Transactions on Computer Science and Intelligent Systems Research
ISSN: 2960-1800, eISSN: 2960-2238 | Volume 7
TPCEE 2024

The Analysis of Time Series Forecasting Based on MLP Models

Ruyi Xie *

Faculty of Science and Technology, University of Macau, Macau, China

* Corresponding author: DC12838@um.edu.mo

Abstract. The field of time series forecasting (TSF) increasingly leverages deep learning architectures. This study examines the latest advancements in Multi-layer Perceptron (MLP)-based models for TSF, focusing on the MLP Mixer, MLP Encoder-Decoder, and Frequency-Domain MLP models. Each model's principles are analyzed to identify commonalities and potential areas for improvement. These innovative approaches enhance the ability to capture features and global dependencies in complex and lengthy time series data through techniques such as dimension transformation, channel independence, and residual connections. This results in more accurate and stable predictions. The MLP Mixer alternates between temporal and feature dimensions; the MLP Encoder-Decoder emphasizes intensive information processing during encoding and decoding; and the Frequency-Domain MLP focuses on processing data from the time domain to the frequency domain. Experimental results reveal that the Frequency-Domain MLP model exhibits superior performance, though the choice of lookback window significantly affects results. Future research will aim to optimize these models further, exploring innovations in mixing, residual structures, and large-scale models. Enhancing the generalization ability and computational efficiency of these models will advance the field of TSF.

Keywords: Time Series Forecasting; Multi-layer Perceptron; Frequency-Domain; Deep Learning.

1. Introduction

Many real-world problems, including those in finance, sales, environment, and healthcare, commonly use time series forecasting (TSF) to forecast the future based on historical data from the past. Traditional time series models, such as the Auto-Regressive Moving Average Model (ARIMA) and Vector Autoregression (VAR), are appropriate for univariate and multivariate time series modelling. However, because of the usual non-stationarity of real-world data, these traditional models have difficulty capturing complex patterns in the data. In recent years, deep learning models have gradually become mainstream methods due to their feature extraction capabilities, especially in capturing covariate information and non-linear relationships. However, they also face issues with overfitting and computational complexity. Therefore, the development of innovative methods has become a trend for the future. This paper reviews recent advances in TSF, focusing on the development of Multi-layer Perceptron (MLP) variants and their performance in practical applications.

Long-term time series forecasting (LTSF) based on deep learning has shown significant improvement in accuracy and efficiency in univariate and multivariate models. Multivariate TSF refers to the simultaneous prediction of future values for multiple interdependent variables, given their historical values and other relevant factors. Conversely, univariate TSF solely utilizes the historical data and covariate characteristics without considering the other historical data as inputs. Transformers have gradually demonstrated good scalability in multivariate TSF, because of their ability to model complex and lengthy sequence data between covariates. One such model, Autoformer [1], solved time complexity and high memory usage by learning each time point through a modular self-attention mechanism. However, Delinear used decomposition to get the temporal feature components and combined it with a linear layer to treat multivariate data as a collection of univariate sequences [2]. This linear model works better than most transformer methods. It inspired PatchTST to create the crucial concept of channel independence, which defines continuous time series as patch inputs to a self-attentive mechanism [3]. In addition, Cross former established the hierarchical encoder-decoder, emphasizing the need to take into account not only the temporal dependence but also the dependence

of the different variables [4]. Recently, iTransformer proposed the necessity of adding feedforward networks to capture multivariate correlations and learn non-linear properties using variable labelling [5].

This paper investigates the development of MLP models that have recently excelled in LTSF. The study begins by describing the role of MLP models in linear time series analysis, followed by a detailed examination of three MLP-based representation architectures: the MLP Mixer, the MLP Encoder-Decoder, and the Frequency-Domain MLP. These architectures demonstrate the ability to capture and decompose complex temporal patterns and effectively address the representational limitations present in MLP models. The paper also identifies common innovative ideas across these architectures, highlighting potential directions for future innovation and improvement. Performance comparisons are made using a classical long-term time series dataset, and the underlying causes of observed performance differences are investigated. The study concludes with a summary of findings, emphasizing the substantial feasibility and creativity of MLP-based models in time series analysis. This research aims to guide newcomers in the field by providing insights into groundbreaking research directions and advancements.

