Frontiers in Economics and Management Volume 6 Issue 7, 2025
ISSN: 2692-7608 DOI: 10.6981/FEM.202507_6(7).0005

Research on the Integration of Artificial Intelligence (AI) and Big Data in Corporate Financial Forecasting

Jing He*

Nanning College of Technology, Nanning, 530100, China

*w235663380307@163.com

Abstract

Traditional financial forecasting methods seem unable to cope with the challenges driven by artificial intelligence (AI). This study explores the integration of artificial intelligence and big data, builds a combined model of LSTM and XGBoost, optimizes data processing, feature engineering, and feedback, and improves the forecast accuracy and response speed of industries such as manufacturing, retail, and technology. This study focuses on key scenarios and influencing factors and provides suggestions for data foundations, talent, and management. The results show that fusion technology expands forecasting boundaries, shifts from passive accounting to active data-driven decision-making, and promotes enterprise intelligence and management upgrades.

Keywords

Artificial Intelligence (AI); Big Data; Financial Forecasting; Long Short-term Memory Networks (LSTM); Extreme Gradient Boosting (XGBoost).

1. Introduction

As enterprises advance in digitalization, financial forecasting plays an increasingly important role in strategies, budgeting, and risk management. Accurate forecasting helps optimize resources, rationally allocate funds, and enhance risk response capabilities (Ren, 2022) [1]. However, traditional forecasting methods rely on historical data and linear models, such as trend analysis, regression, and AutoRegressive Integrated Moving Average (ARIMA), making it difficult to capture the nonlinear relationship and real-time changes in financial indicators. In the face of data growth and information diversification, accuracy and flexibility are insufficient (Ahmed, 2024) [2].

In recent years, the rise of artificial intelligence (AI) and big data technology has provided a new development path for financial forecasting. By integrating heterogeneous data from multiple sources, such as enterprise resource planning (ERP) systems, sales platforms, customer relationship systems, social media, supply chains, and macroeconomic indicators, big data technology provides richer information input for financial forecasting, effectively improving the forecasting model's ability to perceive external changes (Adewusi et al., 2024) [3]. AI algorithms, especially machine learning, such as random forests and extreme gradient boosting (XGBoost), and deep learning, such as long short-term memory networks (LSTM) and convolutional neural networks (CNN), have shown strong capabilities in processing complex data relationships, identifying underlying patterns, and dynamically adjusting model parameters, and can break through the technical bottlenecks of traditional models (Abir et al., 2025) [4].

This study aims to analyze the integration path and mechanism of AI and big data in corporate financial forecasting, explore their potential in improving forecast accuracy, adaptability, and scalability, identify challenges such as data silos, insufficient algorithm interpretability, and talent gaps, and propose optimization suggestions.

ISSN: 2692-7608 DOI: 10.6981/FEM.202507_6(7).0005

Based on the framework of "multi-source heterogeneous data + algorithm fusion," this study proposes a forecast data flow and model integration path suitable for enterprises, enriching financial forecasting theory in a dynamic and complex environment. Simultaneously, it explores the applicable conditions and interpretability of the fusion model, providing practical guidance for the application of AI in enterprise management. The fusion model can be applied to short- and medium-term forecasts of key indicators, such as cash flow and revenue, helping enterprises identify risks in advance.

2. Theoretical Framework and Literature Review

2.1. Application of Big Data in the Financial Field

The core value of big data in financial forecasting does not lie in the huge size of data, but in its ability to provide rich variable dimensions and vertical capture of changes in the external environment, making up for the prediction blind spots caused by a single information source (Warren et al., 2015) [5]. In a data-driven framework, the model structure is no longer preset, and the rules are automatically extracted from multi-source data using machine learning. This method enhances the ability to cope with high-frequency market fluctuations and irrational behaviors, allowing financial forecasting to gradually break away from the limitations of traditional, linear, and static models.

