



International Journal of Information and Communication Technology

ISSN online: 1741-8070 - ISSN print: 1466-6642
<https://www.inderscience.com/ijict>

An intelligent decision support framework for financial market risk using big data and optimised XGBoost

Sihao Zheng

DOI: [10.1504/IJICT.2026.10076768](https://doi.org/10.1504/IJICT.2026.10076768)

Article History:

Received:	12 September 2025
Last revised:	14 November 2025
Accepted:	28 November 2025
Published online:	18 March 2026

An intelligent decision support framework for financial market risk using big data and optimised XGBoost

Sihao Zheng

China University of Geosciences (Wuhan),
College of Computer,
Wuhan, 430074, China
Email: 20241003253@cug.edu.cn

Abstract: Risk forecasting in financial markets is essential in management and engineering choices. In this paper, a more refined XGBoost model that uses big data analytics to identify nonlinear patterns that are usually not detected by the conventional technique is developed. The process can be optimised with a high feature selection, hyperparameter optimisation, as well as advanced preprocessing of the data. The model has been tested and proven to have better performance than the current methods in the market and helps in efficient risk management, resource deployment, and timely risk mitigation. The model is responsive to market changes and is ethical on AI operations in real-time. Combined with live data feeds, it will be able to respond more effectively and promptly to changes, a key that can serve as a powerful financial risk prediction instrument, driven by AI.

Keywords: financial market risk; big data analytics; XGBoost optimisation; intelligent decision support system; IDSS; risk management.

Reference to this paper should be made as follows: Zheng, S. (2026) 'An intelligent decision support framework for financial market risk using big data and optimised XGBoost', *Int. J. Information and Communication Technology*, Vol. 27, No. 24, pp.1–21.

Biographical notes: Sihao Zheng is a student majoring in Data Science and Big Data at the School of Computer Science of China University of Geosciences (Wuhan), mainly focuses on the practical application of machine learning and deep learning.

1 Introduction

The unstable financial world today makes market risk prediction an inseparable role of organisations that need to have long-term survival and viability. The financial markets are dynamic every day due to the numerous events that are dependent on each other, including global economic trends, political unrest, investor mood, and the rapid change of technology (Wang et al., 2023). Conventional risk assessment tools, although still considered as a cornerstone, fail to reflect the complexity, velocity, and magnitude of the contemporary financial framework (Zhang and Zhang, 2025). The spread of information across various sources, including both structured and unstructured forms, such as market

indices, exchange rates, trading volumes and financial news, analyst commentary, or social media, has changed the risk analysis landscape (Nobre and Neves, 2019). Nonetheless, the resulting data explosion is associated with additional heterogeneity, noise, and volume issues that demand more advanced data processing tools (Murugan, 2023). The nonlinear relationships and latent trends related to such large and high-dimensional data are difficult to detect with manual methods and traditional models (Yunsong et al., 2024). In addition, the traditional statistical methods are usually limited by the normality and stationary assumptions, which are hardly applicable to the real-life financial statistics (Wang et al., 2022). That way, there has never been a higher demand than today to create intelligent, adaptive, and scalable solutions to predict risk (Peng et al., 2022). Both financial IDSS and engineers are currently seeking integrated data solutions that do not only understand complex relationships but also respond to sudden market dynamics in real-time to provide a stronger and more visionary IDSS and risk management paradigm (Tan et al., 2023).

An effective way to predict risk in financial markets is suggested in this model by connecting complete data gathering, advanced data pre-processing, feature engineering and a powerful XGBoost algorithm (Li et al., 2022b). It makes use of both formal financial numbers and informal opinions so it can see the whole picture of the market (Liu et al., 2022). Using methods like sentiment analysis, technical indicator generation and dimensionality reduction, the system works to improve data before using it for training (Xia et al., 2023). The approach is further refined using grid search and Bayesian optimisation to ensure both good results and general application (Zeng, 2025). The model that has been applied provides strong risk projections and decision-supporting information, make sound and timely decisions to financial managers and analysts so that they can make decisions fast in any situation (Li et al., 2021).

