A Multi-View Feature Construction and Multi-Encoder-Decoder Transformer Architecture for Time Series Classification

Zihan Li[1][0000−1111−2222−3333], Wei Ding[1,3][1111−2222−3333−4444], Inal Mashukov[1][2222−−3333−4444−5555], Scott Crouter[2][0000−0003−1297−9538], and Ping Chen[1][0000−0003−3789−7686]

[1] University of Massachusetts Boston, Boston MA 02125, USA
[2] University of Tennessee-Knoxville, Knoxville, TN 37996, USA
[3] Paul English Applied AI Institute, University of Massachusetts Boston, Boston MA 02125, USA

Abstract. Time series data plays a significant role in many research fields since it can record and disclose the dynamic trends of a phenomenon with a sequence of ordered data points. Time series data is dynamic, of variable length, and often contains complex patterns, which makes its analysis challenging especially when the amount of data is limited. In this paper, we propose a multi-view feature construction approach that can generate multiple feature sets of different resolutions from a single dataset and produce a fixed-length representation of variable-length time series data. Furthermore, we propose a multi-encoder-decoder Transformer (MEDT) architecture to effectively analyze these multi-view representations. Through extensive experiments using multiple benchmarks and a real-world dataset, our method shows significant improvement over the state-of-the-art methods.

Keywords: Multivariate Time Series Classification · Multi-view Learning, Multi-Encoder-Decoder Transformer.

1 Introduction

Time series data provides a clear and dynamic way to record the evolution of different variables of a phenomenon over time. In dynamic scenarios, time becomes a crucial dimension for complete recording, and time series data can represent an ordered sequence of features based on time, efficiently captures the dynamic relationship within evolving phenomena. Currently, time series data is widely used in various fields, such as Biology, Physics, Meteorology and Sport Medicine, there are still a few significant challenges in time series data analysis. Although many studies providing strong solutions to deal with time series data, such as [7, 2, 4, 9], there are still some limits and challenges in the field:

Time series analysis is very sensitive to the quality of data. Since the prediction is based on the dynamic trend of the relationship between features and time, it would have a significant impact to the model if erroneous data points

existed. To solve this, we assume that any time series events consist of a set of time units with similar characteristics, we define them as atomic units of the data (more details in Section 4.1). These units are always repeated among the whole sequence of time series data since they are basic and key building blocks constituting the data. With a representation using a set of basic atomic units, noise and outliers can be reduced significantly, and only key characteristics of data can be kept in the new representation. The proposed method can extract and construct atomic units from the original sequences of time series data[4], which can significantly reduce the negative impacts caused by the data quality and make the prediction processing more robust compared to the current methods.

Time series data is often recorded in variable length. Considering the characteristics of time series data, it is hard to begin and finish the recording of phenomena concurrently for different individuals. For example, when we monitor people's activities during the day, we don't know how long each activity will last, and it is impossible to perform all actions in the same time period. In such cases, it is incompatible for use with traditional machine learning models due to its variable-length features. To deal with this issue, there are two common techniques used frequently: cutting off or padding the period with the same length for all records [9, 10] or using fixed length sliding windows to represent the original data [23]. However, both of these methods will cause information loss due to the cut-off process. In contrast to these, the proposed method can keep maximum information by constructing a high-level summarization of the original variable-length time series data based on the extracted atomic units contained within the data. The atomic units serve as basic building blocks in data. The summarization of atomic units can provide a fix-length representation of the original variable-length time series data.

Time series data requires high computation cost. To make a time series dataset into a suitable input for a machine learning algorithm, the sliding windows technique is currently the most popular one, however, there is a significant limit of the sliding windows technique - the amount of data would increase rapidly depending on the width of a window and the length of a moving step, which means the cost of computation also would increase rapidly. Our proposed atomic based method can provide an adaptive choice of the size to fix the resource of computation. We are able to serve more flexibility by extracting fixed number of atomic units and building a new representation based on it, for example, we can reducing the number of features to a fixed number and decrease the complexity of computation significantly while producing competitive prediction results.

