BUILD SUSTAINABILITY OF DATA SHARING BASED ON DEEP LEARNING MODELS IN THE ESG ENVIRONMENT

HAN JIANG

MASTER'S THESIS

LUND UNIVERSITY

DATA ANALYSIS AND BUSINESS ECONOMICS

BUILD SUSTAINABILITY OF DATA SHARING BASED ON DEEP LEARNING MODELS IN THE ESG ENVIRONMENT

HAN JIANG

JH01171009@163.COM

AUGUST 22, 2024

SUPERVISOR: BEHNAZ PIRZAMANBEIN

Abstract

Against the backdrop of increasing global emphasis on environmental, social, and corporate governance (ESG), improving the efficiency of ESG data analysis and the sustainability of its sharing has become particularly important. This study proposes a comprehensive research framework, first demonstrating the advantages of deep models for ESG data analysis, and then discussing the combination of ESG data sharing and blockchain technology, aiming to achieve sustainable development of ESG data sharing.

Firstly, we collected and processed ESG data from multiple companies, including annual reports, third-party ratings, and survey data, and standardized and extracted features. In terms of ESG data processing, K-nearest neighbor (KNN) models are used as benchmarks, while multi-layer perceptrons (MLP) and long short-term memory networks (LSTM) are adopted due to their excellent ability to handle complex ESG data. Due to the excellent performance of LSTM in time series analysis, MLP can effectively extract nonlinear features, thus outperforming KNN in ESG score prediction. Next, we are considering using blockchain technology to design a decentralized ESG data sharing platform to ensure data transparency, security, and privacy.

This study is based on a comprehensive framework that combines deep learning models with data sharing platforms. Through deep modeling, dynamic data updates, and feedback mechanisms, continuously optimize model and platform design to ensure that data sharing has practical significance for improving corporate ESG. The results of model analysis can provide specific improvement suggestions for enterprises, forming a virtuous cycle and promoting continuous improvement of ESG performance.

Key words: ESG Data Analysis, Deep Learning Models, KNN Model, MLP, LSTM, Blockchain Technology, Data Sharing Platform, Sustainable Development, ESG Score Prediction

Context

Chapter 1

Introduction

1.1 Implications and Background of the Study

1.1.1 Background

Against the backdrop of increasing global attention to environmental, social, and corporate governance (ESG), how to improve the efficiency of ESG data analysis and the sustainability of data sharing has become a key issue[1][2]. However, the existing ESG data sources are scattered and have inconsistent standards, which poses significant challenges for analysis and sharing[3]. Traditional data analysis methods are unable to fully explore the deep level information in ESG data, which limits the decision-making ability of investors and companies in the field of sustainable development[4].

Deep learning techniques, especially Long Short Term Memory (LSTM) and Multi Layer Perceptron (MLP) models, have the potential to overcome the limitations of traditional methods due to their advantages in processing complex and multidimensional data[5]. By combining blockchain technology, this study aims to build a decentralized ESG data sharing platform that ensures data transparency and security, thereby achieving sustainable development of ESG data sharing. Through this continuous comprehensive research framework, we expect to provide more reliable support and decision-making basis for the sustainable development of ESG data for enterprises and society[6].

1.1.2 Research Implications

The meaning of this study being done is to find how deep learning models and blockchain technology can be applied in ESG data analysis and sharing. The goal is more efficient mechanisms for ESG data processing and sharing that are sustainable. LSTM models, known for their good performance in time series forecasting by capturing long-term dependencies and trends in ESG data, are useful. MLP models are also helpful but in extracting nonlinear features from multidimensional data. When comparing with traditional machine learning models, KNN for instance, this study tries to show how LSTM and MLP bring benefits of higher prediction accuracy and toughness.

A transparent and secure ESG data sharing platform should be built. ESG data, consisting of sensitive information from corporations, must be secure to prevent data breaches, which can cause false analysis results, financial risks, and loss of trust from stakeholders[7]. Ensuring the safety and integrity of this data is important for it to be dependable and for supporting decision-making that is well-informed[8]. Blockchain technology is looked into for meeting these security needs as it offers a decentralized and unchangeable solution[9]. By creating a data sharing platform that uses blockchain technology along with secure encryption methods and strong access control systems, the privacy and security of ESG data can be protected during the entire sharing process[10].

The sharing of ESG data is necessary for fostering transparency and accountability among businesses, investors, and other stakeholders. Making ESG data accessible allows businesses to show their dedication to sustainable practices, helping stakeholders to make better decisions and comparisons.

1.2 Objectives

The main goal of this study is to build an efficient, transparent and sustainable ESG data analysis and sharing framework through the detailed design and implementation of deep learning models and the application of blockchain technology in data sharing platforms. Specific objectives include the following:

1. Develop an efficient ESG data analysis model:
- Select and train deep learning models (such as LSTM and MLP) suitable for ESG data analysis, and compare them with traditional machine learning models (such as KNN).
- Verify the advantages of deep learning models in ESG data prediction and feature extraction, and improve the accuracy and robustness of ESG score prediction.

2. Discuss the design of a decentralized data sharing platform:
- Discuss the decentralization, immutability and transparency characteristics of blockchain technology, and design the architecture of an ESG data sharing platform.
- Ensure the security and privacy of information during data sharing.

3. Design dynamic data update and feedback mechanism:
- Design of dynamic data update mechanism, regular collection and update of the company's ESG data, to ensure the timeliness and accuracy of data.
- Through the analysis of the results of the deep learning model, we provide improvement suggestions to the enterprise, and optimize the ESG strategy and practice of the enterprise based on the feedback, forming a virtuous circle of continuous improvement.

4. Optimize the design and management of data sharing platforms:
- Using the analysis results of deep learning models, we will discuss the design and management of an optimized data sharing platform to ensure that the platform can effectively support ESG improvements for enterprises and society.
- Through model-driven data analysis, it provides scientific basis and technical support for the continuous improvement of the platform (only for design level discussion, not specific implementation).

5. Promote the sustainable development of business and society:

- Through the detailed design and implementation of deep learning models, we can improve the ESG performance of enterprises and promote the sustainable development of society.
- The results of the study will serve as a reference for policymakers, investors, and corporate managers, and promote the progress of research and practice in the field of ESG.

Chapter 2

Literature Review

2.1 The necessity and security of ESG data sharing

The sharing of ESG data plays an important role in enhancing corporate transparency, supporting investment decisions, promoting industry standardization, and driving social development. Gonzalez and Schmidt (2021) point out that by publicly disclosing ESG information, companies can strengthen their social responsibility and provide a more comprehensive perspective for stakeholders[11]. However, Ma's (2023) research suggests that existing ESG data has a significant impact on investors' sustainable investment decisions, particularly when evaluating a company's sustainability capabilities[12]. Li and Li (2023) emphasized the importance of standardizing ESG data, pointing out that through data sharing, companies can compare on the same basis, thereby promoting industry transparency and fairness[13].

These studies emphasize the necessity of ESG data sharing, but in reality, the main challenges faced by ESG data sharing lie in data security and privacy protection. Taylor and Ling (2023) pointed out that ESG data may contain sensitive information, and if not properly protected, these data breaches may have a negative impact on a company's market position and reputation[14]. Sfetcu (2022) studied the role of blockchain technology in ensuring data integrity and immutability, and proposed using encryption technology to ensure the reliability of ESG data during the sharing process[15]. Current research mainly focuses on the theoretical exploration of blockchain technology for data security[16], lacking specific application cases and empirical analysis. Therefore, this study combines deep learning models and blockchain technology to discuss more efficient privacy protection and data security in ESG data sharing.

2.2 The Application of Deep Learning Models in ESG Data Analysis

In recent years, deep learning models have shown great ability in processing complex datasets, especially those involving multidimensional data or time series data. Lee et al. (2022) proposed a method that combines machine learning with deep learning to analyze ESG data and demonstrated its effectiveness in predicting corporate ESG ratings[17]. Gamlath et al. (2023) further developed an automated system that utilizes financial and textual data to generate ESG ratings, demonstrating the potential of integrating multi-source data into ESG ratings[18].

These studies demonstrate the potential of deep learning in ESG data analysis, but most research is still limited to theoretical model validation and lacks large-scale, cross industry data analysis. This study will further expand upon this by utilizing LSTM and MLP models to analyze ESG data from multiple industries. In addition, we will further optimize model performance through hyperparameter tuning and model integration techniques to address the complexity of ESG data in practical applications.

On ESG data management, Munappy and colleagues in 2019 pointed out how important it is to have data consistency, accuracy, and completeness. Their insights are giving guidance for data preprocessing in this paper, ensuring data has quality and consistency before it gets trained with deep learning models, and this step aims to improve prediction reliability[19]. Also, Sokolov and others in 2021 showed methods to boost ESG scoring systems' relevance and accuracy using models like BERT in deep learning. Even though BERT wasn't used here, Sokolov et al.'s findings still provide significant background support, pointing towards exploring various deep learning models' applicability in ESG data study[20].Chen and Liu in 2020, together with Franco et al. also in 2020, they examined analyzing ESG data through machine learning and deep learning methods. They explored how effective these methods are in practical scenarios[5][6]. Their studies validate deep learning models' effectiveness for ESG data, giving important background support to further delve into different models' applications in ESG data analysis.

2.3 The Application of Blockchain Technology in ESG Data Sharing

Due to its decentralization, transparency, and immutability, blockchain technology has become an innovative hotspot in the field of ESG data sharing in recent years. White and Heckman (2023) discussed how blockchain can enhance trust in ESG data sharing systems and ensure data integrity and verifiability by establishing decentralized ledgers[9]. These studies mostly focus on the technological advantages of blockchain and lack in-depth exploration of its application scenarios in actual ESG data sharing.

Priya et al. (2021) proposed the use of public-private key systems in blockchain to ensure the security and anonymity of data transmission in terms of data security[16]. Existing research is mostly theoretical and has not fully explored how to implement these security mechanisms in practical applications. On the basis of existing research, this study designs a decentralized ESG data sharing platform that combines blockchain technology, explores data privacy and security issues in practical applications, and achieves automation and compliance management of data sharing through smart contracts.

Taylor and Ling (2023) dived into computational ways for ensuring ESG data security when deep learning applied[8]. Their findings lay theoretical groundwork for security considerations in deep learning models' setup in this paper. Furthermore, Sfetcu (2022) looked into blockchain’s role in maintaining data sameness and immutability by way of encryption, crucial for sharing and checking ESG data[17]. Those studies underpin this paper’s combination of blockchain tech with deep learning models aiming at securing data.

Mafakheri et al. (2018) inquired how blockchain could ease direct data exchanges among institutions, getting rid of the need for third-party agents[20]. Such strategy delivers a decentralized architecture to the data sharing scheme in this paper, affirming data sameness and openness. Anderson and Zion (2022) examined the use of secure multi-party computing for ESG data scrutiny, stressing on the need for data privacy safeguards[13]. Similarly, this paper adapts a privacy guard mechanism in its data sharing setup ensuring participants’ data remains safe.

2.4 The combined application of deep learning and blockchain technology

Combining models of deep learning with the technology of blockchain has the potential to be of significant benefit to the efficiency and security of data analysis concerning ESG, and sharing that data. Emphasized by Hiroshi Hirobumi (2022), the potential which exists for the integration was noted, involving blockchain technology within deep learning models, for improvements in data processing and for prediction accuracy[12]. The research which was conducted by these individuals is providing theoretical support for the combination of the two technologies, which is discussed in this article at the technical level.

In a practical setting, the research by Martin Reza (2022) showcased the benefit of combining blockchain technology with deep learning models, a conclusion was demonstrated where it enhances significantly the functionality, application breadth also of the models[10]. Besides this, studied was the application in sustainable manufacturing by Sahu Et al. (2021) where deep learning showed potential in managing complex types of data as well as optimizing operational processes[11].
This study will explore the deep integration of deep learning models and blockchain technology, and design a comprehensive framework that can utilize the powerful analytical capabilities of LSTM and MLP models while ensuring data security and transparency through blockchain technology. Through this technology combination, this study aims to provide an efficient, transparent, and sustainable solution for ESG data sharing, and hypothesize application scenarios to explore feasibility in practical scenarios.

