

# Shapelets-Enriched Selective Forecasting using Time Series Foundation Models

Shivani Tomar<sup>1</sup>, Seshu Tirupathi<sup>2</sup>, Elizabeth Daly<sup>2</sup>, Ivana Dusparic<sup>1</sup>

<sup>1</sup>Trinity College Dublin

<sup>2</sup>IBM Research, Dublin

tomars@tcd.ie, seshutir@ie.ibm.com, elizabeth.daly@ie.ibm.com, ivana.dusparic@tcd.ie

## Abstract

Time series foundation models have recently gained a lot of attention due to their ability to model complex time series data encompassing different domains including traffic, energy, and weather. Although they exhibit strong average zero-shot performance on forecasting tasks, their predictions on certain critical regions of the data are not always reliable, limiting their usability in real-world applications, especially when data exhibits unique trends. In this paper, we propose a selective forecasting framework to identify these critical segments of time series using shapelets. We learn shapelets using shift-invariant dictionary learning on the validation split of the target domain dataset. Utilizing distance-based similarity to these shapelets, we facilitate the user to selectively discard unreliable predictions and be informed of the model's realistic capabilities. Empirical results on diverse benchmark time series datasets demonstrate that our approach leveraging both zero-shot and full-shot fine-tuned models reduces the overall error by an average of 22.17% for zero-shot and 22.62% for full-shot fine-tuned model. Furthermore, our approach using zero-shot and full-shot fine-tuned models, also outperforms its random selection counterparts by up to 21.41% and 21.43% on one of the datasets.

## Introduction

Time series forecasting has seen tremendous advancements using general purpose pre-trained time series foundation models (TSFMs) such as Chronos (Ansari et al. 2024), Tiny Time Mixers (TTM) (Ekambaram et al. 2024), Moirai (Woo et al. 2024), TimesFM (Das et al. 2023), LLMTime (Gruver et al. 2023), and GPT4TS (Zhou et al. 2023). These models exhibit strong benchmark performances in zero-shot forecasting with further improvements gained with fine-tuning. However, these models report their performance using average error but often fail to provide reliable predictions in specific segments characterized by patterns of significant or abrupt changes (Qiu et al. 2024). This affects the reliability of such models in real-life applications. The technique of selective forecasting was introduced a long time ago (Chow 2009; El-Yaniv et al. 2010). It involves learning predictive models that are allowed to abstain from predicting if they are not confident in their predictions. There are

two ways of building selective predictors : First, add a selection mechanism to a trained prediction model, and second, jointly train the prediction model along with the selection function (Cortes, DeSalvo, and Mohri 2016b). In order to address the shortcoming associated with unreliable predictions, in this paper, we adopt the first method where we use shapelet-based distance similarity as the selection criteria for discarding model predictions. Shapelets were introduced by (Ye and Keogh 2009) as a time series primitive in the area of time series data mining particularly suited to the downstream task of time series classification, clustering and anomaly detection. Shapelets are time series subsequences that are maximally representative of a class. They are often used in time series classification problems due to their inherent interpretability. We use these shapelets to capture the local patterns in time series segments where the TSFM provides potentially unreliable predictions that are prone to high error. They serve as the basis of the sample selection mechanism used as part of our approach for selective forecasting. We employ a light-weight time series foundation model, Tiny Time Mixer (TTM) (Ekambaram et al. 2024) as the pre-trained prediction model leveraging its robust zero-shot forecasting performance with minimal computational overhead. Our approach allows the user to reduce the overall error by discarding unreliable predictions.

The main contributions of this paper are as follows:

- Proposes a selective forecasting framework, utilizing time series shapelets to identify the occurrence of specific patterns in data which lead to potentially high prediction errors on the target domain datasets.
- We introduce a sample selection module using distance-based similarity with learned shapelets to perform selective forecasting.
- Extensive experiments using TTM, a light-weight time series foundation model on benchmark datasets prove the efficacy of our framework to improve the overall forecasting performance on the target domain ensuring flexibility to discard unreliable predictions. Our approach leads to significant reduction in mean squared error (MSE), averaging 22.17% for zero-shot and 22.62% for full-shot fine-tuned models. It also outperforms random selection baseline using zero-shot and full-shot fine-tuned model by significant margins of 21.41% and

21.43% respectively on one of the datasets.

## Related Work

Recent efforts in the field of time series forecasting using deep learning (Wu et al. 2021; Zhou et al. 2022, 2021; Wu et al. 2022) have led to improved long-term forecasting accuracy with reduced latency. Furthermore, TSFMs have been gaining popularity for their extraordinary zero-shot performance on forecasting time series. There is ongoing research to further improve their domain-specific forecasting capabilities while still retaining the generalized representations learned during pre-training phase. That led to the adoption of parameter efficient fine-tuning (PEFT) techniques (Balne et al. 2024) focused on updating only a fraction of weights to learn domain-specific features while keeping the remaining weights frozen. LoRA (Hu et al. 2021) has emerged as an effective PEFT technique resulting in superior fine-tuned performance especially in the case of scarce datasets. Another work (Qiao et al. 2025) aims to improve the naive fine-tuning process by analyzing it through the causal perspective. They explicitly model multiple scales encountered in time-series target domain task during fine-tuning to better capture the temporal patterns. Another interesting aspect captured by Zhao et al. in (Zhao et al. 2025) reveals the sparsity in hidden states and redundancy in parameters inherent in TSFMs. They use these substructures to prune the TSFM before fine-tuning leading to further performance improvements during inference time. Tomar et al. (Tomar et al. 2025) developed a novel framework named AT4TS, for automated fine-tuning of TSFMs combining parameter efficient fine-tuning methods with state-of-the-art hyperparameter optimization techniques leading to significant improvement for out-of-domain datasets. While improved results have been demonstrated by the above approaches, current methods do not explore the application of selective forecasting in relation to TSFMs.

