


Lightweight Design and Implementation of Machine Learning Models in Time Series Forecasting

Hongwen Pan 

Xi'An Mingde Institute of Technology, Xi'An, 710199, China

Received: 13 Jun 2025

Revised: 14 Jun 2025

Accepted: 18 Jun 2025

Published: 21 Jun 2025

Copyright: © 2025 by the authors. Licensee ISTAER.

This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).



Abstract: With the surge in data volume and the continuous growth of computing requirements, the application of machine learning in time series forecasting faces the challenges of computing resource consumption and real-time requirements. In order to meet this demand, lightweight design has become a key technology to improve the efficiency of time series forecasting models. This paper deeply explores the lightweight design and implementation of machine learning models in time series forecasting, focusing on the application of lightweight technologies such as pruning, quantization, distillation, miniaturized neural networks, and hardware acceleration. By optimizing the network structure and reducing computing resource consumption, the lightweight model can not only improve real-time performance and inference speed, but also ensure high prediction accuracy. Studies have shown that lightweight technology has broad application prospects in fields such as finance, meteorology, and retail. This paper also proposes future research directions for lightweight design, including adaptive lightweight models, the combination of quantum computing and artificial intelligence, and efficient prediction on low-power devices. Finally, this paper looks forward to the optimization and application promotion of lightweight models. It is expected that with the development of technology, lightweight design will be widely used in more emerging fields.

Keywords: Time series prediction; Machine learning; Lightweight design; Model compression; Real-time prediction

1 INTRODUCTION

With the development of big data and intelligent technology, the application demand of time series prediction in many fields is growing, especially in the financial, medical, energy and manufacturing industries. In the financial industry, time series prediction is widely used for trend analysis of stock, foreign exchange, futures and other markets; in the medical field, time series prediction helps monitor the health of patients and predict the development trend of diseases; in the energy industry, load prediction in smart grids and power generation prediction of renewable energy are also important applications of time series prediction; and in the manufacturing industry, inventory management and production plan optimization also rely on accurate time series prediction [1]. The common characteristics of these fields are huge data volume, rapid data changes and strong seasonality or periodicity. Therefore, how to accurately predict future trends has become the key to improving industry efficiency and reducing costs.

However, although traditional statistical methods such as ARIMA and seasonal exponential smoothing methods have been successful in some simple scenarios, these methods perform poorly when dealing with complex and nonlinear time series data. In contrast,

machine learning methods, especially deep learning models such as long short-term memory networks (LSTM) and gated recurrent units (GRU), have become mainstream tools for dealing with complex time series prediction tasks. Machine learning models can better capture the nonlinear characteristics and long-term dependencies of data through their powerful modeling capabilities, thereby improving the accuracy of predictions [2]. However, these deep learning models often require a large amount of computing resources and storage space, which has become a significant bottleneck in practical applications with limited resources, especially in edge devices and real-time applications.

This study aims to design and implement lightweight machine learning models suitable for practical application scenarios. With the continuous update of computing power and hardware devices, the demand for lightweight models has become more urgent. Lightweight design not only requires the model to reduce the consumption of computing resources while ensuring sufficient prediction accuracy, but also requires the model to be deployed on different platforms. Especially in scenarios such as edge computing and mobile devices where there are strict restrictions on computing and storage resources, how to find the best balance between model size and prediction accuracy has become the core goal of this study [3].

In practical applications, especially in mobile devices and embedded systems, computing resources are usually very limited due to hardware performance limitations, which requires the design of lighter, faster and more efficient models. Time series data usually contains a lot of information and the data changes quickly. Therefore, how to deal with fast-changing dynamic data, handle high-frequency prediction tasks, and how to reduce model calculation delays and resource consumption while ensuring real-time performance have become the main challenges in actual deployment. In the face of these challenges, this study will explore lightweight model designs suitable for different scenarios and propose specific solutions to promote the widespread application of machine learning models in actual production environments.