Chapter 3

Research Methods

3.1 Data Collection and Processing

In this study, LSTM and MLP models were selected as the deep learning models to analyze and predict the ESG ratings of enterprises. The selection of these models is based on their theoretical advantages in processing specific types of data (refer to Appendix 2 for detailed content)

3.1.1 Data Source

The data which was used for this study derives from Kaggle platform, it is named "Public Company ESG Ratings Dataset" and got provided by Alistair Kings[1]. Collected data encompasses ESG scores, it's environmental, social, corporate governance scores from many listed companies globally. All are on New York Stock Exchange (NYSE). Various time periods it covers and company info spans multiple industries and geographies broadly.

Details in dataset include company name, scores of ESG and its sub-scores, industry classifications, places where companies operate, etc. Link to dataset is "https://www.kaggle.com/datasets/alistairking/public-company-esg-ratings-dataset?resource=download".

3.1.2 Dataset Description

For effective analysis and to ensure the decision-making process remains uncompromised, maintaining the security of ESG data is crucial. The dataset includes sensitive information such as company identifiers (e.g., ticker symbols, names, and CIK numbers), industry classifications, and ESG performance scores across environmental, social, and governance dimensions.

These scores indicate a company's performance in sustainable development. Any tampering or damage to this data could lead to inaccurate analyses[30], which would compromise the accuracy of company assessments. Furthermore, the misuse of this information by malicious actors or competitors, particularly in revealing flaws in governance or environmental management, poses significant risks[31]. To protect the interests of enterprises, prevent the misuse of information, and ensure its effective use in evaluation and investment decisions, strict security measures for data sharing are imperative. These measures will be discussed in detail in the following chapters.

The dataset includes the following key characteristics:

- Basic company information: stock code, company name, currency, exchange, industry, logo URL, website URL
- Environmental score and rating: environment_store environment_grade 、 environment_level
- Social score and rating: social-score social_grade、social_level
- Governance score and rating: governance score, governance level, governance level
- Overall ESG score and rating: total score total_grade、total_level
- Final processing date for ESG data
- CIK identifier

Environment, society, governance, and overall score are all numerical values, while the corresponding levels are letter ratings (such as AAA, BB, etc.), and levels are classified (such as high, medium, low).Please refer to Appendix 1 for complete data features.

Ticker	dis	gm	gww	mhk	lyv	lvs
Name	Walt Disney Co	General Motors Co	WW Grainger Inc	Mohawk Industries Inc	Live Nation Entertainment Inc	Las Vegas Sands Corp
environment_grade	A	A	B	A	BBB	A
environment_level	high	high	medium	high	high	high
social_grade	BB	BB	BB	B	BB	BB
social_level	medium	medium	medium	medium	medium	medium
governance_grade	BB	B	B	BB	B	BB
governance_level	medium	medium	medium	medium	medium	medium
environment_score	510	510	255	570	492	547
social_score	316	303	385	298	310	318
governance_score	321	255	240	303	250	313
total_score	1147	1068	880	1171	1052	1178
last_processing_date	19-04-2022	17-04-2022	19-04-2022	18-04-2022	18-04-2022	18-04-2022
total_grade	BBB	BBB	BB	BBB	BBB	BBB
total_level	high	high	medium	high	high	high
cik	1744489	1467858	277135	851968	1335258	1300514

Table3.1 Dataset feature table

*(Note: Please refer to the appendix 1 for the complete table)

In this study, we selected the above variables and features for model training and prediction, mainly based on the following considerations:

The criticality of ESG scoring:
ESG ratings (including environmental, social, and governance ratings) are important indicators of a company's sustainable development capabilities, and therefore they directly affect the prediction of overall ESG scores. Correlation analysis shows a significant positive correlation between environmental, social, and governance scores and overall ESG scores.

Introduction of differential mean feature:
In order to better capture the trend of changes in the company's ESG ratings, we introduced differential mean features (such as diff_cean_denvironment_stcore), which help the model understand the dynamic changes in ratings over time and improve the accuracy of predictions.

The value of qualitative variables:
Qualitative variables such as ESG rating and level provide insights beyond quantitative data for the model, which can help evaluate a company's ESG performance more comprehensively and improve the accuracy of predictions.

- Target variable

In this study, the target variable is the overall ESG score (total score) of the company. This variable represents the comprehensive performance of the company in the three dimensions of environment, society, and governance, and is the core indicator that we ultimately need to predict and analyze. By predicting the total score, we can evaluate a company's sustainability capability and provide improvement recommendations.

- Interesting variables

Variables related to environmental rating:
Environment Score
Environment grade
Environment Level
These variables directly reflect the company's performance in environmental protection, and research has shown a strong correlation between them and the overall score, especially the environmental score (environmentally score) which has a significant impact on predicting the overall score.

- Social rating related variables:

Social_Score
Social_grade
Societal level
These variables reveal the company's performance in social responsibility and make a

significant contribution to the overall ESG rating.

- Variables related to governance rating:

Governance_Score (governance score)
Governance_grade (governance level)
Governance Level
These variables reflect the company's performance in terms of governance structure and transparency. Good governance often means higher ESG scores, so these variables are also key predictors.

Reason for modeling:
LSTM model: This model is particularly suitable for capturing long-term dependencies and trends in time series data. Given that ESG performance may vary significantly over time, using LSTM can help predict future ESG ratings more accurately based on historical data.
MLP model: This model is used to handle nonlinear relationships and interactions between multiple features. Due to the complex interrelationships between ESG factors, MLP is suitable for capturing these interactions and providing robust predictions.
KNN model: Although used as a benchmark model, KNN provides a simple comparison for evaluating the performance of more complex models.

3.1.3 Data Cleansing and Standardization

After data collection, we carry out detailed data cleaning and normalization . The specific steps are as follows:
1. Handle missing values:
We have identified a total of 12 missing values.For the logarithmic column(environment_score, social_score, governance_score, total_score), the median is used to fill in the missing values.
2. Calculate the differential mean:
The main purpose of adding differential means is to better capture the trend of changes in a company's ESG rating. By calculating the differential mean of data, the rate of change and volatility of ratings can be more accurately reflected,
Calculate the differential mean for the numeric column and add it to the dataset.

$$\text{Diff\_Mean}[i] = x[i] - \frac{x[i-1] + x[i+1]}{2} \tag{3.1}$$

where $x[i]$ is the i data point, $x[i-1]$ is the previous data point, and $x[i+1]$ is the next data point.
In this way, local changes and trends in the data can be better captured. In the code[32], we define an differential_mean function to calculate the differential mean and apply it to numeric columns such as Environmental Score, Social Score, Governance Score, and Total Score.

3.1.4 Feature Extraction

After data cleansing and normalization, in order to further analyze and build the model, we performed feature extraction on the data and analyzed the correlation between features using a heat map. Here's a detailed description of the feature extraction section:

1. Feature selection and extraction

In feature selection and extraction, we focused on environmental scores, social scores, governance scores, and their associated ratings and levels. In addition, we introduce differential mean features (e.g., diff_mean_environment_score, diff_mean_social_score, diff_mean_governance_score) to capture local changes and trends in the data.

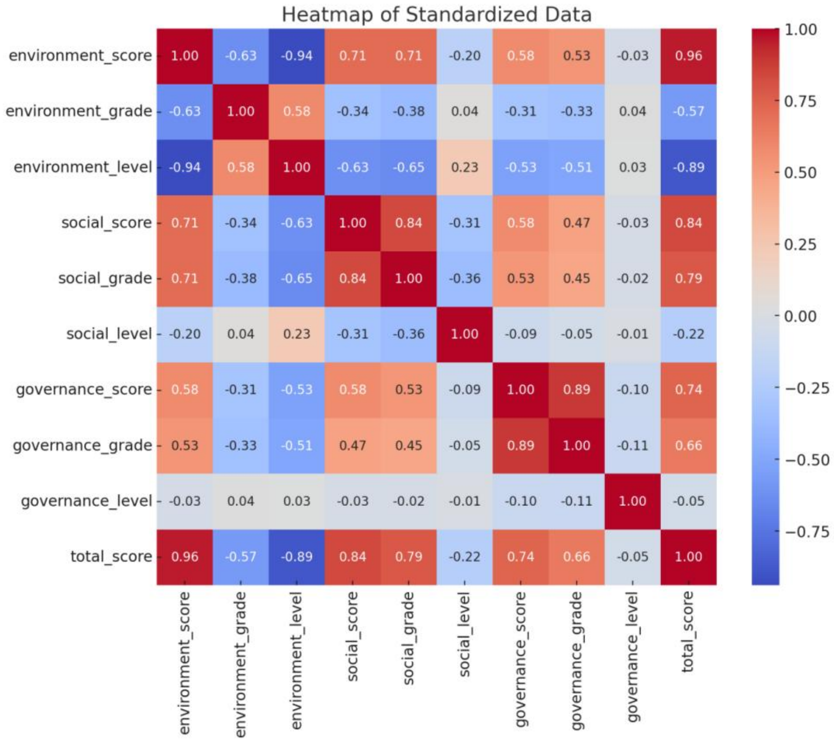
We selected the following characteristics for analysis:

- Environment-related features: environment_score, environment_grade, environment_level
- Socially relevant characteristics: social_score, social_grade, social_level
- Governance-related characteristics: governance_score, governance_grade, governance_level
- Comprehensive scoring characteristics: total_score, total_grade, total_level
- Differential mean features: diff_mean_environment_score, diff_mean_social_score, diff_mean_governance_score, diff_mean_total_score

2. Heatmap correlation analysis

To understand the correlation between these features, we plotted a heat map for correlation analysis[33]. The heat map indicates the degree of correlation between features by the shade of color, where red indicates a positive correlation, blue indicates a negative correlation, and darker the color, the stronger the correlation.

The following Graph A illustrates the correlation heat map between features:



Graph3.1 This heatmap illustrates the correlation between standardized ESG (Environmental, Social, and Governance) data.

From the heat map, we can observe the following:

(1) Correlation of the overall score with the environmental, social and governance score: total_score had a high positive correlation with environment_score, social_score, and governance_score, which were 0.96, 0.84, and 0.74, respectively. This suggests that the overall score is heavily influenced by the environmental, social, and governance score.

(2) Correlation of environmental scores with their ratings and levels:

The correlation between environment_score and environment_grade and environment_level was -0.63 and -0.94, respectively. This suggests that there is a strong negative correlation between environmental scores and their grades and levels.

Correlation of social scores with their grades and levels:

The correlation between social_score and social_grade and social_level was 0.84 and 0.23, respectively. Social scores have a strong positive correlation with their grades, but a weak correlation with levels.

Correlation of governance scores to their rating and level:

The correlation between governance_score and governance_grade and governance_level was 0.89 and 0.03, respectively. Governance scores have a strong positive correlation with their rating, while a weak correlation with their level.

Through these correlation analyses, we can better understand the relationship between different features and provide guidance for subsequent model training. Considering the relevance and practical significance of these features, we selected the features that have a significant impact on the total score for model training and prediction.'environment_score', 'environment_grade','environment_level','social_score','social_grade', 'social_level','governance_score', 'governance_grade', 'governance_level'

3.2 Deep Learning Model Design and Implementation

3.2.1 Multilayer Perceptron (MLP) Model Design

1.Introduction to MLP Model

The Multi-Layer Perceptron (MLP) is a feed forward neural network suitable for a variety of regression and classification. Has at least three layers (input, hidden, and output)[34].

- Input Layer:

Enter a feature vector X= $\left[x_1, x_2, \ldots, x_n\right]$

- Hidden Layer:

The output of each hidden layer is transformed nonlinearly by activating the function ReLU:

$$H^{(l)} = \mathrm{Re}LU(W^{(l)}H^{(l-1)} + b^{(l)}) \tag{3.2}$$

where $H^{(0)} = X$, $W^{(l)}$ and $b^{(l)}$ are the weight matrix and bias vector of layer l, respectively.

- Output Layer:

The output layer directly performs a linear transformation to output the predicted value:

$$\hat{y} = W^{(L)}H^{(L-1)} + b^{(L)} \tag{3.3}$$

2.Construction of MLP Models

1) Data Preprocessing

First, the dataset is divided into a training set (80%) and a test set (20%),The normalized data is used for training and testing of MLP models.

2) The Architecture of The MLP Model

- First Dense Layer:

Units=64: Using 64 neurons. This is the best configuration found in hyperparameter tuning(Theoretical support for hyperparameter tuning can be found in Appendix 3), indicating that the model performs best under this configuration.

