A DECOMPOSITION BASED DUAL PROJECTION MODEL FOR MULTIVARIATE TIME SERIES FORECASTING AND ANOMALY DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Efficient anomaly detection and diagnosis in multivariate time series data is of great importance for various application areas. Forecasting of long-sequence time series is an important problem to prepare for future changes. An accurate prediction can help to detect anomaly events beforehand and make better decisions. It seems that one has to use more complex structures for deep learning models to get better performance, e.g., the recent surge of Transformer variants for time series modeling. However, such complex architectures require a large amount of training data and extensive computing resources. In addition, many of the considerations behind such architectures do not hold for time series applications. The objective of this study is to re-consider the effectiveness of deep learning architectures for efficient and accurate time series forecasting and anomaly detection. A model with direct projections is proposed, and it outperforms existing Transformer based models in most cases by a significant margin. The new decomposition based dual projection (DBDP) model consists of an anchored global profile and a varied number of decomposed seasonal local profiles of the time series for better forecasting performance. In addition to forecasting, a non-contrastive self-supervised learning approach, we propose to include a contrastive learning module in the DBDPC model for better forecasting performance and robustness. Finally, we apply the DBDP and DBDPC models to forecasting based time series anomaly detection and achieve superior performance over the latest SoTA models. These results demonstrate the effectiveness of the several key considerations behind the DBDP and DBDPC models, which also encourages the development of new architectures for time series applications.

1 INTRODUCTION

A time series is defined as an ordered sequence of continue values arranged in chronological order. It is a measure of the numerical properties of a system as it evolves over time. A time series contains a lot of system information that can be used to ensure the proper operation of the system.

Time-series forecasting has been an critical ingredient of many applications, such as climate modeling (Mudelsee, 2019), biological analyzing (Stoffer & Ombao, 2012), high-performance medicine (Topol, 2019), demand forecasting in retail (Böse et al., 2017), economic forecasting (Andersen et al., 2005) and power generation forecasting (Antonanzas et al., 2016; Foley et al., 2012). It can be categorized into one-step-ahead and multi-horizon time series forecasting (forecasting on multiple steps in future time), while the latter can be further divided empirically into short sequence and long sequence time series forecasting (LSTF) (Zhou et al., 2021). The latter, LSTF, has been the recent focus of many works, especially for the application of Transformer (Zhou et al., 2021; Wu et al., 2021; Liu et al., 2022; Cirstea et al., 2022; Zhou et al., 2022).

On the other hand, time series anomaly detection is a key research topic in many fields, such as the maintain and management of the modern IT systems, Internet of Things (IoT), robotics and urban resource (Audibert et al., 2020; Thudumu et al., 2020). Anomaly detection in large-scale databases is increasingly challenging due to the increasing complexity of the data modes (Tuli et al.,

2022). Many data-driven industries are now adopting deep learning based unsupervised methods for anomaly detection.

The time series anomaly detection methods can be categorized into the reconstruction methods and the forecasting methods. The latter do not require the current and future information, and thus can be utilized online for the stream data. As the modern society has seen more and more extensive utilization of the stream data, we focus on the forecasting methods in this study.

Recently, the Transformers have achieved superior performances in many tasks in natural language processing and computer vision, which also stimulated great interests in time series applications. Some notable models for the time series forecasting tasks include: Informer Zhou et al. (2021), Autoformer Wu et al. (2021), Pyraformer Liu et al. (2022), Triformer Cirstea et al. (2022) and FEDformer Zhou et al. (2022). One of the main benefits of the Transformers is the efficient capturing of the long-term dependencies with an optimal path length of one. Essentially, the attention module can be viewed as a fully connected layer with the weights that are dynamically generated based on the pairwise similarity of input patterns. Another efficiency gain is from the output method. The sequential output method is not efficient and there may be error accumulation. Informer (Zhou et al., 2021) started to use the parallel output method that all the elements of the forecasting sequence are generate at one time, which is more efficient and eliminates the error accumulation. However, many recent works also started to doubt about the effectiveness of the Transformers for time series applications (Zeng et al., 2022; Zhang et al., 2022). Time series data is a special kind of data that the step-wise pairwise relation may be not proper to capture the temporal dependencies. In addition, the efficiency gain can be achieved with direct feed forward networks instead of the attention mechanism.

Actually, the recent progresses on time series modeling mainly due to the introduction of multi-resolution representation or time series decomposition to the deep learning architecture (Wu et al., 2021; Zhou et al., 2022; Woo et al., 2022). Pyraformer (Liu et al., 2022) and Triformer (Cirstea et al., 2022) introduce hierarchical structures to get the multi-scale features. Whereas historically and naturally, the seasonal-trend decomposition may be a more suitable way for time series analysis. Such a decomposition block is a crucial part of the Autoformer (Wu et al., 2021) and FEDformer (Zhou et al., 2022).

Our main contributions are:

- We propose a model that does not contain the attention module while inheriting all the key performance contributing blocks, such as the channel projection that serves as the value embedding, and the sequence projection that generates the multi-step outputs directly at one time. To this end, our work is similar to DLinear (Zeng et al., 2022), but we are different that we explicitly propose channel projection for variable correlation learning.
- We include the time series decomposition as a basic building block. The proposed Decomposition Based Dual Projection (DBDP) model has decomposition and dual projection conducted alternatively to capture the total variable-temporal dependencies effectively.
- We further propose the DBDPC model with a contrastive learning module attached to DBDP, inspired by the self-supervised learning.

Our experimental results successfully demonstrated the superior performance of the DBDP and DBDPC models on multi-variate time series forecasting and anomaly detection. The remainder of this paper is organized as follows. Section 2 presents the related works on the key considerations of our model. Section 3 explains the methods and section 4 gives the typical experimental results. Section 5 concludes the paper.

2 Related works

2.1 The Transformer and attention

The vanilla Transformer Vaswani et al. (2017) has an encoder-decoder structure that is commonly adopted in most sequence models. Both encoder and decoder are composed of multiple multi-head self-attention modules, and a cross-attention module is used to connect the two. The position-wise

feed forward network (FFN) is used to transform the dimensionality of input to the model for the intermediate attention layers and finally to the output.

Various variants of Transformers have been proposed for LSTF tasks, with the focus on solving the quadratic time complexity and memory consumption Zhou et al. (2021); Wu et al. (2021); Cirstea et al. (2022); Liu et al. (2022). It is well recognized that the Transformers are capable of capturing the long-range dependencies and interactions, which is especially attractive for time series modeling. This is realized with the attention mechanism which uses the paired information in a sequence, and is permutation-invariant, i.e., regardless of the order. However, the most important information for time series analysis is the temporal dynamics among a continuous set of time steps. Transformers use only the position encoding added in the input embeddings to model the sequence information, which makes people doubt about their effectiveness for time series analysis (Zeng et al., 2022).

In addition, capturing the correlation among multiple variables is another key issue for multi-variate time series modeling. The attention mechanism in the Transformers essentially does not contribute to this point. Instead, the value embedding before the attention layers, and the final point-wise projection to the output target time series are the main contributors.

2.2 Time series decomposition

Recently, people began to realized that time series is a special type of sequence data that down-sampling of time series often preserves most of the information, while this is not true for general sequence data such as text sequences and DNA sequences Liu et al. (2021); Zhang et al. (2022). Various down-sampling techniques were proposed to capture the multi-resolution or long-short term information.

SCINet Liu et al. (2021) mentioned that existing deep learning techniques using generic sequence models all ignore such unique properties of time series. It proposed a hierarchical framework to capture the temporal dependencies at multiple temporal resolutions. LightTS Zhang et al. (2022) proposed two delicate down-sampling strategies, including interval sampling and continuous sampling, inspired also by the fact that down-sampling time series often preserves the majority of its information.

Instead of using down-sampling to construct the multi-resolution architecture, some people tend to the seasonal-trend decomposition for the decomposed representation of time series, which is related to ideas from the Bayesian Structure Time Series models Qiu et al. (2018).

N-BEATS Oreshkin et al. (2019) was among the first to explore the seasonal-trend decomposition of time series to construct an stacked interpretable architecture to make the stacked outputs more easily interpretable. Autoformer Wu et al. (2021) proposes to use the seasonal-trend decomposition as a builtin block for both encoder and decoder. The encoder eliminates the long-term trend part and focuses on seasonal patterns modeling. The decoder accumulates the trend parts extracted from hidden variables progressively. FEDformer Zhou et al. (2022) proposes to combine Transformer with the seasonal-trend decomposition in which the decomposition method captures the global profile of time series while Transformers capture more detailed structures.

2.3 Time series representation learning

In additional to the previous models that are directly applied to the target tasks like time series forecasting, some works propose a two-step approach, that is, learning a good representation of the time series as the first step, and then utilizing the learned representation for various following tasks.

From a broader view point of learning, forecasting is a kind of unsupervised learning, specifically, the self-supervised learning, as no additional label information is involved during the process. Training with forecasting helps to learn the temporal dependencies. However, from the view point of learning the temporal dependencies, there are many other effective self-supervised learning methods except for forecasting. Among all, there is the contrastive learning that is extensively investigated and adopted in various fields.