The proposed framework clearly specifies the task of financial risk prediction, has explicit train/test partitions, dimension selection in folds, random seed management, hyperparameter optimisation, and ethical and responsible AI, which ought to ensure robust and real-world usage in IDSS and risk management. The given approach is built on the combination of integrated data collection, data pre-processing and feature building based on a stable XGBoost algorithm to predict the risk of financial markets. It considers conventional figures and also evaluates remarks to ensure that it encounters market changes entirely. Using the best sentiment analysis, the creation of indicators and feature reduction, the system refines the indicators that are trained on. Next, XGBoost is enhanced with the assistance of such methods as grid search and Bayesian optimisation to obtain not only accuracy but also a generalising chance. The model provides proper risk forecasts and feasible advice, such that on the one hand the financial managers and analysts make informed decisions in time and thus make prompt and relevant decisions in a changing environment.

2 Related works and contemporary models

Studies on predicting risks in financial markets use a mixture of statistical and machine learning tools. Initially, methods including ARIMA and GARCH were meant for simple, straight-line motions which did not allow them to deal well with new, complex changes in financial markets. Because of advances in big data and AI, researchers are now more often using random forests, support vector machines (SVM), LSTMs, CNNs and

XGBoost in their analysis. Such methods have clearly led to better prediction accuracy, more flexibility and the ability to scale up for different uses. A small group of researchers has combined sentiment analysis and uncategorised data sources like financial news and popular social media to increase prediction potential. Even though they help people learn patterns and respond in real-time, the techniques can still be hard to understand, might need exact data and are often complex to compute. Key studies are outlined in Table 1, focusing on the techniques applied, the pros of each method and its weaknesses which support the introduction of the proposed enhanced XGBoost algorithm.

Table 1 Comparative review of recent studies on financial risk prediction using big data and XGBoost-based techniques

<i>Author(s)</i>	<i>Techniques involved</i>	<i>Advantages</i>	<i>Disadvantages</i>
Karami and Igbokwe (2025)	Big data analysis	Highlights big data impact on credit risk	Lacks model implementation
Theodorakopoulos et al. (2025)	PySpark + XGBoost + CatBoost	Scalable and efficient fraud detection	High resource use; complex tuning
Su et al. (2024)	XGBoost + ESG analysis	Integrates ESG with Industry 5.0 insights	Limited generalisability
Yu et al. (2024)	XGBoost + vector optimisation	Improved loan approval accuracy	Sensitive to data quality
Qin (2022)	XGBoost for risk modelling	Strong predictive ability for finance	Weak with unstructured data

Karami and Igbokwe (2025) reviewed the impact of big data aspects such as volume, velocity and variety on the process of assessing credit risk. The authors pointed out that these features can make predictive modelling models in finance either more effective or less reliable. Although their study helps explain big data's influence on financial risk analysis on paper, it does not examine specific machine learning models.

Theodorakopoulos et al. (2025) suggested a big data system based on distributed machine learning for identifying credit card fraud using PySpark, XGBoost and CatBoost. Their method works well with lots of transactions, letting them spot threats quickly and deal with them efficiently. This machine learning model works efficiently with parallel data processing, but it requires many resources and a lot of ongoing testing to tune the hyperparameters.

Su et al. (2024) designed a big data analysis framework using the XGBoost algorithm to measure both Industry 5.0 integration and ESG performance. Their method improves how resources are assigned and links environmental performance to financial risks. Still, its usefulness tends to be limited to the area it was built for and it does not always apply well to other parts of finance.

Yu et al. (2024) presented a new XGBoost-based approach for choosing which loans to approve by using the four-vector optimisation algorithm to enhance it. Doing so in this manner results in better predictions and ensures the model will remain stable. Although the model can do many good things, it requires extensive pre-processing work and is easily affected by data problems.

Qin (2022) designed a risk prediction model using XGBoost, hoping it would assist proactive financial choices by the company. The method used on structured company data showed accurate identification of financial management risks. Because it relies only on structure input, it is not as responsive to sudden moves in the market.

With the constraints observed in the researched models, it is still necessary to have a more flexible, scalable and complete financial risk prediction approach. While prior studies have used both big data and machine learning approaches, including XGBoost, most of them cannot handle unstructured information, grow fast in multiple financial contexts, require a lot of computation and lack quick adaptability. For this reason, the model was designed to use structured indicators along with text data to look at market activities in a detailed way. With Bayesian optimisation, it improves prediction by choosing the most important features, cleans the data correctly and adjusts parameters. Additionally, the model's design helps it to adjust and respond as needed for real-time data use. Using e-commerce data, this approach helps financial managers and engineers take more effective, data-driven actions as the market changes quickly. Due to this reason, the proposed framework is explicit on the prediction tasks, employs structured and unstructured indicators, reproducible train/validation/test split, controlled seed, and optimises and selects features to facilitate better accuracy and generalisation. It also promotes responsible and ethical AI practices and aids in the real-world decision-making situations.

