Science and Technology and Social Development Proceedings Series

Article

2024 International Conference on Education, Economics, Management, and Social Sciences (EMSS 2024)

ESG Performance Prediction and Driver Factor Mining for Listed Companies Based on Machine Learning: A Multi-Source Heterogeneous Data Fusion Analysis

Weiyan Tan [1,*]

[1] Guangdong Veterans Service Center, Guangzhou, 510000, China
* Correspondence: Weiyan Tan, Guangdong Veterans Service Center, Guangzhou, 510000, China

Abstract: With the acceleration of global economic integration and the growing focus on sustainable development, Environmental, Social, and Governance (ESG) factors have become key standards for evaluating a company's long-term value and risk. However, accurately measuring the ESG performance of listed companies and identifying the underlying driving factors remains a significant challenge. This paper proposes a Transformer-based multi-source heterogeneous data fusion model, MSformer, which analyzes diverse data, including financial reports, news, social media comments, and government announcements. It categorizes the data into three types: time-series structured data, time-series structured mapped data, and textual data. The model enhances feature extraction using the Spatial Frequency-coordinated Attention Mechanism (SFHA) and employs Support Vector Regression (SVR) for prediction. Experimental results show that MSformer outperforms other advanced models, achieving an outstanding 87.4% multi-class accuracy and 0.517 average prediction error, proving its effectiveness and advantage in ESG prediction.

Keywords: ESG; multi-source heterogeneous; transformer; deep learning; SFHA

Published: 31 December 2024

1. Introduction

With the acceleration of global economic integration and the increasing focus on sustainable development, Environmental, Social, and Governance (ESG) factors have become key standards for assessing a company's long-term value and risk. Investors, regulators, and the public are placing greater emphasis on corporate performance in these areas. However, accurately measuring the ESG performance of listed companies and identifying the key drivers behind their performance remains a significant challenge in both academic research and practice.

Traditional ESG evaluation methods, primarily relying on surveys and publicly disclosed reports, are costly, lack timeliness, and fail to capture multidimensional information comprehensively. Recently, the rapid development of big data technologies and machine learning algorithms has provided new approaches to tackle these challenges. By analyzing data from diverse sources — such as financial reports, news, social media comments, and government announcements — and integrating it through advanced data processing techniques, it is now possible to predict the future ESG performance of listed companies more accurately and uncover the underlying factors that drive it.

To address the limitations of existing ESG prediction methods, this study designs a Multi-Source Fusion ESG prediction model based on Transformer，named MSformer. Specifically, we utilize three types of data: time-series structured data (D1), time-series structured mapping data (D2), and textual data (D3). D1 consists of annual company-specific indicators, such as historical ratings, composite scores, and total market value. D2 represents the transformation of D1 using a two-dimensional interaction mapper, while D3 includes ESG-related news reports retrieved via keyword searches.

In detail, D1 is processed through two Transformer blocks and an LSTM block, yielding a K dimensional feature vector. D3 is embedded using a large language model and mapped into sentiment and discriminative spaces, generating a 2 × K-dimensional vector. Meanwhile, D2 undergoes processing by a two-dimensional interaction mapper, producing an H × K dimensional vector. These three vectors are concatenated along the first dimension, forming an input sample of size (H + 3) × K. To further enhance feature extraction, we propose a Spatial-Frequency Harmonized Attention (SFHA) mechanism replace the original attention in Transformer to capture spatial-frequency and interaction features from the input. Finally, a support vector regression (SVR) model is employed to generate the ESG prediction results.

2. Related Works

2.1. ESG Factor Analysis

In recent years, empirical research on ESG (Environmental, Social, and Governance) investment in China has expanded, with studies demonstrating that higher ESG ratings are associated with lower market risks and improved financial performance. ESG-based investment strategies have been explored in both equity and bond markets, with findings indicating that portfolios incorporating ESG factors tend to outperform those with lower ESG ratings. Some studies have integrated ESG factors into multi-factor models using machine learning techniques, such as XGBoost, to enhance stock selection strategies.