As introduced in Balestriero & LeCun (2022), on contrast to the unsupervised learning that relies on a collection of inputs (**X**) and the supervised learning that relies on inputs and outputs (**X, Y**), self-supervised learning relies on inputs and some inter-sample relations (**X, G**). The inter-sample

relations can be the future information to be predicted in time series forecasting, or some component similarity that works as a kind of weak-supervision. For the latter, the relation matrix $\mathbf{G}$ is often constructed by augmenting the input through data-augmentations that preserve the input semantics, e.g., the flip for an image that will not change the prediction category of the image.

Contrastive Predictive Coding (CPC) (van den Oord et al., 2018) compresses high-dimensional data into a much more compact latent embedding space in which conditional predictions are easier to model, and then using a powerful autoregressive model in this latent embedding space to predict many steps in the futures. Compressing high-dimensional data can encode the underlying shared information between different parts of the high-dimensional signal. By predicting many steps in the futures, it can discard low-level information and noise that is more local while capturing more global structure. A probabilistic contrastive loss is used to induce the latent space to capture information that is maximally useful to predict future samples.

Time-Series representation learning via Temporal and Contextual Contrasting (TSTCC) (Eldele et al., 2021) learns the transformation invariant representation by augmenting the samples to two different yet correlated views. A temporal contrasting module is proposed to learn robust temporal representations by designing a tough cross-view prediction task. To further learn discriminative representations, a contextual contrasting module built upon the contexts from the temporal contrasting module is proposed to maximize the similarity among different contexts of the same sample while minimizing similarity among contexts of different samples.

TS2Vec (Yue et al., 2022) learns the temporal and contextual contrastive representations like TSTCC (Eldele et al., 2021), but in a different way. It proposes a new augmentation method which is used for multi-scale temporal contrastive learning. It randomly crops the sub-segment from a very long time series, and then mask some part of the cropped sub-segment time series to construct the positive samples. This augmentation method is very different with others like TSTCC (Eldele et al., 2021), which uses the jitters, crop, or permutation. The temporal contrastive learning is also different with TSTCC. The temporal contrastive in TS2Vec learns directly from the two augmentation, which the temporal contrastive in TSTCC learns from the prediction many steps like CPC (van den Oord et al., 2018). It finally gets an compressed latent space vector, which is used for the following tasks.

3 METHOD

We propose a model that is aimed to provide efficient and accurate multi-variate time series forecasting, which can be further applied to online anomaly detection. For the requirement of efficiency, no recurrent network or attention module would be used. Instead, the direct channel-wise sequence projection would be adopted to capture the temporal dependencies, which could be achieved with multi-layer perceptions (MLPs) or conv1D layers. To capture the correlations among the variables, we just borrow the embedding layers of the common sequence models like Transformers, which form the step-wise channel projection. The time series decomposition would be adopted to further improve the accuracy.

3.1 DBDP: DECOMPOSITION BASED DUAL PROJECTION NETWORK

The proposed Decomposition Based Dual Projection (DBDP) model is illustrated in Fig. 1. It is mainly composed of the time series decomposition block and the dual projection block. For time series decomposition, a moving average layer is used to capture the *slow stream* of the input sequence, which is regarded as the *Trend* component. Subtracting this part from the input sequence would produce the *fast stream*, which is regarded as the *Seasonal* component. The dual projection block consists of the channel-wise sequence projection and the step-wise channel projection block which can be stacked in any order.

At the beginning, a decomposition block is applied to the input sequence, and the seasonal component is selected for the following operations. The trend component is discarded, as it does not contain much detail information. The same practice has been also adopted in Autoformer (Wu et al., 2021). We think the selection of the seasonal component is like a sequence-wise scaling that makes the series all in a similar level with less changes in the level and the magnitude of the variations.

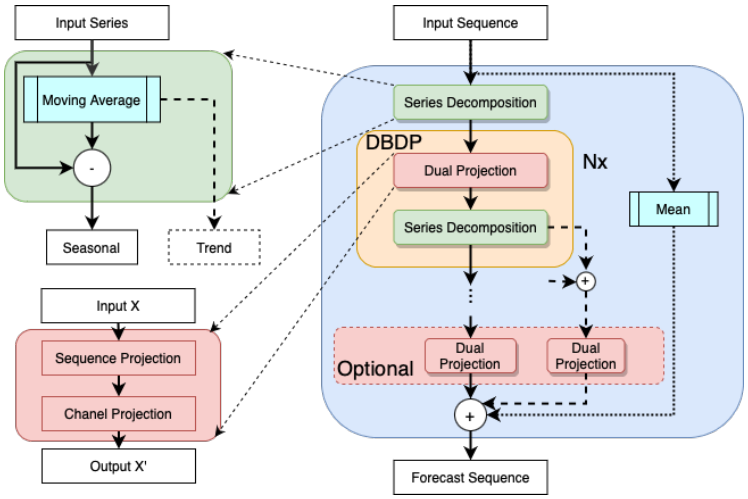

Figure 1: The architecture of the DBDP model.

After the first decomposition block, the stacked dual projection layers are applied, which captures the temporal dependencies and variable correlations and transforms the sequence to a latent representation. After that, the stacked decomposition blocks are applied, with the seasonal component as he input for the next decomposition block while the decomposed trend components are all aggregated together. Finally, optional dual projections are applied to the seasonal component of the last decomposition block and the aggregated trend component separately. The output of the two are added with the mean value of the original input series to give the final forecasting results.

To recover the global information, two architectures are constructed in the model. First, the *Mean* value of the input time series is recorded and added to the final forecasting results. We also tried to utilize the trend component of the beginning decomposition block for this purpose, but the performance is not improved when compared with the mean value. Second, the trend component of the following decomposition blocks are aggregated finally to give the output, which is also adopted in Autoformer (Wu et al., 2021).

3.2 DBDPC: Extension of the DBDP model with Contrastive learning

Data augmentation is a key component of the contrastive learning. For time series augmentation, masking, cropping or jitters may change the temporal variations of the time series. Analysis of the DBDP model shows that the "Mean" value of the original time series plays an important part. Therefore, to extend the DBDP model with contrastive learning (DBDPC), the proposed augmentation method is that, instead of masking some part of the cropped sub-series, the mean value of the input time series is used to randomly replace some selected sub-series. It will keep the mean value of the positive samples the same. The architecture of the DBDPC model is illustrated in Fig. 2. The contrastive learning is conduct on the seasonal components, which contain most of the local details.

DBDPC uses the similar temporal contrastive and instance contrastive learning as in TS2Vec Yue et al. (2022). The forecasting loss of the DBDP and contrastive loss are combined for the DBDPC model. Contrastive losses measure the similarities of the sample pairs in a representation space. Instead of matching an input to a fixed target, in contrastive loss formulations the target can vary on-the-fly during training and can be defined in terms of the data representation computed by a network. The contrastive loss is defined as:

$$L_c = -log \frac{exp(\frac{q \cdot k_+}{\tau})}{\sum_{i=0}^{K} exp(\frac{q \cdot k_i}{\tau})}, \tag{1}$$

where q and k_+ correspond to the original and the augmentation of a sample, respectively. $\{q, k_+\}$ is the positive pair, and τ is the temperature coefficient.

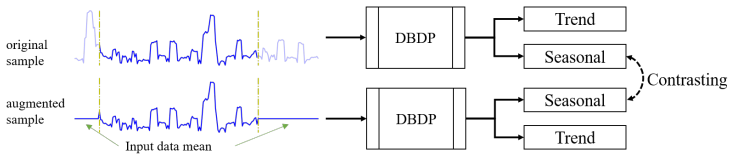

Figure 2: The architecture of the DBDPC model. Since the seasonal components contain the main local details, we conduct the contrastive learning process with the seasonal components.

4 EXPERIMENTS

For all the experiments in this paper, the DBDP model we applied contains only one DBDP module, so that the whole model is very simplified.

4.1 EXPERIMENTS ON LSTF PROBLEMS

Datasets Extensive experiments have been conducted on datasets from various domains. The basic characteristics of the datasets that have been used in the time series forecasting experiments are listed and compared in Table 1, which are ordered with the dimensionality of the channels.

Baselines For a fair comparison, all the experiments are done following the protocol of Informer (Zhou et al., 2021), which is widely used as a baseline for LSTF. The other baselines we reported here are all the latest SoTA, including two Transformer variants, Autoformer (Wu et al., 2021) and FEDformer (Zhou et al., 2022), and a non-Transformer, DLinear (Zeng et al., 2022). In addition, to clearly demonstrate the performance contribution of the decomposition block, we include a Dual Projection (DP) model that is a modification of DBDP by deleting the decomposition block.

Experiment settings For a fair comparison, all the experiments are done following the protocol of Informer (Zhou et al., 2021). All the experiments are conducted on a single GeForce GTX TITAN X GPU.

Experiment results Table 2 summarizes the results of our models and the top baselines for the multi-variate LSTF. The results of FEDformer and DLinear are directly copied from their papers, and so there are missing data for some settings. We run the other models by ourselves. The results show that: (i) The performance of DP is superior to Informer, especially for the longer forecasting sequences. This implies that the step-wise attention mechanism does not capture the temporal dependencies effectively, and the direct sequence projection can do the job better. (ii) DBDP generally performs the best with datasets of relatively lower dimensionality among the four models that utilize time series decomposition, Autoformer, FEDformer, DLinear and DBDP. For the two datasets of higher dimensionality, ECL and Traffic, DBDP performs on par with FEDformer and DLinear.