In financial forecasting, big data rely on distributed storage and parallel computing technologies. Hadoop distributed file system (HDFS) in the Hadoop ecosystem supports large-scale data storage. MapReduce implements distributed computing. Spark provides high-speed in-memory computing capabilities that are suitable for real-time or near-real-time data analysis. NoSQL databases, such as MongoDB and Cassandra, are suitable for storing unstructured or semi-structured data. These tools have improved the ability of enterprises to handle complex financial data, making model building and reasoning more comprehensive (Sheta, 2020) [6].

2.2. Application of AI Algorithms in Financial Forecasting

The application effect of AI in financial forecasting is significantly affected by the scenario structure and interpretability of the target variable. Research has mainly focused on two types of models: machine learning (ML) and deep learning (DL). The first type is ML models, such as XGBoost, light gradient boosting machine (LightGBM), random forest, and support vector machine, which are good at processing structured small- and medium-sized data (such as corporate revenue, profit, inventory, and cash flow), especially in variable screening, classification, and regression tasks. For example, XGBoost improves the overall prediction ability by integrating multiple weak learners and is widely used in corporate profit forecasting and default risk assessment (Liu et al., 2022) [7].

The second category includes DL models, such as LSTM and CNN, which emphasize automatic feature extraction and context understanding. LSTM are suitable for time-series tasks and are used for cash flow and revenue trend forecasting. CNN perform well in text analysis and can extract sentiment and semantic information from unstructured data, such as news and comments, to assist in market risk judgment. These models help to use large amounts of semi-structured or unstructured data (such as news headlines, customer comments, and social media content) for market sentiment analysis and risk warning.

2.3. Research Review

In recent years, the academic community has conducted multi-angle discussions on the application of AI and big data in financial forecasting, laying a theoretical foundation for the promotion of financial digital transformation. Warren et al. (2015) pointed out that big data and intelligent technology will reshape the accounting information system, shifting from data

ISSN: 2692-7608 DOI: 10.6981/FEM.202507_6(7).0005

recording to analysis and value creation [5]. Goodfellow et al. (2016) emphasized the advantages of deep learning in sequence modeling, providing theoretical support for financial data modeling [8]. Kumar et al. (2024) investigated the operating models of financial shared centers of many large enterprises and found that enterprises have begun to introduce machine learning models into budget preparation and rolling forecasting, realizing the automation and visualization of some forecasting processes [9]. Rauf et al. (2024) further pointed out that to achieve efficient integration of AI and big data, enterprises need to systematically build a data governance architecture, open up cross-system and cross-departmental data circulation mechanisms, and introduce explainability mechanisms to enhance management's understanding and trust in model outputs [10].

Despite this, current research still lacks a systematic summary in terms of model integration, data platform adaptation, and system scalability, and strategies for coping with problems such as data silos, algorithm "black boxes," and a shortage of professional talent are also relatively scattered. Based on this, this study takes the integration path as the main line, systematically analyzes the synergy mechanism of AI and big data in corporate financial forecasting, evaluates its potential in improving forecast accuracy, adaptability, and scalability, and attempts to solve key obstacles in practice, providing a reference for enterprises to build an intelligent financial forecasting system.

3. Fusion Path and Model

3.1. Overall Process of Fusion Technology

The application of AI and big data in financial forecasting is not a single point of intervention but runs through the entire data-driven forecasting process. The typical fusion process includes four stages. Figure 1 illustrates the end-to-end integration of data collection, feature engineering, AI model training, and result feedback in the context of financial forecasting using AI and big data technologies.