Activation='tanh ': Use tanh as the activation function. This activation function helps to handle non-linear relationships and maintain good gradients during the training process.

- Second Dense Layer:

Units=32: Use 32 neurons. This smaller number of neurons is used to extract deeper level features of the input.
Activation='tanh ': Continue using tanh as the activation function to maintain model consistency.

- Output layer:

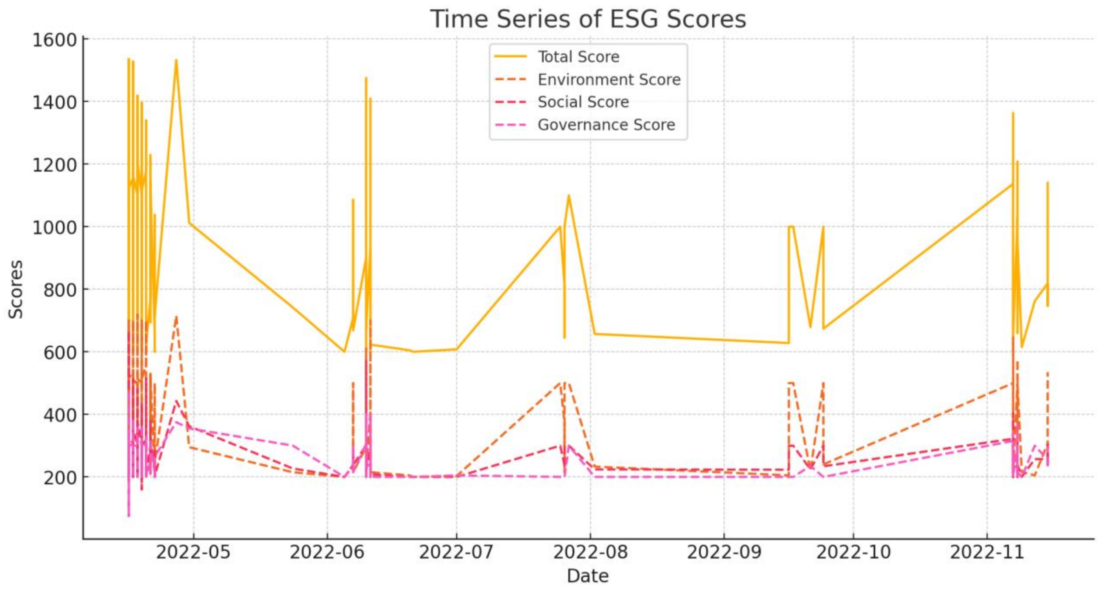
A single neuron is used for the output of regression tasks.

- Model compilation:

Using the Adam optimizer for optimization, the loss function is mean squared error, which is a common choice for regression tasks.
Training configuration:
The optimal configuration of epochs=100 and batch size=16 is determined through hyperparameter search, which determines the convergence speed and batch size of the training.This model uses RandomizedSearchCV for hyperparameter tuning
3) Model Evaluation
The performance of the model on the test set was evaluated by calculating the MAE and MSE.

3.2.2 Long Short-Term Memory Network (LSTM) Model Design

1.Time Series Analysis
Before establishing the LSTM model, I first conducted time series analysis, which is a powerful tool for understanding and predicting patterns and trends in data over time. By analyzing the long-term trends, seasonal variations, and cyclical fluctuations of data, it can help identify the temporal relationships between variables and determine whether models like LSTM are suitable for processing data. The LSTM model excels at capturing long-term dependencies in time series, so if time series analysis shows clear trends, seasonal or periodic variations, and lag effects in the data, then the LSTM model may be a suitable choice for processing these data[35].



Graph 3.2 From this Graph B, we can see that the overall score showed significant fluctuations throughout the entire time period, and sometimes significant peaks and valleys appeared. For environmental score, social score, and governance score,These three scores have relatively stable changes over time, but there may also be some obvious peaks. These changes may correspond to events at specific time points or strategic adjustments of the company. The fluctuation of the overall rating is very obvious, especially at certain time points where there are drastic fluctuations. In contrast, the volatility of environmental ratings, social ratings, and governance ratings is relatively small, but there are still some prominent points of change.

Based on the above analysis, the LSTM model is suitable for analyzing and predicting this time series data.

2.Introduction to LSTM Model

Long Short-Term Memory (LSTM) is a specific type of recurrent neural network (RNN) that is particularly well-suited for processing and predicting time-series-based data. Compared with traditional RNNs, LSTMs are better able to capture and retain long-term dependency information through their unique gating structure[35].

- Input Gate:

Calculation formula：

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \tag{3.4}$$

The input gate determines how much of the information at the current time step needs to be retained in the memory cell.

- Forget Gate:

Calculation formula：

$$f_t = \sigma\left(W_f \cdot \left[h_{t-1}, x_t\right] + b_f\right) \tag{3.5}$$

The Oblivion Gate controls how much of the memory of the previous step in time is preserved.

- Candidate Memory Cell:

Calculation formula：

$$\widetilde{C}_t = \tanh(W_C \cdot \left[H_{t-1}, x_t\right] + b_C) \tag{3.6}$$

Candidate memory units generate new candidates.

- Memory Cell Update:

Calculation formula：

$$C_t = f_t * C_{t-1} + i_t * \widetilde{C}_t \tag{3.7}$$

The memory cell state is updated by a combination of a forgotten gate and an input gate.

- Output Gate:

Calculation formula：

$$o_t = \sigma\left(W_o \cdot \left[h_{t-1}, x_t\right] + b_o\right) \tag{3.8}$$

The output gate determines which part of the memory cell state will be the output.

- Hidden State Update:

Calculation formula：

$$h_t = o_t * \tanh(C_t) \tag{3.9}$$

The hidden state is an output based on the state of the current memory cell.

3.Construction of LSTM Model

1) Data preprocessing

Firstly, the dataset is divided into a training set and a test set, in which the training set accounts for 80% and the test set accounts for 20%. Then use StandardScaler to normalize the numeric columns. The standardized formula is as follows:

$$X' = \frac{X - \mu}{\sigma} \tag{3.10}$$

The LSTM model requires the input data to be in 3D format [samples, timesteps, features], so after standardization, the data needs to be reshaped to a format suitable for the LSTM input.

2) The Architecture of The LSTM Model

- Configuration of LSTM layer:

Units=150: Select 150 LSTM units, which is the optimal configuration determined through hyperparameter tuning, indicating that the model performs well under this configuration.

- Dropout layer configuration:

The dropout rate of Dropout is set to 0.2, which means that 20% of neurons will be randomly discarded during each training session to reduce overfitting. This value was also determined through tuning.

- Dense layer (output layer):

The output layer uses Dense (units=1) to indicate that this is a regression problem that requires outputting a continuous value, i.e. predicting the target variable"Overall ESG score (total score)".

- Model compilation:

Using Adam optimizer and mean_Squared_error as the loss function is a standard choice in regression problems.

overall ESG score (total score)

Optimal parameter selection

This model used RandomizedSearchCV for hyperparameter tuning.

These parameter combinations exhibit good generalization ability on both the validation and test sets.

3) Model Training and Evaluation

The model is trained on 100 epochs in the training set, and the batch size is 32. Then, the test set was used for prediction, and the mean absolute error (MAE) and mean square error (MSE) were used to evaluate the performance of the model.

3.3 K-nearest neighbor (KNN) Regression Model

The K-Nearest Neighbors (KNN) algorithm, a non-parametric supervised learning method, is commonly used in classification and regression tasks[37]. In this study, KNN was employed as a benchmark model for regression analysis to predict ESG scores.

1. Data segmentation and standardization

Firstly, the dataset is divided into a training set and a test set, with the training set accounting for 80% of the total data and the test set accounting for 20%[39]. Use train_test_split functions to segment the dataset. Then, the numerical features are normalized using StandardScaler to eliminate dimensional differences between different features, and the normalized formula is:

$$X' = \frac{X - \mu}{\sigma} \tag{3.11}$$

where X' is the normalized eigenvalue, X is the original eigenvalue, μ is the mean of the eigen, and σ is the standard deviation of the eigen.

2. KNN regression algorithm

The basic formula for KNN regression is:

$$\hat{y} = \frac{1}{K}\sum_{i=1}^{K} y_i \tag{3.12}$$

where $\hat{y}$ is the predicted value and y_i is the known output value of the selected K neighbors.

In this study, the initial model chose a K value of 100 is to smooth the prediction results by considering more neighboring data points, thereby reducing the variance of the model's predictions.

3.4 Model Comparison and Analysis

3.4.1 Model Evaluation Indicators

1.Mean Absolute Error (MAE)

Mean Absolute Error (MAE) is the mean absolute value used to measure the difference between the predicted and actual values of the model. The smaller the MAE, the closer the prediction result of the model is to the actual value. It is calculated as follows:

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |y_i - \hat{y}_i| \tag{3.13}$$

where y_i is the actual value, $\hat{y}i$ is the predicted value, and n is the sample size.

MAE is highly interpretable because it directly reflects the average size of the prediction error, regardless of the direction of the error. Therefore, MAE is often used to measure prediction accuracy in practical applications.

2.Mean Square Error (MSE)

Mean Squared Error (MSE) is another commonly used regression model evaluation metric that measures the accuracy of a model by calculating the average of the sum of squares of the prediction error. The smaller the MSE, the better the prediction result of the model. It is calculated as follows:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{3.14}$$

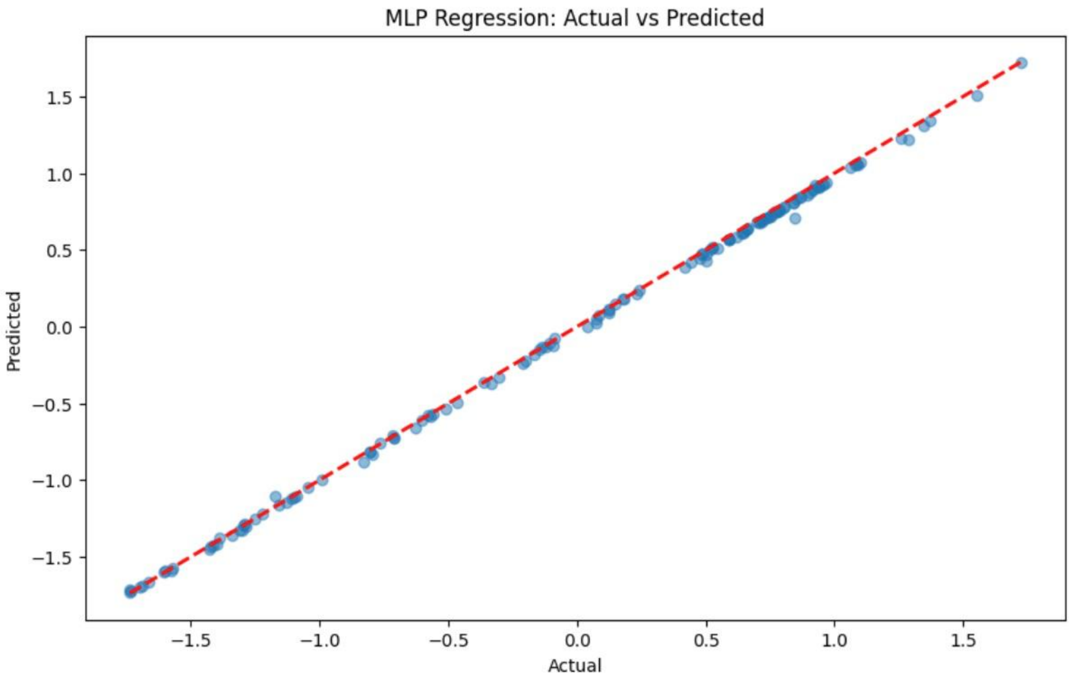
MSE differs from MAE in that it focuses more on larger errors, because when the error is squared, larger differences have a greater impact on the results. As a result, MSE is very useful when dealing with scenarios that are particularly sensitive to error.