Table 1: Statistics of the dataset for forecasting experiments, where T is the length of time series, D is the number of features and R is the recording/sampling rate.

Dataset	T	D	R
$ETTh_1$	17420	7	1 hour
$ETTh_2$	17420	7	1 hour
$ETTm_1$	69680	7	15 minutes
$ETTm_2$	69680	7	15 minutes
Exchange	7588	8	1 day
ILI	983	11	1 week
Weather	52534	30	10 minutes
ECL	26304	213	1 hour
Traffic	17544	862	1 hour

Table 2: Multivariate LSTF results on the electricity datasets for different prediction lengths. The Informer and Transformer results are reproduced with the source code.

Models		Informer		Autoformer		FEDformer		DLinear		DP		DBDP	
Metric		MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
$ETTh_1$	24	0.577	0.549	0.384	0.425	-	-	-	-	0.468	0.480	**0.313**	**0.371**
	48	0.685	0.625	0.392	0.419	-	-	-	-	0.506	0.504	**0.349**	**0.393**
	96	0.865	0.713	0.449	0.459	0.376	0.419	0.386	0.400	-	-	**0.373**	**0.406**
	168	0.931	0.752	0.490	0.481	-	-	-	-	0.646	0.591	**0.428**	**0.444**
	192	1.008	0.792	0.500	0.482	0.420	0.448	0.437	0.432	-	-	**0.414**	**0.432**
	336	1.107	0.809	0.505	0.484	0.459	0.465	0.481	0.459	0.952	0.783	**0.449**	**0.453**
	720	1.181	0.865	0.498	0.500	0.504	0.506	0.519	0.516	0.798	0.679	**0.464**	**0.481**
$ETTh_2$	24	0.72	0.665	0.261	0.341	-	-	-	-	0.477	0.553	**0.225**	**0.318**
	48	1.457	1.001	0.312	0.373	-	-	-	-	0.724	0.689	**0.267**	**0.342**
	96	3.755	1.525	0.358	0.397	0.376	0.419	0.295	0.352	-	-	**0.301**	**0.362**
	168	3.489	1.515	0.457	0.455	-	-	-	-	2.320	1.254	**0.388**	**0.414**
	192	5.602	1.931	0.456	0.452	0.429	0.439	0.452	0.462	-	-	**0.409**	**0.427**
	336	4.721	1.835	0.471	0.475	0.496	0.487	0.504	0.490	2.488	1.300	**0.437**	**0.456**
	720	3.647	1.625	0.474	0.484	**0.463**	**0.474**	0.577	0.538	2.509	1.293	0.466	0.471
$ETTm_1$	24	0.323	0.369	0.383	0.403	-	-	-	-	0.301	0.366	**0.252**	**0.330**
	48	0.494	0.503	0.454	0.453	-	-	-	-	0.404	0.435	**0.316**	**0.372**
	96	0.678	0.614	0.481	0.463	0.379	0.419	**0.345**	**0.372**	0.456	0.473	0.348	0.394
	192	0.795	0.669	0.553	0.496	0.426	0.441	0.380	0.389	-	-	**0.373**	**0.390**
	288	1.056	0.786	0.634	0.528	-	-	-	-	0.528	0.525	**0.404**	**0.425**
	336	1.212	0.871	0.621	0.537	0.445	0.459	0.413	0.413	-	-	**0.408**	**0.416**
	672	1.192	0.926	0.606	0.542	-	-	-	-	0.644	0.597	0.477	0.467
	720	1.166	0.823	0.671	0.561	0.543	0.490	0.474	0.453	-	-	**0.471**	**0.445**
$ETTm_2$	24	0.173	0.301	0.153	0.261	-	-	-	-	0.201	0.331	**0.120**	**0.226**
	48	0.303	0.409	0.178	0.280	-	-	-	-	0.293	0.406	**0.159**	**0.256**
	96	0.365	0.453	0.255	0.339	0.203	0.287	**0.183**	**0.273**	0.367	0.460	0.212	0.295
	192	0.533	0.563	0.281	0.340	0.269	0.328	0.260	0.325	-	-	**0.243**	**0.310**
	288	1.047	0.804	0.342	0.378	-	-	-	-	0.783	0.704	**0.305**	**0.347**
	336	1.363	0.887	0.339	0.372	0.325	0.366	0.336	0.367	-	-	**0.308**	**0.353**
	672	3.126	1.302	0.434	0.430	-	-	-	-	1.353	0.933	**0.401**	**0.402**
	720	3.379	1.338	0.433	0.432	0.421	0.415	0.415	0.423	-	-	**0.409**	**0.411**
Exchange	96	0.847	0.752	0.197	0.323	0.148	0.278	0.078	0.197	0.178	0.343	**0.077**	**0.200**
	192	1.204	0.895	0.300	0.369	0.271	0.380	0.159	0.292	0.516	0.606	**0.153**	**0.289**
	336	1.672	1.036	0.509	0.524	0.460	0.500	0.274	0.391	0.762	0.737	**0.259**	**0.393**
	720	2.478	1.310	1.447	0.941	1.195	0.841	0.558	0.574	1.201	0.929	**0.536**	**0.594**
ILI	24	4.388	1.360	3.483	1.287	3.228	1.260	2.398	1.040	3.075	1.167	**2.092**	**0.951**
	36	4.651	1.391	3.103	1.148	2.679	1.080	2.646	1.088	2.937	1.075	**2.013**	**0.899**
	48	4.581	1.419	2.669	1.085	2.622	1.078	2.614	1.086	2.603	1.048	**2.139**	**0.959**
	60	4.583	1.432	2.770	1.125	2.857	1.157	2.804	1.146	2.676	1.030	**1.793**	**0.869**
Weather	96	0.300	0.384	0.266	0.336	0.217	0.296	0.196	0.255	0.250	0.314	**0.160**	**0.226**
	192	0.598	0.544	0.307	0.367	0.276	0.336	0.237	0.296	0.300	0.352	**0.207**	**0.280**
	336	0.578	0.523	0.359	0.395	0.339	0.380	0.283	0.335	0.341	0.375	**0.256**	**0.314**
	720	1.059	0.741	0.419	0.428	0.403	0.428	0.345	0.381	0.449	0.439	**0.330**	**0.370**
ECL	96	0.274	0.368	0.201	0.317	**0.193**	**0.308**	0.194	0.276	0.325	0.409	0.199	0.312
	192	0.296	0.386	0.222	0.334	0.201	0.315	**0.193**	**0.280**	0.285	0.383	0.202	0.314
	336	0.300	0.394	0.231	0.338	0.214	0.329	**0.206**	**0.296**	0.300	0.397	0.208	0.318
	720	0.373	0.439	0.254	0.361	0.246	0.355	0.242	0.329	0.312	0.401	**0.237**	**0.342**
Traffic	96	0.719	0.391	0.613	0.388	**0.587**	**0.366**	0.650	0.396	0.759	0.456	0.602	0.378
	192	0.696	0.379	0.616	0.382	**0.604**	**0.373**	0.598	0.370	0.713	0.437	0.607	0.391
	336	0.777	0.420	0.622	0.337	0.621	0.383	0.605	0.373	0.693	0.409	0.617	0.388
	720	0.864	0.472	0.660	0.408	**0.626**	**0.382**	0.645	0.394	0.713	0.425	0.637	0.397

Table 3: Further improvement for multivariate LSTF on the ECL and Traffic datasets for different prediction lengths.

Models		FEDformer		DLinear		DBDP		DBDPC	
Metric		MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
Traffic	96	0.587	0.366	0.650	0.396	0.602	0.378	**0.455**	**0.330**
	192	0.604	0.373	0.598	0.370	0.607	0.391	**0.539**	**0.370**
	336	0.621	0.383	0.605	0.373	0.617	0.388	**0.548**	**0.365**
	720	0.626	0.382	0.645	0.394	0.637	0.397	**0.591**	**0.384**
ECL	96	0.193	0.308	0.194	0.276	0.199	0.312	**0.164**	**0.273**
	192	0.201	0.315	0.193	0.280	0.202	0.314	**0.182**	**0.291**
	336	0.214	0.329	0.206	0.296	0.208	0.318	**0.194**	**0.296**
	720	0.246	0.355	0.242	0.329	0.237	0.342	**0.224**	**0.328**

For the high dimensional time series data, ECL and Traffic, the variable correlations become more important. As the DBDP model we adopted for the experiments contains only one DBDP block, it may be not enough to capture the total variable-temporal dependencies. The DBDPC model is thus applied to these two datasets to demonstrate the performance of the contrastive learning module. The results are listed in Table 3, with DBDPC showing obvious performance improvement.

4.2 Experiment on multi-variate time series anomaly detection

We focus on the forecasting based time series anomaly detection. The forecasting loss is the natural choice of the anomaly score. Furthermore, it is also possible to have the anomaly scores for each timestamp and each variable in the case of multivariate timeseries anomaly detection. As is common in prior works (Su et al., 2019; Boniol et al., 2020; Tuli et al., 2022), for a fair comparison, the Peak Over Threshold (POT) (Siffer et al., 2017) method is adopted to choose the threshold automatically and dynamically. The current time step is labeled anomalous if any of the variate is labeled anomalous.