In this work, we focus on these small percentage of time series segments to reason and improve the model’s overall performance. A lot of research has been conducted in the field of selective prediction also known as reject option techniques mainly using algorithms like SVMs, boosting and nearest neighbours (Wiener and El-Yaniv 2015; Cortes, De-Salvo, and Mohri 2016a). Recently, these approaches have been extended to deep learning algorithms (Geifman and El-Yaniv 2017). SelectiveNet (Geifman and El-Yaniv 2017) was the first work which trains to optimize for both the classification loss and the reject function during the training process. However, in case of TSFMs which are already pre-trained, simultaneous training of a selective reject function is not feasible. To this effect, we use a shapelet module to selectively reject predictions for samples during inference time which are similar to the shapelets learned on the validation split. In the past decade, Shapelets (Ye and Keogh 2009) have garnered increasing attention in the context of time series classification. Shapelets are discriminative subsequences of time series that maximally distinguish among different classes by discovering local shape features inherent in each class. This makes them suited to downstream tasks of classification, clustering and anomaly detection. How-

ever, their use in the context of forecasting remains under-explored.

## Proposed Methodology

In this section, we describe the detailed methodology adopted in this paper for selective forecasting using TSFMs guided by shapelets. As shown in Figure 1, we split the target domain dataset into train, validation and test split and use a pre-trained TTM for both zero-shot predictions and fine-tuning. Algorithm 1 outlines the steps involved in our approach.

**Zero-shot performance on held-out Validation set :** We first use a pre-trained TSFM, in our case, TTM, to make zero-shot predictions on the validation split of the dataset,  $X_{val}$ . We formally define the task of time series forecasting below. Let  $X \in \mathbb{R}^{c \times sl}$  be a multivariate time series of length  $sl$  and number of channels  $c$ . The forecasting task involves predicting the future values  $Y \in \mathbb{R}^{c' \times fl}$  given the history/context  $X$ . Here,  $fl$  denotes the forecast length/horizon, and  $c'$  denotes number of forecast channels, where  $c' \leq c$ . In our case, we set the number of channels,  $c$  to 1 for univariate forecasting. The predictions from the model are denoted by  $\hat{Y} \in \mathbb{R}^{c' \times fl}$ .

We use a pre-trained time series foundation model, TTM, and make predictions  $\hat{Y}$  on the held-out validation split of the unseen target domain dataset. We use  $MSE$ , as the evaluation metric used by the original paper (Ekambaram et al. 2024). Based on the average MSE obtained on the entire  $X_{val}$ , we set the error threshold  $\tau$  to identify the segments of the time series where the zero-shot performance of the model is unreliable. We set  $\tau = (err + \delta * std)$  using  $\delta$  as the multiplier to vary the threshold as shown in Step 4 of Algorithm 1. This step involves analyzing the performance of the pre-trained TSFM on the unseen target domain downstream task without any kind of adaptive fine-tuning. Therefore, it enables us to establish the average baseline performance on the target domain dataset.

**Full-shot Fine-tuning TSFM on Train set :** As shown in step 1 of Algorithm 1, we use the train split of the target data to fine-tune the pre-trained TSFM. We later use this fine-tuned model ( $M^*$ ) as the baseline to compare against our approach.

**Shapelet module :** This module is central to our framework as it defines the criteria to filter the test samples for making selective predictions using zero-shot/full-shot fine-tuned TSFM. We utilize the shift invariant dictionary learning approach (Zheng, Yang, and Carbonell 2016) to learn a set of basis elements from the validation set. These basis elements serve as shapelets. We would like to clarify that in the time series literature, shapelets (Ye and Keogh 2009; Grabocka et al. 2014; Ye and Keogh 2011) are defined as time series subsequences in the training data that are useful in discriminating among classes. However, in this work, we do not use shapelets in the context of identification of time series classes but purely to uncover representative patterns in time series data. We first identify samples of time series sequences in the validation set that correspond to high zero-shot prediction error. The error threshold is de-

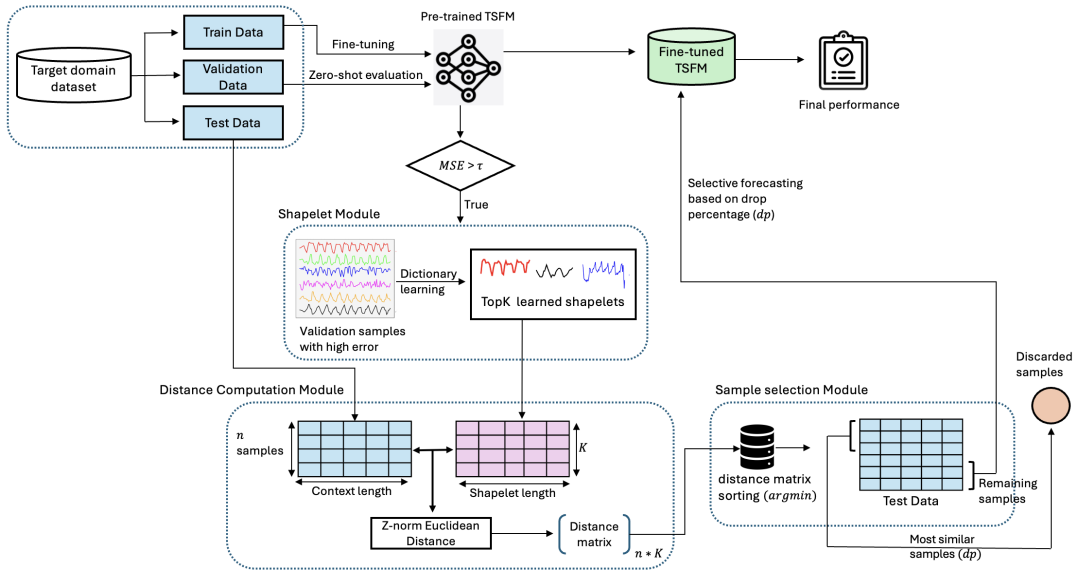


Figure 1: Workflow diagram showing the selective prediction framework guided by shapelets.

cided based on the average zero-shot error on the validation split of the target domain dataset. Shapelets are then learned from these curated high error-causing validation samples. The number ( $K$ ), length ( $q$ ) of shapelets and sparsity regularization parameter  $\lambda$  are selected using grid search over 3-fold cross-validation on the validation split. The property of shift invariance makes the application of dictionary learning useful for extracting representative patterns even in the absence of class labels. For a time series  $\{\mathbf{x}_i \in \mathbb{R}^p\}_{i=1}^n$  with  $n$  input data points and dimension  $p$ , authors in (Zheng, Yang, and Carbonell 2016), relaxed the length constraint of  $K$  basis vectors such that length of each basis or dictionary element  $q \leq p$ . Note that  $K$  basis atoms comprise the dictionary  $D = [d_1, \dots, d_K] \in \mathbb{R}^{q \times K}$  and for each  $x_i$ , the sparse coding  $\alpha_i \in \mathbb{R}^K$  is learned along with the dictionary. In order to achieve the shift invariance property, a new variable  $t_{ik}$  is introduced, which specifies the offset where  $d_k$  can be overlaid to match  $x_i$  such that  $t_{ik}$  can only take integer values in the range  $[0, p - q]$ .