In summary, we propose an innovative feature construction approach that can generate multiple feature sets of different resolutions from a single dataset and produce a fixed-length representation of variable-length time series data. By applying the extracted atomic units from the original data, it allows the proposed model providing a competitive solution of noise, variable-length and computation cost. Our model, Multi-Encoder-Decoder Transformer (MEDT),

attempts to encapsulate and utilize all global information about the time series data. These multiple feature sets provide multiple views on the same data and are fed into a multi-encoder-decoder Transformer architecture, which is inspired by multilingual neural translation. Our main contributions are as follows:

- We propose a novel multi-view feature construction approach to deal with variable-length time series data and noise. In order to keep as much as global information from the original time series data, our method can construct multiple sets of fixed length features representation based on atomic units of data. These multi-view representations capture information of different granularity from original data and produce more robust results.

- We develop a multi-encoder-decoder Transformer model to effectively analyse these multiple feature sets for time series classification since these multiple views describe the same underlying phenomenon, inspired by multilingual neural translation (e.g., different languages encode the same semantics).

- We provide more flexibility in the number of features, which can help reduce the complexity of computation and provide a more efficient method for time series prediction. By constructing selected number of new features summarizing original data, computation could decreases with a simple fixed length data input, while prediction quality improves.

2 Related Works

Numerous time series data analysis methods have been proposed. Recent methods include HIVE-COTE (Lines et al., 2018) [1], ROCKET (Dempster et al.,2020) [2], and TS-CHIEF(Shifaz et al., 2020) [3], which are considered to be the state-of-the-art when tackling time series classification problems. Other popular methods include CNN-based deep learning models such as REsNet (Fawaz et al., 2019b) [11] and InceptionTime (Fawaz et al., 2019a) [8]. However, these methods are computationally costly and complex and often fail to produce good results for datasets containing numerous samples of lengthy time series data. Dempster et al. (2021) [28] introduce a fast MiniRocket method which is 75 times faster than original Rocket. Gao et al. (2022) [29] provides a reinforcement learning framework for multivariate time series classification which can identify interpretable patterns without using neural network. Although these methods show improvements either on speed or accuracy. However, it is hard to find a method which can provide consistent and competitive accuracy versus other SOTA methods.

Currently, there are a lot of transformer based publications of time series analysis. Transformer, although initially proposed for natural language translation and having demonstrated remarkable results in various NLP tasks, have found applications in time series tasks while also providing for efficient computation. In contrast to other popular sequential data classification methods, the classical Transformer model presented by Vaswani et al. (2017) [5] is based exclusively on the attention mechanism. The attention mechanism tends to global

dependencies between input and output, while the architecture of the Transformer model allows for greater parallelization, resulting in a significantly more efficient and accurate classification [5]. Li et al. (2019) [13] and Wu et al. (2020) [14] have employed full encoder-decoder transformer architectures for univariate time series forecasting, outperforming traditional statistical methods and RNN-based models. Ma et al. (2019) applied transformers for the imputation of missing values in multivariate time series, showcasing their effectiveness in handling the data. Hao et al. (2020) [27] provides a 2-step attention-based CNN model which designs an attention mechanism to extract memories across all time steps and then applies another attention mechanism for variable selection. The work presented by Zerveas et al. (2021) [7] introduces a transformer-based framework for unsupervised representation learning of multivariate time series. This methodology leverages unlabeled data by pre-training a transformer model with an input denoising objective. Zhou et al. (2020) [12] introduced a Transformer-based architecture with two symmetric language-specific encoders. This Multi-Encoder-Decoder Transformer architecture effectively captures individual language attributes and employs a language-specific multi-head attention mechanism in the decoder module.

Our approach shows significant improvement over the state-of-the-art methods, as it performs feature extraction on variable-length time series data, learning to construct multiple robust fixed-length representations of the original information, with the different representations serving as inputs to the Multi-Encoder-Decorder Transformer model, leveraging the architecture's efficiency to capture maximal information from each feature set.