It should be noted that we do not apply the *point adjustment* that was adopted in some works (Qi et al., 2022; Audibert et al., 2020) to get relatively higher Recall and F1-score.

Datasets Five publicly available datasets are tested in this study. Table 4 summarizes the dataset characteristics.

Baselines DBDP is compared with five latest SoTA multivariate timeseries anomaly detection baselines, including TranAD (Tuli et al., 2022), GDN (Deng & Hooi, 2021), OmniAnomaly (Su et al., 2019), MSCRED (Zhang et al., 2019), and MAD-GAN (Li et al., 2019).

Experiment settings Generally, all the experiments are done following the protocol of TranAD for a fair comparison. The hyper-parameters are chosen based on the baseline models as presented in their respective papers.

Experiment results Table 5 lists the results of six models on five datasets. DBDP generally performs superior over the other models. For the SMAP dataset with much more test than train data and the two datasets of higher dimensionality, SWaT and WADI, the DBDPC model is tested. The results shown in Table 6 confirmed the effectiveness of the contrastive learning module.

5 Conclusion

We propose a novel DBDP model for efficient and accurate multi-variate time series forecasting and anomaly detection. Instead of the attention mechanism, the direct sequence projection is adopted to learn the short and long-term temporal dependencies. Similarly, the channel projection is adopted to capture the variable correlations. Experiment results show that this dual projection architecture already surpass the step-wise attention on the LSTF tasks.

Table 4: Dataset statistics.

Dataset	train	Test	Dimension	Anomalies(%)
MBA	100,000	100,000	2	0.14
SMAP	135,183	427,617	25	13.13
SMD	708,405	708,420	38	5.37
SWaT	496,800	449,919	51	11.98
WADI	784,571	172,803	123	5.99

Table 5: Performance comparison of DBDP with several latest SoTA baseline methods on the five datasets. P: Precision; R: Recall; AUC: Area under the ROC curve; F1: F1 score with the complete training data. The best F1 and AUC scores are highlighted in bold.

Models	DBDP				TranAD				GDN			
Metric	P	R	AUC	F1	P	R	AUC	F1	P	R	AUC	F1
MBA	0.990	1.000	**0.997**	**0.995**	0.957	1.000	0.989	0.978	0.883	0.989	0.953	0.933
SMAP	0.828	1.000	0.990	**0.906**	0.804	1.000	**0.992**	0.892	0.748	0.989	0.986	0.852
SMD	0.999	1.000	**1.000**	**0.999**	0.926	0.997	0.997	0.961	0.717	0.997	0.992	0.834
SWaT	0.942	0.890	**0.941**	**0.915**	0.976	0.700	0.849	0.815	0.970	0.696	0.846	0.810
WADI	0.614	0.827	**0.898**	**0.705**	0.353	0.830	0.897	0.495	0.291	0.793	0.878	0.426
Models	OmniAnomaly				MSCRED				MAD-GAN			
Metric	P	R	AUC	F1	P	R	AUC	F1	P	R	AUC	F1
MBA	0.856	1.000	0.957	0.923	0.927	1.000	0.980	0.962	0.940	1.000	0.984	0.969
SMAP	0.978	0.696	0.847	0.813	0.813	0.942	0.989	0.873	0.816	0.922	0.989	0.865
SMD	0.888	0.999	0.995	0.940	0.728	0.997	0.992	0.841	0.821	0.922	0.992	0.868
SWaT	0.978	0.696	0.847	0.813	0.999	0.677	0.843	0.807	0.959	0.696	0.846	0.807
WADI	0.316	0.654	0.820	0.426	0.251	0.732	0.841	0.374	0.223	0.912	0.803	0.359

Table 6: Performance comparison of DBDPC with latest state-of-art baseline method TranAD on three datasets. P: Precision, R: Recall, AUC: Area under the ROC curve, F1: F1-score with complete training data. The best F1 and AUC scores are highlighted in bold.

Dataset	DBDP				DBDPC				TranAD			
	P	R	AUC	F1	P	R	AUC	F1	P	R	AUC	F1
SMAP	0.828	1.000	0.990	0.906	0.892	1.000	**0.994**	**0.943**	0.804	1.000	0.992	0.892
SWaT	0.942	0.890	0.941	0.915	0.964	0.929	**0.962**	**0.946**	0.976	0.700	0.849	0.815
WADI	0.614	0.827	0.898	0.705	0.718	0.894	**0.936**	**0.797**	0.353	0.830	0.897	0.495

Encouraged by the recent progresses of time series decomposition on LSTF, the newly proposed DBDP model includes the decomposition as a basic building block. Experiment results show that this combination can surpass the recent Transformers on LSTF tasks.

Inspired by the self-supervised learning, we further attach a contrastive learning module to the DBDP model, together with a newly proposed time series augmentation method. Experiment results demonstrate that it can effectively improve the forecasting performance, especially for the data with higher dimensionality and thus more complex variable correlations.

We also conducted multi-variate time series anomaly detection with forecasting based approach, which is beneficial for online stream applications. The superior performance of DBDP and DBDPC again confirm the effectiveness of our proposed architecture.

REFERENCES

Torben Andersen, Tim Bollerslev, Peter Christoffersen, and Francis Diebold. Volatility forecasting. NBER Working Papers 11188, National Bureau of Economic Research, Inc, 2005.

Javier Antonanzas, Natalia Osorio, Rodrigo Escobar, Ruben Urraca, Francisco J Martinez-de Pison, and Fernando Antonanzas-Torres. Review of photovoltaic power forecasting. *Solar energy*, 136: 78–111, 2016.

Julien Audibert, Pietro Michiardi, Frédéric Guyard, Sébastien Marti, and Maria A Zuluaga. Usad: Unsupervised anomaly detection on multivariate time series. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 3395–3404, 2020.

Randall Balestriero and Yann LeCun. Contrastive and non-contrastive self-supervised learning recover global and local spectral embedding methods. *arXiv preprint arXiv:2205.11508*, 2022.

Paul Boniol, Themis Palpanas, Mohammed Meftah, and Emmanuel Remy. Graphan: Graph-based subsequence anomaly detection. *Proceedings of the VLDB Endowment*, 13(12):2941–2944, 2020.

Joos-Hendrik Böse, Valentin Flunkert, Jan Gasthaus, Tim Januschowski, Dustin Lange, David Salinas, Sebastian Schelter, Matthias Seeger, and Yuyang Wang. Probabilistic demand forecasting at scale. *Proceedings of the VLDB Endowment*, 10(12):1694–1705, 2017.

Razvan-Gabriel Cirstea, Chenjuan Guo, Bin Yang, Tung Kieu, Xuanyi Dong, and Shirui Pan. Triformer: Triangular, Variable-Specific Attentions for Long Sequence Multivariate Time Series Forecasting. *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence*, 2022. URL https://www.ijcai.org/proceedings/2022/277.

Ailin Deng and Bryan Hooi. Graph neural network-based anomaly detection in multivariate time series. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 4027–4035, 2021.

Emadeldeen Eldele, Mohamed Ragab, Zhenghua Chen, Min Wu, Chee Keong Kwoh, Xiaoli Li, and Cuntai Guan. Time-series representation learning via temporal and contextual contrasting. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pp. 2352–2359. International Joint Conferences on Artificial Intelligence Organization, 8 2021.

Aoife M Foley, Paul G Leahy, Antonino Marvuglia, and Eamon J McKeogh. Current methods and advances in forecasting of wind power generation. *Renewable energy*, 37(1):1–8, 2012.

Dan Li, Dacheng Chen, Baihong Jin, Lei Shi, Jonathan Goh, and See-Kiong Ng. Mad-gan: Multivariate anomaly detection for time series data with generative adversarial networks. In *International conference on artificial neural networks*, pp. 703–716. Springer, 2019.

Minhao Liu, Ailing Zeng, Qiuxia Lai, and Qiang Xu. Time Series is a Special Sequence: Forecasting with Sample Convolution and Interaction. *arXiv:2106.09305*, 2021.

Shizhan Liu, Hang Yu, Cong Liao, Jianguo Li, Weiyao Lin, Alex X Liu, and Schahram Dustdar. PYRAFORMER: LOW-COMPLEXITY PYRAMIDAL AT-TENTION FOR LONG-RANGE TIME SERIES MODELING AND FORECASTING. In *ICLR 2022*, 2022. URL https://github.com/alipay/Pyraformer.

Manfred Mudelsee. Trend analysis of climate time series: A review of methods. *Earth-science reviews*, 190:310–322, 2019.

Boris N. Oreshkin, Dmitri Carpov, Nicolas Chapados, and Yoshua Bengio. N-BEATS: Neural basis expansion analysis for interpretable time series forecasting. *arXiv preprint arXiv:1905.10437*, 2019.

Zhi Qi, Hong Xie, Ye Li, Jian Tan, FeiFei Li, and John C. S. Lui. LPC-AD: Fast and Accurate Multivariate Time Series Anomaly Detection via Latent Predictive Coding. *arxiv:2205.08362*, 2022.

Jinwen Qiu, S. Rao Jammalamadaka, and Ning Ning. Multivariate Bayesian Structural Time Series Model. *Journal of Machine Learning Research*, 19(68):1–33, 2018. URL http://jmlr.org/papers/v19/18-009.html.

