Transforming Time-Series Data for Improved LLM-based Forecasting through Adaptive Encoding

Vladimir Ceperic and Tomislav Markovic´

University of Zagreb Faculty of Electrical Engineering and Computing
Unska 3, Zagreb, Croatia.
e-mail: vladimir.ceperic@fer.hr

***Abstract* - The advent of Large Language Models (LLMs) has sparked significant interest in their application across various domains, including time-series forecasting. This paper introduces an encoding strategy designed to bridge the gap between the inherently quantitative nature of time-series data and the primarily textual processing capabilities of LLMs. By leveraging an innovative combination of adaptive segmentation and tokenization, inspired by the fast Brownian bridge-based aggregation (fABBA) algorithm, our method transforms time series data into a format conducive to LLM analysis. Through evaluation on diverse datasets (DARTS series), we demonstrate that our approach, on average, improves time-series forecasting accuracy.**

Keywords - LLM, time-series, encoding, fABBA.

I. INTRODUCTION

Large Language Models (LLMs) have become a cornerstone of modern natural language processing (NLP). These models, characterized by their expansive parameter counts, have been developed to capture complex patterns and understandings of human language [1]. The advent of models like GPT-3 [2], BERT [3], and their variants have not only enabled breakthroughs in tasks such as language translation, text summarization [4], and question answering [5], but have also extended their impact beyond conventional NLP tasks.

One of the critical characteristics of LLMs is their transfer learning capabilities, where a model pre-trained on a vast corpus of text can be fine-tuned for specific tasks with relatively small datasets [6]. This property has allowed for the democratization of state-of-the-art results, making them accessible to a broader range of researchers and practitioners.

In addition to NLP, LLMs have shown promise in diverse fields such as biology, where they help in understanding protein structures [1]. They have also been instrumental in generating creative content, ranging from poetry to code, thus blurring the lines between human and machine generated content.

Time-series data, characterized by its sequential nature, poses unique challenges for LLMs, primarily due to the need for specialized encoding techniques that can capture temporal dependencies and patterns. Progress in this area could uncover new insights and forecasting capabilities across a multitude of application domains, including finance, healthcare, and climatology.

The primary objectives of this paper are to introduce an encoding strategy that combines adaptive segmentation and tokenization for enhancing LLM-based time-series forecasting and to demonstrate its effectiveness through empirical evaluation on diverse datasets. By bridging the gap between the quantitative nature of time-series data and the textual processing capabilities of LLMs, proposed approach aims to unlock new forecasting capabilities and contribute to the advancement of this burgeoning field.

II. REVIEW OF RELATED WORK

A. Analysis of Models in Time-Series Forecasting Utilizing LLMs

Recent explorations into the application of Large Language Models (LLMs) for time-series forecasting have revealed considerable promise in harnessing their sophisticated pattern recognition capabilities for this domain. Among the notable developments, LLMTime [7] stands out for proving the viability of employing pre-trained LLMs, such as GPT-4 and LLaMA-2, in zero-shot forecasting scenarios. The methodology employed by LLMTime, which entails representing time-series as sequences of numerical digits and approaching forecasting as a task of predicting the next token, exemplifies the flexibility of LLMs to adapt to tasks extending beyond their original scope of natural language processing. The effectiveness of this technique in either matching or surpassing dedicated time-series models without requiring explicit fine-tuning highlights the potential role of LLMs within this sphere.

PromptCast [8] introduces an innovative prompt-based learning paradigm tailored for time-series forecasting, effectively altering the conventional approach towards interpreting and generating predictions from numerical sequences. By converting numerical inputs and outputs into text-based prompts, PromptCast capitalizes on the natural language generation prowess of LLMs for forecasting purposes. This method not only validates the adaptability of LLMs in processing numerical data but also pioneers new avenues for incorporating contextual and semantic insights into forecasting endeavors. Preliminary results from PromptCast indicate a promising research trajectory, underscoring the model's proficiency in generalizing effectively in zero-shot conditions, and hinting at its potential to significantly impact the domain of time-series analysis.