3 Problem Formulation

In this section, we formulate the problem solved in this paper in a mathematical way. Time series features record the data in ordered sequences of points over time, such as the gait force and the moving activities being recorded based on a sequence of time slots. We assume a time series sample X can be represented by $|T|$ ordered data points where T is the full-time period. In this case, each sample of records can be written as $X = \{x_{t_1}, x_{t_2}, \dots x_{t_i}; t_i \in T\}$, where x_{t_i} represents all features of X at time t_i, and a time series data set D with N features can be written as $D = \{X_1, X_2, \dots X_n; n \in N\}$. So the main problem can be formulated as follows [4]:

$$f(X_1, X_2 \dots, X_n) : \mathcal{D} \to \mathcal{C} \tag{1}$$

where D is the input time series data and C is the class labels.

For the length of a sample $X_i \in D$, in most cases, it will vary individually, which means $|X_n| \neq |X_m|$ where $n, m \in N$. It is impossible to find a model f to handle inconsistent dimensional data.

Our solution is to extract a set of atomic units A by applying a data-driven summarized method $E(X) : D \to A$, where $A = \{a_1, a_2 \dots a_k\}$ and k will be a fixed value. Given the set A, it allows us to construct a new fixed-dimensional

summarized data D' based on the atomic units contained in it. The samples X' in the new data D' can be represented by the set of atomic units A. Now, it will be able to apply a clustering method f on the summarized data D'.

4 Methodology

In contrast to the existing methods, the proposed algorithm is applicable to both fixed and variable length time series data. The algorithm consists of the following phases:

- Feature Construction: Summarizing the variable-length time series data and constructing a fixed-length representation based on its atomic units.
- Multi-view Representation: Constructing multiple representations of the original dataset, each with varying number of features.
- Classification: Applying our Multi-Encoder-Decoder Transformer (MEDT) model to classify multivariate time series data instances.

4.1 Feature Construction

In time series data, there are always some repeated events that happen over time. The biggest challenge in the field is how to find patterns and relationships between these repeated events and time steps. In our study, we introduce the term of atomic units (Definition 1) which constitutes these events [4].

Definition 1. ***Atomic Unit***
Suppose a time series data is split into a sequence of small time periods, we assume that there are some repeated common characteristics among the sequence, this kind of common characteristics are named as atomic units.

Atomic units are the basic building block of a time series data. They appear in time series data repeatedly and play a significant role in representation. According to the different resolutions (number of features) required, we apply the Gaussian Mixture Model to extract a set of atomic units for each resolution. The new representations are ratio features built by each set of atomic units.

Gaussian Mixture Model (GMM) is a probability-based unsupervised clustering technique which is formed by several single Gaussian distributions[19]. With these individual Gaussian distributions, we can simply simulate the time series events by clustering the atomic units set A over time steps (T_k). Suppose we have time series data D and features X_n, the GMM clustering method is applied to find the best individual distribution of each atomic unit, then we are able to construct new ratio features based on the GMM clustering results.

Considering the time steps as the basic elements of a time series event, we can define a set of the distributions $p_k(x)$ over all time steps. And clustering

each step into one of our designed atomic units following the distribution $p_k(x)$. We calculate the probability μ of each time step belongs to a atomic unit[4], where

$$\mu_k(x) = \frac{\pi_k N(x\|\mu_k, \sigma_k)}{\sigma_l \pi_l N(x\|\mu_l \sigma_l)} \tag{2}$$

The best matching atomic unit will be assigned based on the calculation of $argmax\mu_k(x)$ over all potential atomic units.

After pairing the time steps and atomic units by calculating $argmax\mu_k(x)$, we can construct a fix-length ratio feature to represent the original high dimensional time series events, where

$$r_k = \frac{|A_k|}{|A|} \tag{3}$$

The ratio can be calculated with the number of the atomic unit over the total number of atomic units. Then the time series D and be summarized as a vector $[\frac{|A_1|}{|A}], \frac{|A_2|}{|A|} \cdots \frac{|A_k|}{A}]$.

4.2 Multi-view Representation

Our feature construction approach can represent time series data with an arbitrary number of features. Since the created ratio features have a very low level of collinearity, we can generate multiple summarizations of the source data with different resolutions, which convert the source data into a set of multi-view representations based on atomic units.