2 APPLICATION BACKGROUND AND DEMAND OF TIME SERIES PREDICTION

Time series prediction is widely used in various industries and has become an important tool to promote intelligent decision-making. Especially in the financial field, time series prediction is widely used to predict price fluctuations in stocks, foreign exchange, futures and other markets. The complexity and uncertainty of the financial market make accurate predictions a core requirement in financial analysis and risk management. By analyzing the trends, fluctuations and cycles of historical data, machine learning models can capture the potential laws of the market and help investors make more accurate decisions. In addition, in the field of meteorology, time series prediction also occupies an important position. Weather forecasts and climate change predictions rely on the processing and analysis of large-scale meteorological data [4]. The accuracy of predictions is directly related to disaster warning, resource allocation and the operation of daily life. With the intensification of climate change, the accuracy and real-time requirements of meteorological forecasts are increasing, and machine learning models have shown great potential in this regard.

In addition, in industries such as retail and e-commerce, sales forecasting and inventory management are one of the important applications of time series prediction. By analyzing historical sales data and predicting future market demand, enterprises can optimize inventory management, reduce excess or shortages, and improve the efficiency and response speed of the supply chain. With the diversification of e-commerce platforms and retail formats, the amount of data faced by demand forecasting is growing exponentially, and the computational complexity and real-time performance of the model have become challenges that cannot be

ignored. Therefore, how to design and implement efficient time series prediction models has become the key to improving industry competitiveness [5].

However, although time series prediction has shown broad application prospects in these fields, it still faces many challenges in practical applications. First, the diversity and complexity of time series data make modeling and prediction more difficult. The data is not only seasonal and trendy, but also contains a large amount of noise and outliers, which pose challenges to the stability and prediction accuracy of the model. Traditional time series analysis methods often have difficulty in effectively processing these complex features. Although machine learning methods can better adapt to these changes, they also require the model to have strong computing power and good generalization ability.

Secondly, real-time requirements are an important challenge in many current applications. In high-frequency trading in financial markets, meteorological warning systems, and dynamic inventory management of e-commerce platforms, prediction models need to process and respond to large amounts of data in a very short time [6]. Low latency and fast response are essential for real-time decision-making, which puts higher requirements on the computing speed and deployment platform of the model.

Finally, with the popularity of the Internet of Things, edge computing, and mobile devices, the limitation of computing resources has become a bottleneck for many time series prediction applications. Edge computing devices and mobile devices usually have limited processing power and storage space, which makes it impossible to directly apply traditional deep learning models to these devices. How to achieve efficient and high-precision time series prediction in a resource-constrained environment has become a major problem in current research and application. Especially in real-time applications that need to run for a long time, lightweight design is particularly important. Only by optimizing the computing and storage overhead of the model can efficient time series prediction be ensured under limited resources.

3 CORE TECHNOLOGIES FOR LIGHTWEIGHT MACHINE LEARNING MODELS

Lightweight design is an important research direction for machine learning model optimization in recent years. It aims to improve its operating efficiency in low-resource environments by reducing the computing overhead, storage requirements, and power consumption of the model. In practical applications such as time series prediction, lightweight design not only needs to ensure high prediction accuracy, but also requires better real-time and deployability in different hardware environments. With the widespread application of the Internet of Things, edge computing, and mobile devices, traditional deep learning models are often difficult to adapt to these low-resource devices due to their huge computing resource consumption and storage requirements [7]. Therefore, designing lightweight models has become one of the core technologies for achieving efficient prediction. By reducing the size and amount of computation of the model, the response speed and operation efficiency of the model can be greatly improved, so that it can work more flexibly and efficiently in scenarios such as real-time prediction and large-scale data processing.

The lightweight technology of machine learning models can be achieved through multiple channels, mainly including model compression, network structure simplification and hardware acceleration. Model compression is one of the most commonly used technologies in lightweight design. Through pruning, quantization and distillation, the number of model parameters and computational complexity are reduced, thereby reducing computing resource

consumption. The pruning method reduces the size of the model by removing unimportant connections or neurons, quantization reduces storage requirements by reducing the accuracy of weight values, and distillation technology migrates the knowledge of large models to small models, thereby maintaining high prediction accuracy while reducing computational overhead and storage requirements [8]. These technologies have been successfully implemented in many applications, especially in tasks that need to run on resource-constrained devices, which can significantly improve the performance and efficiency of the model.