Despite the strides made by LLMTime and PromptCast towards integrating LLMs into the realm of time-series forecasting, several challenges persist. Both

methodologies rely on converting numerical values into textual or tokenized forms, a process which may not completely encapsulate the subtleties inherent in time-series data or perfectly align with the original training paradigm of LLMs. Moreover, the constrained context window available to contemporary LLMs presents difficulties in thoroughly processing lengthy sequences or analyzing multivariate time-series data.

Notwithstanding these obstacles, the advancements realized through LLMTime and PromptCast underscore the yet-to-be-fully-exploited potential of LLMs as a versatile instrument for the analysis of sequential data. These models facilitate a convergence between natural language comprehension and quantitative analysis, setting the stage for the development of more sophisticated and universally applicable forecasting methodologies. Future investigations could concentrate on refining the congruence between time series data and LLM architectures, broadening the context window, and incorporating external variables, all aimed at enhancing the precision and relevance of forecasting outputs.

B. Challenges in encoding time-series data for LLM processing

Encoding time-series data for processing by Large Language Models (LLMs) presents several challenges that stem from the intrinsic differences between sequential data and the natural language data LLMs are typically trained on. This subsection outlines the main hurdles encountered in adapting time-series data for LLM processing. First, time-series data is inherently quantitative and often lacks the semantic richness found in textual data. This discrepancy requires innovative encoding strategies that can imbue numerical data with contextual meaning understandable by LLMs, which are fundamentally designed to process and generate text [9].

Second, time-series data frequently exhibits complex temporal dependencies that traditional natural language processing techniques may not adequately capture. Effective encoding must, therefore, preserve these temporal relationships to enable LLMs to forecast future values or identify patterns within the series [10].

Another significant challenge is the high dimensionality and variability of time-series datasets. Different series may vary in length, scale, and frequency, necessitating flexible encoding mechanisms that can adapt to these variations without loss of information or predictive power [11]. Lastly, integrating external variables or covariates that influence the behavior of time-series (e.g., weather conditions for energy consumption data) into the encoding process adds another layer of complexity. These variables must be encoded in a way that LLMs can exploit to improve prediction accuracy, further complicating the challenge of encoding [12].

Addressing these challenges is critical for leveraging the full potential of LLMs in time-series analysis, pushing the boundaries of what can be achieved in fields ranging from finance to climate science. Encoding techniques that can bridge the gap between the quantitative nature of time series data and the linguistic structure of LLMs are essential to unlock the power of these models in sequential data analysis.

C. Survey of Time-Series Encoding Techniques

The task of encoding time-series data for assimilation by Large Language Models (LLMs) plays a pivotal role in influencing the overall efficacy of these models. Historically, a variety of time-series encoding methodologies have been developed (but not specifically for LLMs or applied to LLMs), each exhibiting unique advantages and certain limitations. This subsection presents some of the more distinguished techniques prevalent within this domain.

Symbolic Aggregate approXimation (SAX), heralded as one of the initial and most prevalently employed encoding techniques for time-series, transmutes time-series into sequences of symbols [13]. Despite its noted simplicity and operational efficiency, the SAX method encounters challenges in adequately capturing intricate temporal dynamics and in effectively managing noise.

Extended SAX (eSAX) advances upon the foundational SAX by incorporating a variety of transformations designed to more accurately reflect the dynamics inherent within time-series data. These modifications include adjustments aimed at accounting for trends and seasonal variations, thereby enhancing the representation of the data's fundamental patterns [14].

Adaptive Brownian Bridge-Based Aggregation (ABBA) introduces an alternative strategy by concentrating on the cumulative sum of deviations from a mean trajectory, thereby encapsulating the essence of time-series data through a piecewise linear representation [15]. ABBA has demonstrated its potential in diminishing the dimensionality of data whilst retaining its distinctive attributes.