Inspired by the idea of multilingual model [24], we create multiple representations with different numbers of features to describe single time series data. These representations are considered as multiple views of the data. We believe multi-view representations can help capture more information under different granularity [25, 26]. To take as much as patterns in multi-view representations, we proposed a multi-encoder-decoder transformer model which uses multiple views as inputs.

4.3 Multi-Encoder-Decoder Transformer (MEDT) Classification

We propose a modified architecture of the model - one where each encoder-decoder pair will take as input two different representations of the same dataset. After constructing new features, we may choose to represent the original dataset with two new datasets of n and m features, passing each individually into either a single encoder or an additional decoder. For example, we may choose to construct two representations - one with 10 and another 20 with features. Thus, we would feed either the 10 or the 20 feature dataset into either the encoder or the decoder. If we decide to construct 8 different representations of the original dataset, we may choose to have 4 encoder-decoder pairs. The output of an encoder, along with the output of a decoder, passes through dense layers and a softmax layer, producing the final output of the Multi-Encoder-Decoder Transformer (MEDT).

The detailed architecture of the model is presented in Figure 1.

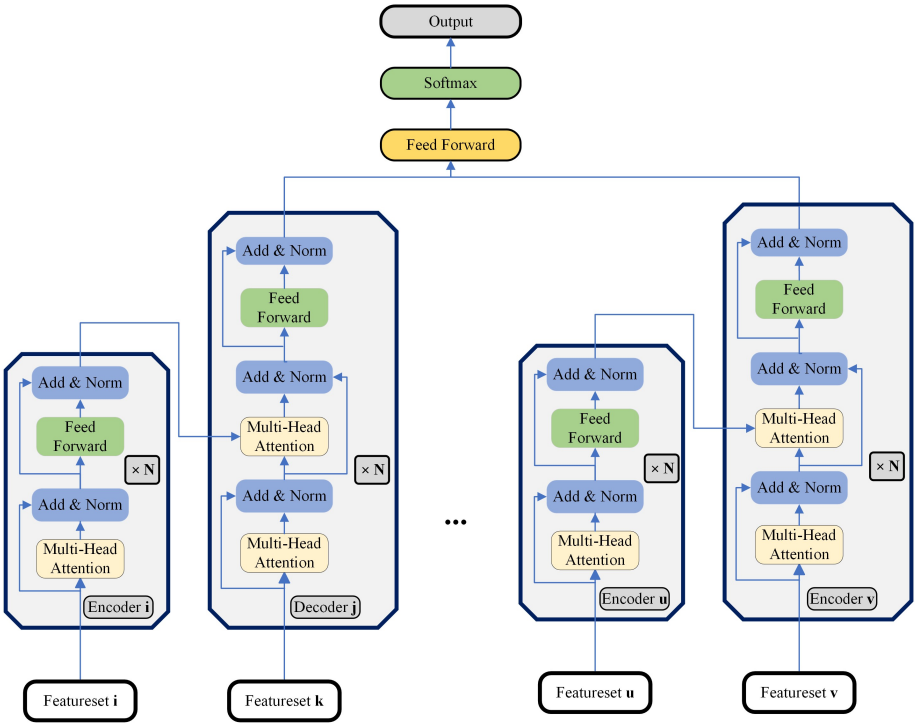

Fig. 1. Multi-Encoder-Decoder Transformer (MEDT) Architecture

5 Experiments

Our model is evaluated on the physical activities dataset [16] and 5 real-world datasets from the UEA multivariate time series classification (MTSC) archive [15]. We compare the performance of our model to the other four state-of-art methods.

5.1 Experiments using Multivariate Time Series Data Benchmarks

We evaluate our method on 5 multivariate time series datasets from the University of East Anglia (UEA) Multivariate archive. We select the data from different domains, such as human activity, motion classification, and audio spectra classification, with various dimensions (from 3 to 1345), length (from 29 to 1751), and number of classes (from 2 to 26). We use 4 fixed-length datasets and 1 variable-length dataset in this study to show the robust and competitive performance of our method. The details are shown in Table 1.