Alban Siffer, Pierre-Alain Fouque, Alexandre Termier, and Christine Largouet. Anomaly detection in streams with extreme value theory. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1067–1075, 2017.

David S. Stoffer and Hernando Ombao. Editorial: Special issue on time series analysis in the biological sciences. *Journal of Time Series Analysis*, 33(5):701–703, 2012.

Ya Su, Youjian Zhao, Chenhao Niu, Rong Liu, Wei Sun, and Dan Pei. Robust anomaly detection for multivariate time series through stochastic recurrent neural network. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pp. 2828–2837, 2019.

Srikanth Thudumu, Philip Branch, Jiong Jin, and Jugdutt (Jack) Singh. A comprehensive survey of anomaly detection techniques for high dimensional big data. *Journal of Big Data*, 7, 2020. doi: 10.1186/S40537-020-00320-X/TABLES/6.

Eric J Topol. High-performance medicine: the convergence of human and artificial intelligence. *Nature medicine*, 25(1):44–56, 2019.

Shreshth Tuli, Giuliano Casale, and Nicholas R Jennings. Tranad: Deep transformer networks for anomaly detection in multivariate time series data. *arXiv preprint arXiv:2201.07284*, 2022.

Aäron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *ArXiv*, abs/1807.03748, 2018.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pp. 5998–6008, 2017.

Gerald Woo, Chenghao Liu, Doyen Sahoo, Akshat Kumar, and Steven Hoi. Cost: Contrastive learning of disentangled seasonal-trend representations for time series forecasting. In *International Conference on Learning Representations*, 2022.

Haixu Wu, Jiehui Xu, Jianmin Wang, and Mingsheng Long. Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting. In *Advances in Neural Information Processing Systems*, pp. 22419–22430, 2021.

Zhihan Yue, Yujing Wang, Juanyong Duan, Tianmeng Yang, Congrui Huang, Yunhai Tong, and Bixiong Xu. Ts2vec: Towards universal representation of time series. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 8980–8987, 2022.

Ailing Zeng, Muxi Chen, Lei Zhang, and Qiang Xu. Are Transformers Effective for Time Series Forecasting? *arXiv preprint arXiv:2205.13504*, 2022.

Chuxu Zhang, Dongjin Song, Yuncong Chen, Xinyang Feng, Cristian Lumezanu, Wei Cheng, Jingchao Ni, Bo Zong, Haifeng Chen, and N. Chawla. A deep neural network for unsupervised anomaly detection and diagnosis in multivariate time series data. In *AAAI*, 2019.

Tianping Zhang, Yizhuo Zhang, Wei Cao, Jiang Bian, Xiaohan Yi, Shun Zheng, and Jian Li. Less Is More: Fast Multivariate Time Series Forecasting with Light Sampling-oriented MLP Structures. *arXiv preprint arXiv:2207.01186*, 2022.

Haoyi Zhou, Shanghang Zhang, Jieqi Peng, Shuai Zhang, Jianxin Li, Hui Xiong, and Wancai Zhang. Informer: Beyond efficient transformer for long sequence time-series forecasting. In *Proceedings of AAAI*, 2021.

Tian Zhou, Ziqing Ma, Qingsong Wen, Xue Wang, Liang Sun, and Rong Jin. Fedformer: Frequency enhanced decomposed transformer for long-term series forecasting. *arXiv preprint arXiv:2201.12740*, 2022.

A APPENDIX

Ablation with the anchor: Based on previous experiments, we additionally verified the effect of anchors on the model. When our model has no anchors, our experimental results are comparable to Dlinear. However, when our experimental results have anchors, our experimental results are greatly improved as shown in Table 7.

Table 7: Anchor or without anchor.

Models		FEDformer		Dlinear		DBDP w/o anchor		DBDP w/ anchor	
Metric		MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
$ETTh_1$	96	0.376	0.419	0.386	0.400	0.386	0.411	**0.373**	**0.406**
	192	0.420	0.448	0.437	0.432	0.432	0.441	**0.414**	**0.432**
	336	0.459	0.465	0.481	0.459	0.468	0.467	**0.449**	**0.453**
	720	0.504	0.506	0.519	0.516	0.506	0.514	**0.464**	**0.481**
$ETTh_2$	96	0.376	0.419	0.295	0.352	0.307	0.363	**0.301**	**0.362**
	192	0.429	0.439	0.452	0.462	0.387	0.412	**0.409**	**0.427**
	336	0.496	0.487	0.504	0.490	0.533	0.506	**0.437**	**0.456**
	720	**0.463**	**0.474**	0.577	0.538	0.619	0.553	0.466	0.471
$ETTm_1$	96	0.379	0.419	0.345	0.372	**0.333**	**0.373**	0.348	0.394
	192	0.426	0.441	0.380	0.389	0.374	0.400	**0.373**	**0.390**
	336	0.445	0.459	0.413	0.413	0.400	0.411	**0.408**	**0.416**
	720	0.543	0.490	0.474	0.453	0.465	0.450	**0.471**	**0.445**
$ETTm_2$	96	0.203	0.287	**0.183**	**0.273**	0.177	0.270	0.212	0.295
	192	0.269	0.328	0.260	0.325	0.253	0.326	**0.243**	**0.310**
	336	0.325	0.366	0.336	0.367	0.316	0.370	**0.308**	**0.353**
	720	0.421	0.415	0.415	0.423	0.484	0.477	**0.409**	**0.411**
ILI	24	3.228	1.260	2.398	1.040	2.697	1.115	**2.092**	**0.951**
	36	2.679	1.080	2.646	1.088	2.781	1.087	**2.013**	**0.899**
	48	2.622	1.078	2.614	1.086	2.873	1.121	**2.139**	**0.959**
	60	2.857	1.157	2.804	1.146	2.377	1.041	**1.793**	**0.869**

Comparing with the contrastive based method for LSTF problems: During the writing of this paper, there is another work coming out, the CoST model Woo et al. (2022) which applies contrastive learning methods to learn disentangled seasonal-trend representations for time series forecasting. Some of its reported results are even better than the latest SoTA baseline Autoformer. The DBDP model is compared with the CoST model following the settings of the latter, and the results are shown in Table 8 and Table 9. It can be seen that DBDP performs better for most of the cases while on par with CoST for the rest cases.

Anomaly Diagnosis: Anomaly diagnosis is the natural extension of anomaly detection, and it is important for anomaly locating, root cause identification and anomaly severity interpreting. The anomaly scores produced for each time step and each feature variable can be used for such diagnosis.

First, the distribution of the detected anomalies along the temporal dimension can be analyzed. As shown in Fig. 3, for the SMD dataset, nearly all the anomalies have been correctly detected by the DBDP model, except for very few false positives.

The predicted anomaly labels of each feature variable can be further analyzed for better anomaly diagnosis. Taking the SMD as an example, most of the anomalies can be revealed in the feature dimension 12 and 0, as shown in Fig. 4 and Fig. 5. For the other feature dimensions, only part of the anomalies can be revealed, such as those shown in Fig. 6, Fig. 7 and Fig. 8.

The same anomaly diagnosis methodology can be applied to the other datasets. Here the temporal distribution of the detected anomalies for the SWaT, SMAP and WADI datasets are shown as some of the examples, as shown in Fig. 9, Fig. 10, Fig. 12 and Fig. 14.

Table 8: Multivariate LSTF results on various datasets for comparison of the DBDP model and the CoST model.

Models		DBDP		CoST	
Metric		MSE	MAE	MSE	MAE
$ETTh_1$	24	**0.313**	**0.371**	0.386	0.429
	48	**0.349**	**0.393**	0.437	0.464
	168	**0.428**	**0.444**	0.643	0.582
	336	**0.452**	**0.454**	0.812	0.679
	720	**0.466**	**0.480**	0.970	0.771
$ETTh_2$	24	**0.225**	**0.318**	0.447	0.502
	48	**0.267**	**0.342**	0.699	0.637
	168	**0.388**	**0.414**	1.549	0.982
	336	**0.437**	**0.456**	1.749	1.042
	720	**0.466**	**0.471**	1.971	1.092
$ETTm_1$	24	**0.237**	**0.315**	0.246	0.329
	48	**0.316**	**0.372**	0.331	0.386
	96	**0.348**	**0.394**	0.378	0.419
	288	**0.404**	**0.425**	0.472	0.486
	672	**0.477**	**0.467**	0.620	0.574
WHE	24	**0.145**	**0.230**	0.298	0.360
	48	**0.191**	**0.277**	0.359	0.411
	168	**0.273**	**0.345**	0.464	0.491
	336	**0.291**	**0.333**	0.497	0.517
	720	**0.370**	**0.380**	0.533	0.542
ECL	24	0.149	0.270	**0.136**	**0.242**
	48	0.175	0.292	**0.153**	**0.258**
	168	0.194	0.310	**0.175**	**0.275**
	336	0.208	0.318	**0.196**	**0.296**
	720	**0.229**	0.338	0.232	**0.327**