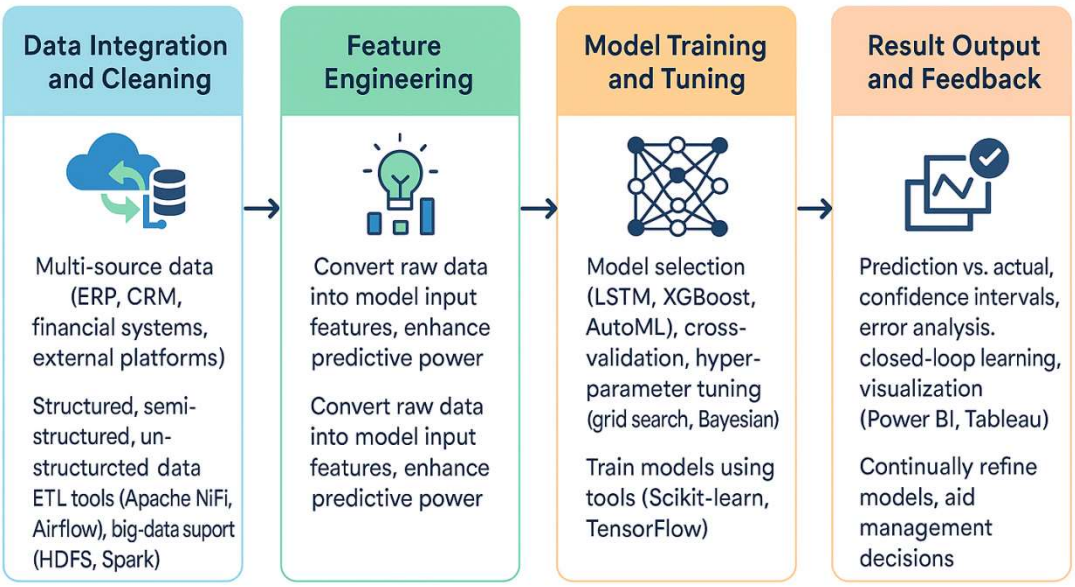

Figure 1. Fusion path and model process
Source: Author's work

Data integration and cleaning

ISSN: 2692-7608 DOI: 10.6981/FEM.202507_6(7).0005

Corporate financial data are usually distributed across multiple systems, including ERP systems, financial sharing systems, sales systems, customer management platforms (CRM), and external platforms (such as the Statistics Bureau and financial news interfaces). These data types include structured tables (such as income statements and cash flow statements), semi-structured data (such as log data and order information), and unstructured data (such as text, images, and comments), and their formats, sources, and update frequencies vary significantly.

In the data fusion stage, we first need to build unified data standards and interface specifications and use extract-transform-load (ETL) tools to clean, transform, and integrate the original data. Common tools include Apache NiFi and Airflow. In big data scenarios, distributed frameworks such as the HDFS and Apache Spark provide strong support.

Feature engineering

After data preparation is completed, feature engineering is required to convert the raw data into input variables suitable for AI model training. Common features include historical financial indicators (such as sales, gross profit margin, and accounts receivable turnover), business behavior variables (such as inventory change rate, order frequency, and customer return rate), and external influencing factors (such as market price index, consumer confidence index, and keyword search popularity).

In addition, unstructured data are converted into numerical features that can be used for modeling through natural language processing (NLP) methods, such as term frequency–inverse document frequency (TF-IDF), sentiment analysis, and bidirectional encoder representations from transformers (BERT). The embedding proportion of negative emotions in customer reviews can be used as an important correction variable for sales forecasting. Good feature engineering not only determines the predictive ability of the model but also significantly affects the subsequent training efficiency and interpretability.

Model training and parameter adjustment

During the modeling stage, enterprises must select appropriate AI algorithms based on the type of prediction task (such as time series, classification, or regression prediction). During the training process, the ratios of the training, validation, and test sets need to be set, cross-validation should be used to avoid model overfitting, and hyperparameters should be adjusted in combination with grid search, Bayesian optimization, and other technologies. Typical training tools include Scikit-learn, XGBoost, Keras, and TensorFlow. Enterprises can also use AutoML platforms (such as Google AutoML and H2O.ai) to improve modeling efficiency.

Result output and feedback

After the prediction was completed, the model output was compared with the actual value. Common outputs include prediction intervals, confidence analyses, and error boundary maps. The system should also have a feedback mechanism, that is, the results with large prediction deviations are re-flowed as training samples to continuously correct the model parameters, thereby achieving "closed-loop learning". In addition, the results can be visualized through visualization tools (such as Power BI, Tableau, or Plotly), which helps managers understand the model logic and output meaning and enhances their willingness to apply it.