Fast ABBA (fABBA) represents an evolution of the ABBA technique, with optimizations to the segmentation algorithm that significantly bolster computational efficiency without detracting from the quality of the data representation [16].

Notwithstanding the progress signified by these methodologies, each harbors its inherent drawbacks, especially when contemplating their integration with LLMs. The principal challenge lies not solely in the reduction of dimensionality and the encapsulation of critical patterns but also in encoding time-series data in a manner that resonates with the operational paradigms of LLMs, which are configured for processing textual content.

III. PROPOSED METHOD

Our approach leverages Large Language Models (LLMs) for time-series forecasting by innovatively incorporating adaptive segmentation through the fast Brownian Bridgebased Aggregation (fABBA) algorithm [16]. This method optimizes the conversion of time-series

data into a format amenable to LLM processing, thereby enhancing forecasting accuracy.

The core of our method is the adaptive segmentation of time-series into linear segments using fABBA, which selects segment boundaries to capture the inherent dynamics of the series efficiently. Following segmentation, we standardize each segment based on its statistical properties and discretize it into tokens. These tokens are then assembled into a sequence formatted for LLM input.

Algorithm 1 Time-Series Encoding Process

Input: Time series $X = \{x_1, x_2, \ldots, x_n\}$
Output: Tokenized representation $T = \{t_1, t_2, \ldots, t_m\}$
Use fABBA to partition X into segments $\{s_1, s_2, \ldots, s_k\}$
for each segment s_i **do**
    Standardize and discretize s_i into tokens $\{t_{i1}, t_{i2}, \ldots, t_{il}\}$
    Append tokens to T
end for
Format T for LLM input

This process ensures that the time-series data is appropriately scaled and tokenized, facilitating its analysis by LLMs. Our method thus bridges the gap between the quantitative nature of time-series data and the textual processing strengths of LLMs, paving the way for improved forecasting models.

A. Benefits of the Proposed Encoding Strategy

The encoding framework we propose manifests several distinct advantages over the conventional methodologies presently employed for time-series forecasting with Large Language Models (LLMs). The advantages are elaborated below:

• Adaptive segmentation: our method leverages the fABBA algorithm for adaptive segmentation, enhancing the ability to delineate the fundamental structure of the time-series data more adeptly than static-length segmentation techniques. This refinement facilitates a denser and more insightful representation, concentrating on pivotal patterns whilst mitigating irrelevant noise.