2. Methodology

2.1. Dataset Description and Preprocessing

This study evaluates the predictions using five popular multivariate datasets [1], widely used for LTSF benchmarking. Electricity Transformer Temperature (ETT) dataset is collected from two power transformers at different sampling rates, 15 minutes for ETTm1 and 1 hour for ETTh1, using seven variables for each timestamp. Every 10 minutes, the Weather dataset collects 21 columns of meteorological indicators. For 862 motorway lanes, the Traffic dataset records road occupancy per hour. The electricity dataset contains hourly records of power use for 321 consumers.

2.2. Proposed Approach

This study aims to discuss the characteristics of MLP Mixer (TSMixer [6]), MLP Encoder-Decoder (TiDE [7]), and Frequency-Domain MLP (FreTS [8]) in the context of LTSF. Fig. 1 illustrates the entire process. First, this paper will discuss the unique advantages and problems of MLP-based linear models, as well as some ideas and methods on how to deal with multivariate information, indicating the latest development direction. Then, three pioneering and representative MLP variant models will be characterized and illustrated in dimension transformation, channel independence, and residual connection. Subsequently, a few points for improvement are suggested. The MLP mixer could try to increase the number of layers in the feature mixing module to fuse features from different time scales. The MLP encoder-decoder could attempt to implement various residual connections within the residual block to eliminate extraneous input features. Frequency-domain MLPs could use the frequency domain to design cross-modal large models of time and context. This paper provides a general framework for future researchers to work on model refinement.

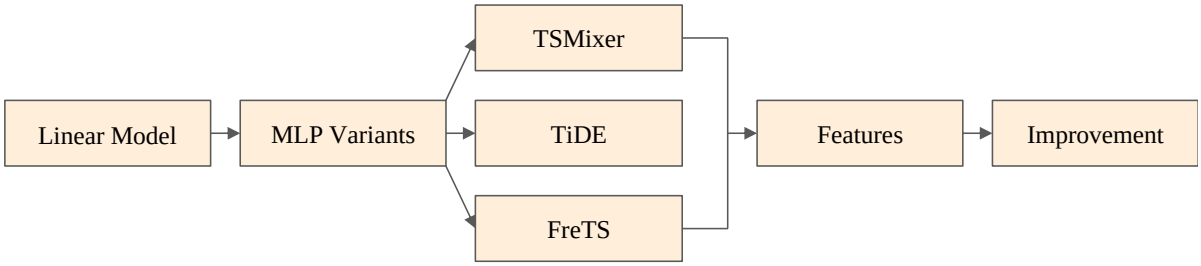

Figure 1. The pipeline of the study

2.3. Introduction of Linear Models

Linear models are superior to other complex architectures because they can learn an appropriate representation of the temporal dependence of univariate time series and capture smooth and complex periodic temporal patterns, effectively preventing model overfitting. Specifically, a linear combination of the preceding time steps determines the next time step instead of a linear combination of the functions of the preceding time steps, thereby preserving time step independence. However, the point-wise mapping method has difficulty resolving global covariate dependencies, consequently impacting the quality of the time series data. Real-world time series data exhibit significant volatility, so it is not enough to rely only on the time patterns observed in the past. It is important to first focus on the decomposition of multivariate time patterns to explore the dependencies between multiple variables and improve forecast reliability. Nowadays, the more popular approach is to extract the features of each univariate using simple independent channels and add residual connections from them. In addition, if the real-world dataset has additional auxiliary variables, including static and future features, this is also a factor that the study needs to be aware of. In conclusion, LTSF models show a trend towards simplicity, and this idea influences other research areas as well.

2.4. Introduction of Individual Models

2.4.1. TSMixer.

The model draws on the MLP-Mixers model in computer vision, applying MLP alternately in the time and feature domains by stacking linear models with nonlinear characteristics. When only considering historical data, input will go through the time mixing, feature mixing and temporal projection in order. In the time mixing process, the historical input matrix is normalized, transposed, fed into a time-domain MLP, and transposed again. The output is normalized and employs a feature-domain MLP in the feature mixing process. Two processes are connected with global residuals once. Following these mixing stages is the temporal projection process. The historical output matrix is obtained by transposing the earlier output and passing it through the fully connected layer. Time-domain MLP consists of a fully connected layer, an activation function, and a dropout layer, while feature-domain MLP consists of two fully connected layers, an activation function, and two dropout layers. In addition, when static and future features also need to be considered, the static and future feature matrices are first aligned to the same shape as the historical output matrix and then operated through multiple mixing layers. Finally, the final output matrix is obtained by the fully connected layer. Fig. 2 shows the framework of TSMixer without auxiliary information.