3.2. LSTM + XGBoost Combination Model

In the context of multi-source data, a single algorithm often cannot consider both sequence pattern recognition and nonlinear variable analysis simultaneously; therefore, enterprises tend to adopt hybrid modeling strategies in actual applications. Among them, the LSTM and XGBoost combined model is a representative integrated structure (see Figure 2).

ISSN: 2692-7608 DOI: 10.6981/FEM.202507_6(7).0005

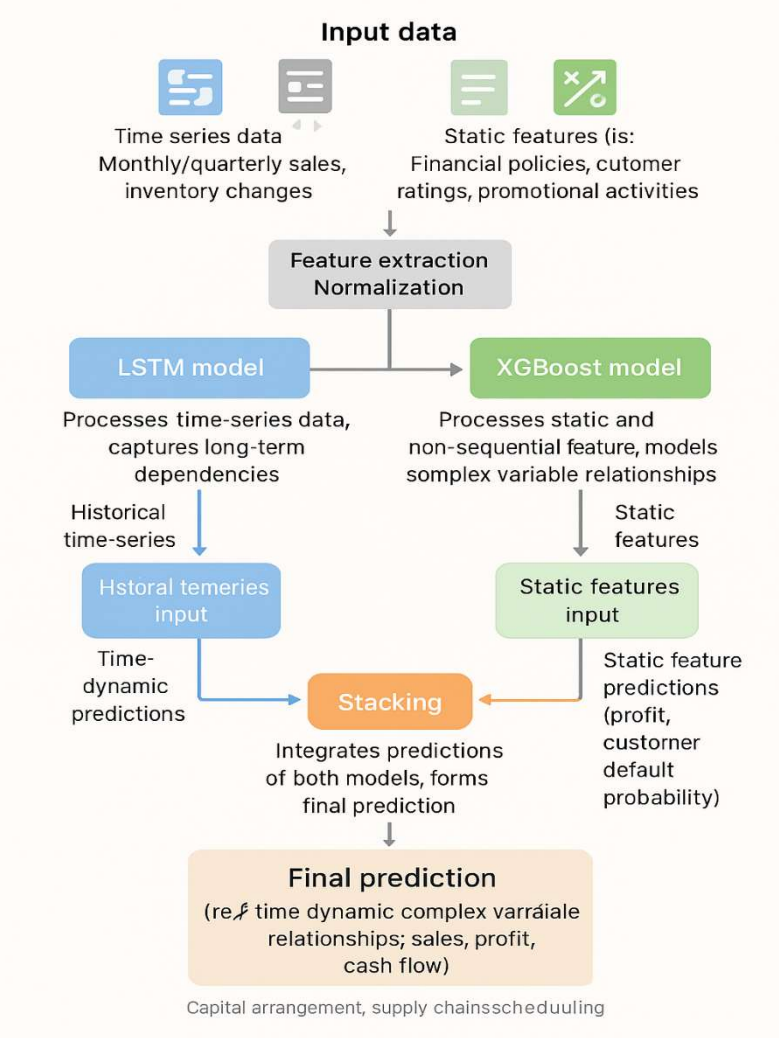

Figure 2. LSTM + XGBoost combination model
Source: Author's work

LSTM is a special recurrent neural network (RNN) that is suitable for processing time-series information, particularly for capturing long-term dependencies. LSTM memorizes and forgets key information in the sequence through a gating mechanism (input gate, forget gate, output gate) and is widely used to predict financial indicators with time continuity, such as cash flow fluctuations and cyclical sales revenue (Weytjens et al., 2021) [11]. Enterprises usually input historical monthly/quarterly sales, inventory changes, raw material price fluctuations, etc., into the LSTM network and output prediction curves to support capital arrangements and supply chain scheduling.

XGBoost is a tree model that is based on gradient boosting. It is effective at processing high-dimensional sparse data and is suitable for prediction tasks with complex interactions between variables, such as profit margins, customer default probability, and the impact of macro indicators on cost structure. This model improves the overall prediction accuracy using a weighted combination of multiple weak learners (decision trees). It has the advantages of fast convergence, strong generalization, and flexible parameter adjustment (Demir & Sahin, 2023) [12]. Enterprises can use XGBoost to handle static variables that LSTM cannot handle (such as financial policies, customer ratings, and promotional activities) and perform weighted fusion with LSTM output results.