The model was developed to consider structured indicators and text information to examine market activities in a fine manner. In case of Bayesian optimisation, it enhances predictive behaviour because it selects the most significant features, cleans the information appropriately and modifies the parameters. Also, the design of the model allows it to be adjusted and react as it is required to use real-time data. This method will also enable financial managers and engineers to make more effective and data-driven decisions because the market is rapidly changing using e-commerce data. In this paper, the concept of applying the latest computational techniques and big data analytics in terms of real-time financial risk management is identified. Using high-dimensional financial and textual data, the proposed framework reveals how intelligent algorithms especially optimised XGBoost can be used to support decision making in dynamic markets. The methodology depicts how the information and communication technologies can be practically applied to process heterogeneous data, extract actionable information, and make adaptive and data-driven strategies that can be used to assess risks effectively.

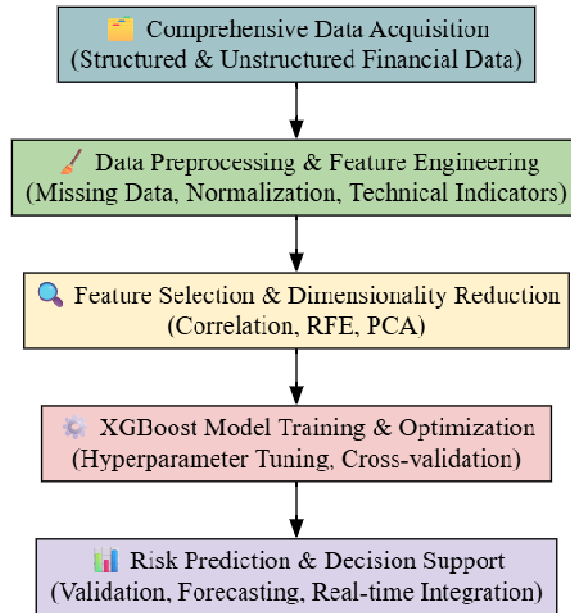
3 Design of the XGBoost-based risk prediction system

The proposed framework is structured and provides the collection of diversely-widely-different financial data which consists of time-series data of stocks, trading volumes and indications, news and social media data. The dataset is then thoroughly cleaned to correct any missing records, remove outliers and normalise all variables and any other data enriching features like technical indicators and scores of the current sentiment among investors. This integrated system can not only process both past and present data, but also have an ability to dynamically adapt to new financial events which means that predictions keep on being updated with live data integration. The entire architecture of the proposed framework as shown in Figure 1 illustrates each of the stages including data acquisition and pre-processing, model training, and real-time risk prediction indicating a data-to-decision seam.

Increase performance and reduction of noise we employ feature selection, which is dependent on correlation whereby features are ranked in terms of the predictions of the model and that feature is eliminated one at a time to ensure the most useful features are

used in the predictive model. The main idea of the framework is training and optimising an XGBoost model that will sequentially create a collection of decision trees that will be able to model phenomena that are difficult to predict in the data; the optimal hyperparameters are found through grid search or Bayesian optimisation, and their correctness and applicability to new data are evaluated using cross-validation. A well-optimised model is carefully examined using major metrics and then put to use for providing accurate risk predictions and helpful information, aiding quick financial decisions. To remain flexible and adaptable, the architecture is set up to integrate live data feeds so that risk assessment is immediate and response to market changes can be quick.

Figure 1 Proposed flow diagram (see online version for colours)

