likelihood-based decisions [22].

Nonparametric density estimation methods, such as *Kernel Density Estimation* (KDE), offer greater flexibility by allowing unknown distributions to be modeled without assuming specific functional forms. In time series analysis, recursive variants enable adaptive updating of estimates over time [1]. KDE has also been employed in tasks such as anomaly detection and learned representation modeling [29]. Its application in temporal classification problems reinforces the potential of this method for characterizing probabilistic structures in sequential data [25].

In this work, these approaches were used to model the distribution of latent representations extracted from the forecasting model. It was assumed that such a distribution could be approximated by a finite Gaussian mixture in the case of GMM, allowing the capture of possible multimodal structures. Alternatively, KDE was employed to estimate the local density of latent activations in a non-parametric manner, with the respective scores defined from the observations' negative log-likelihood (NLL).

In contrast to these probabilistic approaches, the proposed geometric metric is based exclusively on the magnitude of latent activations, enabling the investigation of whether the aggregated intensity of the internal representation contains informative signals regarding forecasting difficulty, independently of explicit assumptions about data distribution.

### 3.6 Input-Space Baselines

To verify whether the discriminative signal associated with forecasting difficulty originates from the latent representation or from properties of the original signal, baselines computed directly from the historical input window used by N-BEATS were added.

More specifically, two simple statistical descriptors of the input sequence $X$ were considered: signal variance and signal amplitude (difference between maximum and minimum values). These baselines allow the comparison between information derived from the original input space and that extracted from the latent representation learned by the model.

It is important to highlight that these metrics are computed from the same historical window responsible for generating the latent coefficients $\theta$, enabling a direct comparison between input-space and latent-space methods.

### 3.7 Forecasting Difficulty Definition

Forecasting difficulty was defined from the continuous error associated with each temporal instance. Binary scenarios based on percentile thresholds were also considered, particularly $P90$ and $P95$.

### 3.8 Evaluation Strategy

The relationship between the scores and forecasting error was analyzed using the Spearman correlation coefficient. Additionally, discriminative capability was evaluated using the area under the ROC curve (AUC), considering as positive class the observations belonging to the upper error quantiles.

## 4 Experimental Configuration

This section presents the experimental setup used to evaluate the relationship between latent representations of the N-BEATS model and forecasting difficulty

in time series. Different seasonal frequencies, model configurations, and latent-space density modeling strategies were considered. The evaluation was based on correlation and discriminative performance metrics, allowing the analysis of the potential of geometric signals derived from interpretable coefficients as indicators of predictive difficulty.

### 4.1 Dataset

The experiments were conducted using time series from the M4 dataset, widely adopted as a benchmark for forecasting tasks. Monthly, weekly, and yearly categories were considered, since they exhibit distinct seasonal patterns and forecasting horizons.

The interpretable version of N-BEATS (*NBEATS-I*) was adopted, investigating the impact of different configurations defined by the combinations *repeat*=0, *lookback* $\in \{2, 4, 6\}$, and *loss* $\in \{\text{MAPE}, \text{MASE}\}$.

### 4.2 Observation Extraction and Selection

During training, the interpretable coefficients produced by the blocks were stored for each pair (time series, forecasting instant). Initially, approximately 15.36 million observations were obtained for monthly series, 512 thousand for weekly series, and 15.36 million for yearly series.

To focus on representations learned after training stabilization, only the last 50 iterations were considered, resulting in approximately $51,200$ observations per category. Repeated pairs were then removed by keeping only the last occurrence, reducing each set to approximately $45,000$ unique observations.

### 4.3 Latent Representation

Each observation is associated with six interpretable blocks of the model, including three trend blocks and three seasonality blocks. For each block $b$, the Euclidean norm of the forecast coefficients was computed as

$$s_b = \|\theta_b^f\|_2, \tag{11}$$

defining the latent activation vector

$$\mathbf{s} = [s_1, s_2, \ldots, s_6], \tag{12}$$

which summarizes the intensity of activations produced by the architecture.

### 4.4 Preprocessing and Data Splitting

For each experiment and series type, the data were randomly divided into fitting (70%) and evaluation (30%) sets, without replacement and using a fixed random seed.