The fusion method of the two types of models usually adopts the stacking strategy: LSTM and XGBoost are used as the first-level models, and their respective subset variables are trained.

ISSN: 2692-7608 DOI: 10.6981/FEM.202507_6(7).0005

Finally, the output results are integrated through a fusion model (such as linear regression or a neural network) to form the final prediction value. This type of structure retains LSTM's ability of LSTMs to recognize time dynamics and that of XGBoost to model complex variable relationships, realizing the three-dimensional prediction perspective of "structure + behavior + environment."

3.3. Effectiveness Evaluation

After the model is built, its effect must be systematically evaluated using multidimensional indicators. Commonly used evaluation indicators include the following: Root mean square error (RMSE) measures the degree of deviation between the predicted value and the actual value, the smaller the better; Mean absolute error (MAE) evaluates the absolute average of the error, which is easy to interpret; R^2 (coefficient of determination) reflects the model's ability to explain the total variation of the sample, the closer to 1 the better. In addition to numerical accuracy, a horizontal comparison is required from the perspective of model stability and adaptability. Stability refers to whether the model performs consistently under different datasets or time windows, and adaptability focuses on the model's ability to respond to new variables and emergencies (such as epidemics and policy adjustments).

By comparing the fusion model with traditional methods (such as ARIMA and OLS regression), it can be found that in the case of processing high-frequency, multi-source, and unstructured data, the RMSE of the AI fusion model usually decreases by more than 15%, the response speed is accelerated, the prediction error is significantly reduced, and it has stronger scalability and real-time feedback capabilities (Goodfellow et al., 2016) [8]. In summary, the LSTM+XGBoost fusion model shows good performance and application prospects in corporate financial forecasting and provides a feasible technical path for realizing data-driven financial intelligent management.

4. Applicability of Technology Integration

4.1. Scenario Analysis

The effective application of AI and big data in financial forecasting is highly dependent on the operating characteristics, data foundation, and organizational capability structure of the industry in which an enterprise is located. Prioritizing the industry type, data dimension, and quantifiability of the business forecast target is necessary.

There are significant differences in the financial forecasting needs of different industries, which determine the application focus of AI models in the financial field. As manufacturing companies are involved in raw material procurement, production scheduling, and inventory management, their core forecasting tasks are focused on cash flow and supply chain efficiency. In such scenarios, LSTM can handle periodic sales fluctuations and cost changes and model structured data outputs from ERP systems (Krishna Madhav et al., 2023) [13]. The retail and e-commerce industries rely more on user behavior data and real-time sales feedback and are suitable for using stream processing frameworks, such as Spark, combined with deep learning models for short-term sales forecasting, thereby assisting in intensive promotional activities and inventory allocation.

In contrast, technology service companies mainly focus on project revenue forecasting and R&D input-output evaluations. The business cycle is nonlinear and unstable, and the adaptability of traditional linear models is limited (Lee & Kwon, 2023) [14]. In such situations, AI algorithms can integrate variables such as project progress, human factors, historical cases, and personnel deployment to achieve more forward-looking estimates and risk assessments.

AI and big data technologies can be applied to three typical scenarios. First, industries with highly dynamic business environments, such as Internet and new energy vehicle companies,

ISSN: 2692-7608 DOI: 10.6981/FEM.202507_6(7).0005

are affected by frequent policy adjustments and rapid technology iterations. Traditional models are difficult to adapt to nonlinear mutation trends, whereas AI models have self-learning and real-time update mechanisms that can respond quickly to changes. Second, companies that have established data asset systems and rich historical data and external data access capabilities. If sales records, customer behavior, market indicators, and comment texts can be integrated, the model's training capabilities and generalization performance will be significantly enhanced. The third is tasks with a clear causal chain between the predicted results and the input variables, such as "order-growth revenue-growth cash flow-improvement." Such scenarios make it easier for AI models to output explainable results, which are easier for managers to understand and adopt.