3.4.2 Comparison of Model Results

1. MLP Model

1)Results Display

- Figure C shows the comparison between the predicted and actual values of the MLP model on the test set. The MLP model is trained on a dataset containing multidimensional data such as environmental ratings, social ratings, and governance ratings. In the figure, the horizontal axis represents the actual ESG total score, and the vertical axis represents the predicted score of the MLP model.
- From the graph, it can be seen that most of the data points are closely distributed around the red dashed line, indicating that the MLP model has high prediction accuracy when processing these multidimensional ESG data. Overall, the model still performs very well in fitting data, demonstrating its outstanding performance in processing multi-dimensional ESG data.



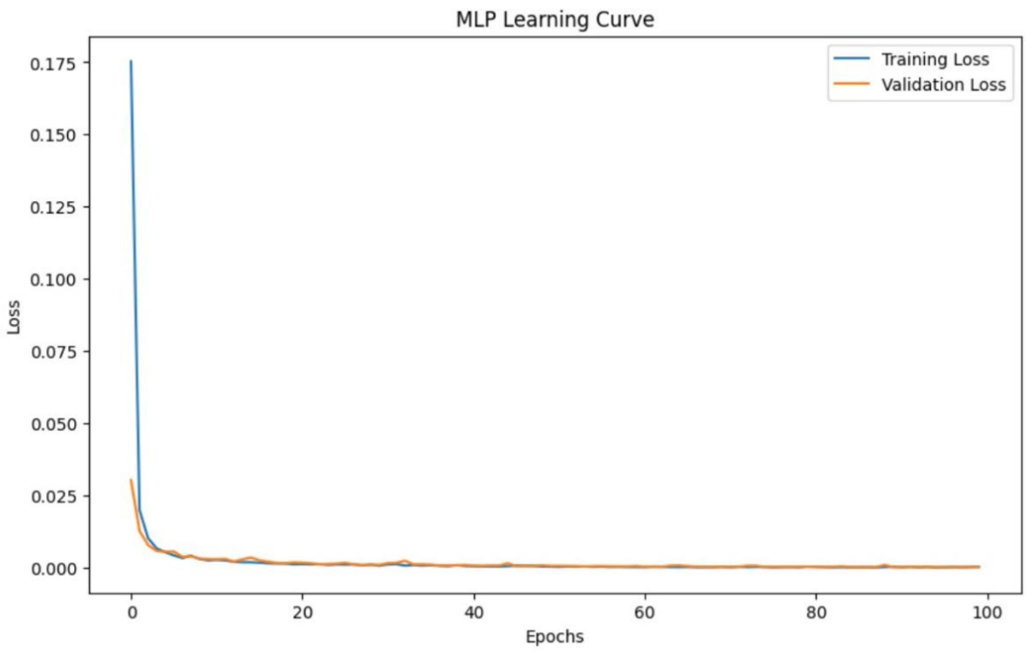
Graph3.3 A scatter plot that shows the relationship between the actual and predicted

values from a Multilayer Perceptron (MLP) regression model. The x-axis represents the actual values, and the y-axis represents the predicted values.

- Evaluation of MLP Model Results:
  - **MAE**: 0.0137
  - **MSE**: 0.0003

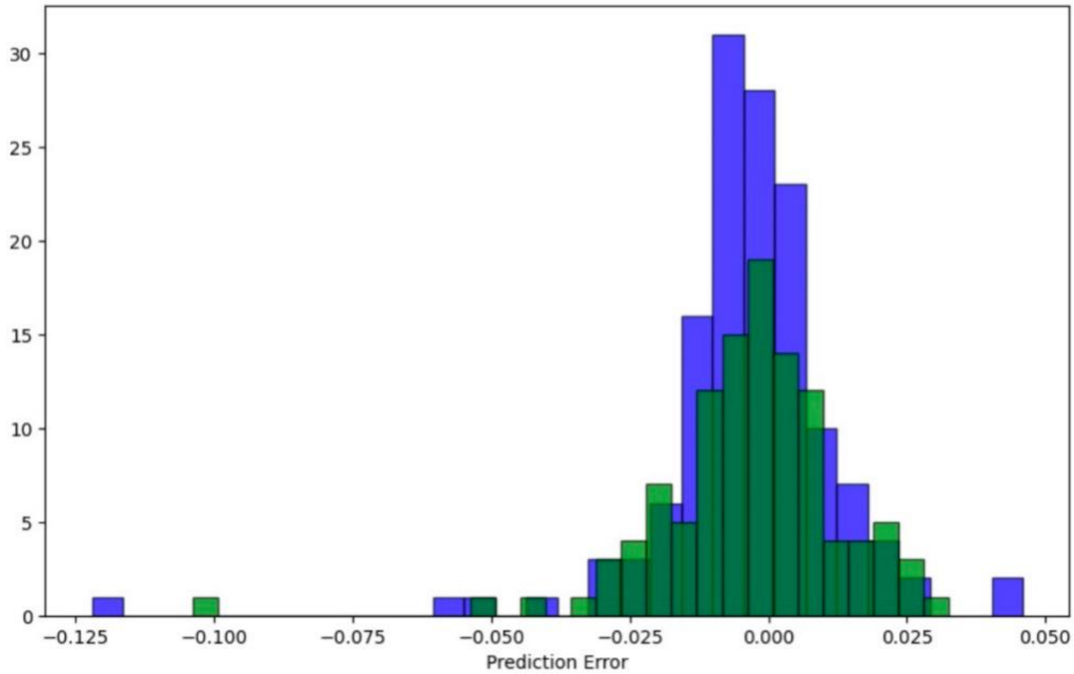
2)Determine if there is overfitting

- Learning curve analysis



Graph3.4 The learning curve graph shows the variation of the model's loss on the training and validation sets over training epochs (epochs). It can be observed that both the training loss (blue) and validation loss (orange) decrease rapidly in the initial stage, then tend to stabilize, and their curves almost overlap.Indicating that the model has not experienced overfitting.

- Prediction Error Distribution

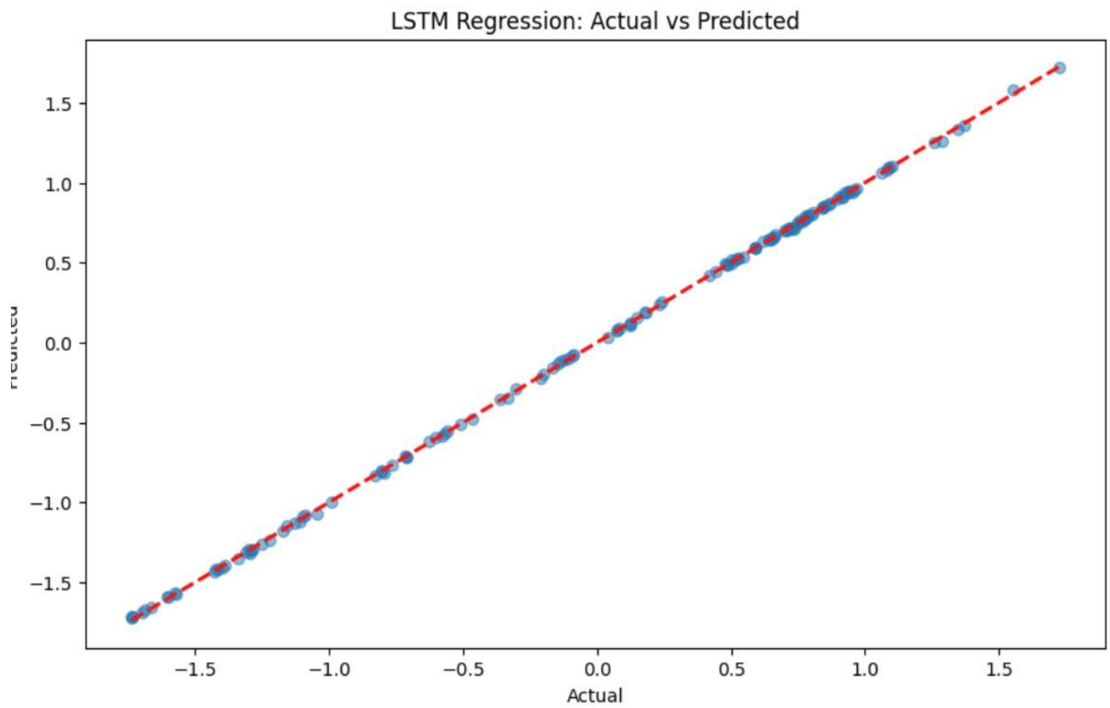


Graph 3.5 In this figure, the test set errors and validation set errors are compared. The error distribution chart shows the prediction error distribution of the model on the test set (blue) and validation set (green). Most of the errors are concentrated around zero, and the distribution of errors is relatively symmetrical and concentrated, indicating that the overall prediction error of the model is relatively small.Overall, the model did not overfit.

1. LSTM Model

1)Results Display

- Figure D shows the scatter plot of the true and predicted values of the LSTM model, where most of the scatter points are closely aligned with the reference line. This alignment indicates a strong correlation between the predicted results and the actual values. The majority of points fall precisely on the 45-degree reference line, suggesting minimal prediction error and high prediction accuracy. Only a few outliers deviate slightly from this line, reinforcing the LSTM model's effectiveness in capturing time series patterns. This result demonstrates the LSTM model's robust ability to accurately predict ESG data, highlighting its potential for reliable forecasting in this domain.

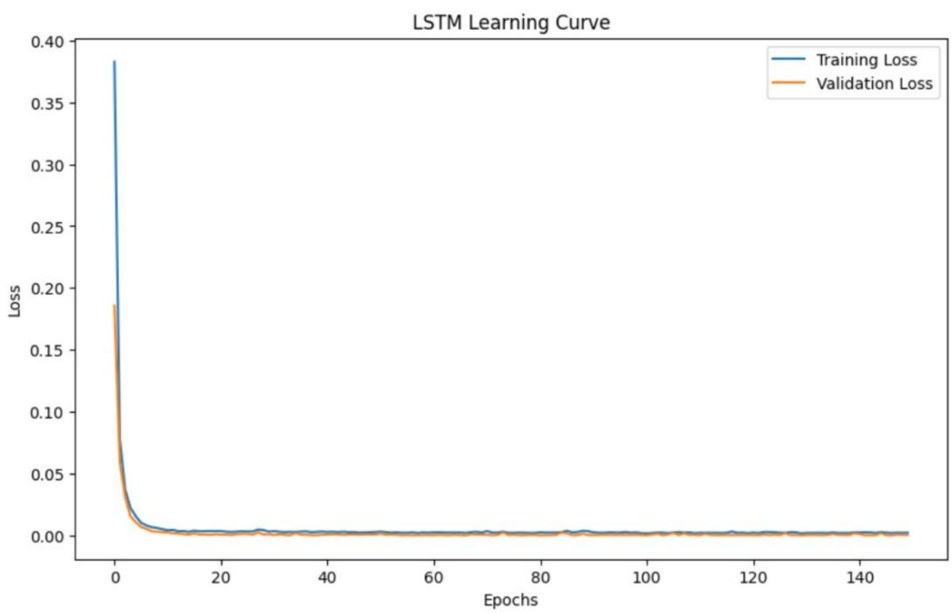


Graph3.6 A scatter plot that shows the relationship between the actual and predicted values from a Multilayer Perceptron (MLP) regression model. The x-axis represents the actual values, and the y-axis represents the predicted values.