Therefore, problem of learning both the sparse coding and dictionary elements or shapelets is given by

$$\arg \min_{\substack{D \in \mathbb{R}^{q \times K} \\ \alpha_i \in \mathbb{R}^K \\ t_{ik} \in [0, p-q]}} \frac{1}{2} \sum_{i=1}^n \left\| \mathbf{x}_i - \sum_{k=1}^K \alpha_{ik} T(\mathbf{d}_k, t_{ik}) \right\|_2^2 + \lambda \sum_{i=1}^n \|\alpha_i\|_1 \quad (1)$$

$$\text{s.t. } \|\mathbf{d}_k\|^2 \leq c, \quad \text{for } k = 1, \dots, K \quad (2)$$

Similar to the classical dictionary solving approach, the above problem is also solved using gradient descent approach.

Based on the number of shapelets ( $K$ ) learned, we further reduce the dictionary size to avoid redundant/self-similar

Table 1: Selected datasets

Dataset	Resolution	Length
ETTh1	1 hour	17420
ETTh2	1 hour	17420
ETTm1	15 min	69680
ETTm2	15 min	69680
Exchange Rate	1 day	7588
Traffic	1 hour	17544

shapelets by selecting *TopK* shapelets by using the ranking obtained from sparse coding  $\alpha$  matrix.

**Distance Computation Module:** Post shapelet learning module described above, we take the *TopK* shapelets which basically represent the patterns that coincide with higher than average prediction error on the validation split. We then use z-normalized euclidean distance-based similarity matching on the test data samples which is implemented as DISTANCE\_COMP procedure in Algorithm 1. This yields a distance matrix ( $D_{mat}$ ) as shown in Step 11.

**Sample Selection Module:** The distance matrix ( $D_{mat}$ ) computed in the above module is then sorted in ascending order to retrieve the indices of test samples that closely match the shapelets. The DISCARD procedure mentioned in Step 14 is used to drop the test samples and their corresponding prediction error is marked as 0. The drop percentage  $dp$  is a user-defined threshold to specify the ratio of predictions that are allowed to be discarded from the final set of predictions made by the TSFM on the test set to ensure adequate coverage.

**Selective forecasting of target domain:** Based on the user-defined drop percentage, we selectively filter the predictions of those test samples that show close proximity to the shapelets learned from the validation set.

---

Algorithm 1: Selective Forecasting using TSFM guided by Shapelets

---

**Input:** Pre-trained TSFM,  $TTM$ , target dataset  $X$  split into  $X_{\text{train}}$ ,  $X_{\text{val}}$ , and  $X_{\text{test}}$ , user-defined drop percentage  $dp$ , multiplier  $\delta$  for error threshold

**Output:** Model predictions based on  $dp$

---

```

1:  $M^* \leftarrow \text{FineTuneModel}(X_{\text{train}})$ 
2:  $y_{\text{preds}} \leftarrow \text{EvaluateModel}(M, X_{\text{val}})$ 
3:  $\text{err} \leftarrow \text{MSE}(y_{\text{preds}}, y_{\text{actual}})$ 
4:  $\tau \leftarrow (\text{err} + \delta * \text{std})$ 
5: for each value  $x \in X_{\text{val}}$  do
6:   if  $\text{err} > \tau$  then
7:      $X_{\text{val\_small}} \leftarrow x$ 
8:   end if
9: end for
10:  $S_k \leftarrow \text{LearnShapelets}(X_{\text{val\_small}})$ 
11:  $D_{\text{mat}} \leftarrow \text{DISTANCE\_COMP}(X_{\text{test}}, S_k)$ 
12:  $X_{\text{selected}} \leftarrow \text{DISCARD}(dp, X_{\text{test}}, D_{\text{mat}})$ 
13:  $y_{\text{selective}} \leftarrow \text{EvaluateModel}(M^*, X_{\text{selected}})$ 
14: procedure DISCARD( $dp, X_{\text{TEST}}, D_{\text{mat}}$ )
15:    $\text{Indices\_to\_drop} \leftarrow \text{argmin}(D_{\text{mat}})$ 
16:   return  $X_{\text{remaining\_test}}$ 
17: end procedure
18: procedure DISTANCE\_COMP( $X_{\text{TEST}}, S_k$ )
19:   for each value  $x \in X_{\text{test}}$  do
20:     for each value  $s \in S_k$  do
21:        $D_{\text{mat}}[x, s] \leftarrow \text{znorm\_ED}(x, s)$ 
22:     end for
23:   end for
24:   return  $D_{\text{mat}}$ 
25: end procedure

```

---

## Experimental Setup

This section describes the datasets used, the baselines adopted to compare against our framework as well as our choice of the model and its hyperparameters.

### Baseline models

In this paper, we use TTM as the time series foundation model owing to its lightweight architecture that makes it easier to fine-tune on resource-constrained environments. Furthermore, despite its compact model size, it has shown to outperform popular benchmarks in zero-shot forecasting (Ekambaram et al. 2024; Ye et al. 2024). For robust evaluation of our proposed approach, we compare against the following baselines:

- Zero-shot model (Drop percentage - 0%).
- Full-shot Fine-tuned model (Drop percentage - 0%).
- Random Selection using zero-shot model (Drop percentage - 20%).
- Random Selection using Full-shot Fine-tuned model (Drop percentage - 20%).