Before density modeling, the latent vectors were transformed using the *Yeo-Johnson* transformation, a generalization of the Box-Cox transformation that allows handling both positive and negative values and is commonly used for variance stabilization and approximation to normality [26], followed by standardization using parameters estimated only from the fitting set.

### 4.5 Density Modeling and Scores

Three types of latent-space scores were considered: a parametric Gaussian mixture model with diagonal covariance and automatic component selection via BIC (*gmm_bestK*); a nonparametric *Kernel Density Estimation* model with Gaussian kernel (*kde*); and a deterministic geometric metric based on latent activation magnitude (*radial_norm*).

For *gmm_bestK*, the number of components was selected within the range $k \in \{2, \ldots, 9\}$, and the score was defined as the observations' negative log-likelihood (NLL) on the evaluation set. For KDE, a fixed bandwidth equal to 1.0 was adopted, also using NLL as a surprise measure.

### 4.6 Difficulty Definition and Evaluation

Forecasting difficulty was defined from the final forecasting error. Both a continuous measure (*difficulty_continuous*) and a binary definition based on the 90th percentile of the error distribution (*difficulty_high_p90*) were considered.

The relationship between latent scores and forecasting error was evaluated using Spearman correlation, while discriminative capability in identifying difficult forecasts was measured through the area under the ROC curve (AUC).

## 5 Results and Discussion

This section presents and discusses the results obtained from the analysis of the relationship between N-BEATS latent representations and forecasting difficulty, considering geometric analyses, correlation with continuous forecasting error, and performance in difficult forecast detection tasks. :contentReference[oaicite:0]index=0

### 5.1 Latent Space Visualization

Figure 3 presents two-dimensional projections of the latent space obtained through t-SNE applied to the *block_l2_vector* vectors, constructed from the norms of the N-BEATS latent coefficients. Before applying t-SNE, a preliminary dimensionality reduction step using PCA was performed.

Each panel represents a specific experimental configuration and temporal frequency. Points were colored according to forecasting difficulty level, defined from the continuous forecasting error: *Low* ($\leq P50$), *Medium* ($P50$–$P90$), and

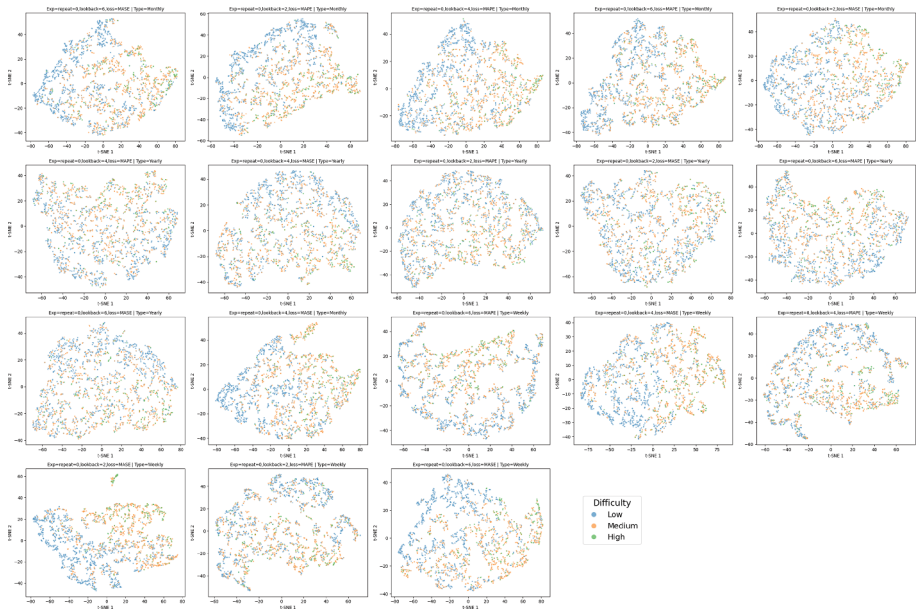


**Fig. 3.** t-SNE projections of the latent space for different experimental configurations and temporal frequencies. Points represent samples projected from the *block_l2_vector* vectors after preliminary dimensionality reduction using PCA. Colors indicate forecasting difficulty levels defined from the continuous forecasting error: *Low* ($\leq P50$), *Medium* ($P50$–$P90$), and *High* ($> P90$).