- Evaluation of LSTM Model Results:
  - **MAE**: 0.0079
  - **MSE**: 0.0001

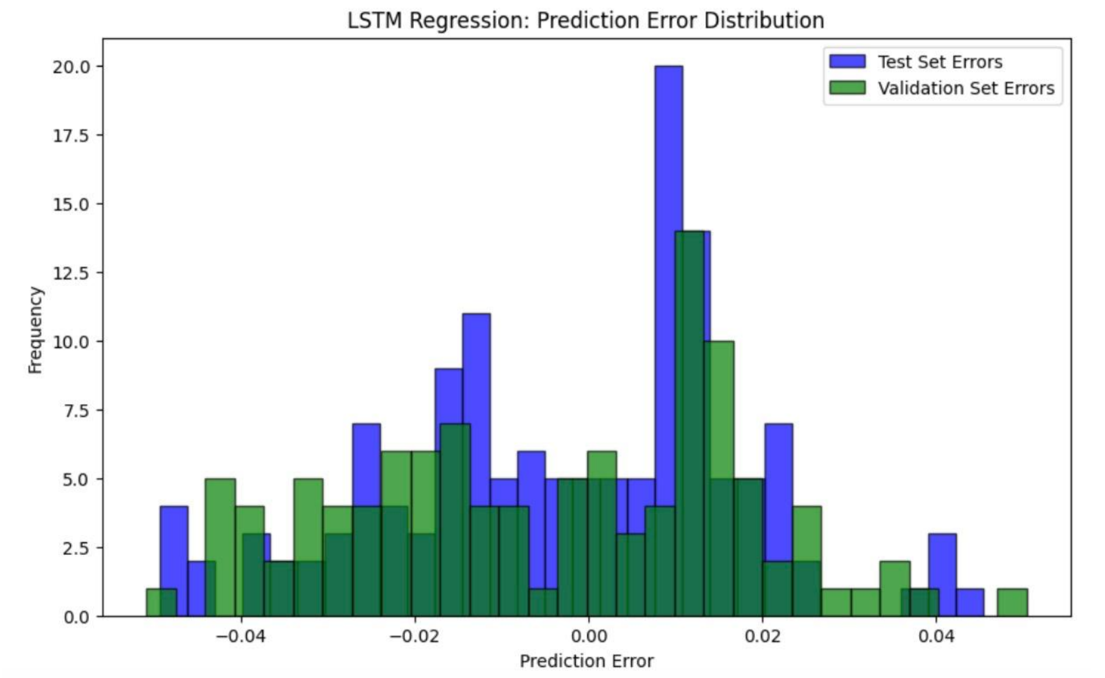
2)Determine if there is overfitting

- Learning curve analysis



Graph 3.7 Stability: The learning curve shows that both the training loss (blue) and validation loss (orange) gradually decrease during the training process, and after nearly 20 epochs, both tend to stabilize with relatively low loss values.
The training and validation losses are close: the curves of training loss and validation loss almost overlap, indicating that the model performs similarly on both the training and validation sets. Indicating that the model is not overfitting.

- Prediction Error Distribution

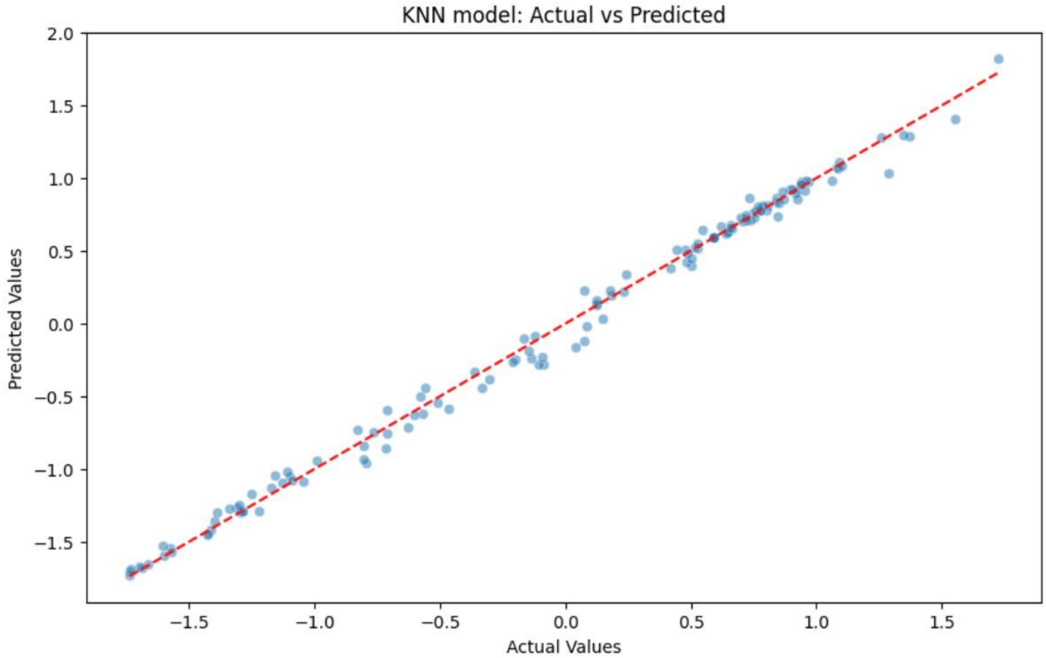


Graph3.8 Distribution overlap: From the error distribution map, there is a significant overlap in the distribution of test set errors (blue) and validation set errors (green), indicating that the error distribution of the model is similar on these two datasets.
Error concentration: Most errors are concentrated between 0 and 0.1, but there are also some points with larger errors, especially in the test set.

Overall, the model performs well on both the training and validation sets, with no obvious signs of overfitting.

3.KNN Model
1)Graphical results display
Graph E displays the scatter plot of the actual versus predicted values using the KNN model. While many points are relatively close to the 45-degree reference line, there is a noticeable deviation for some points, indicating that the KNN model's predictions are not perfectly aligned with the actual values.
Despite the fact that the majority of the points are near the reference line, the accuracy

of the KNN model appears to be lower compared to the LSTM or MLP models. This deviation may highlight the limitations of the KNN model, particularly in handling ESG data[41].



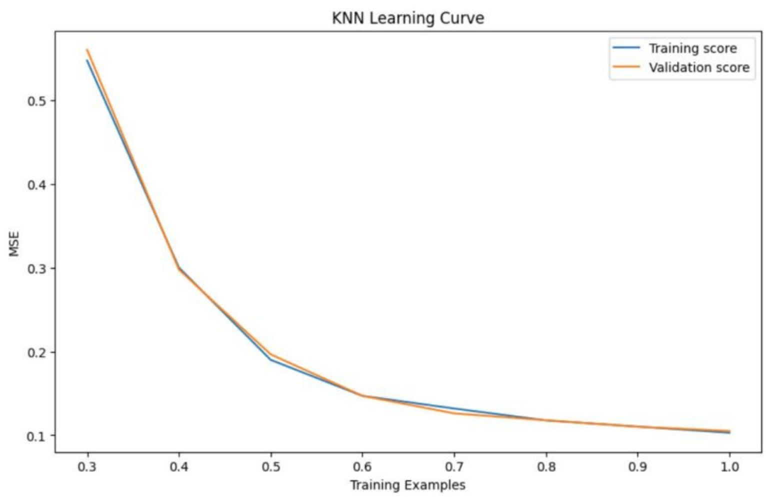
Graph3.9 A scatter plot that shows the relationship between the actual and predicted values from a Multilayer Perceptron (MLP) regression model. The x-axis represents the actual values, and the y-axis represents the predicted values.

2)Evaluation of KNN Model Results

- **MAE**: 0.0490
- **MSE**: 0.0048

3)Determine if there is overfitting

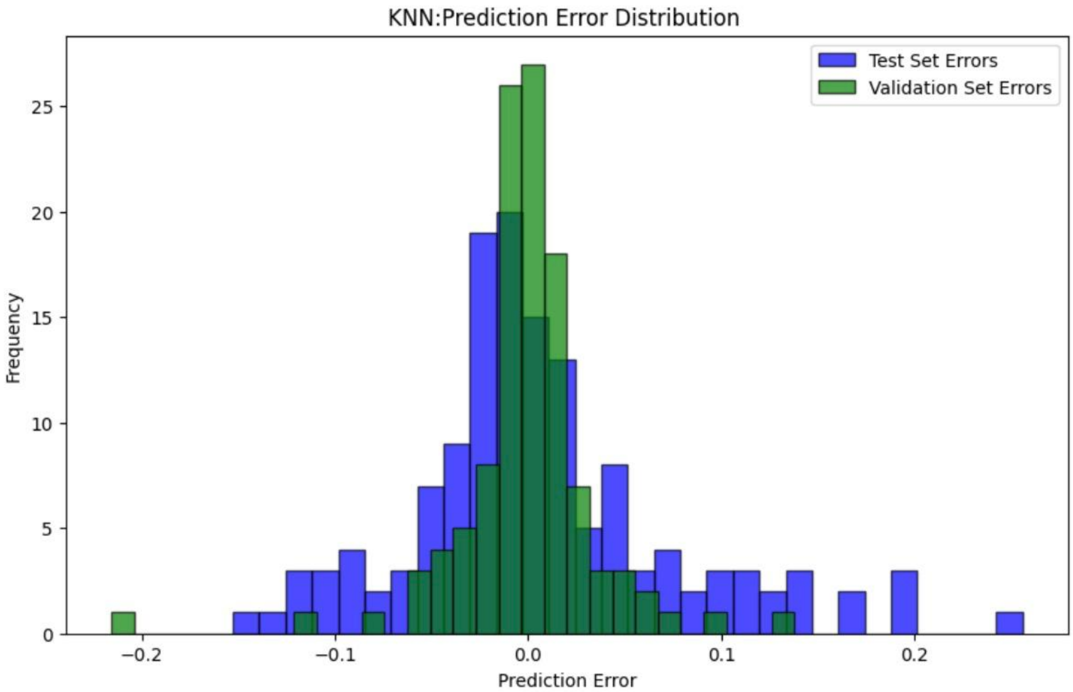
- Learning curve analysis



Graph3.10 The learning curve shows that as the number of training samples increases, the training error (blue) and validation error (orange) gradually decrease and eventually stabilize.

The two curves almost overlap, which means that the performance of the model on the training and validation sets is very similar, indicating that the model has good generalization ability and there is no obvious overfitting or underfitting.

- Prediction Error Distribution

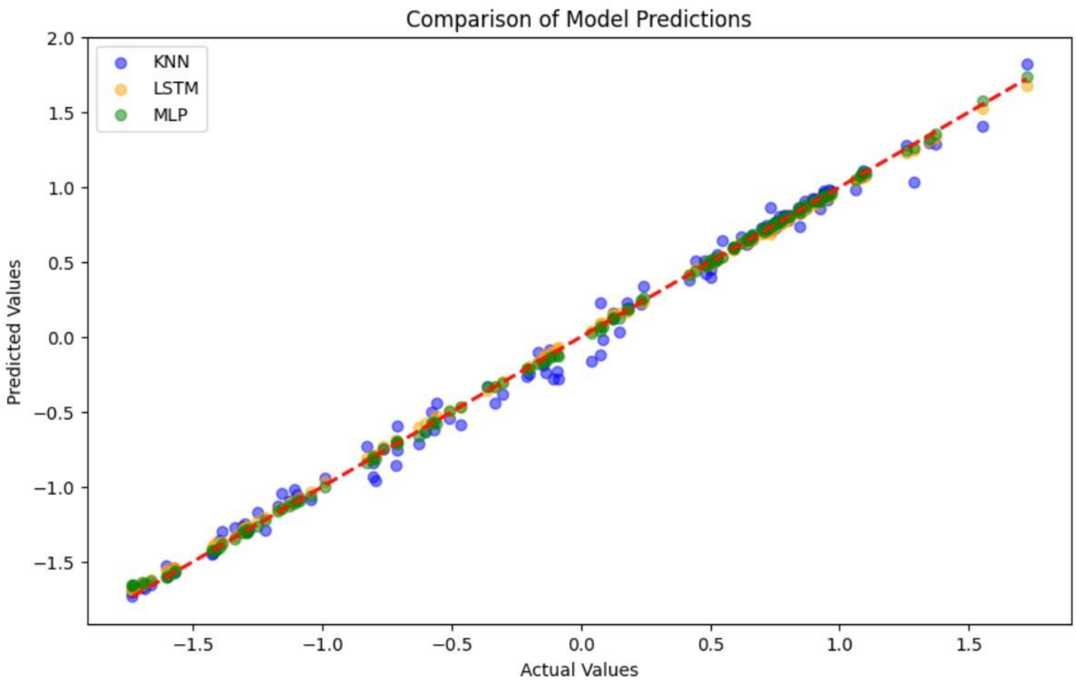


Graph3.11 The error distribution chart shows the prediction error distribution of the test set (blue) and validation set (green).

The error distribution roughly follows a normal distribution, with the majority of errors concentrated around zero, indicating a high prediction accuracy of the model.
The similarity in error distribution between the test set and the validation set further indicates that the performance of the model is consistent between the two.
Therefore, combining these two charts, the KNN model did not exhibit overfitting in predicting this ESG data, and the performance of the model on the training and validation sets was very similar and stable.