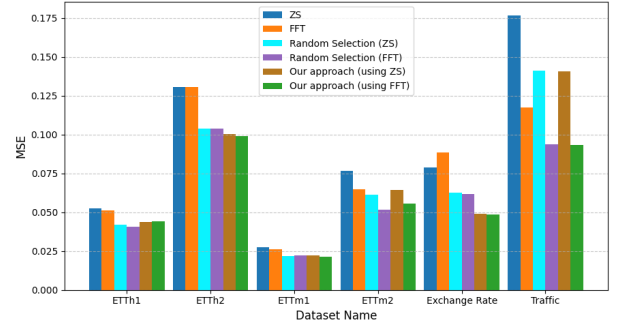


Figure 2: Performance comparison of Selective Forecasting using shapelets against baselines.

### Datasets

To evaluate the performance of our framework, we use six open benchmark time-series datasets. Table 1 enlists the datasets along with the resolution or sampling frequency and the length of the dataset. The Electricity Transformer Temperature datasets: ETTh1, ETTh2, ETTm1, ETTm2 (Zhou et al. 2021) contain 2 years of measurements from two electricity transformers in separate Chinese counties, each with 7 sensor features. ETTh1 and ETTh2 are collected at hourly interval whereas ETTm1 and ETTm2 are collected at every 15 minute interval. All four datasets have 7 channels. However, we only use one channel as part of our univariate forecasting setting. The Exchange Rate dataset contains daily exchange rates for 8 different currencies against USD from 1990 to 2016, with XRP/USD as the target variable for forecasting. Next, we also use Traffic dataset which records the hourly rates of road occupancy on the San Francisco Freeways using 862 sensors. We split the datasets into train, validation and test splits in the same ratio as described in the original TTM paper (Ekambaram et al. 2024).

### Experiment details

We use the TTM-Base model with the context length equal to 512 and forecast length set to 96. The full-shot fine-tuned version of the model involves fine-tuning the model head while keeping the backbone parameters frozen. For each dataset, we fine-tune the model on the train split of the data to get the full-shot fine-tuned version of the model. The hyperparameters for fine-tuning namely head dropout, batch-size and learning rate are based on the validation performance as reported in (Ekambaram et al. 2024). We run all the experiments for three random seeds and present the average results across runs. We set the drop percentage ( $dp$ ) to 20% and  $\delta$  to 2 to derive the error threshold for all our experiments.

## Results and Discussion

In this section, we present the results obtained on running the experiments using the models and datasets discussed in the previous section. We evaluate the effectiveness of our approach by calculating the mean squared error (MSE) on

Table 2: MSE obtained using Selective Forecasting guided by shapelets

Dataset	Zero-shot(ZS)	Full-shot Fine-Tuned (FFT)	Random Selection (ZS)	Random Selection (FFT)	Selective Forecasting with shapelets (using ZS)	Selective Forecasting with shapelets (using FFT)
ETTh1	0.0524	0.0512	0.0418	<b>0.0408</b>	0.0439	0.0443
ETTh2	0.1306	0.1304	0.1038	0.1039	0.1002	<b>0.0992</b>
ETTm1	0.0275	0.0264	0.0220	0.0221	0.0221	<b>0.0215</b>
ETTm2	0.0765	0.0646	0.0615	<b>0.0518</b>	0.0644	0.0555
Exchange rate	0.0790	0.0776	0.0626	0.0616	0.0492	<b>0.0484</b>
Traffic	0.1766	0.1173	0.1412	0.0936	0.1407	<b>0.0933</b>

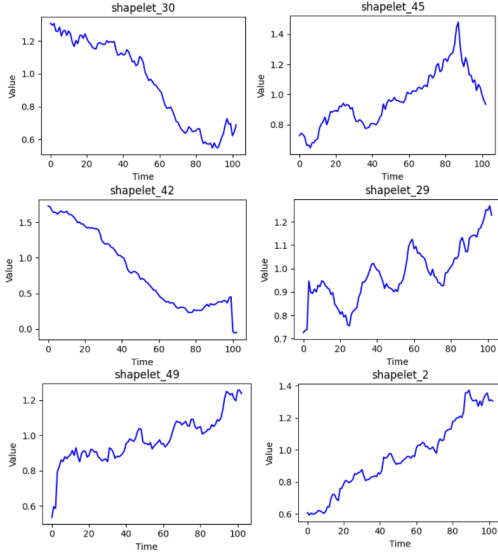


Figure 3: The plot showing a subset of shapelets learned from selected validation samples of Exchange Rate.

the test split of the data. Table 2 shows the average MSE obtained for each dataset. We use selective forecasting guided by shapelets to discard the test samples that are prone to high error. We use shapelets learned from the validation set to establish distance based similarity with the test samples using the distance metric: z-normalized euclidean distance. The closest matching test instances are selected to be discarded based on the user-defined drop percentage. From the MSE scores shown in Table 2, we observe that for four out of six datasets, our approach of selective forecasting yields lowered error improving the overall predictive performance while discarding unreliable predictions. Although, random selection using full-shot fine tuned model baseline performs slightly better for two datasets, it leads to significantly higher error for the remaining datasets. Our approach, on the other hand, provides a visual insight into the selection process and only closest matching test samples are chosen to be discarded at the inference time.

Figure 2 presents the results from Table 2 in a bar plot which clearly shows green and purple bars lead to the lowest errors across all datasets. These results exhibit significant improvement in predictive performance specifically for exchange rate and traffic datasets. This shows that informed se-

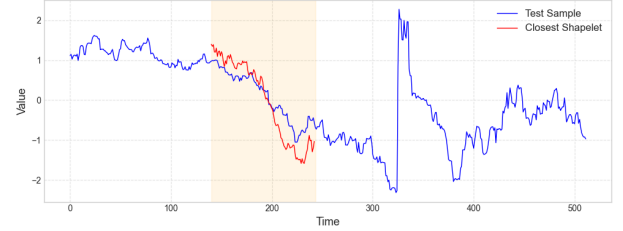


Figure 4: A test sample selected for discarding predictions with the best matching shapelet for Exchange Rate.