In addition to model compression, simplification of network structure is also an important aspect of lightweight design. By designing miniaturized neural network architectures, such as MobileNet and EfficientNet, the computational complexity and number of parameters of the model can be significantly reduced without sacrificing too much prediction accuracy. These miniaturized network architectures optimize the network's hierarchical structure and parameter allocation, allowing the model to adapt to the deployment requirements of different hardware platforms while ensuring efficient performance [9]. Especially in mobile and embedded systems, these simplified network architectures can effectively reduce memory usage and computing latency, thereby improving real-time prediction capabilities and deployment convenience.

Hardware acceleration is another important technical means in the process of achieving model lightweighting. By using hardware acceleration devices such as FPGA (field programmable gate array) and ASIC (application-specific integrated circuit), the calculation speed and efficiency of the model can be greatly improved, especially in time series prediction tasks that require a lot of calculations. FPGA and ASIC can make the calculation process more efficient, reduce power consumption and latency through customized hardware design, thereby improving real-time processing capabilities. In some special application scenarios, the combination of hardware acceleration and lightweight models can give full play to the advantages of hardware and provide ultra-high prediction efficiency and accuracy, especially in the applications of the Internet of Things and edge computing, where the demand for hardware acceleration is increasing.

In summary, the lightweight design and implementation of machine learning models requires the comprehensive use of multiple technical means. Through methods such as model compression, network structure simplification and hardware acceleration, the computing resource consumption of the model can be effectively reduced, and the real-time and deployability performance can be improved, so that the machine learning model can better adapt to various practical application scenarios, especially perform efficient time series prediction tasks on resource-constrained devices.

4 LIGHTWEIGHT DESIGN AND IMPLEMENTATION METHODS IN APPLICATIONS

In practical applications, lightweight design and implementation are the key to improving the efficiency of time series prediction models, especially when running on resource-constrained devices. Pruning and quantization technology is one of the most common lightweight methods and is widely used in time series prediction. Pruning methods reduce the

size of the model by pruning unnecessary neural network connections or nodes. According to the characteristics of time series data, pruning strategies can selectively select important features and connections to retain and remove redundant calculation parts. For example, when processing time series data with seasonal changes, pruning can focus on retaining parameters in the model that have a greater impact on periodic changes and remove dependence on non-critical factors, thereby improving prediction efficiency and real-time performance. Quantization methods reduce the storage requirements of the model by reducing numerical precision, which is particularly important for platforms with limited computing power such as embedded systems and mobile devices [10]. By converting weights represented by floating-point numbers into low-precision integers, quantization not only significantly reduces the size of the model, but also speeds up the calculation process and improves the ability of real-time prediction. On mobile devices, the optimization of time series prediction models often relies on the combination of these two technologies. Through pruning and quantization, the model can run efficiently under lower hardware requirements and meet the performance requirements of practical applications.

Model distillation and transfer learning are two other effective lightweight techniques, which are particularly suitable for dealing with complex and large-scale time series problems. Knowledge distillation guides small models through large models, so that small models can retain the accuracy of large models while reducing the amount of calculation. In tasks such as financial market forecasting, it is usually necessary to deal with complex market behaviors and highly uncertain data. Using large models to distill small models can reduce the computational complexity while maintaining high prediction accuracy. Transfer learning uses models that have been pre-trained on large data sets to transfer their knowledge to new fields or tasks. In time series forecasting, transfer learning can help models quickly adapt to data in different fields, such as migrating weather forecast models to financial market forecasting, or migrating retail sales forecasting models to e-commerce platforms. This method not only reduces the need for large-scale training data, but also greatly improves the adaptability and performance of models on new tasks. The application case of distillation technology in financial market forecasting is particularly typical. The use of pre-trained deep learning models, such as LSTM, can reduce the consumption of computing resources while maintaining high prediction accuracy after distillation.

In terms of miniaturized neural network design, in recent years, miniaturized deep learning models combined with lightweight structures have become an important research direction. Network structures such as LSTM variants and GRU significantly reduce computational complexity by simplifying network layers and the number of parameters while ensuring high accuracy. Combined with lightweight network architectures such as MobileNet, the efficiency of time series prediction models can be further improved. These network architectures optimize the design of convolutional layers and fully connected layers, so that the network can significantly improve computational efficiency and reduce memory usage when processing large-scale time series data. These miniaturized neural networks have shown obvious advantages when deploying time series prediction models in smart hardware devices. For example, in smart watches or IoT devices, the use of optimized miniaturized neural network models can achieve efficient real-time data processing on low-power hardware

platforms to meet the long-term operation requirements of the device.

