Scholar Publishing Group
Machine Learning Theory and Practice
https://doi.org/10.38007/ML.2025.050105
ISSN 2790-0983 Vol. 5, Issue 1: 44-52

Construction and LSTMM Modelling of Financial Indicator Glittering Dataset for Financial Sector from Machine Learning Perspective

Xinyi Jiang

School of Economics and Management, Nanjing University of Technology, Nanjing 210000, Jiangsu, China

Keywords: Financial whitewash; machine learning; financial indicators; dataset construction; LSTMM models

Abstract: Financial whitewash, as the behaviour of enterprises to conceal their true operating conditions through manipulation of financial data, not only undermines market fairness, but also increases the systemic risk faced by investors. The traditional method of relying on financial ratio analysis has obvious limitations in dealing with the complex, changing, and hidden means of financial modelling. For this reason, this paper constructs a financial index whitewash identification dataset applicable to the financial sector based on the machine learning perspective, and proposes an LSTMM model that integrates the long and short-term memory network (LSTM) and the multilayer perceptual machine (MLP) for the identification study.In terms of data construction, this paper is based on public violation cases and multi-dimensional financial data, screening out the whitewash and non-whitewash samples, and combining the enterprise history and industry comparison data to form the indicator dataset with time-series characteristics. In terms of modelling, the LSTMM model integrates the temporal feature extraction capability of LSTM with the nonlinear representation advantage of MLP, which significantly improves the recognition accuracy and stability. The experimental results show that the proposed method outperforms the traditional model in a number of performance indicators, which validates its application value in financial whitewash identification.This study provides an effective method for intelligent analysis and anomaly detection of financial data, which is of practical significance for improving the level of financial regulatory technology and market risk prevention and control capability.

1. Introduction

As the capital markets continue to develop, the role of corporate financial information in investment decision-making and risk assessment has become increasingly prominent. However, some listed companies, in order to evade regulation, embellish performance, or mislead investors, frequently manipulate financial statements through methods such as inflating revenue and concealing liabilities, resulting in serious financial falsification issues. Such behaviors not only disrupt market order and harm investors' interests but also place higher demands on regulatory systems. In recent years, the outbreak of several financial fraud cases both domestically and internationally has further sparked ongoing academic and practical attention to methods for identifying financial falsification.

Traditional identification methods rely largely on financial ratio analysis and audit experience, but as the methods of falsification become more covert and complex, the ability of these methods to detect such behavior has become increasingly limited. Against this backdrop, machine learning technologies, with their strong feature extraction and pattern recognition capabilities, offer new pathways for identifying financial falsification. However, existing research mainly focuses on the overall company or statement-level judgment of falsification and lacks a systematic identification and analysis of falsification behavior at the level of specific financial indicators.

This study addresses the practical needs of financial regulation and risk control, using data from listed companies in the Chinese capital market to construct a dataset for identifying falsified financial indicators. It proposes an LSTMM model, which integrates Long Short-Term Memory Networks (LSTM) and Multi-Layer Perceptron (MLP) for multi-label classification tasks to identify falsified indicators. Through comparison experiments with various mainstream machine learning models, this method has demonstrated superior performance in terms of identification accuracy and practicality.

This study not only enriches the technical framework for identifying financial falsification but also provides more targeted and forward-looking technological support for regulatory agencies, investors, and auditors, making a significant contribution in both theoretical and practical terms.

2. Related Work

Mahesh B emphasized that as a versatile technology, machine learning demonstrates unique advantages in financial forecasting, risk management, and other fields through automatic feature extraction, promoting the intelligent development of the financial sector. Kelly B **错误!未找到引用源。**’s review summarized the early results of machine learning in financial market research, presented several outstanding cases, and pointed out future research directions. This review targets both financial economists and machine learning experts, fostering interdisciplinary communication and collaboration. Nazareth N, based on the PRISMA systematic review methodology, revealed the trends in the application of machine learning across six major financial domains, particularly highlighting cross-market sensitivity analysis as a key research direction for the future.