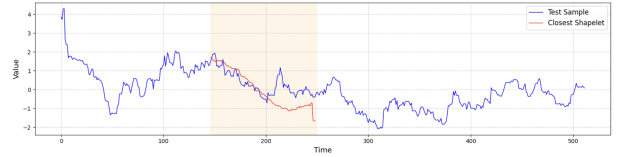


Figure 5: A test sample selected for making predictions with the best matching shapelet for Exchange Rate.

lective forecasting can significantly contribute to improved forecasting performance for TSFMs.

Figure 3 illustrates a subset of shapelets learned from the selected validation samples using Exchange Rate dataset. The mean error on validation set was 0.4506 and  $\delta$  was set to 2 which resulted in the error threshold = 1.6764. These samples correspond to high prediction errors highlighting distinct patterns where the TSFM is most likely to lead to unreliable predictions. In Figure 4, we show an example of a test sample from Exchange rate dataset that was discarded from the predictions due to a close match (z-norm euclidean distance = 4.46) with the shapelet highlighted in red. However, another test sample that was selected to make predictions is shown in Figure 5. In this case, the closest matching shapelet shows a high z-norm euclidean distance of 7.44. We also observe that this shapelet does not match the upward peak in the second half of the shaded region of the plot. This further validates our approach, which effectively utilizes shapelet-based similarity for selective forecasting.

Figure 6 illustrates the average error obtained on the test set using zero-shot model, full-shot fine-tuned and our approach using both zero-shot and full-shot fine-tuned model on the Exchange rate dataset. We observe that using our approach guided by shapelets, we are able to identify test samples which are similar to the patterns captured by shapelets.

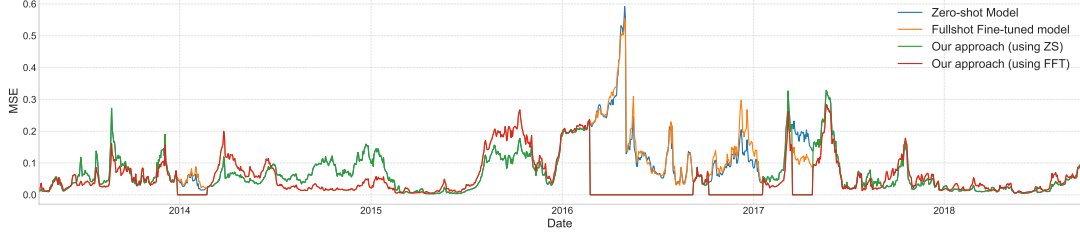


Figure 6: Error comparison using our approach with the baselines for Exchange Rate.

Table 3: Ablation studies by varying the error threshold for filtering validation samples.

Error Threshold (mean error + $\delta$ * std)	MSE using Selective Forecasting	
	Exchange Rate	ETTh2
$\delta = 1.0$	0.0526	0.1013
$\delta = 1.5$	0.0516	0.1040
$\delta = 2.0$	0.0484	0.0992
$\delta = 2.5$	0.0516	0.1028
$\delta = 3.0$	0.0507	0.1017

These test samples correspond to higher than average error as shown by peaks during the year 2016 in figure 6. We then selectively drop these samples at inference time such that the corresponding prediction error falls to zero as shown in the plot using the red line. Figure 4 depicts an example test sample showing a steep downward trend just before a spike which closely matches to the learned shapelet in red. Since this shapelet characterizes a pattern which is correlated to high error in the validation set, we selectively drop the predictions for this sample at test time.

### Ablation Study

We also performed an ablation study by varying the error threshold for filtering the validation samples used for shapelet learning for two datasets, ETTh2 and Exchange Rate. We add multiples of standard deviation to the mean error to increase the error threshold and observe the resulting shapelets learned from the validation data samples. This analysis shows that as the threshold increases, fewer validation samples are selected for shapelet learning. Figure 7 shows the unique set of shapelets and their corresponding closest matching test samples which are selected to be discarded while making predictions. We can clearly observe from the plot that, for error thresholds upto 1.5 times the standard deviation, more diverse shapelets are learned capturing both upward and downward trends. However, as the threshold increases, there are fewer validation set samples available to learn the shapelets which risks missing the patterns that might lead to high errors. This can be validated by the MSE reported in Table 3.

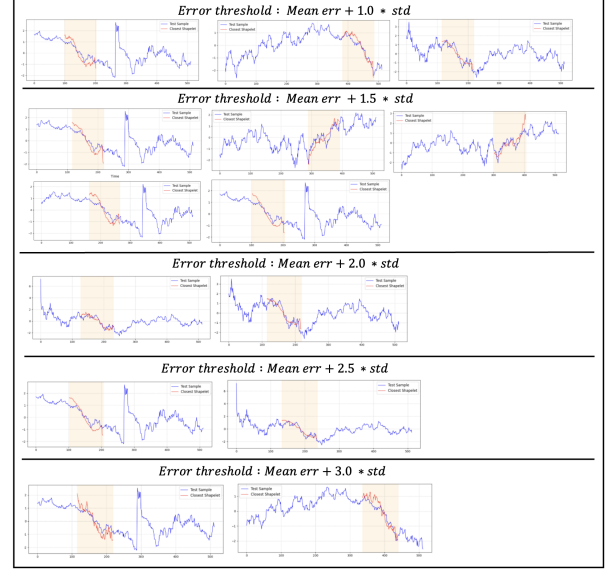


Figure 7: Matching shapelets to the selected test samples for dropping predictions for different error thresholds.

### Conclusions and Future Work

Pre-trained TSFMs have shown to provide robust zero-shot downstream performance for time series forecasting tasks. However, there is immense scope of improving the average zero-shot performance by specifically focusing at time series segments where the model predictions are highly unreliable. In this paper, we presented our approach to selective forecasting in the context of TSFMs guided by using shapelets. We introduced a shapelet-based sample selection module that helped in discarding high error-prone model predictions. In order to achieve this goal, we used z-normalized euclidean distance metric to establish distance-based similarity between the learned shapelets and test samples. We demonstrate that our approach reduces the overall error by an average of 22.17% for zero-shot and 22.62% for full-shot fine-tuned models. It also achieves notable gains, improving the performance up to 21.41% and 21.43% over random selection baselines for zero-shot and full-shot fine-tuned models respectively, on one of the datasets.

**Acknowledgments.** This work has been partially supported by the 6G-XCEL project (grant agreement



101139194), funded by the EU Horizon Europe program.