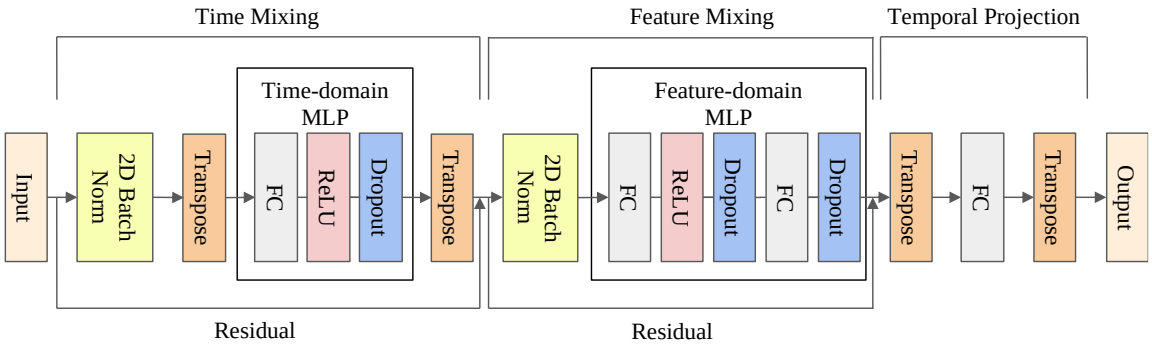

Figure 2. The framework of TSMixer without auxiliary information

2.4.2. TiDE.

The model draws on the encoding and decoding model and embeds it through a dense linear model and embeds it through a dense linear model. The core part is the residual block, which consists of a residual connection and a linear connection. The latter consists of an activation function, a fully connected layer, and a dropout layer. Both connections will pass through a normalization layer. In terms of process, firstly, the dynamic covariates are converted into a low-dimensional space to get

the predicted features in the prediction length by feature projection. Then, predicted features are stacked with historical features and static features. These are encoded through a dense encoder (DE) that maps to the feature vectors in the prediction length using multiple residual blocks. The result goes through a dense decoder (DD) that maps the encoded vector to the decoded vector in the prediction length using multiple residual blocks. Finally, the temporal decoder (TD) combines the decoded vectors with the predicted features in the prediction length to form the prediction output. In addition, there is an external linear residual connection between the historical features and the prediction output. Fig. 3 shows the framework of TiDE.