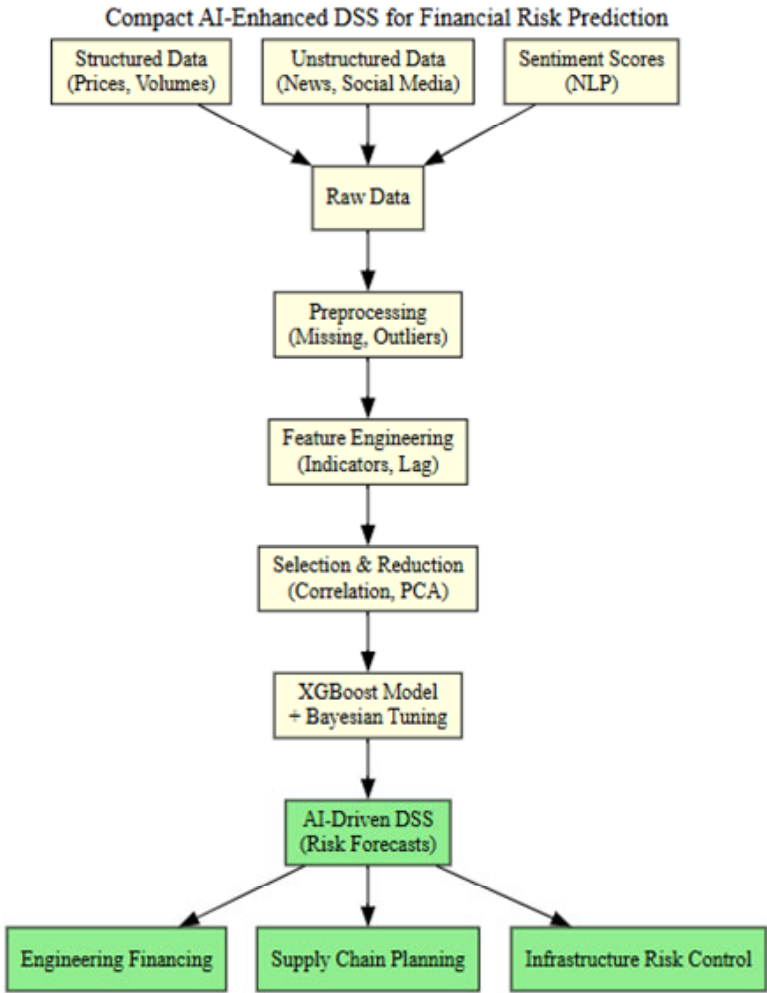


3.1 Comprehensive data acquisition

Gathering extensive, broad data from a variety of solid sources is the basis of creating the proposed risk prediction model. Here, structured and unstructured information is collected to put together a dataset that shows the different aspects of financial market behaviour (Zhou, 2024). Structured data consists of time-series records for stock prices, returns on assets, trading volumes, interest rates, exchange rates, inflation rates and reports from banks or central banks (Andrew, 2018). For unstructured data, information comes from real-time business news, expert opinions, financial reports and conversations on Twitter and Reddit about how investors and the market think. Our algorithms use sentiment analysis to change subjective feelings into objective scores, giving us a better understanding of market psychology (Mohbey et al., 2022). Having quantitative data, signals from the economy and social trends all handled by the model helps it reflect both rationally and emotionally driven market risks. In addition, these systems are set up to

function automatically, support growth and provide live, frequent updates for on-the-spot observation. First, the complete and versatile data is what prepares the rest of the process by allowing the framework to use the most important and timely input for risk analysis in the finance industry (Chen, 2023).

Figure 2 Flowchart of the process (see online version for colours)



3.2 Data pre-processing and feature engineering

Once data are collected, the raw information is processed and changed to get it ready for analysis and to enhance the prediction's results. The first step in data pre-processing is to handle gaps in the data that may be due to unmatched times of collection. The methods for filling these gaps include forward filling, using the average or interpolation, based on the variable being used and its relationship with time (Xu, 2024). If outliers or noise distort our training, we can spot them with statistical methods such as Z-scores or the interquartile range, then either remove them or apply some smoothing to them. It is then

important to put all the features on a common scale by using Min-Max scaling or Z-score normalisation, so no algorithm is biased towards larger ranges of values. New variables are created in the feature engineering stage to improve our grasp of market activity by expanding the data (Liu, 2024). This step involves finding technical indicators such as moving averages, RSI and Bollinger Bands, to see the movement of the price, how fast it is moving and the levels of the market's volatility. In the same way, RSI which examines recent price changes, is calculated using equation (1).

$$RSI = 100 - \left(\frac{100}{1 + \frac{\text{Average gain}}{\text{Average loss}}} \right) \quad (1)$$

Average gain and average loss are usually determined based on observations over a 14-day period. RSI above 70 might indicate an overbought market and a reading below 30 may indicate market oversoldness. Based on the topics written in financial news and social media, NLP models give us sentiments that we transform into meaningful statistics, representing the mood of the market. Features with a lag are designed to consider factors from a past day's activity, including return rates, changes in volatility or overall sentiment (Zhang et al., 2022). These remarkable attributes allow the model to uncover unusual patterns and relationships in time that statistical methods usually miss. Both data pre-processing and feature engineering are needed to link raw financial data to model learning, helping ensure its richness and stability (Nahar et al., 2024).