While international research on ESG investment is often tailored to specific markets, domestic studies on integrating ESG with machine learning for quantitative investment remain limited. Before 2020, the lack of comprehensive ESG disclosures in China posed a challenge for applying machine learning models. However, recent policy advancements and the availability of expanded datasets from sources such as Wind, China Securities Index, and SynTao Green Finance — covering over 4,000 A-share companies — have created new opportunities for research. Despite this progress, few studies have systematically combined ESG factors with machine learning algorithms for quantitative investment strategies. This study addresses this gap by empirically analyzing ESG-based stock selection using Boosting ensemble algorithms.

2.2. Time Series Forecasting Model

Time series analysis has been widely applied in various domains, including signal processing, financial analysis, and biomedical research, attracting significant attention to time series forecasting. Early deep learning models primarily relied on Long Short-Term Memory (LSTM) and Recurrent Neural Networks (RNN), which utilize their recurrent structures to retain past information and incorporate it into current computations. In contrast, the Transformer model processes sequential data without recurrence by leveraging a self-attention mechanism, enabling it to capture dependencies across arbitrary positions within an input sequence. This characteristic allows the Transformer to efficiently handle long time series and facilitates large-scale parallel computation. In recent years, numerous Transformer-based advancements have emerged, such as i-Transformer [1], which enhances variable embedding by transposing input representations; TimesNet [2], which employs a Fourier transform-based decomposition of input tensors into a two-dimensional structure; and PatchTST [3], which segments time series into patches for improved processing. In this study, we adopt the Transformer as the backbone model and leverage

multi-source heterogeneous data to extract the spatial-frequency and spatial-interaction features of ESG data, aiming to enhance predictive performance.

3. Methodology

3.1. Multi-Source Heterogeneous Data Fusion

We categorize the data into three types: time-series structured data (D1), time-series structured mapping data (D2), and textual data (D3). The D1 dataset is extracted from ESG performance reports of various companies, including Wind ESG ratings and scores for listed companies from 2018 to 2023, as well as Menglang ESG ratings for listed companies from 2014 to 2023. The D2 dataset is obtained through a two-dimensional interaction mapper, which will be introduced in B. Two-Dimensional Interaction Mapper. The D3 dataset consists of news texts acquired from sources such as Weibo, news websites, company official websites, and various statistical center websites, based on keywords related to the target listed companies. The overall architecture of Multi-Source Heterogeneous Data Fusion is shown in Figure 1.

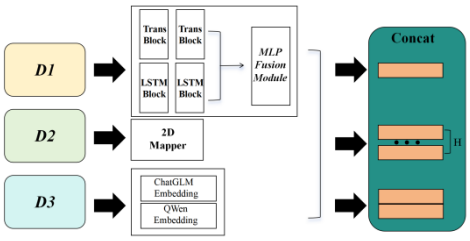

Figure 1. The pipline of Multi-Source Heterogeneous Data Fusion.

For D1, the data is processed sequentially through two Transformer blocks followed by two LSTM blocks:

$$f_{Trans} = Transformer(x_{D1}) \tag{1}$$

$$f_{LSTM} = LSTM(x_{D1}) \tag{2}$$

Among them, f_{Trans} represents the feature space obtained after processing with the Transformer, while f_{LSTM} denotes the feature space derived from the LSTM processing. Next, these two feature vectors are concatenated and fed into a feature fusion module constructed with an MLP:

$$f_{D1} = MLP - 2layers([f_{Trans}; f_{LSTM}]) \tag{3}$$

Where $f_{D1} \in R^{1\times K}$ represents the final encoded representation of the D1 data. Next, all D3 text data is fed into the first embedding layer of the ChatGLM model to obtain the encoded representations. These representations are then passed through a fully connected layer to scale the dimensionality to K:

$$f_{GLM} = Embed_{GLM}(x_{D3}) \tag{4}$$

$$f_{D3,1} = Linear(f_{GLM}) \tag{5}$$

Where $f_{D3,1} \in R^{1\times K}$ is the first-level feature. Similarly, the raw D3 data is fed into the first embedding layer of the QWen model to obtain the second-level feature representation of x_{D3} $f_{D3,2} \in R^{1\times K}$:

$$f_{QWen} = Embed_{QW}(x_{D3}) \tag{6}$$

$$f_{D3,2} = Linear(f_{QWen}) \tag{7}$$

Finally, the two levels of feature representations are concatenated along the first dimension of the tensor to obtain the final feature representation of D3:

$$f_{D3} = Concat(f_{D3,1}, f_{D3,2}) \tag{8}$$

Where $f_{D3} \in R^{2\times K}$ is the final feature space of D3. Finally, the D2 data can be represented as:

$$f_{D2} = 2DMapper(x_{D2}) \tag{9}$$

Where $f_{D2} \in R^{H\times K}$ is the final feature space of D2. A detailed introduction to the 2D Mapper can be found in the next subsection. By concatenating the final representations of the three types of data along the first dimension, we obtain the final representation of each input sample:

$$x_{in} = Concat(f_{D1}, f_{D3}, f_{D2}) \tag{10}$$

Where $x_{in} \in R^{(H+3)\times K}$ represents the final representation of the input tensor.

3.2. Two-Dimensional Interaction Mapper

The 2DMapper is composed of multiple layers with learnable parameters that perform various numerical transformations. Specifically, it includes four types of numerical transformation layers: fully connected layers, Fourier enhancement layers, logarithmic transformation layers, and periodic activation layers. The fully connected layer is a simple linear layer, and its mathematical expression is as follows:

$$f_{L1} = \sigma(W_1 x_{in} + b) \tag{11}$$

Where W is the parameter matrix and b is the bias. σ is the ReLU function. The logarithmic transformation layer applies a logarithmic transformation to each input value and performs residual connections:

$$f_{L2} = log_b(x_{in}) + x_{in} \tag{12}$$

The Fourier enhancement layer is based on the idea of Fourier transform. It applies a cosine activation function to the first half of the values and a sine activation function to the second half of the values. Afterward, the GLU activation function is used to fuse the features:

$$f_{L3} = \sigma(cos(W_{cos} x_{in/2,first} + b) + sin(W_{sin} x_{in/2,second} + b)) \tag{13}$$

$x_{in/2,first}$ represents the values from the first half of the input. $x_{in/2,second}$ represents the second half of the input. The W_{cos} and W_{sin} represent the parameter matrices used for activation with the cosine and sine functions, respectively. Finally, the periodic activation layers refer to replacing the activation function of a fully connected layer with a cosine function:

$$f_{L4} = cos(W_2 x_{in} + b) \tag{14}$$

The four transformed feature vectors are pairwise concatenated and then passed through a multi-layer perceptron (MLP) to reduce their dimensionality and scale them to K-dimensional features, generating interaction features $f_{int,i}$.

$$f_{int,i} = MLP([f_{Lk}; f_{Lj}]) \tag{15}$$

Additionally, the individual transformed features are also passed through an MLP to scale them to K-dimensional vectors. Finally, all interaction features and the four separately transformed features are concatenated together throught the first dimension:

$$f_{D2} = Concat(f_{L1}, \ldots, f_{L4}, f_{int,1}, \ldots, f_{int,i}) \quad (16)$$

Where $Concat(\cdot)$ denoted the operation of concatenation along the first dimension. The dimension of f_{D2} is $R^{H \times K}$.

3.3. MSformer

The MSformer is a specialized architecture that we have designed, based on the Spatial-Frequency Harmonized Attention (SFHA) mechanism, as shown in Figure 2.

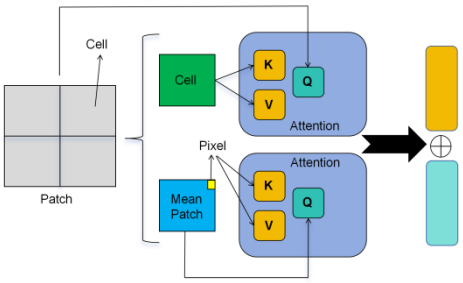

Figure 2. The pipline of MSformer.