*High* ($> P90$). To facilitate visualization, stratified sampling with up to 2000 points per panel was used.

Overall, monthly and weekly series present more organized regions and greater separation between low- and high-difficulty samples. In contrast, yearly series exhibit greater overlap between groups, consistent with the lower AUC and correlation values previously observed.

Although t-SNE does not strictly preserve global distances, the obtained projections provide qualitative evidence that the latent representations learned by N-BEATS partially organize difficult and non-difficult samples into different latent-space regions.

### 5.2 Score Distribution by Difficulty Level

Figure 4 presents the distribution of the evaluated scores considering the binary formulation of forecasting difficulty, in which observations belonging to the upper error percentile are labeled as the positive class.

Overall, a consistent shift in the distributions associated with high-difficulty samples can be observed, reflected by increased median and dispersion of the scores in the positive class. This behavior occurs both in latent-space methods and in baselines computed directly from the input signal.

Among the latent methods, *radial_norm* presents clearer visual separation between difficulty groups when compared to the probabilistic *GMM* and *KDE* models, consistent with the quantitative results previously observed in terms of AUC and Spearman correlation.

Variance and amplitude baselines also exhibit strong separation between groups, especially in monthly series, indicating that simple statistical properties of the original signal are significantly related to forecasting difficulty.

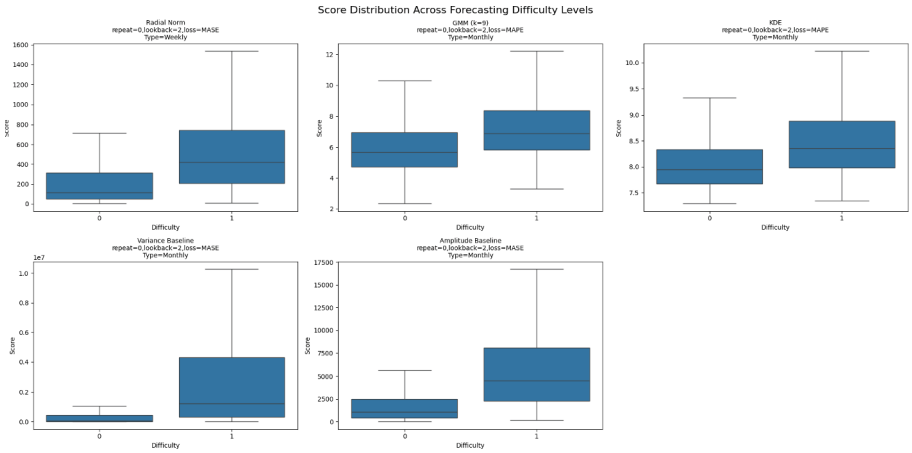


**Fig. 4.** Distribution of the evaluated scores for different forecasting difficulty levels, including latent-space methods (*radial_norm*, *GMM*, and *KDE*) and baselines computed directly from the input space (variance and amplitude).

### 5.3 Performance in Difficult Forecast Detection

The results show that baselines computed directly from the input space, especially amplitude and variance, achieved the highest global AUC and Spearman correlation values, mainly in monthly and weekly series. This behavior suggests that part of the discriminative signal associated with forecasting difficulty is already present in the variability and scale of the temporal signal itself.

Among the latent-space methods, the geometric *radial_norm* approach consistently outperformed the probabilistic *KDE* and *GMM* models while maintaining competitive performance across different experimental scenarios.

Table 1 presents a comparison between the best results obtained by the evaluated methods.

Overall, the results indicate that forecasting difficulty depends both on simple statistical properties of the original signal and on structural information organized in the latent space learned by N-BEATS.