3.3 Feature selection and dimensionality reduction

After the pre-processing and engineering stage, the next important step is to select and reduce the dataset's overall dimensionality. The objective is to find the most valuable features and eliminate features that are repetitive, do not add any value or are too closely connected, to avoid issues with noise or overfitting (Nuthalapati, 2022). Rich information from high-dimensional spaces may cause the model to underperform, complicate calculations and allow it to focus on unimportant patterns in the market. An approach that includes statistics and algorithms is applied to resolve this problem (Peng, 2021). At the start, multi-collinear features are found and removed through correlation analysis, using either Pearson or Spearman correlation values. If a correlation coefficient between two variables is greater than a defined quantity (for example, 0.85), the two should be compared to decide if one can be removed to avoid redundancy (Yi et al., 2023).

At this stage, features are evaluated by their importance, mainly with algorithms like XGBoost and random forests which naturally assign significance to each feature by the role it plays in accuracy from gain, cover and frequency metrics (Petroopoulos and Siakoulis, 2021). The model is improved by eliminating features that have almost no effect on outcome. Besides, the RFE approach is applied to continually train the model and get rid of unimportant features at every iteration. RFE is carried out until the set of features that achieves the best performance is located (Chai et al., 2022). Dimensionality reduction such as principal component analysis (PCA), is often used with data sets that are very large. It converts the closely related original variables into fewer unrelated components that keep a large amount of the data's information. The transformation is shown mathematically in equation (2).

$$Z = XW \quad (2)$$

Here, Z represents all the principal components, X is the standardised matrix containing the study data and W is a collection of the eigenvectors from the covariance matrix of X . It retains the main features of the data and decreases its dimensionality by a great deal (Li et al., 2022a). Because of these processes, the predictive model concentrates on the most important and different variables which helps it train faster, better forecast future events and produce more accurate findings (Han et al., 2023).

3.4 *XGBoost model training and optimisation*

The extreme gradient boosting (XGBoost) algorithm which combines decision trees, is the fundamental element of our proposed architecture (Rao et al., 2023). It is well-known that XGBoost can handle financial risk prediction well since it understands nonlinear connections, detects intricate feature interactions and manages data efficiently (Mittal and Panchal, 2025). The first act in training is to pass the smaller, cleaned dataset to XGBoost. An ensemble of weak learners (decision trees) is built step by step, with each added tree correcting the mistakes left by the previous one (Gao and Xiao, 2021). With this method, the model learns better by reducing the risk of overfitting, since it finds a trade-off between its training results and complexity. The objective function is written in equation (3).

$$\mathcal{L}(\varphi) = \sum_i l(y_i, \hat{y}_i) + \sum_k \Omega(f_k) \quad (3)$$

$\Omega(f_k)$ refers to the measure that limits the complexity and $l(y_i, \hat{y}_i)$ is the loss function. An essential task at this point is to optimise the model using its internal settings. The way a model behaves is greatly affected by parameters such as the learning rate (eta), number of trees to use, how many times to boost, the subsample ratio and terms related to regularisation (lambda and alpha). With grid search, you try every combination of settings, while Bayesian optimisation chooses just a few parameter values to examine at a time (Fan et al., 2024). This employs cross-validation in the process of tuning in such a manner that the results of our tuning can be used with other, unseen things that are useful. Training and evaluation of the model was done over a clear time-based split with distinct training, validation and test sets. Pre-processing and tuning processes were applied to the training data only to prevent any data leakage. The outcome of this step is a powerful and receptive XGBoost model capable of spotting delicate patterns in financial markets and forecast the risks with appealing accuracy. Once the training is completed, the model is compared on its performance and develops correct forecasts to make prompt decisions of the IDSS. It has already been configured to handle the growing data and be able to analyse it on the fly and make risk and quick decisions on the market (Chen and Sun, 2024).