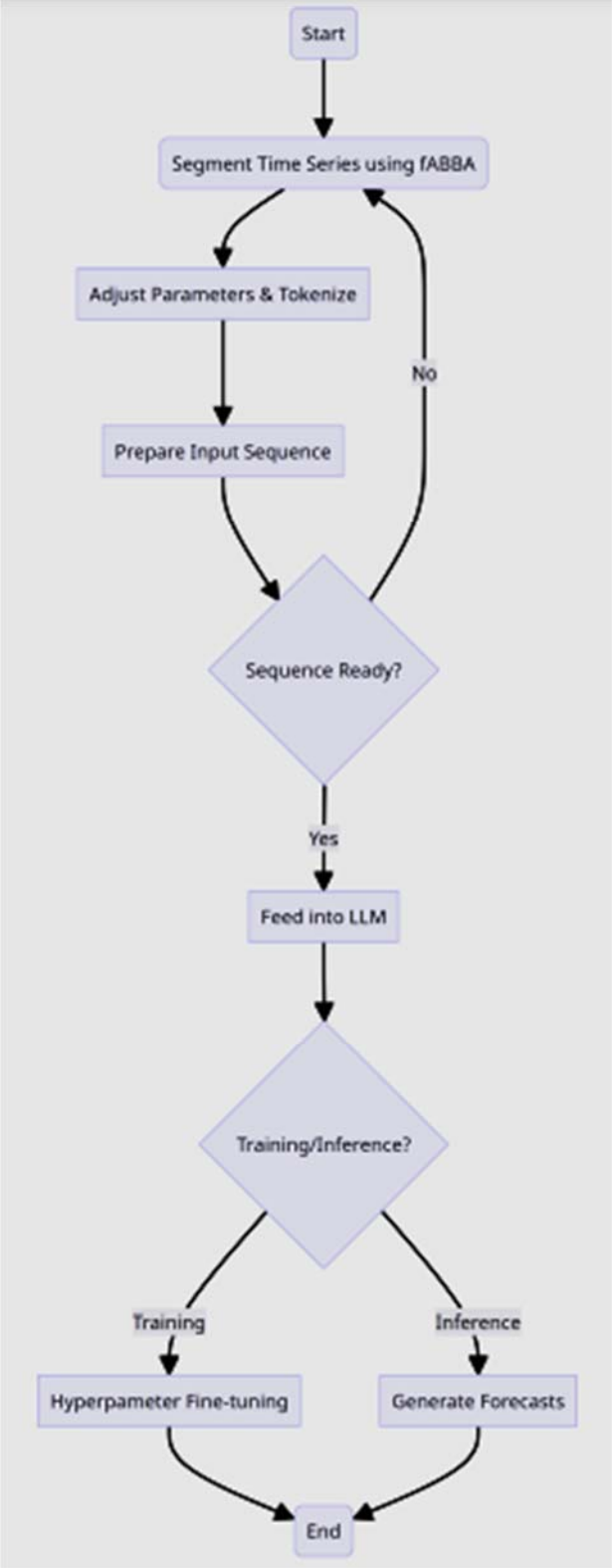

Figure 1. Overview of Adaptive Segmentation and Tokenization for LLM Processing

• Compatibility with LLMs: the tokenization phase transforms the segmented time-series into a lexicon of discrete tokens, aligning seamlessly with the input format predisposed for LLMs.

• Reduced data size: through adaptive segmentation, the method significantly condenses the size of the time-series. Such compression not only preserves but accentuates crucial information, thereby streamlining both training and inference phases, particularly for large time-series and large-scale applications.
• Enhanced forecasting accuracy: by capitalizing on the combined strengths of adaptive segmentation and targeted tokenization, we observed increase in forecasting accuracy, as shown in Section V.

IV. EXPERIMENTAL DESIGN

To assess the efficacy of our proposed encoding framework for time-series data within the context of Large Language Models (LLMs), we utilise DARTS collection [17] which offers a versatile suite for benchmarking models dedicated to time-series forecasting. We specifically used the following datasets from the DARTS collection:

Air Passengers Dataset, Aus Beer Dataset, Gas Rate CO2 Dataset, Monthly Milk Dataset, Sunspots Dataset, Wine Dataset, Wooly Dataset, and Heart Rate Dataset. Encompassing an array of domains such as finance, traffic, and energy demand, these datasets feature both univariate and multivariate time-series. The inherent diversity, marked by seasonal patterns and erratic trends, presents a challenge for forecasting.

In alignment with established experimental protocols, the datasets were segmented into training and testing subsets. Last 20% of the dataset is set for testing and has not been used in training or hyper-parameter optimisation. The effectiveness of our model is gauged using standard metrics:

Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and Mean Absolute Percentage Error (MAPE).

A. Comparative Models

We compared performance of our proposed method, against a spectrum of established time-series forecasting models. Specifically, TCN (Temporal Convolutional Networks) [18], N-BEATS, and N-HiTS [19] were implemented via the DARTS library [17]. For each comparative model (TCN, N-BEATS, N-HiTS), we performed a grid search over a range of hyper parameters on the training data set to find the best-performing configuration for each dataset.

Moreover, our comparison extends to innovative methodologies such as LLMTime and PromptCast, which exemplify the use of LLMs in the realm of time-series forecasting. This comparison underscores the novel contributions our encoding technique introduces, particularly in enhancing the application of LLM-based forecasting strategies.

V. EMPIRICAL EVALUATION

The empirical evaluation conducted as part of this research provides an analysis of the forecasting performance across various datasets, utilizing our proposed encoding strategy and comparing it against established models. The following summarizes our findings, presented in Table I and further visualized in Figure 2, which illustrates the normalized Mean Absolute Error (MAE) across different models and datasets.