Table 1. Multivariate Time Series Datasets

Dataset	Train	Test	Dimensions	Length	Classes
DuckDuckGeese	50	50	1345	270	5
Heartbeat	204	205	61	405	2
Handwriting	150	850	3	152	26
EthanolConcentration	261	263	3	1751	4
JapaneseVowels*	270	370	12	29	9
*: Variable-length dataset					

Baseline:
We compare our performance with the plain transformer using single view and four state-of-art multi-variable time series classification methods: Vanilla Transformer (TF) [5], Rule Transformer (RT) [6], Time Series Transformer (TST) [7], Complexity Measures and Features for Multivariate Time Series (CMFMTS) [9], RandOM Convolutional KErnal Transform (ROCKET) [2, 10].

Results:
In this section, we show the experiment results of our proposed method and compare it with other current classification algorithms. The original training and testing splits are used. Our method presents a competitive performance in most cases and in the first rank of average ranking, the details of results are shown in Table 2.

Our method (MEDT) shows the best performance in DuckDuckGeese and Heartbeat datasets. The DuckDuckGeese has the highest dimensions which is 1345, because of this, some competitors didn't provide results on this dataset, for instance, it is too large to run with the vanilla TF model. For other datasets, although our method stands at the second place among all models, it provides a very close performance to the best one. The JapaneseVowels dataset has variable-length time steps for each sample, and due to this issue, two of the competitors do not include it in their studies. Our proposed MEDT method presents the best overall average rank among all methods listed, which shows a solid good performance of it.

Table 2. Performance in Accuracy for Multivariate datasets

Dataset	TF	RT20%	TST(pre-train)	CMFMTS	ROCKET	MEDT
DuckDuckGeese	–	18.0%	–	51.0%	–	**51.99%**
Heartbeat	72.66%	73.17%	77.6%	76.8%	72.68%	**78.05%**
Handwriting	3.76%	26.24%	**35.9%**	27.4%	21.88%	28.5%
EthanolConcentration	25.81%	**41.44%**	32.6%	26%	27.38%	33.46%
JapaneseVowels*	93.4%	–	**99.7%**	83.7%	–	98.8%
Average Rank	5.25	3	1.75	3.4	4.67	**1.6**
*: Variable-length dataset						

5.2 Experiment using a real-world Physical Activities Dataset

The physical activities dataset includes a variety of activities from 184 child participants between 8 years old and 15 years old. The original data was published in [16]. Table 3 shows the detailed distribution for the related activities of each class.

The raw sensor recordings were cut into 12-second windows and generated model-based features with domain knowledge. Our method is feature-agnostic, so we use one of the most popular features, the percentile features, following the previous study in [17, 18]. The 10th, 25th, 50th, 75th, and 90th percentiles are used in the summarization process to generate more robust features. The minimum and maximum are excluded to reduce potential outliers. Since the window length is 12 seconds and the resolution of the data is 1 second, in this case, the nearest points (2nd, 3rd, 6th, 9th, and 11th) are used for the features[4].

Table 3. The activity classes in each coarse categories
and the number of 12-second windows recorded in the classes and categories.

Category and Class		Number of Windows	
Sedentary	Lying Rest	14755	16475
	Playing Computer Games	860	
	Reading	860	
Light Household and Games	Light Cleaning	840	2505
	Sweeping	865	
	Workout Video	800	
Moderate-Vigorous Household and Sports	Wall Ball	845	1570
	Playing Catch	725	
Walk	Brisk Track Walking	1210	3775
	Slow Track Walking	1000	
	Walking Course	1565	
Run	Track Running	485	485

Results:
To compare the performance of our algorithm, we run the baseline model with identical hyperparameters, single view without summarization, same training and testing split. The results show that our summarized model completely outperforms the baseline classification model. Our model presents a good performance of overall accuracy at 93.24%, while, the baseline model has much lower accuracy at 86.30%. Typically, the baseline model requires much more time in the training. The training process takes 29372 seconds on the baseline model, but only half of that, 14981 seconds, on our proposed model with summarize mechanism. The confusion matrix of both models is presented in Fig. 2 and 3. In the assessment of sensitivity (Sens) and specificity (Spec), our MEDT model consistently outperforms across all sub-categories. Especially, in minor sub-categories

such as LHH, MtV and Run, our method exhibits a superior level of sensitivity performance in comparison to the baseline model.