Specifically, we first divide each input into n patches of equal size:

$$x_{in} -> n * Patch \in R^{(H/n) \times (K/n)} \quad (17)$$

For the low-frequency branch of SFHA, we first map the entire patch into a query vector Q . Then, the patch is further divided into ww equally sized cells, where each cell is individually mapped into a key vector K and a value vector V. Next, the cross-attention is computed between Q and the K and V vectors from the w cells. The output is then concatenated along the final dimension and serves as the Low-frequency Spatial (LfS) output:

$$f_{Lfs} = Concat(Attention(f_{patch}, f_{cell,i})) \quad (18)$$

f_{Lfs} represents the low-frequency output feature space, Attention represents the cross-attention operation, and $f_{cell,i}$ represents the mapping input of the i-th cell. For the high-frequency branch, the average pooling mapping value of each patch is used as Q, and the pixel value mapping is used as K and V. The cross-attention calculation is performed for each one and concatenated to produce the High-frequency Space output:

$$f_{Hfs} = Concat(Attention(f_{meanpatch,i}, f_{pixel,j})) \quad (19)$$

Where $f_{meanpatch,i}$ and $f_{pixel,j}$ represent the average pooling of the i-th patch and the mapping of the j-th pixel, respectively. Finally, the high-frequency and low-frequency outputs are concatenated along the last dimension, fused through an MLP to combine features, and then passed through a support vector machine to output the prediction results:

$$y_{pre} = SVR(MLP([f_{Lfs}; f_{Hfs}])) \quad (20)$$

The model's training is based on the MAE loss function

4. Experiments

In order to fully validate the effectiveness of MSformer and our designed multi-source data heterogeneous method, we designed three experiments. First, a comparison experiment was conducted, comparing with six advanced prediction models. Second, an

ablation experiment on multi-source data was performed, where different feature combinations were gradually eliminated to verify the rationality of each feature selection. Finally, clustering was performed on the feature inputs before SVR, and the driving factors of ESG were elaborated.

4.1. Comparison Experiment

We selected six models for comparison: Transformer [4], RNN [5], LSTM [6], CNN-LSTM, iTransformer, and GRU [7], as shown in Table 1. We evaluate the model's performance using two metrics: Five-class Accuracy, which represents the classification result by mapping the predicted value to [0,1] and dividing it into five equal parts, and the Average Prediction Error. It can be observed that our model achieves 87.4% and 0.517 for the two metrics, respectively. Compared to the worst-performing RNN, which achieves 78.6% and 1.990, our model improves by 8.8% and 1.473. Compared to the best-performing model, iTransformer, it also improves by 2.8% and 0.325. This demonstrates the superiority of our model in ESG prediction and validates the rationality of our designed multi-source heterogeneous approach.

Table 1. In the Comparative Experiment Results. The Best Results are Highlighted in Bold.

Model	Metric	
	Five-class Accuracy	**Average Error**
Transformer	81.5%	1.153
RNN	78.6%	1.990
LSTM	81.1%	1.672
CNN-LSTM	83.1%	0.927
iTransformer	84.6%	0.842
GRU	79.2%	1.833
MSformer(Ours)	87.4%	0.517

4.2. Ablation Experiment

For the feature space of the three types of data, we randomly removed one or two of them, denoted as w/o, and recorded the model accuracy based on different combinations of data types, as shown in Table 2. It can be observed that removing the D2 data features has the greatest impact, indicating that the 2DMapper can effectively extract deep semantic features from the data. The second most significant impact is from the D3 data, which also demonstrates the powerful representational capability of the embedding layer of large language models. Additionally, simultaneously removing both D2 and D3 results in a decrease of 6.2% in accuracy and an increase of 1.247 in the average prediction error.

Table 2. Ablation Experiment Results.

Model	Metric	
	Five-class Accuracy	**Average Error**
w/o f_{D1}	86.7%	0.872
w/o f_{D2}	83.5%	1.392
w/o f_{D3}	85.9%	1.016
w/o $f_{D1,2}$	82.4%	1.585
w/o $f_{D1,3}$	84.5%	1.506
w/o $f_{D2,3}$	81.2%	1.764
MSformer(Ours)	87.4%	0.517

4.3. Driving Factors of ESG

Finally, we performed clustering on the input features of SVR and used explainable machine learning algorithms to extract the feature importance from the clustering results. The results are shown in Figure 3. Based on feature ranking, we identified several key driving factors for ESG performance. These include environmental indicators like carbon emissions, energy efficiency, and renewable energy use; social responsibility factors such as employee satisfaction, workplace safety, and community engagement; corporate governance aspects like board independence and executive compensation transparency. Financial health indicators such as profitability and cash flow stability also play a role, alongside market and industry characteristics, policy compliance, technological innovation, and public opinion. Additionally, the impact of regulations, clean tech R&D, and media exposure further influence ESG outcomes.