In addition, the requirement for real-time prediction makes hardware acceleration technology an indispensable component of lightweight design. With the development of edge computing, how to improve model computing efficiency through hardware acceleration has become the core of achieving real-time prediction. Edge computing devices usually have limitations in computing power and storage space. Therefore, combining model reasoning with hardware platforms such as FPGA or ASIC can greatly improve computing speed and efficiency. This hardware acceleration can not only reduce latency, but also reduce power consumption, so that time series prediction models can run efficiently and meet real-time requirements. The application case of using edge computing devices for real-time prediction of meteorological data shows that hardware acceleration technology can greatly improve model processing capabilities and prediction accuracy in scenarios where multiple data sources are input and require rapid response. In the meteorological prediction system, combining FPGA for real-time data analysis can provide prediction results at the minute level, significantly improving disaster warning and decision-making efficiency.

In summary, the lightweight design and implementation methods in the application include pruning and quantization, model distillation and transfer learning, design and implementation of miniaturized neural networks, and hardware acceleration technology. These methods effectively reduce the computing overhead and storage requirements of time series prediction models through different technical means, which not only improves the prediction accuracy, but also optimizes the performance and deployability of real-time prediction. They are widely used in low-resource environments such as mobile devices, smart hardware, and edge computing, providing more efficient prediction solutions for various industries.

5 EXPERIMENTAL AND APPLICATION CASE ANALYSIS

In the process of lightweight design and implementation of machine learning models, experimental design and application case analysis are important steps to verify the effectiveness of lightweight technology. In order to comprehensively evaluate the performance of lightweight models in practical applications, we selected several typical time series data sets and designed a series of experiments to examine the model's computing resource consumption, prediction accuracy, model loading time, and inference speed in different tasks.

In the experimental design, we first selected financial market data, meteorological data, and retail industry sales data as experimental objects. These data sets are representative and can effectively reflect the complexity and diversity of time series data in practical applications. For financial market data, we selected historical stock price data to predict future price trends. Meteorological data comes from historical weather data from meteorological stations, which is mainly used for predictions of temperature, humidity, precipitation, etc. Retail data includes sales records of e-commerce platforms, which are used to predict product sales and inventory management. These data sets not only contain complex features such as seasonality and trend, but also involve high-frequency data and real-time requirements, which can better test the effectiveness of lightweight models.

In terms of experimental indicators, we mainly focus on four aspects: computing resource consumption, prediction accuracy, model loading time and inference speed. Computing resource consumption measures the model's demand for hardware resources during training and inference, especially on mobile terminals and edge computing devices. Prediction accuracy evaluates the accuracy of the model to ensure that the model's prediction ability is not significantly reduced while lightweight design. Model loading time and inference speed are directly related to the real-time performance of the model. Especially in scenarios that require fast response, how to reduce the amount of calculation while ensuring fast model loading and inference has become the focus of the experiment.

Based on these experimental designs, we will analyze the actual effect of lightweight design through several application cases. In the time series prediction application in the financial market, we designed a lightweight LSTM model specifically for stock market prediction on mobile devices. Through techniques such as pruning and quantization, we significantly reduced the computing resource consumption of the LSTM model while maintaining a high prediction accuracy. In actual deployment, the lightweight LSTM model can be quickly loaded on resource-constrained devices such as smartphones and predict stock prices in real time. Performance analysis shows that the model can complete data processing with low latency while ensuring high prediction accuracy, showing excellent real-time prediction capabilities.

In the prediction and application of meteorological data, we used a lightweight neural network model based on pruning and quantization for weather forecasting. The model removes redundant connections through pruning and reduces the weight accuracy through quantization, thereby reducing the size of the model. By deploying on edge computing devices, the model can respond quickly in real-time meteorological monitoring systems. Experimental results show that despite pruning and quantization, the prediction accuracy of the model is not significantly affected. It can quickly provide accurate weather trend forecasts in meteorological warnings, and its computing resource consumption is greatly reduced compared to traditional models, which adapts to the needs of edge computing platforms.