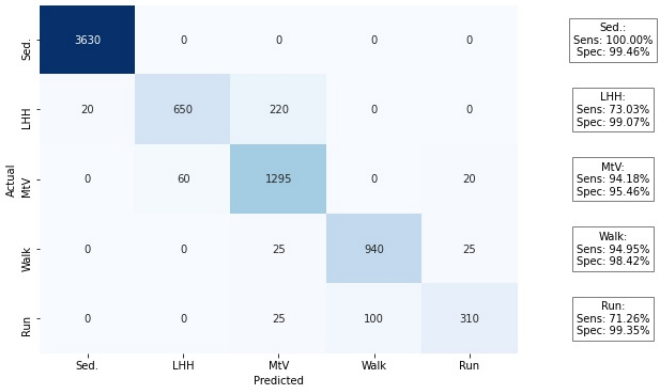

Fig. 2. The Confusion Matrix of MEDT model

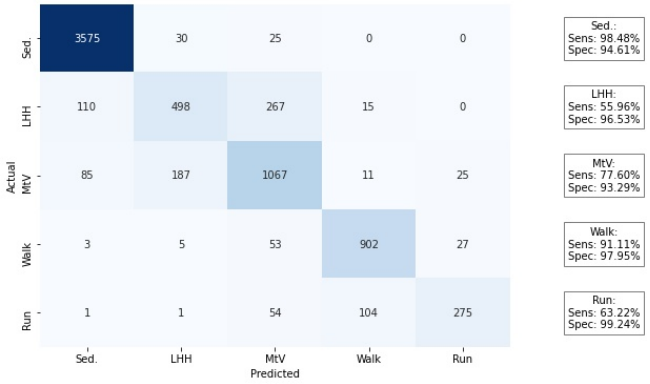

Fig. 3. The Confusion Matrix of Baseline Model

6 Conclusion

Although time series data offers insights into dynamic trends through collections of ordered data points, prevalent machine learning methods for time series are presenting challenges when dealing with noise, variable-length data and large data. We introduce a new approach utilizing GMM that provides fixed-length multi-view representations of variable-length time series data, allowing

for compatibility with any classical machine learning methods. Leveraging this algorithm, we construct multiple representations from the original dataset, applying them to a Multi-Encoder-Decoder Transformer (MEDT) architecture for a comprehensive, multi-view classification approach. Through extensive experiments using multiple benchmarks and a real-world dataset, our method shows significant improvement compared to the state-of-the-art methods.

7 Acknowledgement

This material is based upon work partially supported by the National Institutes of Health under grant NIH 1R01DK129428-01A1 and National Science Foundation under NSF grants 2008202 and 2334665. Any opinion, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the funding agencies.

References

1. Lines, Jason, Sarah Taylor, and Anthony Bagnall. "Time series classification with HIVE-COTE: The hierarchical vote collective of transformation-based ensembles." ACM Transactions on Knowledge Discovery from Data (TKDD) 12.5 (2018): 1-35.
2. Dempster, Angus, François Petitjean, and Geoffrey I. Webb. "ROCKET: exceptionally fast and accurate time series classification using random convolutional kernels." Data Mining and Knowledge Discovery 34.5 (2020): 1454-1495.
3. Shifaz, Ahmed, et al. "TS-CHIEF: a scalable and accurate forest algorithm for time series classification." Data Mining and Knowledge Discovery 34.3 (2020): 742-775.
4. Amaral, Kevin, et al. "SummerTime: Variable-length Time Series Summarization with Application to Physical Activity Analysis." ACM Transactions on Computing for Healthcare 3.4 (2022): 1-15.
5. Vaswani, Ashish, et al. "Attention is all you need." Advances in neural information processing systems 30 (2017).
6. O. Bahri, P. Li, S. F. Boubrahimi, and S. M. Hamdi, "Shapelet-based Temporal Association Rule Mining for Multivariate Time Series Classification," IEEE Xplore, Dec. 01, 2022. https://ieeexplore.ieee.org/stamp/stamp.jsp?tp= & arnumber=10020478 (accessed Sep. 02, 2023).
7. Zerveas, George, et al. "A transformer-based framework for multivariate time series representation learning." Proceedings of the 27th ACM SIGKDD conference on knowledge discovery & data mining. 2021.
8. Ismail Fawaz, Hassan, et al. "Inceptiontime: Finding alexnet for time series classification." Data Mining and Knowledge Discovery 34.6 (2020): 1936-1962.
9. Baldán, Francisco J., and José M. Benítez. "Multivariate times series classification through an interpretable representation." Information Sciences 569 (2021): 596-614.
10. Bier, Agnieszka, Agnieszka Jastrzębska, and Paweł Olszewski. "Variable-Length Multivariate Time Series Classification Using ROCKET: A Case Study of Incident Detection." IEEE Access 10 (2022): 95701-95715.
11. Ismail Fawaz, Hassan, et al. "Deep learning for time series classification: a review." Data mining and knowledge discovery 33.4 (2019): 917-963.
12. Zhou, Xinyuan, et al. "Multi-encoder-decoder transformer for code-switching speech recognition." arXiv preprint arXiv:2006.10414 (2020).