A. Discussion

The data presented in Table I and Figure 2 demonstrate advancements facilitated by time-series encoding strategy in the domain of time-series forecasting with LLMs. Notably, across a spectrum of datasets, our method has demonstrated an ability to enhance predictive accuracy beyond that achieved by both conventional and LLM-based forecasting models.

Our method's superior performance on datasets such as AirPassengers and MonthlyMilk is attributed to its capacity for capturing and leveraging temporal patterns through adaptive segmentation and tokenization. Nevertheless, the relative underperformance on the Wooly Dataset, where it was outperformed by N-HiTS, indicates areas where our methodology may require further refinement to address specific series characteristics more effectively.

B. Directions for Future Research

Despite these encouraging outcomes, this study acknowledges certain limitations that future research must address. The current evaluation is based on a specific selection of datasets (DARTS), and further validation on a wider range of time-series data from various domains is necessary to establish the generalizability of the proposed approach. Moreover, the scalability of our method for handling largescale datasets and its efficacy in real-time forecasting scenarios warrants further examination. Notably, our encoding strategy's data compression capabilities underscore its potential suitability for large-scale applications.

Recognizing the preliminary nature of this research, we envisage several avenues for future exploration:

• Enhancement of encoding strategies: ongoing efforts will be directed towards the refinement and augmentation of our encoding methodology, aiming to accommodate a wider array of time-series data characteristics. This includes tackling datasets marked by high volatility and complex non-linear patterns, thereby broadening the applicability of our approach.

• Integration of multimodal data and insights: the fusion of multimodal data sources and the incorporation of domain-specific knowledge present exciting

• Advancement towards interpretability and adaptability: a critical objective moving forward is to enhance the interpretability and adaptability of LLMbased forecasting models.

• Handle extremely large-scale datasets: The encoding approach we took can also compress time-series data, making it particularly suitable for large-scale datasets.

TABLE I. FORECASTING PERFORMANCE ACROSS DATASETS

Dataset	Model	MAE	RMSE	MAPE
Air Passengers Dataset	TCN	54.960	65.535	11.667
	N-BEATS	97.886	118.072	20.756
	N-HiTS	59.160	79.032	12.546
	PromptCast	56.448	67.988	13.183
	LLMTime	34.373	41.210	7.738
	Proposed Method	**31.536**	**37.089**	**6.964**
Aus Beer Dataset	TCN	30.897	35.883	7.309
	N-BEATS	10.394	14.074	2.395
	N-HiTS	34.229	40.878	7.811
	PromptCast	62.221	75.611	14.602
	LLMTime	16.127	18.928	3.796
	Proposed Method	**14.515**	**17.035**	**3.416**
Gas Rate CO2 Dataset	TCN	2.641	2.981	4.837
	N-BEATS	2.628	3.085	4.806
	N-HiTS	3.854	4.503	7.227
	PromptCast	**2.093**	**2.416**	**3.842**
	LLMTime	3.496	4.214	6.235
	Proposed Method	2.198	2.495	3.900
Monthly Milk Dataset	TCN	70.859	88.781	7.893
	N-BEATS	33.641	40.258	4.064
	N-HiTS	32.726	39.349	3.849
	PromptCast	81.103	90.011	9.344
	LLMTime	9.677	11.938	1.112
	Proposed Method	**8.709**	**10.744**	**1.001**
Sunspots Dataset	TCN	51.816	70.214	264.954
	N-BEATS	73.151	91.815	96.035
	N-HiTS	49.933	68.827	196.231
	PromptCast	61.729	80.228	227.035
	LLMTime	47.339	66.677	136.676
	Proposed Method	**42.606**	**60.009**	**123.212**
Wine Dataset	TCN	3287.137	4559.931	13.980
	N-BEATS	4562.018	6059.011	16.366
	N-HiTS	3909.508	5498.956	14.924
	PromptCast	6789.694	7772.813	29.771
	LLMTime	1569.324	2055.041	6.558
	Proposed Method	**1413.392**	**1850.337**	**5.900**
Wooly Dataset	TCN	1158.795	1279.322	25.356
	N-BEATS	903.013	1034.849	19.535
	N-HiTS	**382.088**	**453.168**	**7.771**
	PromptCast	1949.667	2072.459	42.117
	LLMTime	808.731	877.411	17.322
	Proposed Method	728.258	789.270	15.600
Heart Rate Dataset	**TCN**	**5.493**	**6.599**	**5.953**
	N-BEATS	6.566	7.697	6.934
	N-HiTS	6.098	7.774	6.914
	PromptCast	5.511	7.166	6.143
	LLMTime	6.211	8.012	7.009
	Proposed Method	5.500	6.892	6.800