In the sales and inventory forecasting case of the retail industry, we used miniaturized deep learning models (lightweight models based on GRU) for inventory forecasting. These lightweight models can run on low-power devices by optimizing the neural network structure and reducing the number of parameters, and adapt to real-time demand forecasting in e-commerce platforms. The model processes sales data in real time and predicts future product demand, thereby optimizing inventory management. Application examples show that in the actual deployment of e-commerce platforms, lightweight deep learning models can quickly respond to market changes, provide accurate sales forecasts, and significantly improve the efficiency of inventory management. In addition, due to the small size of the model, the deployment time and computing resource consumption during runtime are greatly reduced, which improves the response speed and stability of the system.

Through these experiments and application case analysis, we verified the feasibility and effectiveness of lightweight design in time series prediction. Whether in real-time prediction of financial markets or in meteorological monitoring and inventory management in the retail industry, lightweight models have shown excellent performance. Through pruning,

quantization, miniaturization of neural networks and hardware acceleration, lightweight design can significantly reduce computing resource consumption while ensuring prediction accuracy, meeting the requirements of real-time prediction and efficient deployment. These experimental results provide strong support for the promotion of lightweight machine learning models in practical applications, and also provide valuable experience and data basis for the deployment of lightweight models in more industries in the future.

6 CHALLENGES AND SOLUTIONS IN ACTUAL DEPLOYMENT

In actual deployment, lightweight time series prediction models face many challenges, which mainly come from resource limitations in the deployment environment, real-time and variability issues of data, and cross-platform and cross-device compatibility issues. In order to solve these problems, corresponding technical means and optimization strategies must be adopted for different application scenarios.

First, resource constraints in the deployment environment, especially resource bottlenecks on edge computing devices and mobile devices, pose great challenges to the deployment of lightweight models. Edge computing devices usually have limited processing power, storage space, and battery life, especially mobile devices. Time series prediction models usually need to process a large amount of historical data and perform inference in real time, which places high demands on computing resources. Therefore, it is crucial to choose the appropriate lightweight method based on hardware resources. On resource-constrained devices, methods such as pruning and quantization have become the most commonly used lightweight strategies, which reduce the burden on the device by reducing the number of parameters and computation of the model. In addition, the use of miniaturized network structures (MobileNet or EfficientNet) can further reduce computing requirements while still maintaining high prediction accuracy. In practical applications, appropriate lightweight methods can be selected for the hardware performance of different devices to ensure that the model can run stably in an environment with high real-time requirements.

Secondly, the real-time and variability issues of data are also challenges that cannot be ignored in the deployment of time series prediction models. Time series data is usually affected by multiple factors such as seasonality, trend, and periodicity. These characteristics often change over time, and the data changes rapidly. How to handle dynamically changing time series data and ensure that the model can capture these changes in real time has become the key to achieving efficient prediction. For applications with high real-time requirements, such as financial market forecasting and meteorological monitoring, lightweight models must be able to complete data processing and reasoning in a very short time. However, the dynamic changes and real-time requirements of data often require the model to have strong adaptability, while ensuring the accuracy of predictions and reducing response time. Therefore, when designing the model, it is necessary to fully consider how to balance between accuracy and speed. According to different data characteristics and application scenarios, suitable real-time update and incremental learning methods are adopted to enable the model to quickly adapt and provide accurate predictions when the data changes.

Finally, cross-platform and cross-device compatibility issues are also a major challenge

faced by lightweight models in actual deployment. Different devices and operating systems usually affect the compatibility and execution efficiency of the model, especially in multi-platform environments such as IoT devices, embedded systems, and mobile devices. Lightweight models need to have strong adaptability. In order to ensure that the model can run smoothly on different platforms, the model needs to be specially optimized. Common solutions include customized optimization for different hardware architectures (CPU, GPU, FPGA, etc.), and the use of platform-independent frameworks (TensorFlow Lite, ONNX, etc.) to achieve cross-platform deployment. In addition, in terms of operating systems, the deployment of lightweight models must also take into account the characteristics and limitations of different operating systems (Android, iOS, Linux, etc.) to ensure that the model can run stably in various environments. These optimization measures can effectively improve the compatibility of lightweight models on multiple platforms and multiple devices, ensuring that they can run efficiently and reliably in different application scenarios.