**Table 1.** Comparison between the best results obtained by input-space baselines and latent-space methods.

| Space | Method | Series | Spearman | AUC |
|---|---|---|---|---|
| Input | amplitude | Monthly | **0.662** | **0.818** |
| Input | variance | Monthly | 0.649 | 0.809 |
| Input | amplitude | Weekly | 0.616 | 0.781 |
| Latent | radial_norm | Weekly | 0.588 | 0.765 |
| Latent | radial_norm | Monthly | 0.528 | 0.745 |
| Latent | radial_norm | Weekly | 0.528 | 0.732 |
| Latent | gmm_bestK | Monthly | 0.413 | 0.699 |
| Latent | kde | Monthly | 0.337 | 0.696 |
| Latent | gmm_bestK | Yearly | 0.251 | 0.686 |

### 5.4 ROC Curves of the Evaluated Models

Figure 5 presents the ROC curves of the best results obtained by the evaluated methods. Baselines based on amplitude and variance achieved the highest AUC values, indicating strong discriminative capability of the input signal itself. Among the latent-space methods, *radial_norm* consistently outperformed the probabilistic *GMM* and *KDE* models, suggesting that simple geometric measures derived from latent activations can capture relevant information associated with forecasting difficulty.

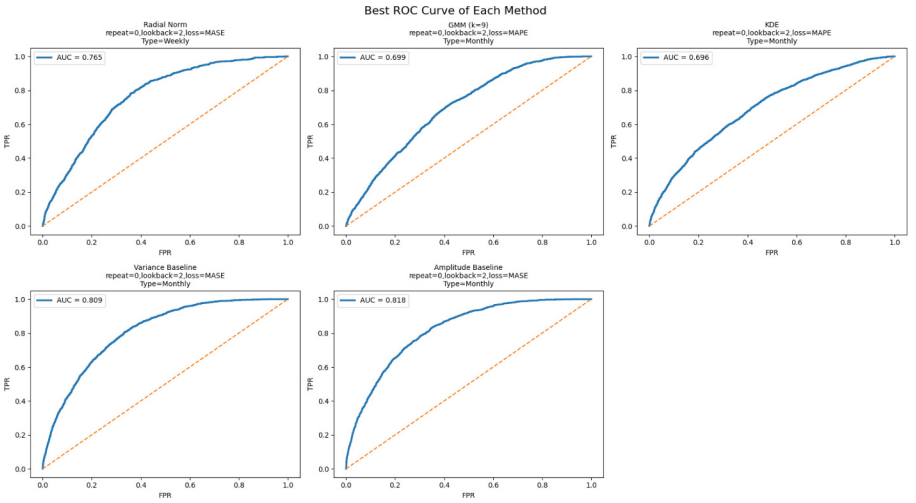


**Fig. 5.** ROC curves of the best results obtained by the evaluated methods in the task of detecting highly difficult forecasts.

### 5.5 Correlation Analysis

Figure 6 presents the Spearman correlation coefficients between the different scores and the continuous forecasting error. The highest coefficients were obtained by baselines computed in the input space, especially for monthly and weekly series. Among latent-space methods, *radial_norm* consistently outperformed *GMM* and *KDE*. In contrast, yearly series presented lower correlations in almost all experimental configurations.