In terms of specific methods, Dudek G pointed out that previous studies significantly improved prediction accuracy by decomposing multi-period load forecasting problems and applying local neural networks, a concept that inspires us to effectively improve the performance of LSTM models by properly partitioning subproblems and optimizing data representation in financial falsification detection. Chen Q, addressing the high noise and nonlinear characteristics of financial time series, developed a hybrid deep learning model combining attention mechanisms, multi-layer perceptrons (MLP), and bidirectional long short-term memory (BiLSTM) networks, which significantly improved forecasting performance and provided valuable insights for modeling complex financial indicators.

Additionally, Horsman G proposed three types of test dataset construction standards in the field of digital forensics, including tool evaluation, behavior, and scenario datasets, and emphasized the critical role of standardized data documentation in ensuring the reliability of experimental results. Paullada A pointed out that there are social bias issues in the process of dataset construction in the field of machine learning, recommending the integration of both qualitative and quantitative methods to improve data documentation standards in order to enhance data fairness and model generalization capabilities.

In economic research related to financial falsification, Dalisova N A found a weak and slightly negative correlation between financial cycles and asset returns. Sun G's study showed that digital finance significantly inhibits corporate financial fraud by alleviating financing constraints, especially for large-scale enterprises and in regions with lower levels of marketization. Kuzmina O, through analyzing historical stock index dynamics, revealed internal and external factors influencing index sensitivity, focusing on the differences in how different markets respond to the same economic variables, which provides theoretical support for understanding the impact of the market environment on financial falsification behavior.

In conclusion, the existing research provides theoretical foundations and methodological references for financial indicator falsification detection based on the LSTMM model from multiple aspects, including machine learning technologies, dataset construction standards, and empirical economic analysis. It also reveals the significant impact of data quality and market complexity on model performance, offering guidance for this research's dataset construction and model design.

3.Construction of the Financial Falsification Indicator Dataset and Modeling Method Design

3.1Dataset Construction: Sample Selection and Feature Label Extraction

In this study, we construct a financial whitewash dataset for the identification of financial indicator whitewash in the financial field, which covers the two core aspects of sample screening and feature label extraction.

We first collect financial data from public databases, and then select those companies with annual financial statements as samples, and then combine them with cases of penalties imposed by securities regulators on financial violations to find out the samples of companies with financial whitewash behaviour.

These samples of companies with financial whitewashing come from different sizes and industries, making the dataset both rich and representative. Each sample of firms with financial contamination has at least five years of financial data, so we labelled them as 'contamination' samples and used them as targets for our analysis. At the same time, we find some samples of firms that do not engage in financial modelling, so that the number of modelling and non-modelling samples in the dataset is more evenly distributed. These non-glittering samples are as consistent as possible in terms of industry, company size, and year of data, so that the samples can be fairly compared with each other. The second step in building the dataset is to extract feature labels. To do this, the research team looks at each financial indicator in the financial statements and combines it with information on past cases of financial irregularities to tag each indicator. If an indicator has been falsified in a certain year, it will be labelled with a '1', which means that the indicator has been falsified; if the indicator has not been falsified, it will be labelled with a '0', which means that it has not been falsified. These labels are based on the analysis of financial irregularities, which ensures that the labels of each indicator in the dataset are plausible.

Finally, after the above steps of screening samples and labelling indicators, we have successfully built a complete dataset with and without samples of financial whitewashing. This dataset can show what the financial whitewashing behaviour is really like, and can provide a lot of difficult and valuable training information for the machine learning model to be used later.

3.2Feature Selection and Data Preparation for Modeling

In this paper, we propose an architecture called LSTMM, which integrates LSTM (Long Short-Term Memory Network) and MLP (Multi-Layer Perceptron) models specifically to deal with the task of multi-label classification of financial indicator whitewash. The main role of this model is to determine whether the indicators in the financial statements have been subjected to whitewash or not, with a particular focus on combining long-term dependencies with classification decisions.

The LSTM part is the key to this model, and it is good at dealing with the kind of data that is temporally sequential. Financial data is often temporally correlated, so LSTM is particularly suitable for this task, which receives as input the batch size, time step and feature dimensions, and outputs a feature representation for each time step. When designing the model, we specify that the shape of the input data for the LSTM layer is (batch_size, time_steps, input_feature), where batch_size is the size of the batch of data to be processed each time, time_steps denotes the length of the time step, and input_feature is the dimensionality of the input features at each time step. time_steps, input_feature is the dimension of the input feature at each time step.