4.Comprehensive Analysis of Model Results
In this study, we predict the ESG (Environmental, Social, and Governance) scores of companies through three models: KNN, LSTM, and MLP. The KNN model is used as a benchmark to demonstrate the basic performance of traditional non-deep learning models, while deep models such as LSTM and MLP demonstrate significant advantages

when processing ESG data. The following is an analysis and synthesis of the performance of each model.



Graph3.12 Comparison of Forecast Results

Model	MAE	MSE
KNN	0.0490	0.0048
LSTM	0.0079	0.0001
MLP	0.0137	0.0003

Table 3.2 Model evaluation value

Compared and analyzed the performance of three different machine learning models (KNN, LSTM, MLP) in predicting ESG data. The results show that although all three models have some performance in prediction accuracy, there are significant differences in processing complexity and prediction accuracy.

- Comparison of Model Performance:

KNN model: The MAE of the KNN model is 0.0490 and the MSE is 0.0048, indicating relatively weak performance among the three models. Although the predicted results are close to the actual values, their errors are relatively large, indicating that KNN has certain limitations in handling ESG data with complex nonlinear relationships.

LSTM model: The LSTM model performs the best with a MAE of 0.0079 and MSE of 0.0001. The predicted results almost completely match the actual values, indicating that LSTM has strong capabilities in capturing time series patterns and processing complex data, and can provide highly accurate prediction results.

MLP model: The MLP model also performs well, with a MAE of 0.0137 and MSE of 0.0003. Although slightly inferior to the LSTM model, it is still better than the KNN model. MLP can effectively process multidimensional data and provide more accurate predictions.

Chart analysis:

- The scatter plot in the figure shows the comparison between the predicted values and actual values of three models. It can be clearly seen that the predicted values of LSTM and MLP models almost completely coincide with the actual values, while the predicted values of KNN model have a larger deviation from the actual values.The scatter distribution of the KNN model deviates significantly from the reference line, indicating that the prediction accuracy of this model is not as good as the other two models when facing complex ESG data.

Research conclusion:

In summary, the LSTM model demonstrates the strongest predictive ability in ESG data prediction tasks, accurately capturing complex patterns in the data. The MLP model ranks second and performs quite well, while the KNN model performs weakly in this task due to its insufficient ability to capture complex relationships.

Therefore, for ESG data prediction tasks, especially when dealing with complex and nonlinear data, LSTM and MLP models are better choices.

Although the KNN model has the advantages of simple calculation and easy understanding, its prediction accuracy is relatively low, and in the case of large data volume and high dimensionality, the computational complexity is high and it is easily affected by noise[38]. These limitations of the KNN model suggest that while it can provide a valuable reference point as a benchmark model, its ability to handle complex ESG data is insufficient.

LSTM models perform well in forecasting ESG data, mainly because they are effective at capturing long-term dependencies in time series. This capability of LSTMs makes them a powerful tool in ESG data sharing and sustainability analysis.

As a general deep learning model, the MLP model does not have the special ability of LSTM to process time series data, but it performs well in the processing of static features. Through the structure of multi-layer neural networks, MLP models can learn complex patterns in data, and can provide more accurate predictions even in the absence of clear time series relationships[40]. This flexibility and high training efficiency of MLP models make it another important tool in ESG data analysis.

Chapter 4

Comprehensive Framework for Data Sharing

4.1 Application of Deep Learning Models in ESG Data Sharing Platform Design

On the basis of the previously mentioned research, in the context of designing an ESG data sharing platform, we consider using LSTM and MLP models in data analysis and prediction. A significant advantage is presented by the LSTM model due to its strong capability in processing time series; thus offering the possibility of dynamically updating ESG scores[35], which in turn grants the platform's users access to evaluations of corporate sustainability performance that are more accurate, as well as timely. The comprehensive analysis provided by the MLP model results from its processing of data across multiple dimensions[41], which assists companies by giving support in multiple areas to optimize their ESG strategies.

4.2 The purpose of building a data sharing framework

The design of this framework aims to realize the advantages of deep learning models in ESG data analysis mentioned in previous research, and to build an efficient, transparent, and sustainable data sharing platform by combining blockchain technology. The combination of deep models and blockchain technology has formed a virtuous cycle for analyzing and sharing ESG data, providing optimization suggestions for enterprise development, and achieving sustainable ESG data sharing[42].

4.2.1 Dynamic Data Update and Time Series Prediction

The use of LSTM model: The above chapters indicate that LSTM model has significant efficiency in predicting ESG data. Therefore, in the design process, this framework combines LSTM model to dynamically analyze and predict ESG data related to enterprises. This model can capture long-term dependencies of data, ensuring that the latest ESG data can be used to update the platform's business performance in a timely manner, and ensuring that users receive predictions of future trends.

4.2.2 Multidimensional Data Processing

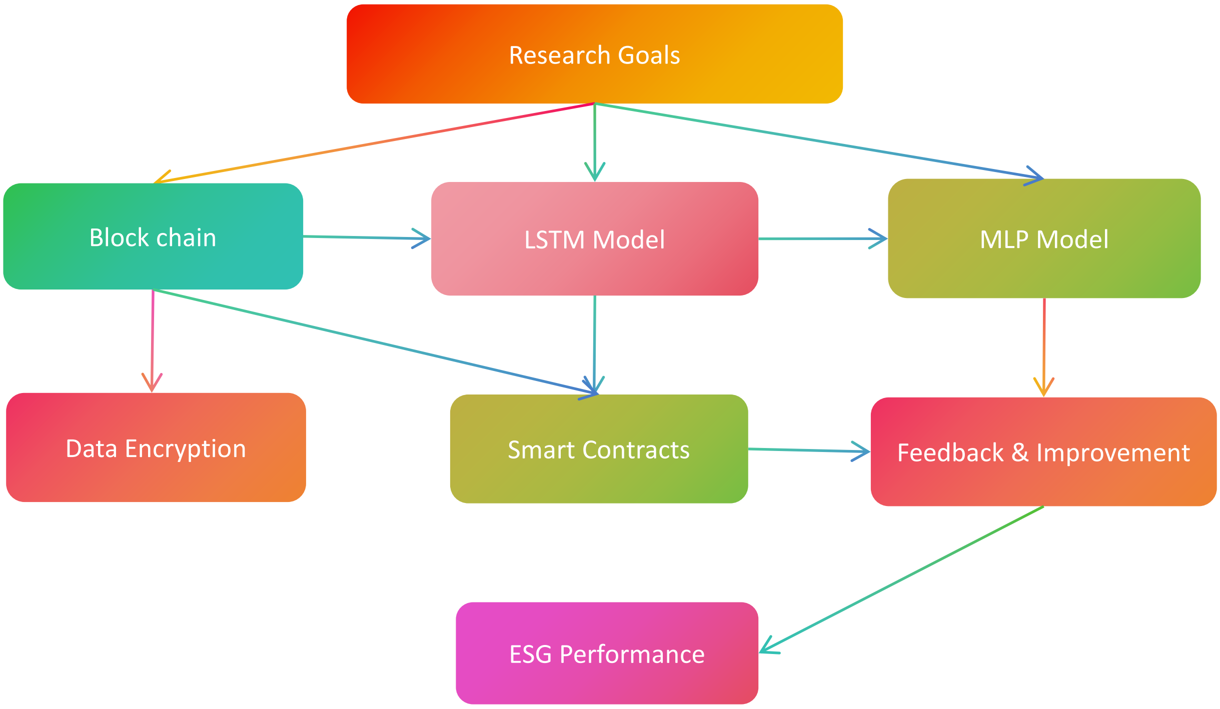
The MLP model, it has shown performance, very excellent, for processing ESG data of multidimensional nature, in studies done before. The framework design, therefore, it integrates the MLP model for analyzing and processing ESG data that enterprises have, in a comprehensive manner[41]. The detailed analysis provided by the platform, it can be through the MLP model, of enterprises in many dimensions, like environment, responsibility social, and governance corporate, helping to identify the areas that have potential for improvement, for those enterprises, and optimizing their ESG strategies[41].

4.2.3 Continuous improvement and feedback mechanism

In order to continuously optimize ESG performance, enterprises have designed a feedback mechanism based on a comprehensive framework of deep learning models and data sharing platforms. Enterprises upload data, the platform analyzes it, generates improvement suggestions, and sends feedback to the enterprise. This mechanism relies on the first analysis results of ESG data by LSTM and MLP models, which help companies continuously improve their ESG strategies and form an optimization cycle[35].

4.2.4 Decentralized data sharing platform

Decentralized architecture: The platform utilizes a distributed framework to guarantee transparent and equitable data sharing among various stakeholders. By eliminating the need for centralized intermediaries typically involved in data exchange, this approach minimizes operational risks and improves the reliability and trustworthiness of the data[43].



Graph4.1 ESG Data Sharing Platform Framework

4.3 Technical Implementation of Integrated Framework

4.3.1Blockchain technology

1. Reasons for ensuring data immutability during ESG data sharing process
- Ensuring data true and credible must be done: avoiding tamper, important it is. Investors and stakeholders, decisions based they are, on real, actual data.
- Decision risks prevented need to be: erroneous investment decisions must be avoided, economic losses due to tamper avoided.
- Trust enhanced must be: parties on platform, more trust should be there, sharing and using data greatly promoted.
- Regulations followed must be: enterprise compliance ensured ought to be, legal risks from data tampering should be avoided.

2. Implementation of blockchain technology

The invariance of data is crucial; Therefore, in order to ensure that ESG information remains unchanged during sharing, the framework utilizes the distributed ledger technology of blockchain. Blockchain technology also helps to improve visibility and make data clearer. All participants on the platform have the ability to view transactions and data on the blockchain; This visibility helps to reduce or even eliminate information asymmetry and enhance the credibility of corporate ESG performance. In addition, using smart contracts in blockchain can automatically execute protocols for sharing data, ensuring efficient and secure transmission of data between authorized users [43].

Blockchain technology also helps to improve data transparency and make data clearer. All participants on the platform have the ability to view transactions and data on the blockchain; This transparency helps to reduce or even eliminate information asymmetry and enhance the credibility of corporate ESG performance. In addition, using smart contracts in blockchain can automatically execute protocols for sharing data, ensuring efficient and secure transmission of data between authorized users [43].

4.3.2 Deep model application

1.Implementation of LSTM model:
In this framework, LSTM models are integrated to process enterprise ESG data with time series features. Firstly, standardize the data and convert it into a three-dimensional format suitable for LSTM model input, namely [sample size, time step size, feature number]. During the model training process, the random search CV method is used for hyperparameter tuning to ensure the optimal performance of the LSTM model in predicting ESG scores. LSTM effectively captures long-term dependencies through its gating mechanism, taking into account the long-term impact of historical data, thereby improving the accuracy of enterprise ESG rating predictions, which is particularly important for dynamic data updates and time series analysis.

2.Implementation of MLP model:
The MLP model can be used on this platform to analyze the static characteristics of enterprises in multiple dimensions such as environmental governance, corporate governance, and social responsibility. Through the application of hyperparameter tuning and dropout mechanism, the platform has optimized the performance of the MLP model while preventing overfitting, enabling it to extract meaningful features from high-dimensional data[40]. This method ensures the accuracy of ESG ratings, helps companies better understand their performance in sustainable development, and provides strong support for developing improvement measures.

3.The specific implementation of data privacy and security protection:
In this framework, to protect the privacy of ESG data during sharing, the platform can use Advanced Encryption Standard (AES) for symmetric encryption. This encryption technology encrypts data before it is uploaded to the platform, ensuring that only authorized users holding the corresponding key can access and decrypt the data, thereby preventing the leakage of sensitive information and safeguarding the company's market position and legal compliance[43]. In addition, in scenarios that require higher security, the platform can also use RSA algorithm to perform asymmetric encryption on transmission

keys, further ensuring the security of data during transmission and storage, preventing data leakage and unauthorized access.

User permission management is implemented through role-based access control (RBAC) mechanism, where the platform assigns different access permissions to users based on their roles (such as data provider, data consumer, platform administrator). Smart contracts are used to automatically record and manage permission changes, ensuring transparency and compliance in data access[43]. This dual mechanism not only protects the privacy and security of data, but also enhances the fairness of the platform and the trust of users, thereby promoting more data contributions.