In general, the challenges in actual deployment are mainly concentrated in resource limitations, real-time requirements, data variability, and cross-platform compatibility. In response to these challenges, the use of appropriate lightweight methods, real-time update mechanisms, and cross-platform optimization strategies can not only improve the computational efficiency and adaptability of the model, but also ensure that the model can run stably in complex application environments and meet actual application needs.

7 FUTURE OUTLOOK AND DEVELOPMENT DIRECTION

With the continuous advancement of technology, the lightweight design and implementation of machine learning models in time series prediction still face many new challenges, but also usher in a broad space for development. In the future, the research on lightweight technology will be more in-depth, especially the introduction of adaptive lightweight models and the combination of quantum computing and artificial intelligence, which may bring revolutionary breakthroughs to lightweight design.

Adaptive lightweight model is one of the important directions for future research. The core idea of this model is to dynamically adjust the complexity of the model according to the characteristics of real-time data during the operation of the model. Traditional lightweight methods, such as pruning, quantization, and miniaturization of network structures, are usually optimized based on fixed preset rules, while adaptive lightweight models can automatically select appropriate network structures and parameter quantities according to different characteristics of input data. For example, when the data has strong seasonal or periodic characteristics, the model can increase the network level and complexity according to these characteristics to ensure prediction accuracy; when the data changes relatively smoothly, the model can automatically reduce complexity and reduce computational overhead. This adaptive ability not only improves the flexibility of the model, but also better adapts to the needs of different application scenarios and improves the effectiveness of the model in actual deployment.

The combination of quantum computing and artificial intelligence, especially in the research of lightweight models, has shown great potential. Quantum computing has parallel

computing capabilities that traditional computers cannot match, and it can provide stronger computing power when processing large-scale data. For time series prediction, a task that is highly dependent on data processing, quantum computing is expected to significantly improve the computing speed and efficiency of the model. In the future, quantum machine learning models may be able to perform more efficient feature extraction and modeling of time series data through quantum algorithms, thereby providing faster computing performance and lower resource consumption for lightweight models. This direction is still in the exploratory stage, but with the continuous development of quantum computing technology, its potential for lightweight design deserves high attention.

The expansion of cross-domain applications is also an important trend in the development of lightweight time series prediction models in the future. Traditional application scenarios, such as finance, meteorology, and retail, have proven the effectiveness of lightweight models, but with the development of technology, more and more emerging fields will benefit from this technology. For example, the demand for real-time data processing and prediction in the fields of smart homes and industrial Internet of Things (IIoT) is growing rapidly. In smart homes, devices need to sense and respond to environmental changes in real time. Lightweight time series prediction models can effectively help devices to perform intelligent control and improve energy efficiency and comfort. In the industrial Internet of Things, a large amount of time series data generated by devices and sensors needs to be processed quickly. Lightweight models can not only improve prediction accuracy, but also reduce computing resource consumption and support low-power devices, thereby reducing overall costs and enhancing the level of industrial automation.

At the same time, with the popularization of IoT devices and edge computing, the application prospects of machine learning models in low-power devices are also very broad. Low-power devices have limitations in battery life and processing power. Therefore, how to design models that can still provide efficient predictions in resource-constrained hardware environments has become a key direction for future research. By combining lightweight design, hardware acceleration technology and adaptive learning algorithms, the real-time prediction capabilities on low-power devices will continue to improve. With the continuous development of edge computing, these low-power devices will be able to break away from cloud computing and independently perform efficient local data processing and prediction, thereby reducing latency and improving the overall efficiency of the system.

In general, lightweight technology will develop in a more intelligent and adaptive direction in the future. The combination of adaptive lightweight models and quantum computing with AI indicates that the field of time series prediction will usher in more efficient and flexible model design. In addition, the expansion of cross-domain applications and the popularization of low-power devices will also promote the widespread application of lightweight models in more emerging fields, which will not only improve the intelligence level of various industries, but also promote the innovation and development of the entire technology ecosystem.

8 CONCLUSION

This study deeply explores the lightweight design and implementation of machine learning models in time series prediction, focusing on the important application significance of lightweight design in improving computing efficiency, reducing resource consumption, and improving model real-time and deployability. Through the application of pruning, quantization, distillation, hardware acceleration and other technologies, we have successfully achieved the lightweight of time series prediction models, so that these models can run efficiently on devices with limited computing resources. T

his design not only meets the application scenarios with strict requirements on real-time and computing power, such as mobile devices and edge computing, but also ensures the effective retention of prediction accuracy, significantly improving the practicality and adaptability of the model.