As shown in Figure 1, the model first stitches together historical data, current sample data, and industry data to form input features. These input features go to the LSTM layer, which can extract the temporal patterns of the data from them. Then the data processed by the LSTM layer will be passed to the MLP layer, which will complete the multi-label classification decision. During the computation process of LSTM, the model can capture those long-lasting correlations in the data, so that it can provide useful feature information for the subsequent classification tasks.

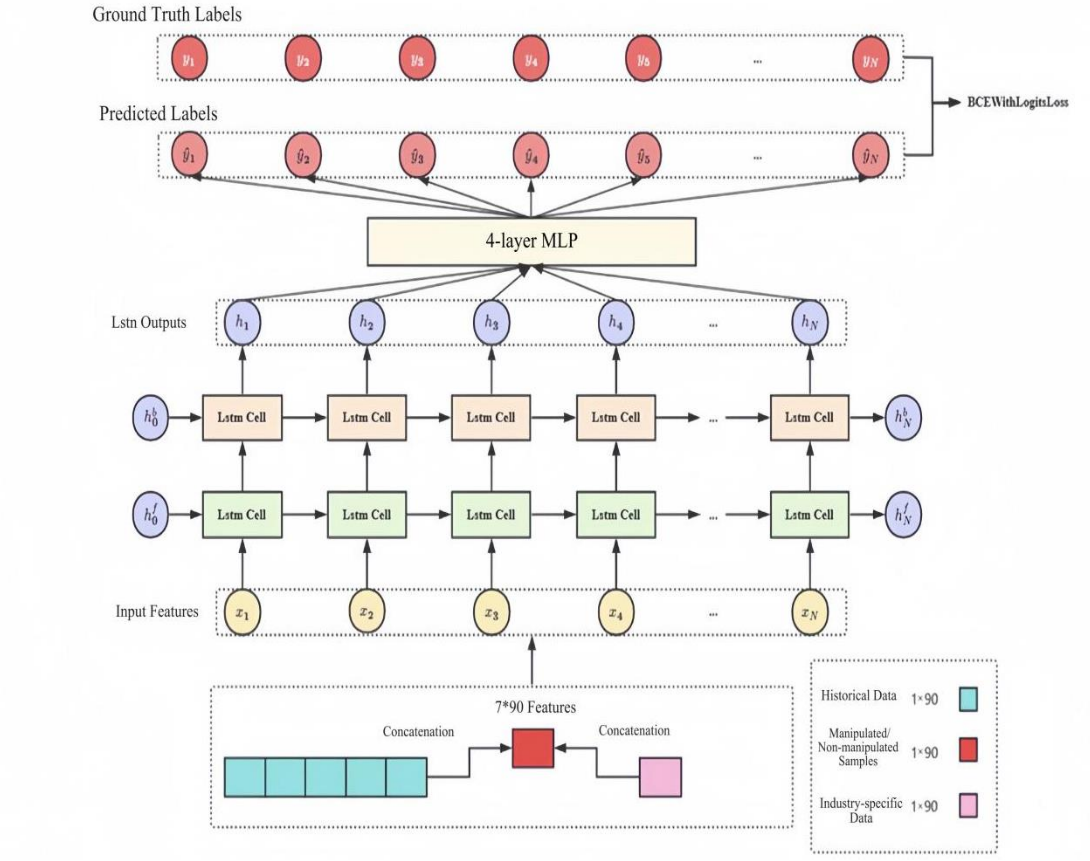

Figure 1 :LSTMM Model Architecture

Then, the MLP layer takes over the processed output of the LSTM layer and uses this data for deeper classification work.The structure of the MLP layer contains four fully connected layers, and the ReLU activation function is used to do nonlinear data transformation between layers.The input data shape of the MLP layer is (batch_size, input_feature), and the output data shape is (batch_size, num_classes), which is the number of classes in the final classification result. After the MLP layer, the model is able to perform complex transformations and combinations of the input features to give a prediction of whether each financial indicator is whitewashed or not.

The model outputs are processed to give a prediction probability corresponding to each label. A specific threshold is set, and if the predicted probability of a metric exceeds this threshold, then the metric is determined to have been subjected to financial whitewash and labelled ‘1’. On the other hand, if the predicted probability does not exceed the threshold, the indicator is considered not to have been whitewashed and is labelled as ‘0’. Through this process, the model is able to accurately classify each financial indicator in a multi-label classification task.