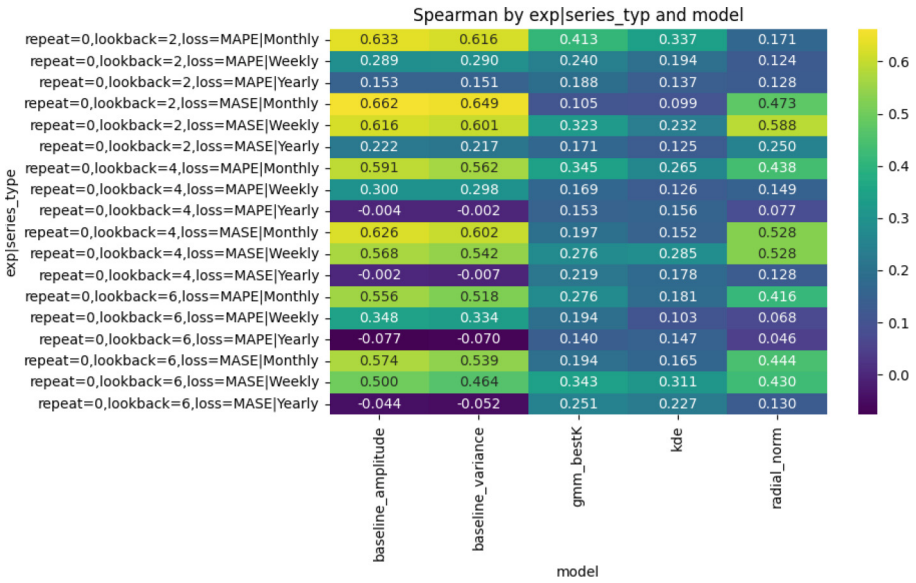


**Fig. 6.** Spearman correlation coefficients between the evaluated scores and the continuous forecasting error.

## 6 Conclusion

This work investigated whether latent representations derived from the interpretable coefficients of the N-BEATS model can provide informative signals about forecasting difficulty in time series. For this purpose, a geometric approach based on the magnitude of latent activations was proposed and compared with probabilistic density models and baselines computed directly from the input signal.

The results showed that baselines based on signal amplitude and variance achieved the highest global AUC and Spearman correlation values, indicating that a relevant part of the discriminative signal associated with forecasting difficulty is already present in the variability and scale of the temporal signal itself.

Despite this, latent-space methods remained competitive across different experimental scenarios. In particular, the *radial_norm* metric consistently outperformed the probabilistic *GMM* and *KDE* models, suggesting that simple geometric measures derived from latent activations can capture additional structural information associated with predictive difficulty.

The analyses based on t-SNE projections also indicated partial organization of samples according to difficulty levels in the latent space, especially for monthly and weekly series. Overall, the results suggest that forecasting difficulty depends both on statistical properties of the original signal and on the latent structure learned by the model.

The results also indicate that internal representations based on the interpretable coefficients of N-BEATS can provide relevant information about predictive difficulty, even without structural modifications to the architecture or the use of complex probabilistic strategies.

As future work, we highlight the investigation of the proposed approach in other forecasting architectures and the exploration of more robust geometric metrics for monitoring and calibration of forecasting models.

**Acknowledgments.** The work of Paulo Novais has been supported by FCT Fundação para a Ciência e Tecnologia within the R&D Unit Project Scope UID/00319/2025 - Centro ALGORITMI (ALGORITMI/UM) https://doi.org/10.54499/UID/00319/2025.

**Disclosure of Interests.** The authors have no competing interests to declare that are relevant to the content of this article.

## References

1. Aboubacar, A., El Machkouri, M.: Recursive kernel density estimation for time series. IEEE Trans. Inf. Theory **66**(10), 6378–6388 (2020)
2. Arvanitidis, G., Hansen, L.K., Hauberg, S.: Latent space oddity: on the curvature of deep generative models. arXiv preprint arXiv:1710.11379 (2017)
3. Bojanowski, P., Joulin, A., Lopez-Paz, D., Szlam, A.: Optimizing the latent space of generative networks. arXiv preprint arXiv:1707.05776 (2017)
4. Casolaro, A., Capone, V., Iannuzzo, G., Camastra, F.: Deep learning for time series forecasting: advances and open problems. Information **14**(11), 598 (2023)
5. Connor, J.T., Martin, R.D., Atlas, L.E.: Recurrent neural networks and robust time series prediction. IEEE Trans. Neural Netw. **5**(2), 240–254 (1994)
6. Freeborough, W., van Zyl, T.: Investigating explainability methods in recurrent neural network architectures for financial time series data. Appl. Sci. **12**(3), 1427 (2022)
7. Gawlikowski, J., et al.: A survey of uncertainty in deep neural networks. Artif. Intell. Rev. **56**(Suppl 1), 1513–1589 (2023)
8. Guo, T., Lin, T., Lu, Y.: An interpretable LSTM neural network for autoregressive exogenous model. arXiv preprint arXiv:1804.05251 (2018)
9. Hyndman, R.J., Athanasopoulos, G.: Forecasting: Principles and Practice. OTexts, 3 edn. (2021). https://otexts.com/fpp3/