## References

- Ansari, A. F.; Stella, L.; Turkmen, C.; Zhang, X.; Mercado, P.; Shen, H.; Shchur, O.; Rangapuram, S. S.; Arango, S. P.; Kapoor, S.; et al. 2024. Chronos: Learning the language of time series. *arXiv preprint arXiv:2403.07815*.
- Balne, C. C. S.; Bhaduri, S.; Roy, T.; Jain, V.; and Chadha, A. 2024. Parameter Efficient Fine Tuning: A Comprehensive Analysis Across Applications. *arXiv preprint arXiv:2404.13506*.
- Chow, C.-K. 2009. An optimum character recognition system using decision functions. *IRE Transactions on Electronic Computers*, (4): 247–254.
- Cortes, C.; DeSalvo, G.; and Mohri, M. 2016a. Boosting with abstention. *Advances in neural information processing systems*, 29.
- Cortes, C.; DeSalvo, G.; and Mohri, M. 2016b. Learning with rejection. In *International conference on algorithmic learning theory*, 67–82. Springer.
- Das, A.; Kong, W.; Sen, R.; and Zhou, Y. 2023. A decoder-only foundation model for time-series forecasting. *International Conference on Machine Learning (ICML)*.
- Ekambaram, V.; Jati, A.; Dayama, P.; Mukherjee, S.; Nguyen, N.; Gifford, W. M.; Reddy, C.; and Kalagnanam, J. 2024. Tiny time mixers (ttms): Fast pre-trained models for enhanced zero/few-shot forecasting of multivariate time series. *Advances in Neural Information Processing Systems*, 37: 74147–74181.
- El-Yaniv, R.; et al. 2010. On the Foundations of Noise-free Selective Classification. *Journal of Machine Learning Research*, 11(5).
- Geifman, Y.; and El-Yaniv, R. 2017. Selective classification for deep neural networks. *Advances in neural information processing systems*, 30.
- Grabocka, J.; Schilling, N.; Wistuba, M.; and Schmidt-Thieme, L. 2014. Learning time-series shapelets. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, 392–401.
- Gruver, N.; Finzi, M. A.; Qiu, S.; and Wilson, A. G. 2023. Large Language Models Are Zero-Shot Time Series Forecasters. In *Thirty-seventh Conference on Neural Information Processing Systems*.
- Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; and Chen, W. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.
- Qiao, Z.; Liu, C.; Zhang, Y.; Jin, M.; Pham, Q.; Wen, Q.; Suganthan, P.; Jiang, X.; and Ramasamy, S. 2025. Multi-Scale Finetuning for Encoder-based Time Series Foundation Models. *arXiv preprint arXiv:2506.14087*.
- Qiu, X.; Chen, P.; Wang, Y.; Cheng, H.; Shu, Y.; Hu, J.; Guo, C.; Zhou, A.; Wen, Q.; Jensen, C. S.; et al. 2024. Foundts: Comprehensive and unified benchmarking of foundation models for time series forecasting. *arXiv preprint arXiv:2410.11802v1*.
- Tomar, S.; Tirupathi, S.; Marinescu, R.; Daly, E. M.; and Dusparic, I. 2025. AT4TS: Autotune for Time Series Foundation Models. *Transactions on Machine Learning Research*.
- Wiener, Y.; and El-Yaniv, R. 2015. Agnostic pointwise-competitive selective classification. *Journal of Artificial Intelligence Research*, 52: 171–201.
- Woo, G.; Liu, C.; Kumar, A.; Xiong, C.; Savarese, S.; and Sahoo, D. 2024. Unified training of universal time series forecasting transformers. *International Conference on Machine Learning (ICML)*.
- Wu, H.; Hu, T.; Liu, Y.; Zhou, H.; Wang, J.; and Long, M. 2022. Timesnet: Temporal 2d-variation modeling for general time series analysis. *arXiv preprint arXiv:2210.02186*.
- Wu, H.; Xu, J.; Wang, J.; and Long, M. 2021. Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting. *Advances in neural information processing systems*, 34: 22419–22430.
- Ye, J.; Zhang, W.; Yi, K.; Yu, Y.; Li, Z.; Li, J.; and Tsung, F. 2024. A survey of time series foundation models: Generalizing time series representation with large language model. *arXiv preprint arXiv:2405.02358*.
- Ye, L.; and Keogh, E. 2009. Time series shapelets: a new primitive for data mining. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, 947–956.
- Ye, L.; and Keogh, E. 2011. Time series shapelets: a novel technique that allows accurate, interpretable and fast classification. *Data mining and knowledge discovery*, 22(1): 149–182.
- Zhao, L.; Shen, Y.; Liu, Z.; Wang, X.; and Deng, J. 2025. Less is More: Unlocking Specialization of Time Series Foundation Models via Structured Pruning. *arXiv preprint arXiv:2505.23195*.
- Zheng, G.; Yang, Y.; and Carbonell, J. 2016. Efficient shift-invariant dictionary learning. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 2095–2104.
- Zhou, H.; Zhang, S.; Peng, J.; Zhang, S.; Li, J.; Xiong, H.; and Zhang, W. 2021. Informer: Beyond efficient transformer for long sequence time-series forecasting. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, 11106–11115.
- Zhou, T.; Ma, Z.; Wen, Q.; Wang, X.; Sun, L.; and Jin, R. 2022. Fedformer: Frequency enhanced decomposed transformer for long-term series forecasting. In *International conference on machine learning*, 27268–27286. PMLR.
- Zhou, T.; Niu, P.; Wang, X.; Sun, L.; and Jin, R. 2023. One Fits All: Power General Time Series Analysis by Pretrained LM. In *NeurIPS*.

## Appendix

We conducted additional experiments including ablation studies by varying the drop percentage (dp) parameter. The results are reported in Table 4. The error threshold  $\delta$  was set to 2 same as the main experiments. We can confirm that

as the drop percentage increases for each dataset, the MSE decreases monotonically.

Table 4: Ablation studies by varying the drop percentage (dp) in the range of  $\{10\%, 20\%, 30\%, 40\%, 50\%\}$  using our approach.