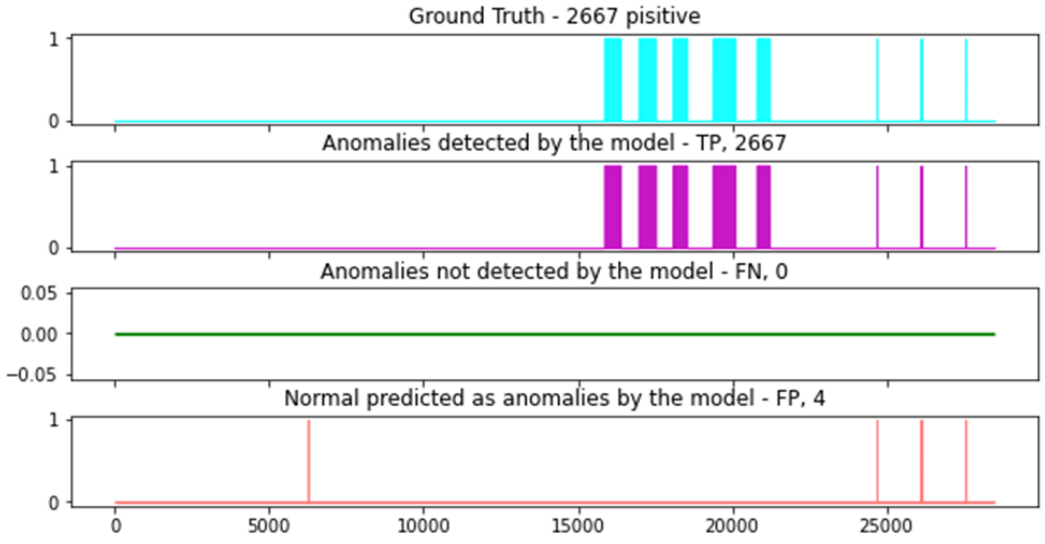

Figure 3: The predicted and ground truth labels for the SMD dataset with the DBDP model. Top row: Ground truth labels of the SMD machine-1-1 Dataset; Second row: the True Positives; Third row: the False Negatives; Fourth row: the False Positives.

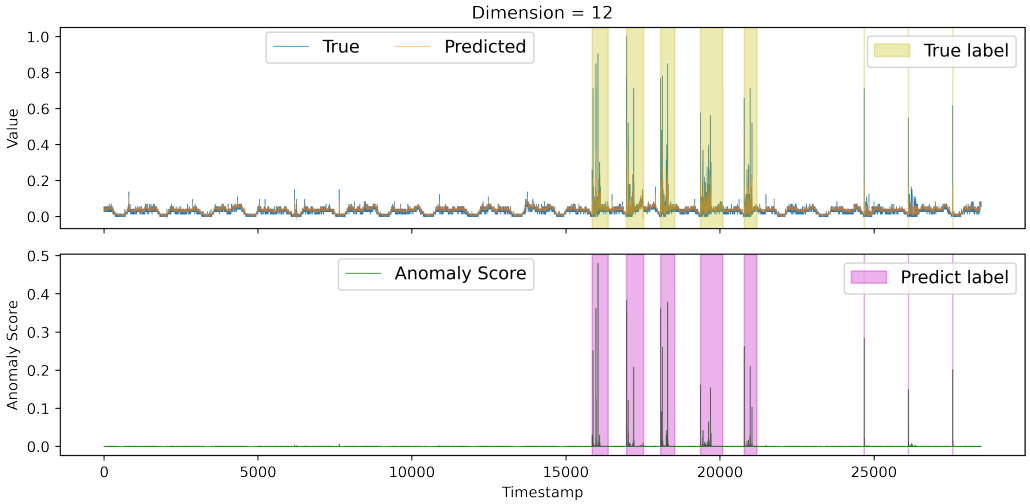

Figure 4: Anomaly scores, the predicted and the ground truth labels of the SMD datset with feature dimension 12.

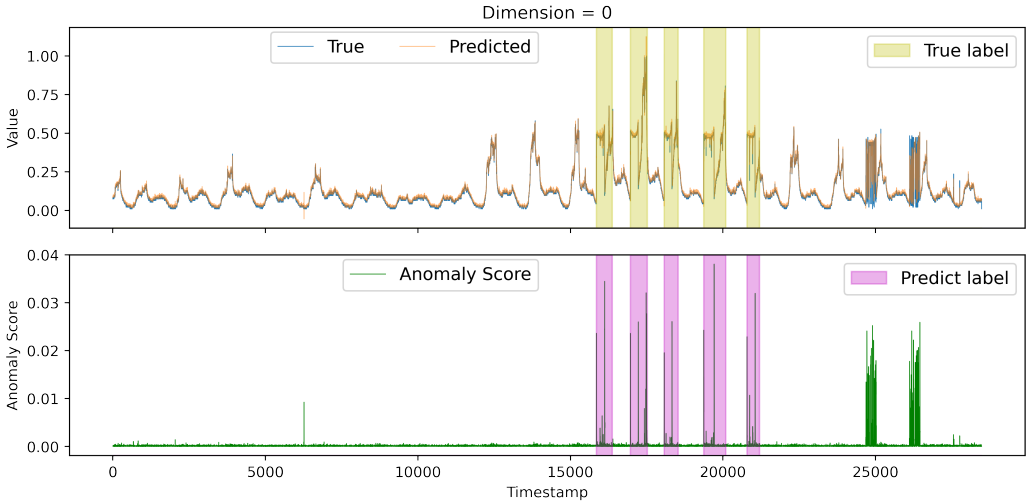

Figure 5: Anomaly scores, the predicted and the ground truth labels of the SMD datset with feature dimension 0.

Table 9: Univariate LSTF results on various datasets for comparison of the DBDP model and the CoST model.

Models		DBDP		CoST	
Metric		MSE	MAE	MSE	MAE
$ETTh_1$	24	0.040	0.155	**0.040**	**0.152**
	48	**0.055**	**0.179**	0.060	0.186
	168	**0.078**	**0.217**	0.097	0.236
	336	**0.090**	**0.238**	0.112	0.258
	720	**0.124**	**0.280**	0.148	0.306
$ETTh_2$	24	**0.075**	**0.208**	0.079	**0.207**
	48	**0.109**	**0.255**	0.118	0.259
	168	**0.181**	**0.331**	0.189	0.339
	336	0.214	0.368	**0.206**	**0.360**
	720	0.234	0.387	**0.214**	**0.371**
$ETTm_1$	24	**0.013**	**0.086**	0.015	0.088
	48	0.025	0.121	**0.025**	**0.117**
	96	**0.038**	**0.147**	0.038	0.147
	288	**0.074**	**0.207**	0.077	0.209
	672	**0.102**	**0.240**	0.113	0.257
WHT	24	**0.016**	**0.101**	0.096	0.213
	48	**0.015**	**0.102**	0.138	0.262
	168	**0.015**	**0.102**	0.207	0.334
	336	**0.028**	**0.133**	0.230	0.356
	720	**0.034**	**0.146**	0.242	0.370
ECL	24	0.305	0.407	**0.243**	**0.264**
	48	0.394	0.466	**0.292**	**0.300**
	168	**0.390**	0.458	0.405	**0.375**
	336	**0.414**	0.474	0.560	**0.473**
	720	**0.460**	**0.514**	0.889	0.645

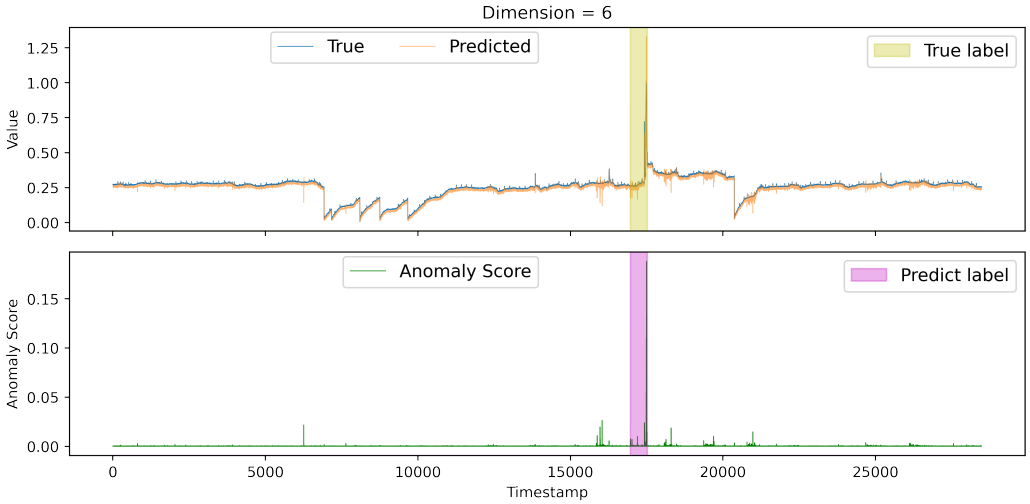

Figure 6: Anomaly scores, the predicted and the ground truth labels of the SMD datset with feature dimension 6.