VI. CONCLUSION

This study introduces an encoding strategy that leverages the strengths of adaptive segmentation and precision tokenization for applying Large Language Models (LLMs) to time-series forecasting. The proposed approach demonstrates the benefits of custom preprocessing techniques in enhancing the predictive performance of LLMs, with notable improvements in accuracy across various time-series datasets.

This work not only promotes the integration of LLMs into time-series forecasting but also highlights the improvements encoding strategies can contribute to this field. The ongoing refinement and exploration of these methodologies are crucial for unlocking the full potential of LLMs in time-series analysis.

ACKNOWLEDGMENT

This research was carried out with FWO's funding for Project G088822N.

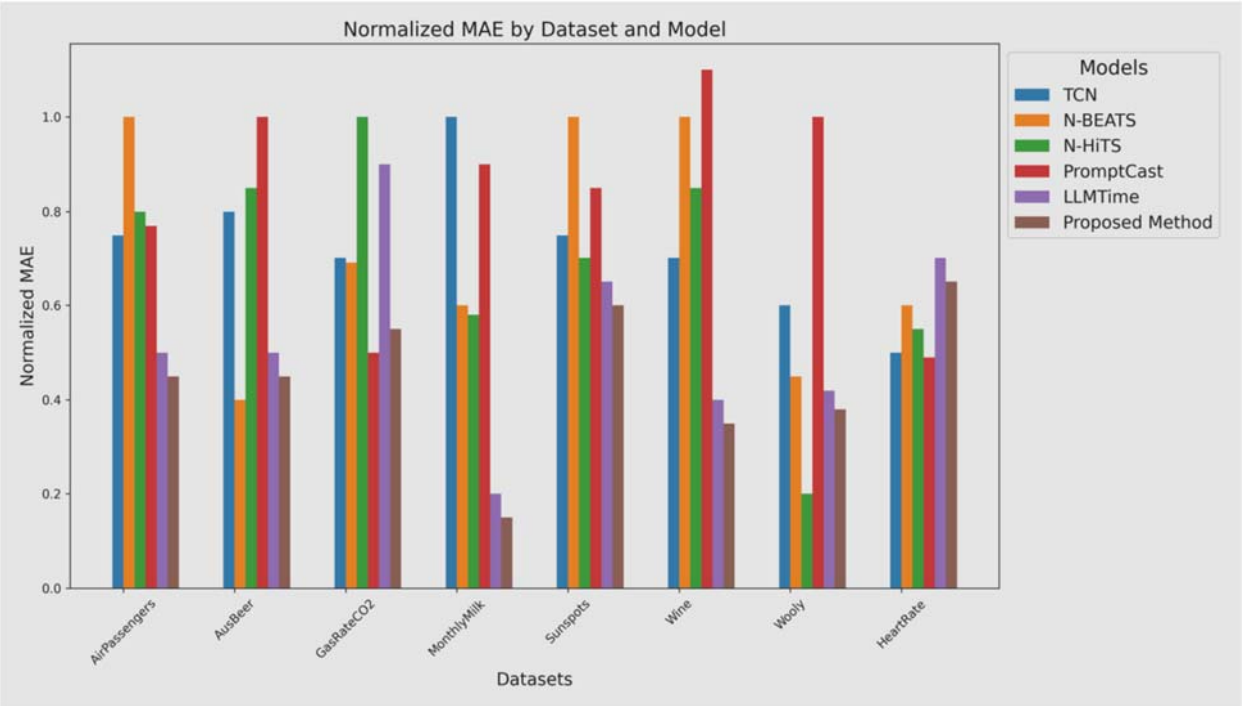

Figure 2. Normalized MAE Analysis of results

REFERENCES

[1] A. Senior et al., "Improved protein structure prediction using potentials from deep learning," Nature, vol. 577, no. 7792, pp. 706–710, 2020.

[2] T. Brown et al., "Language models are few-shot learners," in Advances in Neural Information Processing Systems, H. Larochelle et al., Eds., vol. 33. Curran Associates, Inc., 2020, pp. 1877–1901.

[3] J. Devlin et al., "Bert: Pre-training of deep bidirectional transformers for language understanding," 2018, cite arxiv:1810.04805Comment: 13 pages. [Online]. Available: http://arxiv.org/abs/1810.04805

[4] Z. Chen et al., "Logic2text: High-fidelity natural language generation from logical forms," ArXiv, vol. abs/2004.14579, 2020. [Online]. Available: https://api.semanticscholar.org/CorpusID:216914911

[5] K. Lo et al., "S2orc: The semantic scholar open research corpus," in Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020.

[6] Y. Tian et al., "Understanding user resistance strategies in persuasive conversations," in Findings of the Association for Computational Linguistics: EMNLP 2020, 2020.