13. Li, Shiyang, et al. "Enhancing the locality and breaking the memory bottleneck of transformer on time series forecasting." Advances in neural information processing systems 32 (2019).
14. Wu, Neo, et al. "Deep transformer models for time series forecasting: The influenza prevalence case." arXiv preprint arXiv:2001.08317 (2020).
15. Bagnall, Anthony, et al. "The UEA multivariate time series classification archive, 2018." arXiv preprint arXiv:1811.00075 (2018).
16. Crouter, Scott E., Kurt G. Clowers, and David R. Bassett Jr. "A novel method for using accelerometer data to predict energy expenditure." Journal of applied physiology 100.4 (2006): 1324-1331.
17. Staudenmayer, John, et al. "An artificial neural network to estimate physical activity energy expenditure and identify physical activity type from an accelerometer." Journal of applied physiology 107.4 (2009): 1300-1307.
18. Trost, Stewart G., et al. "Artificial neural networks to predict activity type and energy expenditure in youth." Medicine and science in sports and exercise 44.9 (2012): 1801.
19. Aitkin, Murray, and Granville Tunnicliffe Wilson. "Mixture models, outliers, and the EM algorithm." Technometrics 22.3 (1980): 325-331.
20. Xu, Chang, Dacheng Tao, and Chao Xu. "A survey on multi-view learning." arXiv preprint arXiv:1304.5634 (2013).
21. Dufter, Philipp, Martin Schmitt, and Hinrich Schütze. "Position information in transformers: An overview." Computational Linguistics 48.3 (2022): 733-763.
22. Costa-jussà, Marta R., et al. "No language left behind: Scaling human-centered machine translation." arXiv preprint arXiv:2207.04672 (2022).
23. Hota, H. S., Richa Handa, and Akhilesh Kumar Shrivas. "Time series data prediction using sliding window based RBF neural network." International Journal of Computational Intelligence Research 13.5 (2017): 1145-1156.
24. Devlin, Jacob, et al. "Bert: Pre-training of deep bidirectional transformers for language understanding." arXiv preprint arXiv:1810.04805 (2018).
25. Li, Yingming, Ming Yang, and Zhongfei Zhang. "A survey of multi-view representation learning." IEEE transactions on knowledge and data engineering 31.10 (2018): 1863-1883.
26. Xie, Zhuyang, et al. "Deep learning on multi-view sequential data: a survey." Artificial Intelligence Review 56.7 (2023): 6661-6704.
27. Hao, Yifan, and Huiping Cao. "A new attention mechanism to classify multivariate time series." Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence. 2020.
28. Dempster, Angus, Daniel F. Schmidt, and Geoffrey I. Webb. "Minirocket: A very fast (almost) deterministic transform for time series classification." Proceedings of the 27th ACM SIGKDD conference on knowledge discovery & data mining. 2021.
29. Gao, Ge, et al. "A reinforcement learning-informed pattern mining framework for multivariate time series classification." In the Proceeding of 31th International Joint Conference on Artificial Intelligence (IJCAI-22). 2022.