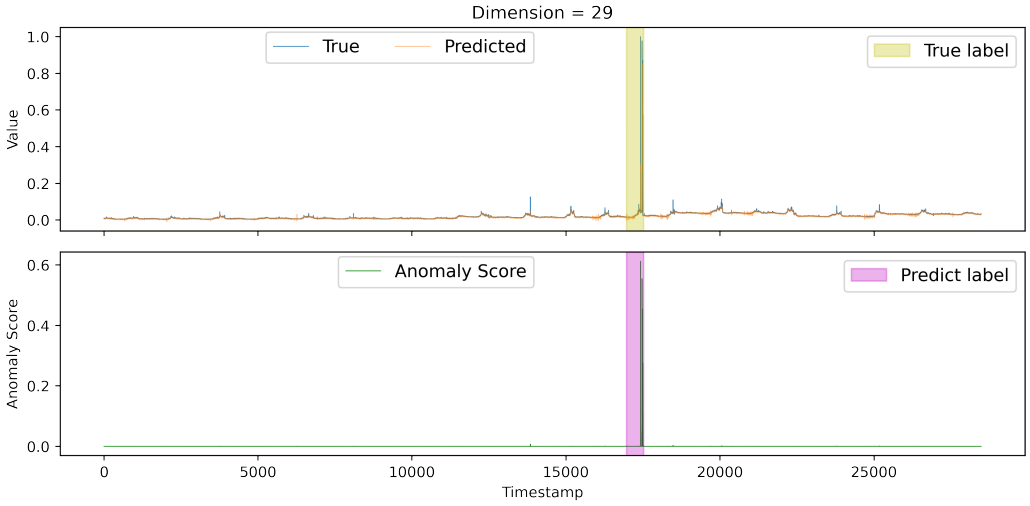

Figure 7: Anomaly scores, the predicted and the ground truth labels of the SMD datset with feature dimension 29.

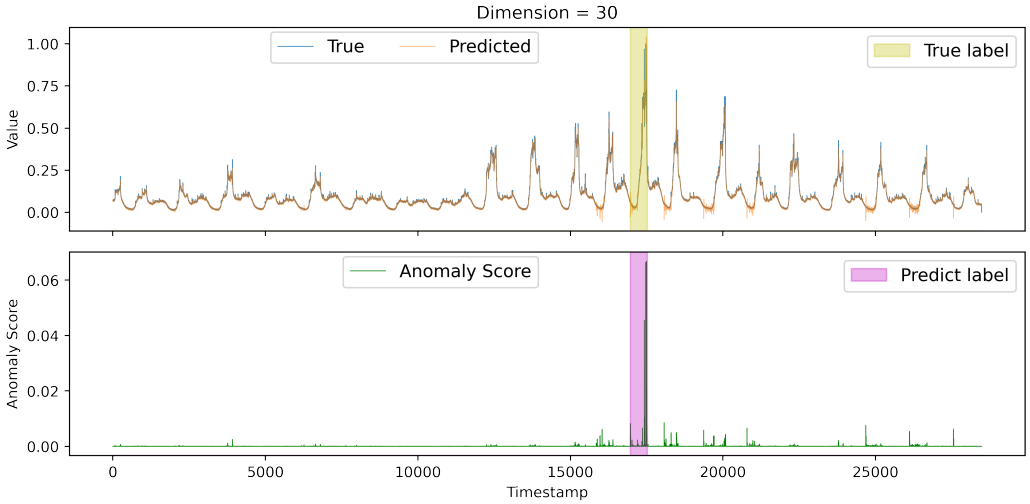

Figure 8: Anomaly scores, the predicted and the ground truth labels of the SMD datset with feature dimension 30.

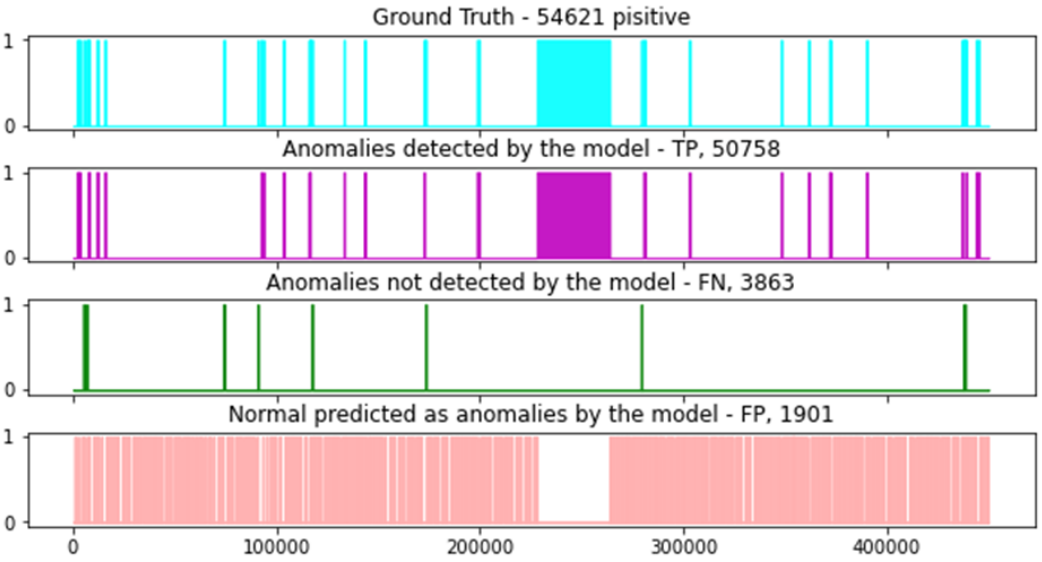

Figure 9: The predicted and ground truth labels for the SWaT dataset with the DBDP model. Top row: Ground truth labels of the SWaT Dataset; Second row: the True Positives; Third row: the False Negatives; Fourth row: the False Positives.

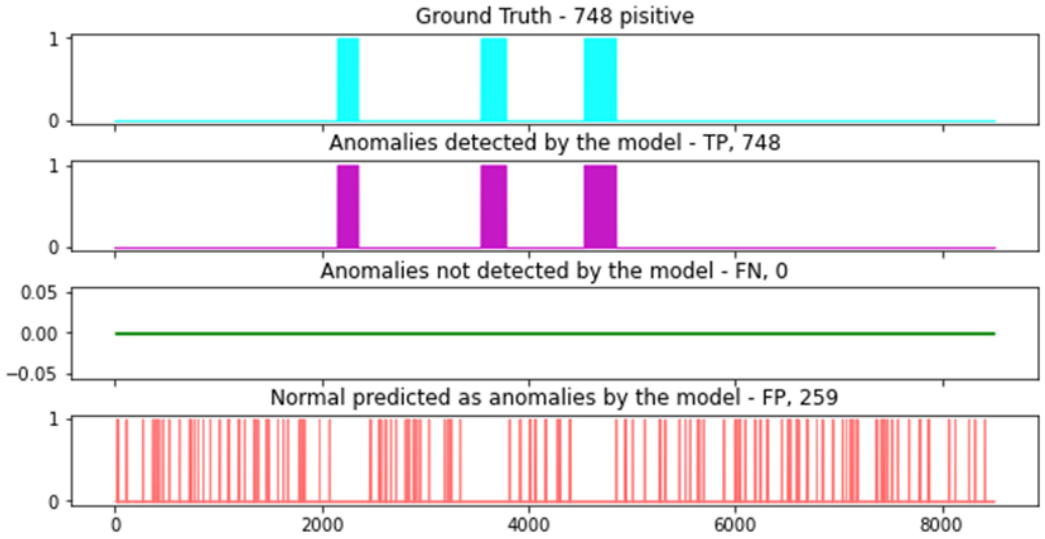

Figure 10: The predicted and ground truth labels for the SMAP dataset with the DBDP model. Top row: Ground truth labels of the SMAP Dataset; Second row: the True Positives; Third row: the False Negatives; Fourth row: the False Positives.

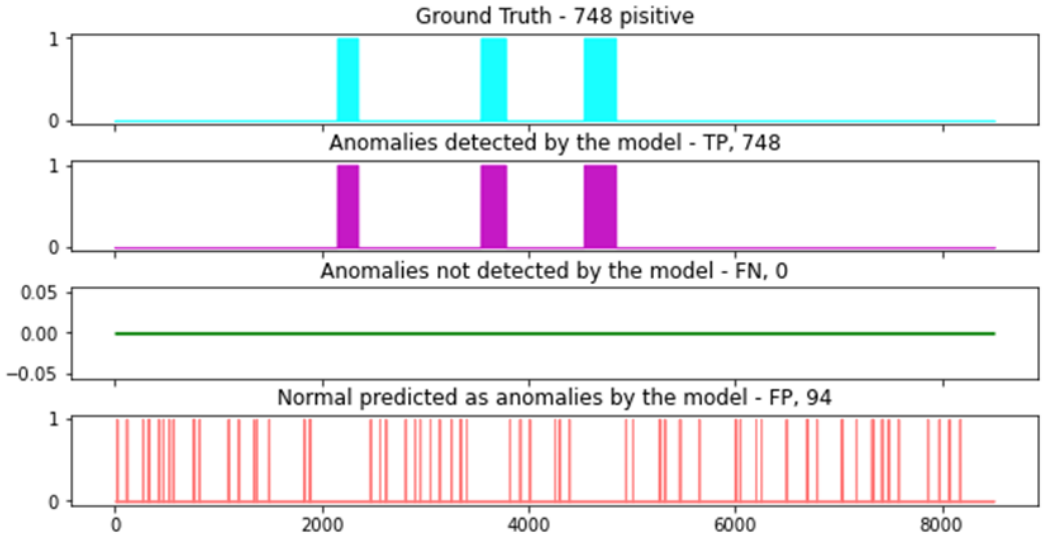

Figure 11: The predicted and ground truth labels for the SMAP dataset with the DBDPC model. Top row: Ground truth labels of the SMAP Dataset; Second row: the True Positives; Third row: the False Negatives; Fourth row: the False Positives.