In short, the LSTMM model combines the strengths of LSTM, which is good at dealing with backward and forward dependencies in time-series data, and MLP, which is good at accurate classification. the LSTMM model is particularly good at identifying financial indicator whitewash. With this model, regulators will be able to more easily detect and respond to financial manipulation, making financial statements more transparent and accurate.

4.Empirical Analysis of the Model and Optimization Strategies

4.1Model Comparison Experiments and Recognition Performance Evaluation

In this section, the paper conducts comparative experiments with different machine learning models with the aim of seeing how effective they are at the task of identifying financial whitewash. The experiments were evaluated using four commonly used metrics, namely accuracy, precision, recall and F1 value, so as to have a more comprehensive look at how good each model is at identifying financial whitewash. The datasets used for the experiments were pre-processed appropriately to ensure that the ratio of the number of samples with and without financial whitewash was balanced. Each model was then trained and tested on the same dataset. As shown in Table 1, the LSTMM models outperform other traditional machine learning models on all metrics.

Table 1: Comparison of Machine Learning Experimental Results

	Decision Tree	Random Forest	XGBoost	TabNet	LSTMM
Accuracy	0.9589	0.977	0.9779	0.875	0.9826
Precision	0.5688	0.9065	0.9467	0.8958	0.9001
Recall	0.566	0.5333	0.5119	0.5624	0.5909
F1	0.5673	0.6715	0.6644	0.6909	0.7127

In detail, the LSTMM model is quite impressive in terms of accuracy, precision, recall and F1 value. It has an accuracy of 0.9826, precision of 0.9001, recall of 0.5909 and F1 value of 0.7127, which demonstrates its strength in identifying financial whitewash. Looking at other traditional machine learning models, like Random Forest and XGBoost, although they also score well in terms of accuracy, with Random Forest accuracy being 0.9770 and XGBoost being 0.9779, they don't score so well in terms of recall and F1 value. This indicates that there are still some deficiencies in their ability to identify financial whitewash indicators as well as in their stability. The performance of the decision tree model is the worst, its accuracy is only 0.9589, precision and recall are also low, and the F1 value is much lower than the other models, which clearly reflects its great limitation in dealing with this complex task.

Additionally, we conducted ablation experiments to verify the advantages of combining LSTM and MLP in the LSTMM model. The experimental results showed that although LSTM and MLP layers contribute individually, the combined LSTMM model performed better in the multi-label classification task, further proving the model's effectiveness in financial falsification recognition.

Through these comparison experiments, we demonstrate that the LSTMM model has significant advantages over traditional machine learning methods in financial falsification recognition, especially in terms of accuracy, stability, and recognition precision. The LSTMM model is better able to capture underlying patterns in financial data, enhancing the effectiveness of financial falsification recognition. As shown in Figure 2, the distribution of financial falsification behaviors in different reports reveals the varying reflections of falsification behaviors across reports. This distribution helps understand the importance of different reports in financial falsification recognition and provides a data foundation for subsequent experiments.

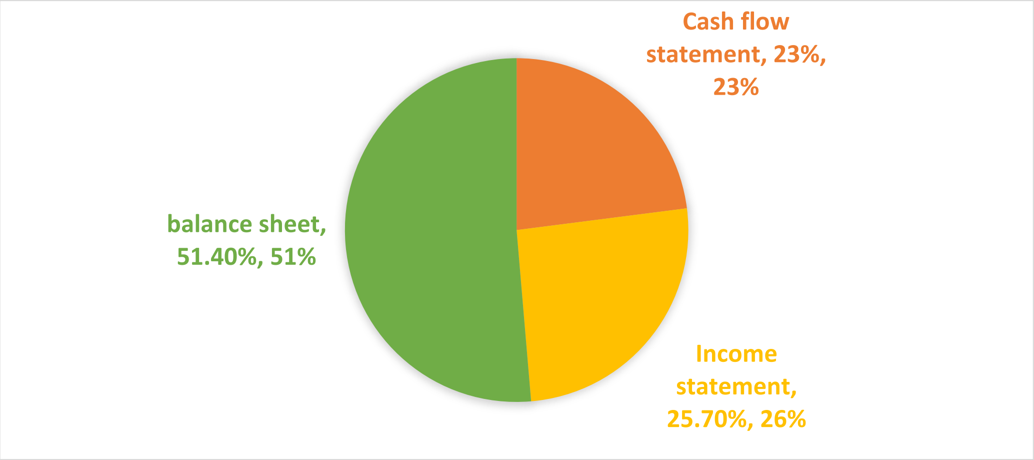

Figure 2:Proportional Distribution of Falsification Indicators in Financial Statements