4.2. Key Factors Affecting the Integration Effect

Although technology is developing rapidly, the actual effect of AI and big data on financial forecasting is still constrained by various non-technical factors. Three key factors significantly affect the final implementation of the integration results.

Data infrastructure maturity

The premise of AI modeling is to have available data resources and efficient data pipelines. If the enterprise has not yet established a data middle platform and lacks unified data standards and interface protocols, data collection, cleaning, and integration will consume a lot of resources, seriously affecting modeling efficiency and result quality. Therefore, infrastructure construction (such as data lakes and master data management systems) is a prerequisite for promoting integration.

Talent structure and organizational collaboration capabilities

AI plus big data projects usually require cross-functional collaboration among data scientists, financial analysts and IT engineers. However, in reality, most corporate financial personnel still rely on traditional accounting or reporting skills and lack modeling capabilities and data awareness. Simultaneously, communication barriers between the IT and financial departments often lead to demand deviations and ineffective results. Therefore, it is of great significance to promote the training of "finance + technology" compound talents and establish specialized financial data analysis positions in the industry.

Management's understanding of and willingness to adopt intelligent predictions

Whether senior leaders trust AI models and are willing to incorporate prediction results into budget management and performance appraisals are key factors that affect the strategic role of integration. Some companies are still accustomed to empirical judgment or linear model output and have reservations about "black box models," which limits the actual application scope of AI. Therefore, strengthening model interpretability and building a visual decision support system are important ways to increase willingness to adopt.

5. Optimization Suggestions for Technology Integration

Combined with the above analysis, the application of AI and big data in financial forecasting should follow the idea of formulating policies according to the industry and taking appropriate measures according to the situation. To improve the actual application effect of fusion technology in enterprises, it is recommended to optimize it from four aspects: infrastructure, modeling strategy, organizational mechanism, and adoption mechanism.

Customize modeling strategies according to industry characteristics

The manufacturing industry should focus on time-series modeling and production and sales collaborative forecasting, prioritize the deployment of models such as LSTM that can process long-term data, and connect with manufacturing execution system (MES) and ERP systems to build an end-to-end forecasting pipeline. Retail and e-commerce companies should strengthen

ISSN: 2692-7608 DOI: 10.6981/FEM.202507_6(7).0005

their real-time processing capabilities to handle unstructured data. It is recommended to build a Spark streaming computing platform combined with NLP tools, such as BERT embedding or sentiment analysis, to capture changes in user behavior and support short-term sales forecasts. Technology service companies are more suitable for building lightweight and highly interpretable forecasting systems, introducing algorithms such as XGBoost, which can model variable interactions to estimate project revenue and research and development (R&D) investment returns.

Strengthen the construction of data platform and feature collection mechanism

Enterprises should prioritize establishing a data platform that integrates "financial data + business data + external data," build a forecasting data warehouse on this basis, and design a feature library according to the forecasting task. It is recommended to include operating indicators (such as inventory turnover and return rate), external variables, such as consumer price index (CPI) and search index, and text features (such as user comments) in the modeling scope, and realize automatic updates through application programming interface (API) or batch interface to improve the richness of data dimensions and the timeliness of updates.

Build a cross-departmental collaborative modeling mechanism

In the process of AI and financial forecasting modeling, it is recommended that the data team be responsible for model building, the financial staff be responsible for feature screening and business interpretation, and the IT department ensure the operation of the platform. A three-person team system of "product manager + data scientist + financial analyst" can be adopted to address these issues. After each round of model updates, the prediction effect is verified in conjunction, and the evaluation indicators are included in the team performance to form a business-driven modeling cycle.

Improve the interpretability of forecast results and management usability

In response to managers' concerns about forecast credibility in high-frequency scenarios, it is recommended to introduce explainable artificial intelligence (XAI) methods, such as SHapley additive exPlanations (SHAP) value decomposition, to show the specific impact of variables on forecast values in the result report. Tools such as Power BI or Tableau can be used to visualize forecast results and output a three-dimensional dashboard of the "forecast-error-contribution variable" to facilitate a high-level understanding and inclusion in budget or performance decision-making systems.