Dataset	Selective Forecasting with shapelets (using FFT)				
	10%	20%	30%	40%	50%
<b>ETTh1</b>	0.0482	0.0442	0.0400	0.0344	0.0312
<b>ETTh2</b>	0.1126	0.0992	0.0872	0.0758	0.0654
<b>ETTm1</b>	0.0253	0.0215	0.0195	0.0165	0.0133
<b>ETTm2</b>	0.0607	0.0555	0.0484	0.0413	0.0341
<b>Exchange rate</b>	0.0622	0.0484	0.0375	0.0293	0.0266
<b>Traffic</b>	0.1031	0.0933	0.0787	0.0669	0.0567

Figure 8 shows the error plot as we increase the drop percentage for each dataset. The reported results correspond to our approach, selective forecasting using full-shot fine-tuned model. We observe that the monotonically decreasing error trend is consistent across all datasets.

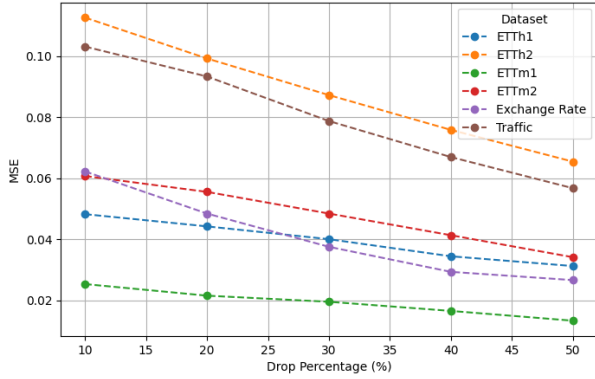


Figure 8: Plot showing the change in MSE as the drop percentage increases using Selective Forecasting (using FFT) for all datasets.

In the following sections, we also illustrate test samples for each dataset showing the best and the worst matching shapelets and the corresponding error curve after using selective forecasting. Figure 9 clearly depicts how shapelet based similarity assists in the selective forecasting process for ETTh1 dataset. The error plot for ETTh1 in figure 11 shows that the shapelet matching helps identify samples leading to high prediction errors for the first two peaks. However, we also observe that test samples leading to high error peaks at the start and end of January 2018 do not get discarded. This is highly likely due to distribution shift between the validation and test set. In such cases, shapelets learned on validation set are unable to represent the changes in the patterns/trends that may arise later. Future work will explore expanding the approach to encompass adaptive shapelet learning, further strengthening its applicability across scenarios.

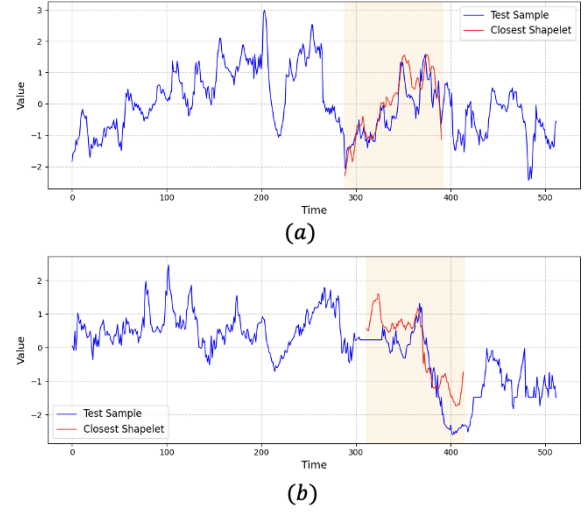


Figure 9: The figure (a) shows a discarded sample with the best closest match shapelet ( $d=5.47$ ) and (b) shows a sample selected for prediction with the worst closest match shapelet ( $d=7.86$ ) for ETTh1.

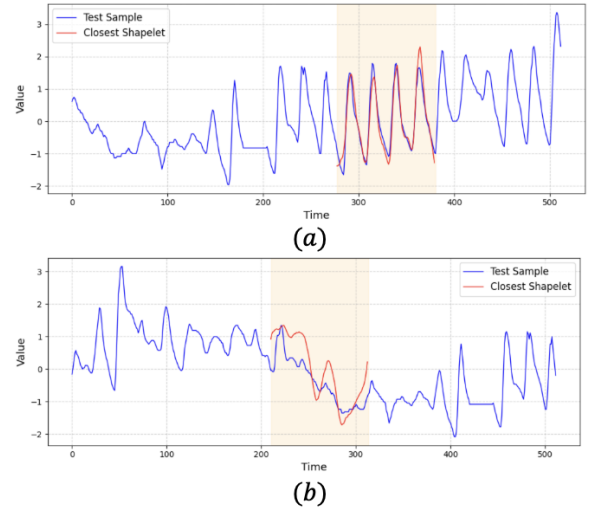


Figure 10: The figure (a) shows a discarded sample with the best closest match shapelet ( $d=2.78$ ) and (b) shows a sample selected for prediction with the worst closest match shapelet ( $d=6.16$ ) for ETTh2.