4.2 Optimization of Multi-Label Classification and Recognition Performance

In this section, the application of the LSTMM model in financial indicator falsification recognition is further explored, with a focus on optimization strategies for the multi-label classification task. To improve recognition performance, especially in addressing the issue of label imbalance, BCEWithLogitsLoss was adopted as the loss function. Compared to the traditional cross-entropy loss function, BCEWithLogitsLoss not only effectively handles cases where each sample has multiple labels but also automatically adjusts label weights, mitigating the impact of label imbalance on model training and ensuring balanced contributions of each label to the loss function.

In addition, the model's training details were optimized in the experiment by incorporating historical and industry data to expand the training samples, thereby improving the model's performance in financial falsification recognition. In processing the data, a 1:1 ratio of whitewashed to non-whitewashed samples is used, and the entire dataset is divided into a training set and a test set according to a ratio of 8:2. This makes the data richer and more diverse during training, and the model can have better generalisation ability to adapt to different situations.

Looking at the model structure, the LSTMM model combines the advantages of both the LSTM layer, which captures the long-term correlations in the time series, and the MLP layer, which converts the output of the LSTM layer into the classification space. With this structure, the LSTMM model is able to efficiently find out the indicators of financial statements with whitewash behaviour, while taking into account the precision and recall of the model. The experimental results show that the LSTMM model performs better than traditional machine learning models on two metrics, F1 value and recall. In particular, the LSTMM model is significantly more capable of identifying samples that are truly guilty of financial whitewashing (i.e., positive examples).

In this section, we optimise the loss function in the multi-label classification task, carefully construct the dataset, and design a reasonable model structure, so that the LSTMM model can perform better in identifying the financial whitewash indicators. This provides useful technical support for auditing and monitoring financial statements in the financial sector.

5. Conclusion and Outlook

This study mainly focuses on the problem of financial indexes' whitewash identification. We first construct a dataset dedicated to financial indicators in the financial field, and propose a solution based on the LSTMM model, through which we want to explore how to use machine learning techniques to improve the accuracy of the identification of financial indicators. Through a large number of experiments, the constructed dataset and model work quite well, and the LSTMM model shows strong recognition ability in the multi-label classification task of financial whitewash. Compared with traditional machine learning methods, the LSTMM model is more adept at capturing the dependencies brought about by the temporal order in financial data, and it can also incorporate static features, which is better than any other algorithms in terms of classification results.

In this study, we construct a dataset of financial indicators for financial sector, which fills the gap of the existing data resources and provides a new reference standard for the problem of financial whitewash identification. The proposed LSTMM model combines the advantages of both LSTM and MLP models, so that it can deal with the features brought by the temporal sequence in financial data, and also extract the static features, which can more accurately find out the financial statements of those whitewash behaviours. The experimental results show that compared with other common machine learning methods, the LSTMM model is more effective in recognition and shows strong potential for application.

Although this study has achieved some preliminary results so far, the study still has some limitations. On the one hand, although the financial data used in this study is relatively comprehensive, the diversity and completeness of the data still need to be further improved. In terms of applicability in different economic environments and industry contexts, the existing data cannot fully meet the needs. On the other hand, the depth and complexity of LSTMM models are high, and the training and computational efficiency may face challenges when dealing with large-scale data. Based on this, future research can focus on data quality optimisation and lightweighting of the model to improve its operational efficiency and enhance its scalability.

Looking ahead, financial markets will continue to grow and data will become more diverse, presenting both opportunities and challenges for financial whitewash identification. Machine learning techniques, especially deep learning, will become more and more common in financial risk prediction and management. The next research can try to mine more useful feature factors and make the dataset richer. It can also combine financial data from different industries and regions, so that the model can be more adaptable and work in a variety of situations. As technology continues to advance, the identification of financial whitewash will not only provide strong support for regulators, but also help investors make more scientific and accurate decisions.