6. Conclusion

This study analyzes the integration path of AI and big data in financial forecasting and believes that multi-source data integration has significantly expanded the basis of forecasting information, and AI models (such as LSTM and XGBoost) have effectively improved the accuracy and adaptability of forecasts, providing a feasible technical framework for companies to achieve dynamic decision-making.

However, integrated applications still face practical challenges, such as weak data governance, insufficient model interpretability, and a lack of compound talent. In the future, we should promote financial forecasting from experience-oriented to intelligence-driven by improving data infrastructure, introducing explainable AI methods, and strengthening organizational coordination mechanisms.

References

[1] Ren, S. (2022). Optimization of Enterprise Financial Management and Decision-Making Systems Based on Big Data. Journal of Mathematics, 2022(1), 1708506.

ISSN: 2692-7608 DOI: 10.6981/FEM.202507_6(7).0005

[2] Ahmed, Q. O. (2024). The future of aerospace research: How data collection systems can advance space exploration. Volume, 9, 360-370.

[3] Adewusi, A. O., Okoli, U. I., Adaga, E., Olorunsogo, T., Asuzu, O. F., & Daraojimba, D. O. (2024). Business intelligence in the era of big data: A review of analytical tools and competitive advantage. Computer Science & IT Research Journal, 5(2), 415-431.

[4] Abir, S. I., Sarwer, M. H., Hasan, M., Sultana, N., Dolon, M. S. A., Arefeen, S. S., ... & Saha, T. R. (2025). Deep Learning for Financial Markets: A Case-Based Analysis of BRICS Nations in the Era of Intelligent Forecasting. Journal of Economics, Finance and Accounting Studies, 7(1), 01-15.

[5] Warren, J. D., Moffitt, K. C., & Byrnes, P. (2015). How big data will change accounting. Accounting horizons, 29(2), 397-407.

[6] Sheta, S. V. (2020). Enhancing data management in financial forecasting with big data analytics. International Journal of Computer Engineering and Technology (IJCET), 11(3), 73-84.

[7] Liu, J., Zhang, S., & Fan, H. (2022). A two-stage hybrid credit risk prediction model based on XGBoost and graph-based deep neural network. Expert Systems with Applications, 195, 116624.

[8] Goodfellow, I., Bengio, Y., Courville, A., & Bengio, Y. (2016). Deep learning (Vol. 1, No. 2). Cambridge: MIT press.

[9] Kumar, N., Agarwal, P., Gupta, G., Tiwari, S., & Tripathi, P. (2024). AI-Driven financial forecasting: the power of soft computing. In Intelligent optimization techniques for business analytics (pp. 146-170). IGI Global.

[10] Rauf, M. A., Shorna, S. A., Joy, Z. H., & Rahman, M. M. (2024). Data-driven transformation: Optimizing enterprise financial management and decision-making with big data. Academic Journal on Business Administration, Innovation & Sustainability, 4(2), 94-106.

[11] Weytjens, H., Lohmann, E., & Kleinsteuber, M. (2021). Cash flow prediction: MLP and LSTM compared to ARIMA and Prophet. Electronic Commerce Research, 21(2), 371-391.

[12] Demir, S., & Sahin, E. K. (2023). An investigation of feature selection methods for soil liquefaction prediction based on tree-based ensemble algorithms using AdaBoost, gradient boosting, and XGBoost. Neural Computing and Applications, 35(4), 3173-3190.

[13] Krishna Madhav, J., Varun, B., Niharika, K., Srinivasa Rao, M., & Laxmana Murthy, K. (2023). Optimising Sales Forecasts in ERP Systems Using Machine Learning and Predictive Analytics. J Contemp Edu Theo Artific Intel: JCETAI-104.

[14] Lee, J., & Kwon, H. B. (2023). Synergistic effect of R&D and exports on performance in US manufacturing industries: high-tech vs low-tech. Journal of Modelling in Management, 18(2), 343-371.