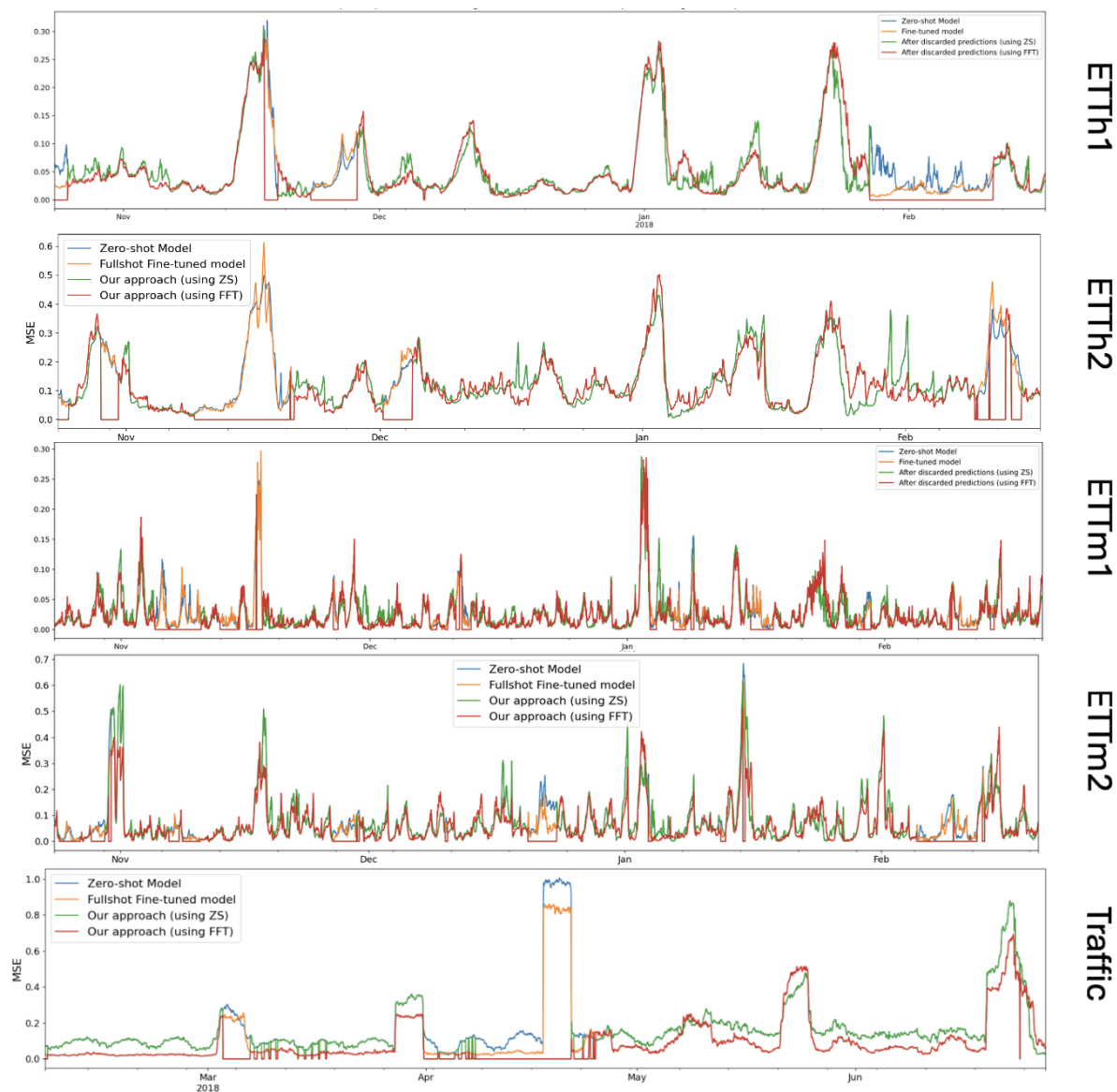


Figure 11: Error comparison using our approach with the baselines for the remaining datasets.

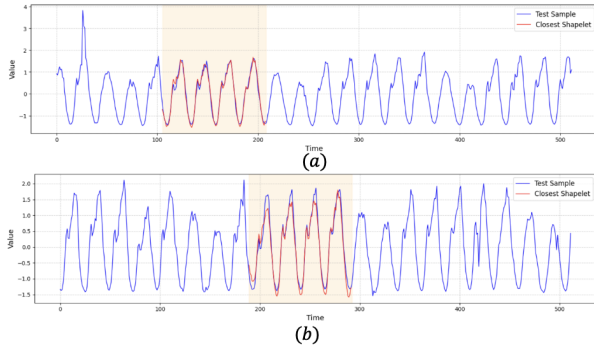


Figure 12: The figure (a) shows a discarded sample with the best closest match shapelet ( $d=1.43$ ) and (b) shows a sample selected for prediction with the worst closest match shapelet ( $d=2.18$ ) for Traffic.

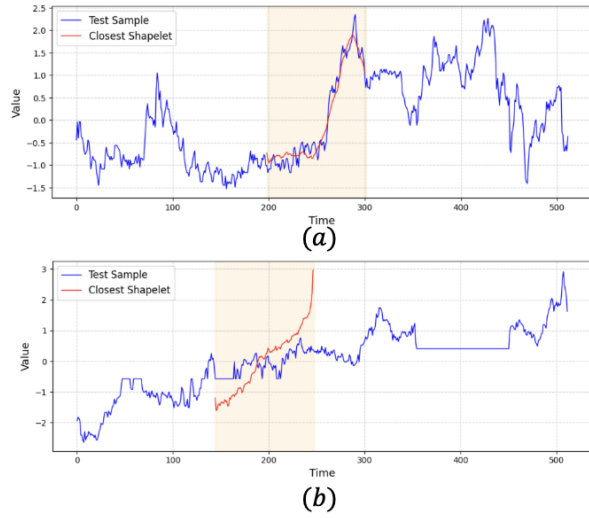


Figure 13: The figure (a) shows a discarded sample with the best closest match shapelet ( $d=2.23$ ) and (b) shows a sample selected for prediction with the worst closest match shapelet ( $d=7.67$ ) for ETTm1.

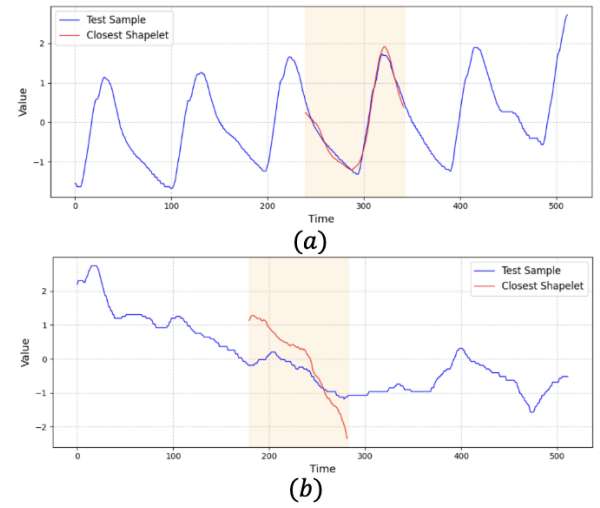


Figure 14: The figure (a) shows a discarded sample with the best closest match shapelet ( $d=1.27$ ) and (b) shows a sample selected for prediction with the worst closest match shapelet ( $d=7.38$ ) for ETTm2.