References

[1] Ren B. Research Progress of Content Generation Model Based on EEG Signals[J]. Journal of Computer, Signal, and System Research, 2025, 2(4): 97-103.

[2] Liu Y. The Impact of Financial Data Automation on the Improvement of Internal Control Quality in Enterprises[J]. European Journal of Business, Economics & Management, 2025, 1(2): 25-31.

[3] Hua X. Optimizing Game Conversion Rates and Market Response Strategies Based on Data Analysis[J]. European Journal of AI, Computing & Informatics, 2025, 1(2): 37-43.

[4] Zhou Y. Research on the Innovative Application of Fintech and AI in Energy Investment[J]. European Journal of Business, Economics & Management, 2025, 1(2): 76-82.

[5] Huang J. Resource Demand Prediction and Optimization Based on Time Series Analysis in Cloud Computing Platform[J]. Journal of Computer, Signal, and System Research, 2025, 2(5): 1-7.

[6] Sheng C. Research on AI-Driven Financial Audit Efficiency Improvement and Financial Report Accuracy[J]. European Journal of Business, Economics & Management, 2025, 1(2): 55-61.

[7] Zhang Q. Research on AI-Driven Advertising Optimization and Automated Decision System[J]. European Journal of Business, Economics & Management, 2025, 1(2): 62-68.

[8] Xu D. Design and Implementation of AI-Based Multi-Modal Video Content Processing[J]. European Journal of AI, Computing & Informatics, 2025, 1(2): 44-50.

[9] Li W. Audit Automation Process and Realization Path Analysis Based on Financial Technology[J]. European Journal of Business, Economics & Management, 2025, 1(2): 69-75.

[10] Liu X. The Role of Generative AI in the Evolution of Digital Advertising Products[J]. Journal of Media, Journalism & Communication Studies, 2025, 1(1): 48-55.

[11] Yang, D., & Liu, X. (2025). Research on Large-Scale Data Processing and Dynamic Content Optimization Algorithm Based On Reinforcement Learning. Procedia Computer Science, 261, 458-466.

[12] Shen, D. (2025). AI-Driven Clinical Decision Support Optimizes Treatment Accuracy for Mental Illness. Journal of Medicine and Life Sciences, 1(3), 81-87.

[13] Liu F. Research on Supply Chain Integration and Cost Optimization Strategies for Cross-Border E-Commerce Platforms[J]. European Journal of Business, Economics & Management, 2025, 1(2): 83-89.

[14] Xiu L. Research on the Design of Modern Distance Education System Based on Agent Technology[J]. Pinnacle Academic Press Proceedings Series, 2025, 2: 160-169.

[15] Chen, X. (2025). Research on the Application of Multilingual Natural Language Processing Technology in Smart Home Systems. Journal of Computer, Signal, and System Research, 2(5), 8-14.

[16] Fu, Yilin. "Design and Empirical Analysis of Financial Quantitative Trading Model based on VMD-DCNN-SGRU Architecture and Integrated System." In 2025 4th International Conference on Distributed Computing and Electrical Circuits and Electronics (ICDCECE), pp. 1-7. IEEE, 2025.

[17] Yang D, Liu X. Collaborative Algorithm for User Trust and Data Security Based on Blockchain and Machine Learning[J]. Procedia Computer Science, 2025, 262: 757-765.

[18] Liu Z. Research on the Application of Signal Integration Model in Real-Time Response to Social Events[J]. Journal of Computer, Signal, and System Research, 2025, 2(2): 102-106.

[19] *Hui, X. (2025). Research on Improving the Matching Efficiency between Cancer Patients and Clinical Trials Based on Machine Learning Algorithms. Journal of Medicine and Life Sciences, 1(3), 74-80.*

[20] *Huang, Jiangnan. "Online Platform user Behavior Prediction and Decision Optimization based on Deep Reinforcement Learning." In 2025 4th International Conference on Distributed Computing and Electrical Circuits and Electronics (ICDCECE), pp. 1-6. IEEE, 2025.*