10. Ismail Fawaz, H., Forestier, G., Weber, J., Idoumghar, L., Muller, P.-A.: Deep learning for time series classification: a review. Data Min. Knowl. Disc. **33**(4), 917–963 (2019). https://doi.org/10.1007/s10618-019-00619-1
11. Kabir, H.D., Khosravi, A., Hosen, M.A., Nahavandi, S.: Neural network-based uncertainty quantification: a survey of methodologies and applications. IEEE access **6**, 36218–36234 (2018)
12. Kalasampath, K., Spoorthi, K., Sajeev, S., Kuppa, S.S., Ajay, K., Maruthamuthu, A.: A literature review on applications of explainable artificial intelligence (XAI). IEEE Access **13**, 41111–41140 (2025)
13. Liu, S., Wu, K., Jiang, C., Huang, B., Ma, D.: Financial time-series forecasting: towards synergizing performance and interpretability within a hybrid machine learning approach. arXiv preprint arXiv:2401.00534 (2023)
14. Miller, J.A., et al.: A survey of deep learning and foundation models for time series forecasting. arXiv preprint arXiv:2401.13912 (2024)
15. Oreshkin, B.N., Carpov, D., Chapados, N., Bengio, Y.: N-beats: neural basis expansion analysis for interpretable time series forecasting (2020). https://arxiv.org/abs/1905.10437
16. Postels, J., Blum, H., Cadena, C., Siegwart, R., Van Gool, L., Tombari, F.: Quantifying aleatoric and epistemic uncertainty using density estimation in latent space. arXiv preprint arXiv:2012.03082 **1** (2020)
17. Povinelli, R.J., Johnson, M.T., Lindgren, A.C., Ye, J.: Time series classification using gaussian mixture models of reconstructed phase spaces. IEEE Trans. Knowl. Data Eng. **16**(6), 779–783 (2004)
18. Rojat, T., Puget, R., Filliat, D., Del Ser, J., Gelin, R., Díaz-Rodríguez, N.: Explainable artificial intelligence (xai) on timeseries data: a survey. arXiv preprint arXiv:2104.00950 (2021)
19. Rongali, S., et al.: Learning latent space representations to predict patient outcomes: model development and validation. J. Med. Internet Res. **22**(3), e16374 (2020)
20. Sha, F., Saul, L.K.: Large margin gaussian mixture modeling for phonetic classification and recognition. In: 2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings, vol. 1, pp. I–I. IEEE (2006)
21. Slack, D., Hilgard, A., Singh, S., Lakkaraju, H.: Reliable post hoc explanations: modeling uncertainty in explainability. Adv. Neural. Inf. Process. Syst. **34**, 9391–9404 (2021)
22. Słoński, M.: Gaussian mixture model for time series-based structural damage detection. Comput. Assist. Methods Eng. Sci. **19**(4), 331–338 (2012)
23. Song, X., Deng, L., Wang, H., Zhang, Y., He, Y., Cao, W.: Deep learning-based time series forecasting. Artif. Intell. Rev. **58**(1), 23 (2024)
24. Van Amersfoort, J., Smith, L., Teh, Y.W., Gal, Y.: Uncertainty estimation using a single deep deterministic neural network. In: International Conference on Machine Learning, pp. 9690–9700. PMLR (2020)
25. Wei, C., Peng, B., Li, C., Liu, Y., Ye, Z., Zuo, Z.: A two-stage optimized robust kernel density estimation for Bayesian classification with outliers. Int. J. Mach. Learn. Cybern. **16**(12), 9999–10023 (2025)
26. Weisberg, S.: Yeo-johnson power transformations. Department of Applied Statistics, University of Minnesota. Retrieved June 1 (2003), 8 (2001)
27. Yan, Y., Zhu, J., Duda, M., Solarz, E., Sripada, C., Koutra, D.: Groupinn: grouping-based interpretable neural network for classification of limited, noisy brain data. In: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pp. 772–782 (2019)

28. Yunita, A., et al.: Performance analysis of neural network architectures for time series forecasting: a comparative study of RNN, LSTM, GRU, and hybrid models. MethodsX **15**, 103462 (2025)
29. Zayara, A., Fandeyev, V.: Kernel probability density estimation in solving the problem of classification of the technical condition of complex systems. J. Commun. Technol. Electron. **70**(5), 226–231 (2025)