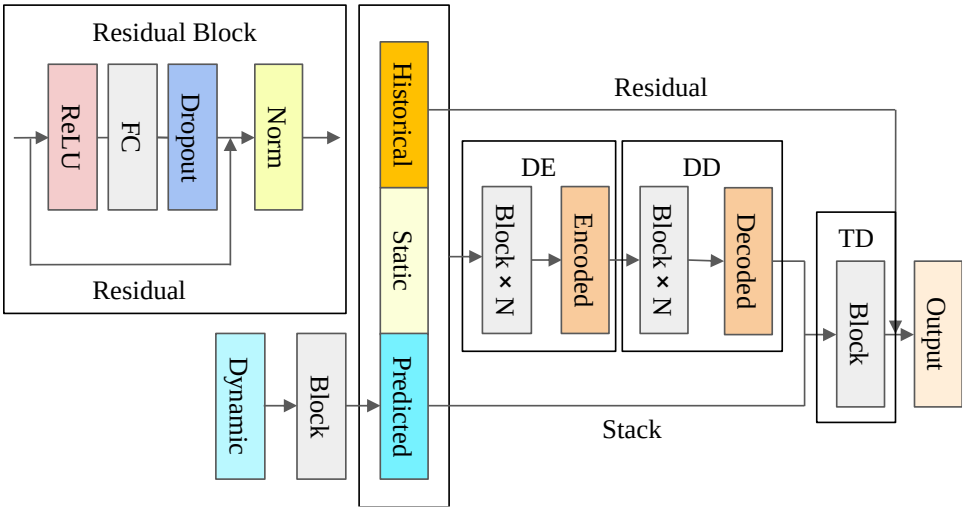

Figure 3. The framework of TiDE

2.4.3. FreTS.

The model draws on a frequency learning architecture and domain transformation. Historical input is divided into channel dimensions and time dimensions. A three-dimensional tensor is obtained by multiplying the historical input with a learnable weight vector. Then, the tensor passes through the frequency channel learner part, which performs a Discrete Fourier Transform (DFT) using the channel dimensions to transform the tensor into the frequency domain with real and imaginary parts. The result goes through a Frequency-Domain MLP sharing the same weights between each time step in the channel dimensions. Finally, the result transforms the updated frequency domain representation through an Inverse Discrete Fourier Transform (IDFT) back to the time domain. The next part is to pass through the frequency temporal learner. The steps are similar to the frequency channel learner, except using the time dimensions. After learning channel and time dependencies, they are used to predict through a two-layer Feed-forward Network (FFN) to obtain prediction values for future time steps. The Frequency-Domain MLP contains several hidden layers. Each layer applies an activation function after linearly transforming the real and imaginary parts using weight matrices with bias terms. Fig. 4 shows the framework of FreTS.

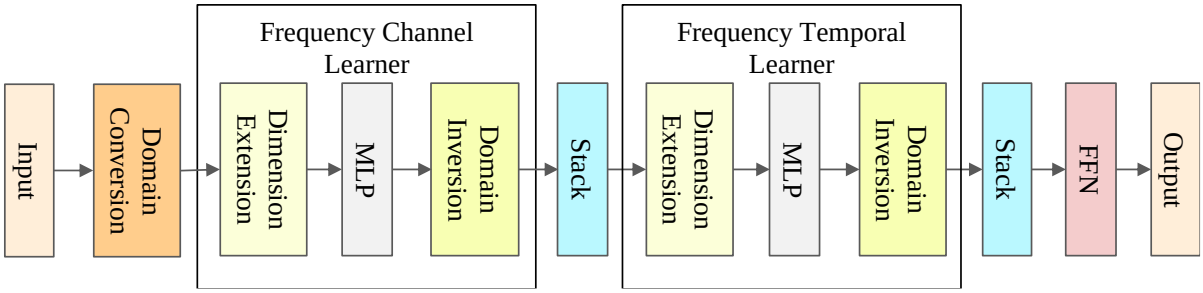

Figure 4. The framework of FreTS

3. Result and Discussion

3.1. Features and Improvement

Apart from the fact that the three models utilize a simple and efficient MLP as their core, they all involve a special treatment of dimensionality. TSMixer extends the feature dimensions, TiDE implements a dimensionality reduction process, and FreTS transforms into a frequency domain dimension. These transformations have the advantage of simplifying the handling of complex features and improving computational efficiency. In addition, the channel independence processing based on a linear model can effectively separate the feature components and reduce the interference between the features. TSMixer applies independent MLP channels alternately in time and feature dimensions. TiDE uses independent channels for feature projection and encoding processes. FreTS creates independent frequency channels that deal with the low-frequency and high-frequency components. Besides, the residual connection is also a crucial component in the design. TSMixer applies residual connections between each time mixing layer and feature mixing layer so that the model can analyze covariate information effectively. Meanwhile, the stacked residual block in TiDE is involved in the feature projection, encoding, and decoding phases, which serve as an information integration and transformation function. Although residual connections are not mentioned in FreTS, the cross-layer processing in the frequency and time domains can be treated as a kind of residual connection, allowing the model to transfer and merge information efficiently between the two domains.

This paper proposes several extensions to improve existing models by enhancing their feature extraction and integration capabilities. For TSMixer, it suggests incorporating multi-scale temporal feature mixing. By using different time window lengths, such as short, medium, and long-term windows, and processing data through separate multi-layer MLPs, features can be extracted at each time scale. These features can then be combined through fully connected layers, using linear regression and Fourier transform to fit trend and seasonal components. Building on this, TiDE can be enhanced by introducing a weighted residual connection, where input features are assigned weight parameters. The outputs are weighted and fused via the FFN, resulting in more accurate feature fusion and prediction. Additionally, for FreTS, cross-modal adapters can be designed for large models. Frequency domain features are extracted from time series data, while contextual features are derived from pre-trained models. These features are then mapped to the same embedding space using MLP, aligned, and stacked to form final multimodal features. By integrating these approaches, the models can achieve better accuracy and efficiency in temporal feature extraction, residual connection utilization, and multimodal feature integration, thereby providing a robust framework for improved performance in various applications.