The balance between lightweight technology and actual needs is a core issue of this study. While ensuring the accuracy of the model, how to reduce computing resource consumption and improve reasoning speed, especially on hardware platforms with limited resources, has always been a challenge for lightweight design. By optimizing the network structure, using adaptive learning mechanisms, and combining hardware acceleration, this study successfully found this balance point in multiple application scenarios. Especially in practical tasks such as financial forecasting, meteorological monitoring, and retail forecasting, lightweight models not only show high computational efficiency, but also maintain prediction accuracy comparable to traditional large models, proving the great potential of lightweight design in practical applications.

In the future, with the continuous development of computing technology, lightweight models will usher in greater optimization space and application prospects. Lightweight design will continue to develop in a more intelligent and adaptive direction, especially in the face of dynamically changing data and complex real-time prediction tasks, how to make the model more flexible and adaptable will become a future research focus. In addition, the combination of quantum computing and artificial intelligence, and further innovations in technologies such as quantization and distillation will bring more efficient computing power and lower resource consumption to lightweight models. With the popularization of the Internet of Things, edge computing, and low-power devices, lightweight models will be widely used in more and more emerging fields, promoting the implementation of intelligent technology in a wider range of scenarios.

In general, lightweight design has become an important direction for the development of machine learning models in time series prediction. In the future, with the advancement of technology and the continuous expansion of application scenarios, lightweight models will play a greater role in all walks of life, and promote the development of intelligent prediction technology towards a more efficient, flexible and low-power direction.

REFERENCES

- [1] Wang, Z., Ruan, S., Huang, T., Zhou, H., Zhang, S., Wang, Y., ... & Liu, Y. (2024). A lightweight multi-layer perceptron for efficient multivariate time series forecasting. *Knowledge-Based Systems*, 288, 111463.

- [2] Ekambaram, V., Jati, A., Nguyen, N., Sinthong, P., & Kalagnanam, J. (2023, August). Ts mixer: Lightweight mlp-mixer model for multivariate time series forecasting. In *Proceedings of the 29th ACM SIGKDD conference on knowledge discovery and data mining* (pp. 459-469).
- [3] Coelho, I. M., Coelho, V. N., Luz, E. J. D. S., Ochi, L. S., Guimaraes, F. G., & Rios, E. (2017). A GPU deep learning metaheuristic based model for time series forecasting. *Applied Energy*, 201, 412-418.
- [4] Mukherjee, A., Mukherjee, P., Dey, N., De, D., & Panigrahi, B. K. (2020). Lightweight sustainable intelligent load forecasting platform for smart grid applications. *Sustainable Computing: Informatics and Systems*, 25, 100356.
- [5] Hissou, H., Benkirane, S., Guezzaz, A., Azrour, M., & Beni-Hssane, A. (2024). A lightweight time series method for prediction of solar radiation. *Energy Systems*, 1-38.
- [6] Torres, J. F., Hadjout, D., Sebaa, A., Martínez-Álvarez, F., & Troncoso, A. (2021). Deep learning for time series forecasting: a survey. *Big data*, 9(1), 3-21.
- [7] Campos, D., Zhang, M., Yang, B., Kieu, T., Guo, C., & Jensen, C. S. (2023). LightTS: Lightweight time series classification with adaptive ensemble distillation. *Proceedings of the ACM on Management of Data*, 1(2), 1-27.
- [8] Lim, B., & Zohren, S. (2021). Time-series forecasting with deep learning: a survey. *Philosophical Transactions of the Royal Society A*, 379(2194), 20200209.
- [9] Lee, M. C., Lin, J. C., & Katsikas, S. (2024, August). Impact of recurrent neural networks and deep learning frameworks on real-time lightweight time series anomaly detection. In *International Conference on Information and Communications Security* (pp. 228-247). Singapore: Springer Nature Singapore.
- [10] Chiu, S. M., Chen, Y. C., & Lee, C. (2022). Estate price prediction system based on temporal and spatial features and lightweight deep learning model. *Applied Intelligence*, 52(1), 808-834.