4.Design and Implementation of Smart Contracts:
In the platform, there used, smart contracts, to automate processes of data protocols for sharing, the operations such as upload of data, storage, verification, access, included are. For instance, a company, when new ESG data it uploads, the smart contract automatically verifies authenticity and it records on blockchain the data. From analysis results deep learning models provided, automatic feedback the smart contracts generate in reports, performance scores of enterprises environmental, in social responsibility, governance aspects are included, additionally identifying needing improvement areas. Analysis according to, smart contracts specific improvement suggestions too they provide[43]; encompassing measures environmental enhancement or governance corporate optimization, to help companies optimize continuously ESG performance.

4.4 Application scenario assumptions

4.4.1 Assuming the nature of the scene

To demonstrate the implementation process of the above comprehensive framework. This study proposes a hypothetical application scenario aimed at demonstrating how the designed ESG data sharing comprehensive framework can help companies predict and optimize their ESG scores in practice. It should be clarified that this scenario is based on a virtual enterprise called "GreenFoods" and aims to clearly illustrate the core functions and potential application methods of the data sharing platform. Through this scenario, we hope to demonstrate the comprehensive performance of the platform in data collection, deep learning model prediction ability, blockchain technology application, automated feedback of smart contracts, and ESG score optimization process.

4.4.2 Rationality and construction basis

1.Scene background
GreenFoods is a global food manufacturing company with multiple production and processing bases located in different countries. As a typical large multinational

corporation, GreenFoods faces complex ESG management challenges. In order to maintain a leading position in the fierce market competition and meet the high requirements of investors and regulatory agencies for sustainable development, GreenFoods has decided to predict and optimize its ESG score by introducing the ESG data sharing platform designed in this study.

Among them, the challenges faced by this company are：

- Environmental protection: It is necessary to predict the carbon emission trends of each production base in the coming years to avoid potential compliance risks.
- Social responsibility: Predicting the impact of employee welfare measures on employee satisfaction and turnover rates to improve social responsibility scores.
- Corporate governance: Predict the impact of governance structure adjustments on corporate governance transparency and shareholder trust, and optimize governance scores.

2. Platform usage and data uploading

GreenFoods has decided to use an ESG data sharing platform to manage and optimize its ESG scores. The company has collected detailed ESG data from various production bases, including:

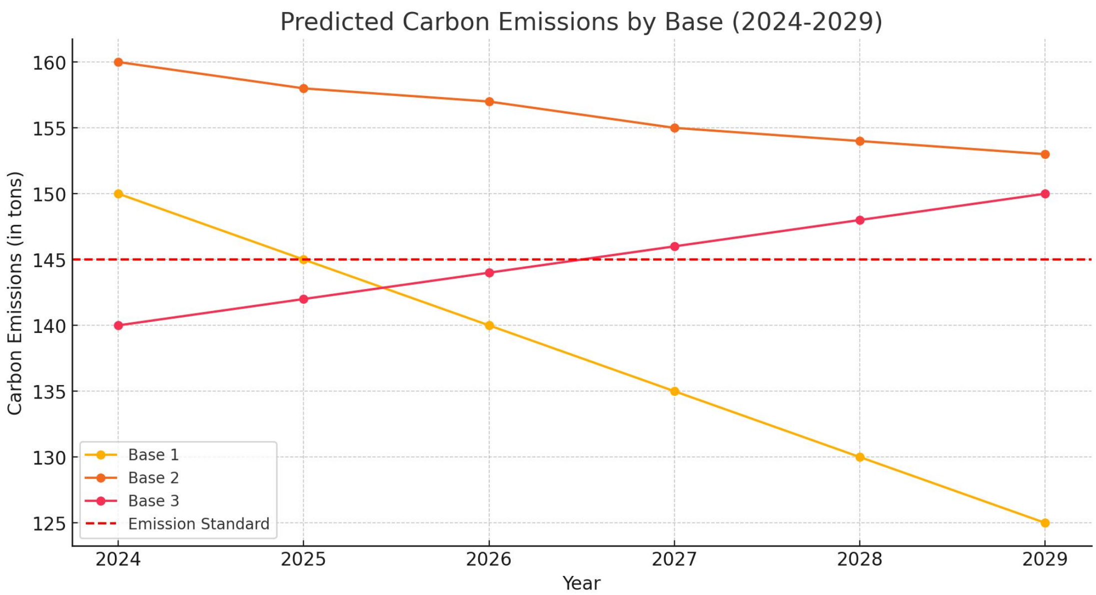
- The annual carbon emissions of each base and the expected environmental measures for the next few years.
- Employee welfare plan, employee satisfaction survey results, and turnover rate data.
- Data related to board structure, transparency reporting, and governance.

All data is encrypted with AES before being uploaded to the platform to ensure data privacy. The uploaded data is recorded on the blockchain, utilizing the immutability and transparency of blockchain technology to ensure the integrity and reliability of the data.

3.ESG score prediction

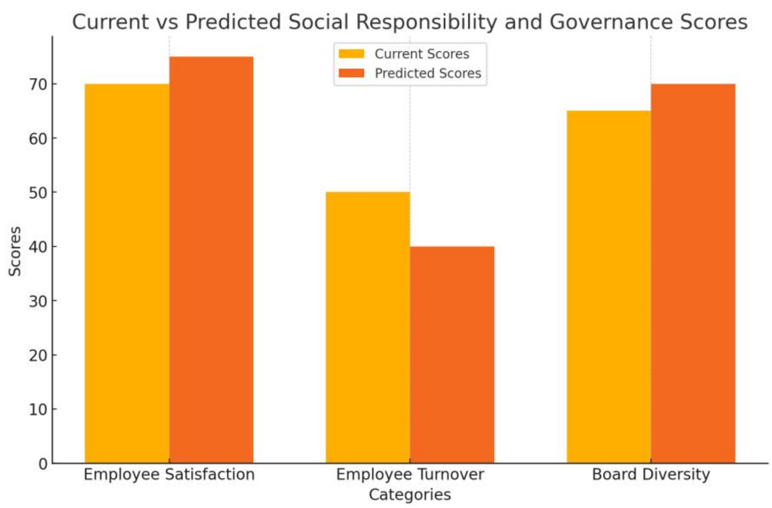
The platform uses deep learning models to predict GreenFoods' ESG scores and provide data support for its optimization measures:

LSTM model analysis: The LSTM model conducted time series analysis on the historical carbon emission data of GreenFoods, assuming the "Predicted Carbon Emissions of Each Base" chart. By combining environmental measures in the plan, the model predicts the future carbon emission trends of each base. The chart visually illustrates the changes in carbon emissions in the coming years and points out that without further measures, the carbon emissions of certain bases may exceed the standard. Based on these predictions, the platform estimated that the environmental scores of these bases may decrease and recommended necessary improvement measures to prevent potential score reductions.



Graph4.2 Assumed 'Predicted Carbon Emissions for Each Base' Chart(The data shown in the figure are all virtual data and do not have authenticity, limited to the assumption of this scenario)

MLP model analysis: The MLP model analyzed GreenFoods' social responsibility and corporate governance data, assuming the current and predicted social responsibility and governance scores as shown in the chart. The model predicts the impact of employee welfare improvement on satisfaction and turnover rates, as well as the impact of governance structure changes on transparency and trust. The chart compares the current score with the predicted future score, showing possible improvements in social responsibility and governance. These predicted results were subsequently integrated to forecast GreenFoods' overall ESG score. The model shows that by strengthening employee training and enhancing governance transparency, individual scores and overall ESG scores can be significantly improved.



Graph4.3 Assuming current and predicted social responsibility and governance scores(The data shown in the figure are all virtual data and do not have authenticity, limited to the assumption of this scenario)

4.Feedback and suggestions generated by smart contracts

The smart contract is based on the prediction results of LSTM and MLP models, automatically generates feedback reports, and provides the following improvement suggestions to GreenFoods:

- Environmental protection: It is recommended to increase the use of green energy in high-risk areas and introduce more efficient wastewater treatment technologies to achieve the platform's predicted carbon emission targets, thereby improving environmental scores.
- Social responsibility: It is recommended to increase employee benefits and training, especially in bases with high predicted turnover rates, to improve social responsibility scores.
- Corporate governance: It is recommended to add more independent directors and diverse members to the board of directors to optimize the corporate governance score.

5. Implementation and continuous improvement

GreenFoods has implemented multiple improvement measures based on feedback and suggestions provided by the platform:

- Solar facilities have been installed in the base where carbon emissions are predicted to exceed the standard, and the water resource management system has been optimized.
- Added employee benefits and training programs, and conducted regular employee satisfaction surveys to ensure the achievement of the predicted social responsibility score.
- Adjusted the composition of the board of directors, increased the proportion of independent directors, and regularly disclosed the minutes of board meetings, in line with the platform's predicted governance optimization goals.

The platform monitors and analyzes the effectiveness of these measures through regularly uploaded new data, updates forecast results in real-time, and ensures that GreenFoods can continuously optimize its ESG performance based on the latest forecasts.

6. Final result

After a year of continuous optimization, GreenFoods' actual ESG score is in line with the platform's predicted upward trend:

- Environment: Carbon emissions have decreased by 15%, and the environmental score has significantly improved.
- Social responsibility: The employee turnover rate has decreased by 10%, and the social responsibility score has increased.

- Corporate governance: The transparency and diversity of the board of directors have been improved, and the governance score has increased.

Through accurate score prediction and continuous optimization, GreenFoods has successfully improved its ESG performance and attracted more green investments.

4.4.3 Assuming rationality of the scenario

It should be clarified that this application scenario is a hypothetical case, designed to demonstrate the potential application and functionality of ESG data sharing platforms in practice. GreenFoods and its related data are all virtual constructions, and the main purpose of the scenario is to demonstrate the actual operation and effectiveness of the platform through theoretical models.

To ensure the rationality of this hypothetical scenario, the data types and operation methods in the scenario are based on real-world business practices and ESG management needs. The application of deep learning models, the immutability of blockchain technology, and the automated feedback mechanism of smart contracts are currently feasible and practical technologies and methods.

4.4.4 Limitations

Although this hypothetical scenario provides an effective framework to showcase the platform's functionality, its virtual nature brings certain limitations. The complexity, quality, and volatility of real enterprise data may affect the platform's performance in practice. The prediction results of deep learning models in ideal data environments may be affected in reality due to incomplete data, noise, or other unpredictable factors. In addition, the effectiveness of implementing suggested measures by enterprises in practice may vary due to changes in internal resources, market conditions, or policy regulations.

These limitations may have an impact on the research conclusions, therefore, further verification and adjustment are needed in practical applications to ensure that the platform can also achieve the expected results in real environments. Future research and practical testing will help further validate the effectiveness of the framework and optimize its practical applications.

Chapter 5

Conclusion

5.1 Research Conclusion

This study proposes and validates an ESG data sharing framework that combines deep learning models (LSTM and MLP) with blockchain technology. Using application scenario assumptions, it demonstrates how these technologies can enhance the efficiency of ESG data analysis and the sustainability of data sharing. In the deep learning model section, the LSTM model performs the best in ESG score prediction by capturing long-term dependencies in time series data, achieving the lowest mean square error (MSE 0.0001) and mean absolute error (MAE 0.0079). This indicates that the LSTM model can accurately capture the complex patterns of ESG ratings changing over time, making it particularly suitable for predicting dynamic data.

The MLP model also performs well in handling multidimensional static data, although its prediction accuracy is slightly inferior to LSTM, it is still superior to traditional KNN models when dealing with complex ESG feature relationships. The MSE of the MLP model is 0.0003 and the MAE is 0.0137, which proves its powerful ability in multidimensional data analysis.

This study also explores the application of blockchain technology in ESG data sharing. Blockchain technology ensures the security and transparency of ESG data during the sharing process within the entire integrated framework through its decentralized and tamper proof features. By combining the use of smart contracts, a sustainable virtuous cycle of ESG data sharing has been achieved.

5.2 Limitations of the research

This study has made significant progress in ESG data analysis and sharing, but there are still some limitations. Firstly, the dataset used in this study mainly comes from the publicly available Kaggle platform. Although it covers multiple industries and regions, it may not fully represent the complexity and diversity of global enterprises.

We only proposed reasonable application scenario assumptions in the comprehensive framework section. In future research, the framework combining deep learning and blockchain proposed in this study can be applied to actual enterprises, and its feasibility and effectiveness can be verified through empirical analysis. This can not only provide support for improving the ESG performance of enterprises, but also provide data support and theoretical basis for industry standardization and policy-making.