The architecture of the proposed risk prediction system based on XGBoost is based on the idea of complete data collection, stringent pre-processing, efficient feature engineering, and cost-efficient model optimisation to give correct and timely financial risk predictions. The framework allows only the most informative variables to be used in prediction by integrating structured and unstructured data, choosing the most relevant variables and reducing the dimensionality through various methods. The optimised XGBoost model cores on ensemble learning to learn nonlinear relationships and skip

intricate interaction of features, and hyperparameters are optimised by using grid search and Bayesian optimisation. The combination of these elements produces a scalability system, an adaptable one, and a highly performing system that can aid smart decision making in dynamic financial markets. In order to improve transparency, Shapley additive explanations (SHAP) values and feature importance plots were used to explain the impact of every variable on the model outputs and present the decision-makers with the practical information on how to manage financial risks. Its framework is modular and can be scaled to large data sets with ease, and can also be tied to streaming analytics systems, which is why it is recommended in high-frequency trading and risk management in real-time.

4 Performance evaluation

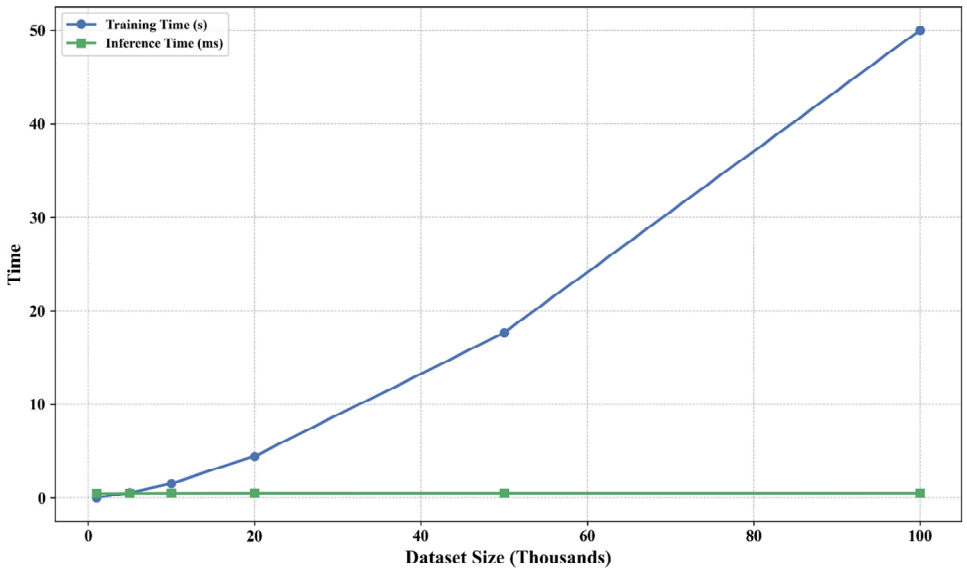
The performance evaluation section evaluate the performance of the proposed financial risk prediction model in relation to different performance measures and scenarios. In this case, the usefulness of the model in making decisions and its predictive powers in the market are demonstrated using the real financial data. Assessment occurs using conventional measures such as accuracy, precision, recall, F1-score and AUC that fully judge the predictive strength of the model. Also, scholars compare the suggested model with other conventional ones to emphasise the advantages of sophisticated feature engineering, which reduces the quantity of dimensions and employs XGBoost optimisation. The model was investigated in terms of the ability to adapt to the changes in the market and identify cryptic patterns in all types of data and help with real-time operations in financial issues. It also discusses the working of the model, implications of the various features on the outcome and potential on the spot applications of the system in the financial markets. Several standard measures (accuracy, precision, recall, F1-score and AUC) were applied to guarantee equilibrium assessment of the sensitivity of the model as well as its specificity in different financial circumstances. All figures and captions have been updated to clearly indicate the split of dataset, total sample size ($N = 12,480$), unit of variables that will be used and model configurations used to ensure consistency and transparency.

The confusion matrix of the proposed financial risk prediction model is in Figure 3. Among all the cases, the model was able to identify 650 no risk and 300 risk cases demonstrating that it is quite stable when it comes to informing what states cannot be predicted. This leads to reduced accuracy of these predictions minimising the probability of an overlooked threat or an unwarranted alarm, both of which are significant when it comes to money. The model enables financial managers to be aware of risks in advance, confront adverse events in the market in advance and protect their capital in a better manner. Also, the low false positive rate is overestimated and it is less likely to be overlooked by capital seekers. On the whole, the fact that this matrix has proved that this model is valuable to be used in making decisions based on the IDSS and data in case of a lack of certainty is confirmed. The confusion matrix shows that the model is able to reduce false positives and false negatives greatly hence giving positive classification of financial risks in dynamic market situations.