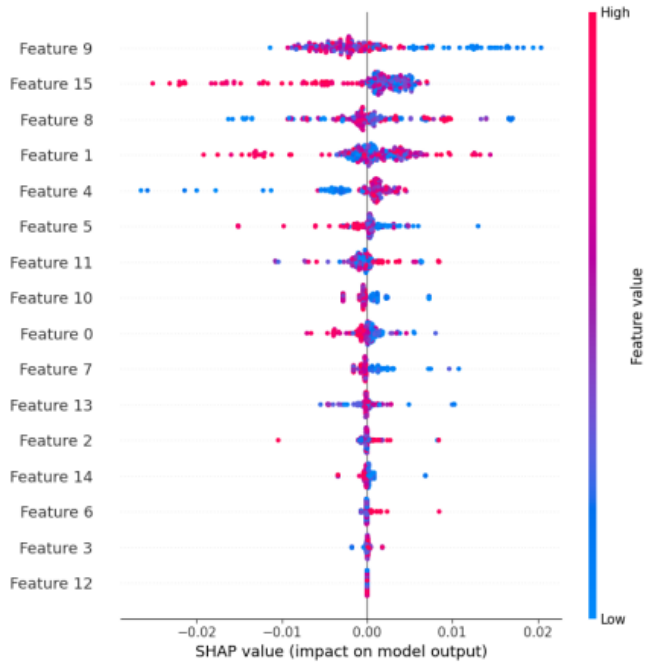

Figure 3. The SHAP of MSformer.

5. Conclusion

This paper presents a Transformer-based multi-source heterogeneous data fusion model, MSformer, aimed at improving the accuracy of ESG performance prediction for listed companies and uncovering driving factors. By integrating time-series structured data, mapped data, and textual data into a two-dimensional tensor, the model employs a Spatial Frequency-coordinated Attention Mechanism (SFHA) to enhance feature extraction, and utilizes Support Vector Regression (SVR) for prediction. Experimental results show that MSformer outperforms other advanced models, achieving 87.4% in multi-class accuracy and 0.517 in average prediction error, confirming its effectiveness. Additionally, key driving factors such as environmental indicators, social responsibility, and corporate governance were identified, offering valuable insights into ESG performance. These findings provide strong support for corporate evaluation and investment decisions.

References

1. Y. Liu *et al.*, "iTransformer: Inverted transformers are effective for time series forecasting," *arXiv preprint arXiv:2310.06625*, 2023, doi: 10.48550/arXiv.2310.06625.
2. H. Wu *et al.*, "TimesNet: Temporal 2D-variation modeling for general time series analysis," *arXiv preprint arXiv:2210.02186*, 2022, doi: 10.48550/arXiv.2210.02186.

3. H. Yi *et al.*, "PatchesNet: PatchTST-based multi-scale network security situation prediction," *Knowl.-Based Syst.*, vol. 299, p. 112037, 2024, doi: 10.1016/j.knosys.2024.112037.
4. K. Han *et al.*, "A survey on vision transformer," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 1, pp. 87-110, 2022, doi: 10.1109/TPAMI.2022.3152247
5. S. Li *et al.*, "Independently recurrent neural network (IndRNN): Building a longer and deeper RNN," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2018, doi: 10.1109/CVPR.2018.00572.
6. R. C. Staudemeyer and E. R. Morris, "Understanding LSTM—A tutorial into long short-term memory recurrent neural networks," *arXiv preprint arXiv:1909.09586*, 2019, doi: 10.48550/arXiv.1909.09586.
7. R. Rana, "Gated recurrent unit (GRU) for emotion classification from noisy speech," *arXiv preprint arXiv:1612.07778*, 2016, 10.48550/arXiv.1612.07778.