References

[1] Friede, G., Busch, T., & Bassen, A. (2015). ESG and financial performance: Aggregated evidence from more than 2000 empirical studies. Journal of Sustainable Finance & Investment, 5(4), 210-233.
[2]Eccles, R. G., Ioannou, I., & Serafeim, G. (2014). The impact of corporate sustainability on organizational processes and performance. Management Science, 60(11), 2835-2857.
[3] Khan, M., Serafeim, G., & Yoon, A. (2016). Corporate sustainability: First evidence on materiality. The Accounting Review, 91(6), 1697-1724.
[4] Sustainalytics. (2020). 2020 ESG trends to watch.
[5] MSCI. (2021). ESG trends 2021. Retrieved from https://www.msci.com/www/blog-posts/esg-trends-2021/02234444434
[6] World Economic Forum. (2020). Measuring stakeholder capitalism: Towards common metrics and consistent reporting of sustainable value creation. Retrieved from https://www.weforum.org/reports/measuring-stakeholder-capitalism
[7]Taylor, J., & Ling, Q. (2023). Ensuring data security in deep learning processes for ESG data. Journal of Secure Computing, 35(1), 15-30.
[8]White, G., & Kumar, S. (2021). Privacy protection in ESG reports using natural language processing (NLP). Journal of Data Privacy and Security, 24(2), 50-65.
[9]White, A., & Heckman, R. (2023). Enhancing trust in ESG data-sharing networks using blockchain technology. Journal of Blockchain Applications, 30(1), 22-37.
[10]Priya, R., Singh, A., & Gupta, K. (2021). Ensuring transaction security and anonymity in blockchain using public and private keys. Journal of Blockchain Technology, 5(3), 155-170.
[11]Gonzalez, R., & Schmidt, T. (2021). The impact of General Data Protection Regulation (GDPR) on European ESG data practices. European Journal of Data Protection, 34(2), 89-103.
[12]Ma, Q. (2023). The role of ESG data in sustainable portfolio management: Insights for investors and businesses. Journal of Portfolio Management, 49(1), 47-61.
[13]Li, Y., & Li, X. (2023). Inconsistent industry standards: The need for standardized ESG metrics and reporting frameworks. Journal of Industry Standards, 22(3), 67-82.
[14]Taylor, J., & Ling, Q. (2023). Ensuring data security in deep learning processes for ESG data. Journal of Secure Computing, 35(1), 15-30.
[15] Sfetcu, N. (2022). Blockchain as a continuously growing list of records: Ensuring data consistency and immutability through encrypted messages. Journal of Cryptographic Engineering, 10(1), 45-59.
[16]Priya, R., Singh, A., & Gupta, K. (2021). Ensuring transaction security and anonymity in blockchain using public and private keys. Journal of Blockchain Technology, 5(3), 155-170.
[17]Lee, H., Kim, J., Park, S., & Choi, K. (2022). An integrated approach using machine learning and deep learning algorithms to analyze ESG data: Methods for predicting

specific firm's ESG rankings. Journal of Sustainable Finance, 15(4), 223-245.
[18]Gamlath, R., Fernando, S., & Jayawardena, N. (2023). Generating ESG ratings from financial and textual data using machine learning models: Towards a comprehensive automated ESG rating system. Journal of Financial Data Science, 18(2), 98-115.
[19]Munappy, A., Loke, S. W., & Palaniswami, M. (2019). Challenges in data management for deep learning: The need for consistency, accuracy, and completeness in ESG data. Journal of Data and Information Quality, 11(3), 56-72.
[20]Sokolov, D., Ivanov, P., & Petrov, A. (2021). Enhancing ESG scoring systems with deep learning models like BERT: Improving relevance and accuracy in ESG data analysis. Journal of Applied AI in Finance, 10(1), 34-49.
[21]Chen, Y., & Liu, X. (2020). An automated machine learning approach using ESG scholar data to quantify and capture the ESG premium of companies. Journal of Financial Engineering, 7(1), 21-35.
[22] Franco, M., Rodriguez, L., & Smith, J. (2020). The importance of ESG factors in investment decisions: Efficiency of machine learning algorithms in linking ESG profiles with financial performance. Investment Analysis Journal, 24(3), 145-162.
[23]White, A., & Heckman, R. (2023). Enhancing trust in ESG data-sharing networks using blockchain technology. Journal of Blockchain Applications, 30(1), 22-37.
[24]Mafakheri, F., Nasiri, F., & Trienekens, J. (2018). How blockchain facilitates direct data sharing between institutions: Eliminating the need for third-party intermediaries. Journal of Information Technology & People, 31(2), 318-336.
[25]Anderson, S., & Zion, P. (2022). Collaborative ESG data analysis using secure multi-party computation techniques. Journal of Sustainable Computing, 45(3), 102-115.
[26]Hiroshi, Y., & Hirobumi, N. (2022). Overcoming ESG data collection challenges through deep learning integration. Journal of Environmental Data Science, 27(4), 58-72.
[27]Martin, L., & Reza, M. (2022). Real-time analysis of ESG data using deep learning. Journal of Sustainable Investment, 12(1), 45-60.
[28] Sahu, R., Verma, K., & Patel, M. (2021). Applications of deep learning in sustainable manufacturing. Journal of Industrial Sustainability, 16(2), 100-115.
[29]Stallings, W., & Brown, L. (2012). Computer Security: Principles and Practice. Prentice Hall.
[30]Jouini, M., Rabai, L. B. A., & Aissa, A. B. (2014). Classification of Security Threats in Information Systems. Procedia Computer Science, 32, 489-496. DOI: 10.1016/j.procs.2014.05.452.
[31]Kshetri, N. (2017). Blockchain's roles in meeting key supply chain management objectives. International Journal of Information Management, 39, 80-89. DOI: 10.1016/j.ijinfomgt.2017.12.005.
[32]Hyndman, R. J., & Athanasopoulos, G. (2018). Forecasting: principles and practice. OTexts.
[33]McKinney, W. (2017). Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython. O'Reilly Media.
[34]Hinton, G. E., & Salakhutdinov, R. R. (2006). Reducing the Dimensionality of Data with Neural Networks. Science, 313(5786), 504-507. DOI: 10.1126/science.1127647.
[35]Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory. Neural

Computation, 9(8), 1735-1780. DOI: 10.1162/neco.1997.9.8.1735.
[36]Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: A Simple Way to Prevent Neural Networks from Overfitting. Journal of Machine Learning Research, 15(1), 1929-1958.
[37]Cover, T., & Hart, P. (1967). Nearest neighbor pattern classification. IEEE Transactions on Information Theory, 13(1), 21-27.
[38]Altman, N. S. (1992). An Introduction to Kernel and Nearest-Neighbor Nonparametric Regression. The American Statistician, 46(3), 175-185.
[39]Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, É. (2011). Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, 12(Oct), 2825-2830.
[40]Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep Learning. MIT Press.
[41]Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. Nature, 323(6088), 533-536.
[42]Bocken, N. M. P., Short, S. W., Rana, P., & Evans, S. (2014). A literature and practice review to develop sustainable business model archetypes. Journal of Cleaner Production, 65, 42-56.
[43] Zheng, Xie, Dai, Chen, & Wang (2018) on blockchain technology and decentralized systems.

Appendix 1

- **The specific content of the data**

Basic company information: The dataset includes basic information such as company name, stock code (Ticker), industry classification, and company website. These pieces of information help researchers identify and classify different companies, and provide a foundation for further analysis.

Environmental dimension data:
Environment Score: quantifies a company's performance in environmental protection and sustainable development.
Environment Grade: A company's environmental performance evaluation expressed in the form of letter grades (such as A, B, C, etc.).
Environment Level: Classification represents the level of environmental performance (such as high, medium, low).

Social dimension data:
Social Score: measures a company's performance in fulfilling its social responsibilities, such as employee benefits, community impact, etc.
Social Grade: Reflects a company's overall performance in social responsibility through letter grades.
Social Level: Refers to the classification level of social responsibility performance.

Governance dimension data:
Governance Score: Evaluating the soundness and transparency of a company's governance structure.
Governance Grade: Displays a company's performance in corporate governance in alphabetical order.
Governance Level: The classification level of governance performance.

Comprehensive ESG score:
Total Score: A comprehensive score that considers the three dimensions of environment, society, and governance, representing the company's overall performance in ESG.
Total Grade: A comprehensive ESG rating represented by letter grades.
Total Level: The classification level of overall performance.
Timestamp information: Each data has a Processing Date, indicating the last update time of the data. This is particularly important for time series analysis, as it can help researchers track and analyze changes in a company's ESG performance over time.

- The role of data in research

This dataset played a central role in this study, supporting the training and validation of deep learning models such as LSTM and MLP. These models can effectively predict a company's future ESG ratings and identify key factors that affect these ratings by analyzing and processing multidimensional information in the dataset. In addition, the dataset provides important basis for exploring the application of blockchain technology in ESG data sharing. By combining deep learning and blockchain technology, the research aims to build an efficient, transparent, and sustainable ESG data analysis and sharing platform. This platform helps companies improve their ESG performance, promote data transparency, and support more accurate investment decisions.

- Reason for choosing this dataset

The main reasons for choosing this dataset over other datasets are as follows:

Wide coverage: This dataset covers listed companies from multiple industries and regions worldwide, with a wide range of data sources and representativeness. This makes the research results have strong universality and reference value, which can provide guidance for companies in different industries and regions.

Time span and data integrity: The dataset contains years of company ESG ratings and related information, making it suitable for time series analysis. Especially in this study, the LSTM model needs to utilize historical data to capture long-term dependencies, and this dataset provides sufficient time span to support such analysis.

Rich feature dimensions: The dataset not only includes ratings for environment, society, and governance, but also classification information such as corresponding levels and levels. These multidimensional features provide rich inputs for the MLP model, which can better capture the complexity of corporate ESG performance.

Openness and Transparency: As the dataset comes from the Kaggle platform, it is open and accessible, allowing researchers to easily access and validate the data, which helps to improve the transparency and reproducibility of research.

- Complete dataset: https://github.com/Alicia-JH/Data-about-Thesis-

Appendix 2

LSTM model: LSTM is a recursive neural network (RNN) designed specifically for processing time series data, with the ability to capture long-term dependencies. LSTM controls the flow of information through its unique gating mechanisms, such as input gates, forget gates, and output gates, enabling it to perform excellently in processing data with long-term dependencies and complex temporal dynamics. In ESG data analysis, LSTM models are particularly suitable for capturing long-term trends and dependencies in time series, as a company's ESG performance changes over time and historical performance may have a significant impact on future ratings.

MLP model: MLP is a classic feedforward neural network suitable for handling complex nonlinear relationships between multi-dimensional features. Due to the fact that ESG scoring involves multiple dimensions (environment, society, governance), there may be complex interactions between these dimensions, and MLP models can capture these interaction effects through multi-layer nonlinear transformations. In addition, MLP also performs well in handling static data and is suitable for feature extraction and classification tasks that do not have significant temporal dependencies. Therefore, the MLP model was used in this study to analyze the various dimensions of corporate ESG ratings and capture the nonlinear relationships within them.

The selection of LSTM and MLP models is not only based on their respective advantages in specific data types, but also to form complementarity between models: LSTM handles temporal dependencies, while MLP handles multidimensional static features, making ESG rating prediction more efficient.

Appendix 3

During the training process of the model, we used the Random Search Cross Validation (Random Search CV) method for hyperparameter tuning. Compared with traditional grid search, random search can more effectively explore high-dimensional hyperparameter spaces, especially exhibiting higher efficiency in situations where computing resources are limited.

- The theoretical basis for hyperparameter selection: In LSTM models, hyperparameters such as the number of LSTM units, learning rate, and dropout rate have a particularly critical impact on model performance. The number of LSTM units determines the capacity and expressive power of the model, while the learning rate affects the convergence speed and stability of the model. Dropout rate is used to prevent overfitting by randomly ignoring the output of some neurons during the training process, thereby improving the model's generalization ability. Similarly, in the MLP model, the number of hidden layers and the number of neurons in each layer directly affect the complexity and predictive performance of the model. By optimizing these hyperparameters through random search, we can maximize the predictive accuracy and robustness of the model.