[7] N. Gruver et al., "Large language models are zero-shot time series forecasters," in Thirty-seventh Conference on Neural Information Processing Systems, 2023. [Online]. Available: https://openreview.net/forum?id=md68e8iZK1

[8] H. Xue and F. D. Salim, "Promptcast: A new prompt-based learning paradigm for time series forecasting," 2023.

[9] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 35, no. 8, pp. 1798–1828, 2013.

[10] Z. Che et al., "Recurrent neural networks for multivariate time series with missing values," Scientific Reports, vol. 8, 04 2018.

[11] H. Ismail Fawaz et al., "Deep learning for time series classification: a review," Data Mining and Knowledge Discovery, vol. 33, no. 4, pp. 917–963, 2019.

[12] G. Lai et al., "Modeling long- and short-term temporal patterns with deep neural networks," 06 2018, pp. 95–104.

[13] J. Lin et al., "Experiencing sax: a novel symbolic representation of time series," Data Mining and Knowledge Discovery, vol. 15, no. 2, pp. 107–144, 2007.

[14] S. Malinowski et al., "1d-sax: A novel symbolic representation for time series," vol. 8207, 10 2013.

[15] S. Elsworth and S. G¨uttel, "Abba: adaptive brownian bridge-based symbolic aggregation of time series," Data Mining and Knowledge Discovery, vol. 34, pp. 1175 – 1200, 2020. [Online]. Available: https://api.semanticscholar.org/ CorpusID:199006026

[16] X. Chen and S. G¨uttel, "An efficient aggregation method for the symbolic representation of temporal data," 2022.

[17] J. Herzen et al., "Darts: User-friendly modern machine learning for time series," J. Mach. Learn. Res., vol. 23, pp. 124:1–124:6, 2021. [Online]. Available: https://api.semanticscholar.org/CorpusID:238419259

[18] C. Lea et al., "Temporal convolutional networks: A unified approach to action segmentation," in Computer Vision – ECCV 2016 Workshops, G. Hua and H. J´egou, Eds. Springer International Publishing, 2016, pp. 47–54.

[19] C. Challu et al., "Nhits: Neural hierarchical interpolation for time series forecasting," in Proceedings of the AAAI Conference on Artificial Intelligence, vol. 37, no. 6. AAAI, 2023, pp. 6989–6997.