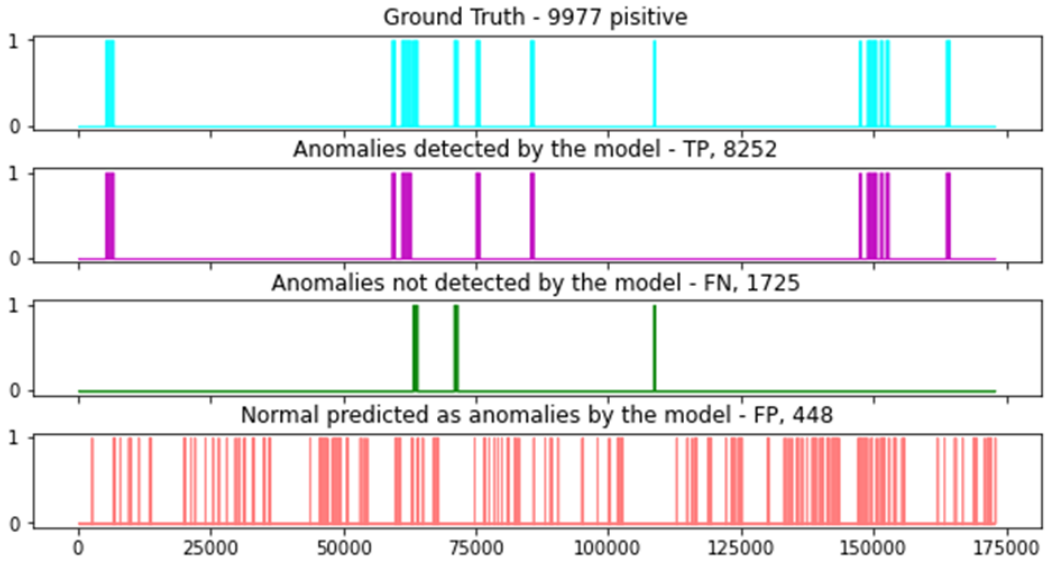

Figure 12: The predicted and ground truth labels for the WADI dataset with 82 filtered features using the DBDP model. Top row: Ground truth labels of the WADI Dataset; Second row: the True Positives; Third row: the False Negatives; Fourth row: the False Positives.

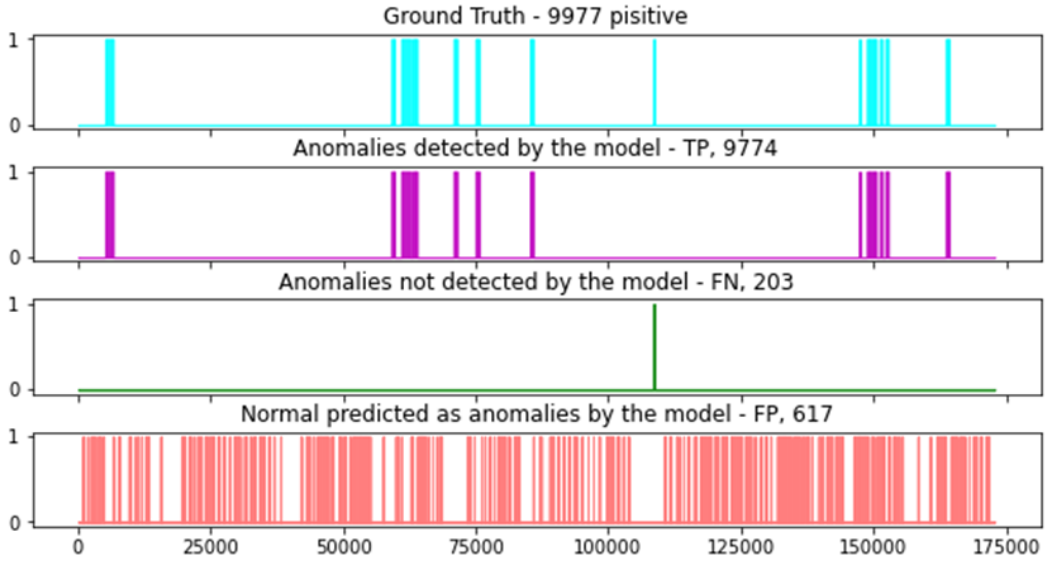

Figure 13: The predicted and ground truth labels for the WADI dataset with 82 filtered features using the DBDPC model. Top row: Ground truth labels of the WADI Dataset; Second row: the True Positives; Third row: the False Negatives; Fourth row: the False Positives.

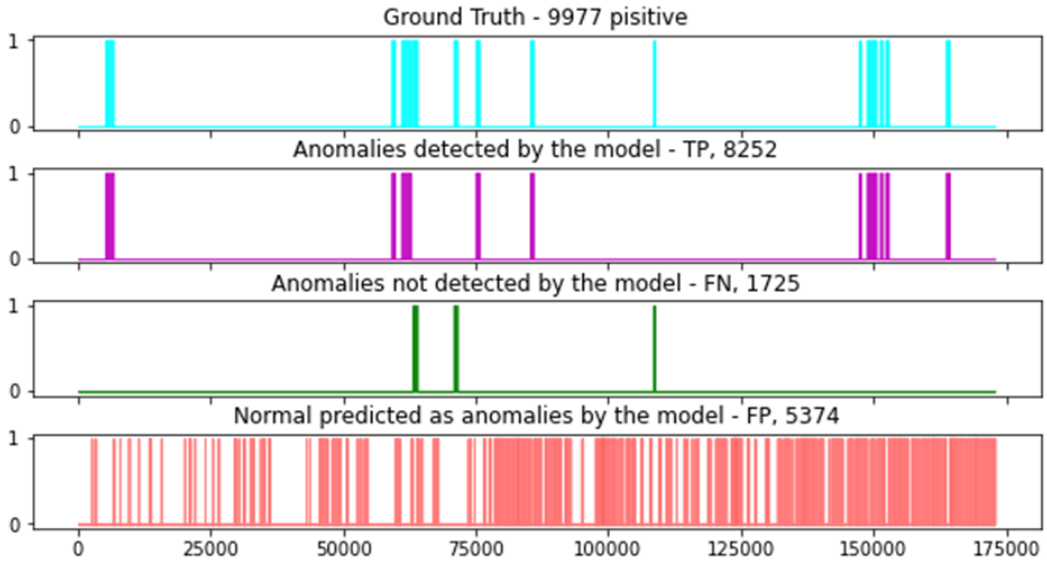

Figure 14: The predicted and ground truth labels for the WADI dataset with all the 123 features using the DBDP model. Top row: Ground truth labels of the WADI Dataset; Second row: the True Positives; Third row: the False Negatives; Fourth row: the False Positives.

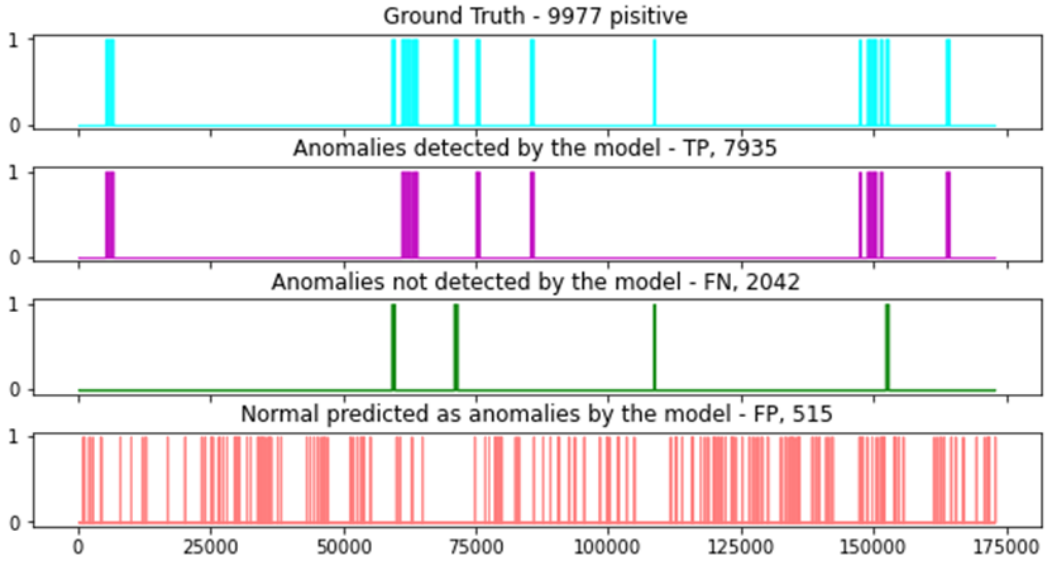

Figure 15: The predicted and ground truth labels for the WADI dataset with all the 123 features using the DBDPC model. Top row: Ground truth labels of the WADI Dataset; Second row: the True Positives; Third row: the False Negatives; Fourth row: the False Positives.