3.2. Discussion

Table 1 presents the performance of three representative models on each dataset, evaluated using Mean Squared Error (MSE) [1]. All models were trained with MSE as the training loss. Experiments were conducted on a standard normalized dataset to ensure consistency with previous studies. Four different prediction lengths—96, 192, 336, and 720—were tested on each dataset. The lookback window sizes varied depending on the model, with the best results being 96, 192, 336, 512, 672, and 720. The results reveal that FreTS outperforms the other two models. This superiority is attributed to FreTS's ability to capture dynamic evolution through the temporal domain and identify periodic patterns through the frequency domain, effectively leveraging the strengths of both domains. Additionally, all three models exhibited decreased performance as the lookback window size increased. This decline could be due to the dilution of relevant temporal information by noise, increasing the risk of overfitting.

Table 1. Evaluation results

Dataset	Prediction Lengths	TSMixer	TiDE	FreTS
ETTh1	96	0.361	0.375	0.174
	192	0.404	0.412	0.182
	336	0.420	0.435	0.192
	720	0.463	0.454	0.216
ETTm1	96	0.285	0.306	0.154
	192	0.327	0.335	0.166
	336	0.356	0.364	0.178
	720	0.419	0.413	0.192
Weather	96	0.145	0.166	0.142
	192	0.191	0.209	0.162
	336	0.242	0.254	0.18
	720	0.320	0.313	0.198
Electricity	96	0.131	0.132	0.13
	192	0.151	0.147	0.128
	336	0.161	0.161	0.144
	720	0.197	0.196	0.158
Traffic	96	0.376	0.336	0.076
	192	0.397	0.346	0.076
	336	0.413	0.355	0.078
	720	0.444	0.386	-

4. Conclusion

This study introduces three innovative approaches based on MLP models for LTSF: the MLP Mixer, the MLP Encoder-Decoder, and the Frequency-Domain MLP. These approaches not only leverage the strengths of linear models in handling time dependencies but also address challenges related to global covariates. By integrating dimension transformation, channel independence, and residual connections into linear models, these approaches enhance the models' ability to comprehend and process complex time series data. They improve the accuracy and richness of feature representations by separating and fusing information across different dimensions, while maintaining the stability of model training and the reliability of prediction results. Experimental results indicate that the Frequency-Domain MLP outperforms the other two models. However, an increased lookback window negatively impacts performance. Future research should focus on developing models with improved performance by exploring multi-scale temporal feature mixing, variant residual blocks, and cross-modal large models to address more complex and diverse TSF tasks.

References

[1] Wu H. Xu J. Wang J. et al. Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting. Advances in neural information processing systems, 2021, 34: 22419 - 22430.

[2] Zeng A. Chen M. Zhang L. et al. Are transformers effective for time series forecasting? Proceedings of the AAAI conference on artificial intelligence, 2023, 37 (9): 11121 - 11128.

[3] Nie Y. Nguyen N.H. Sinthong P. et al. A time series is worth 64 words: Long-term forecasting with transformers. 2022, arXiv preprint: 2211. 14730.

[4] Zhang Y. Yan J. Crossformer: Transformer utilizing cross-dimension dependency for multivariate time series forecasting. The eleventh international conference on learning representations, 2023.

[5] Liu Y. Hu T. Zhang H. et al. itransformer: Inverted transformers are effective for time series forecasting. 2023, arXiv preprint arXiv: 2310.06625.

[6] Chen S.A. Li C.L. Yoder N. et al. Tsmixer: An all-mlp architecture for time series forecasting. 2023, arXiv preprint: 2303. 06053.

[7] Das A. Kong W. Leach A. et al. Long-term forecasting with tide: Time-series dense encoder. 2023, arXiv preprint: 2304.08424.

[8] Yi K. Zhang Q. Fan W. et al. Frequency-domain MLPs are more effective learners in time series forecasting. Advances in Neural Information Processing Systems, 2024, 36.