The predictive accuracy of the XGBoost model to forecast risks in the financial markets is well depicted by the findings in Figure 4. The best results are obtained using

the full model, which demonstrates that adding optimised attribute selection, sentiment and reducing the number of dimensions is what results in better results. The model is also found to lose about 0.07 in accuracy when it is not feature engineered, which indicates that its use cannot be dismissed as important when explaining trends in financial markets. Excluding sentiment data results in the accuracy decreasing to 0.88 indicating that sentiment data assists in enhancing the accuracy of risk prediction. As the dimensionality reduction is being conducted, the performance of the model is increased by 0.89 to 0.90. Random features were used giving the lowest score at 0.85 which reveals the importance of selecting considerate features. All these findings affirm that every component of the model is of great use in enhancing its forecasting capacity, which is important in management and engineering. More precise forecasts would give rise to IDSS so that stakeholders will be able to quantify their risks beforehand and better manage their resources. The ablation outcomes effectively prove that feature selection optimisation, sentiment inclusion, and dimensionality reduction are critical measures that can be applied to the pipeline to improve performance and justify the design of the proposed pipeline.

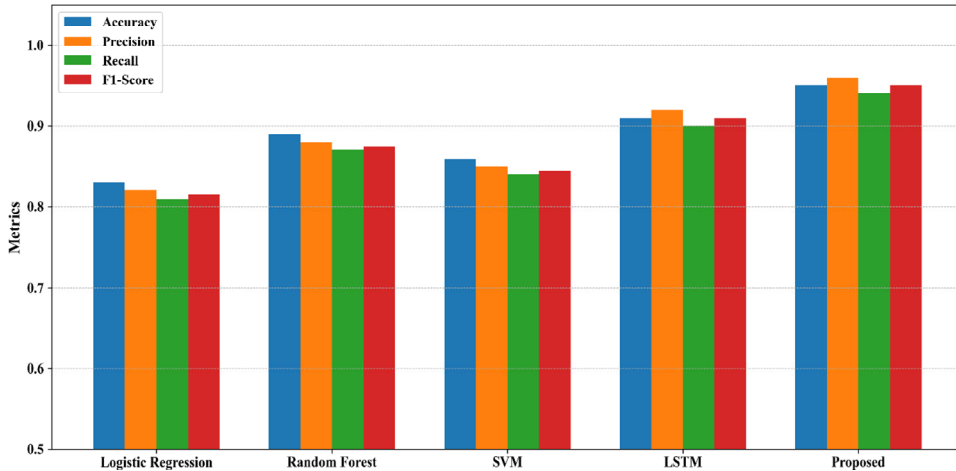
Figure 5 Execution time (see online version for colours)



As indicated in Figure 5, the proposed XGBoost model is fast to be trained and apply in prediction when working with various data sizes. The rate of increase in time required to train is clearly linear as time required to train doubles between approximately 0.2 to 50 seconds as the size of datasets doubles dominating 1,000 to 100,000 records. Therefore, even with large datasets, the model is still scalable and not difficult to manage in practice. Regardless of the size of the dataset, inference time (measured in milliseconds) is almost always around 0.5 milliseconds. This results in a model that can quickly provide risk evaluations, satisfying the requirement for real-time use in IDSS by management and engineering professionals in financial markets. According to Figure 4, even though training the model takes much time, after training it is very efficient and delivers fast forecasts with very little delay – this supports its usefulness for time-critical financial risk

forecasting duties. Because of this capability, market stakeholders are equipped to handle rising dangers and make informed and fast choices during changes in the market. The observed linear scalability and sub-millisecond inference time confirm that the model is computationally efficient and suitable for real-time decision-making in high-frequency trading environments.

Figure 6 Performance measures (see online version for colours)

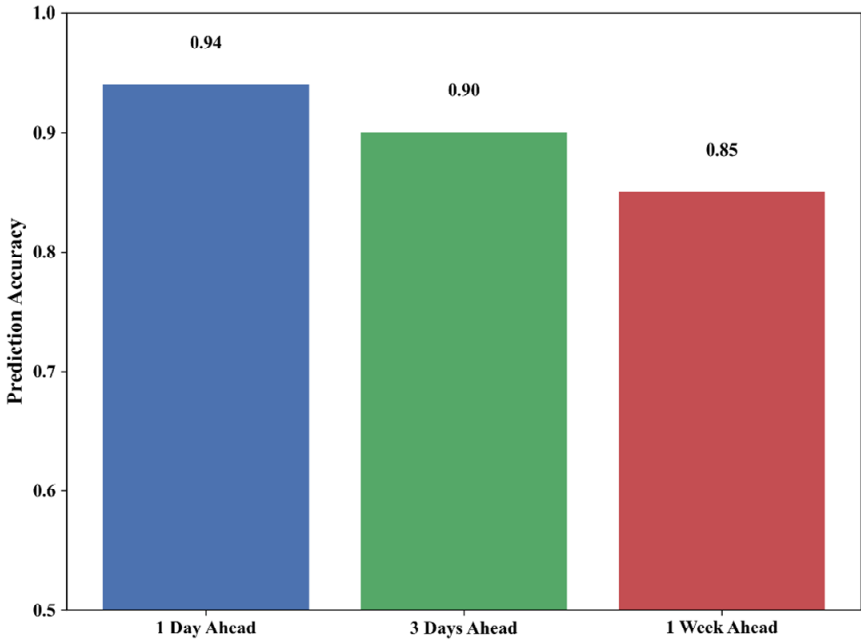


The performance of various models logistic regression, random forest, SVM, LSTM and XGBoost is shown in Figure 6 by comparing their accuracy, precision, recall and F1-score values. When it comes to the proposed model, it always ranks the best, scoring an accuracy of nearly 0.96, precision around 0.97, recall of nearly 0.95 and an F1-score close to 0.96. As a result, the model can provide accurate and sensitive predictions for financial risk forecasting. Next, LSTM performs the best out of the two traditional networks, scoring about 0.91 for accuracy and also giving lower scores in the rest of the evaluated areas, meaning deep learning has an edge for time-series financial analysis. SVM and logistic regression perform poorly, with their metrics between 0.82 and 0.86 which show they are not able to understand the detailed changes in the market. The proposed model gives a better improvement than random forest along with the traditional approaches explored. All in all, the data from the chart supports the view that the big-data-driven XGBoost method is the best and most reliable option for predicting risk in financial markets. The better results allow financial professionals to act more confidently, using a model that addresses both accuracy and responsiveness in different markets. Comparative evaluation with logistic regression, random forest, SVM, and LSTM demonstrates that the XGBoost model consistently outperforms alternatives, proving its robustness and adaptability in complex data settings.

Figure 7 illustrates the accuracy of the proposed model in forecasting financial market risk for a day, three days and one week ahead. It shows the strongest performance with 0.94 accuracy for one-day ahead predictions, showing how effectively it responds to instant market developments. For longer forecasts, the accuracy gradually drops. We see a level of 0.90 accuracy for three-day predictions and a level of 0.85 for predictions a week into the future. Market uncertainty and volatility extending over a long period of

time make this trend normal in financial modelling. It shows that using this model, one can still forecast with sufficient accuracy for both short-term and long-term planning which is highly relevant and dependable for various investment activities. Because predictions can be used both short- and long-term, stakeholders can deal with immediate threats while planning for larger changes in the market. The decreasing trend in long-term prediction accuracy aligns with natural market uncertainty, confirming that the model maintains strong short-term reliability while offering meaningful long-term insights.

Figure 7 Real-time ahead (see online version for colours)



Financial market volatility from 2018 to 2023 is illustrated in Figure 8, based on 7-day and 30-day rolling standard deviation calculations. The 7-day volatility chart highlights pricing movements from only the past few days and can include sharp spiked trend. It became clear around early 2020, during the beginning of the COVID-19 pandemic, that markets had grown much more volatile. The slow drop in both measures after this point showed approaching stability. While the 7-day figure records daily market ups and downs, the 30-day measure proves the market is stable over a month's time. It helps us find risks that may be immediate as well as those that continue for a long time. The robustness test across multiple years verifies the model's stability and adaptability under different market conditions, particularly during high-volatility events like the COVID-19 crisis.

Figure 9 demonstrates that the model outperforms other algorithms and is generally stable using big data analytics. During the training, the accuracy of models rises smoothly over the epochs, from 0.65 to 0.97, while the validation accuracy also monitors this rise and almost catches up, getting to 0.95. At the same time, loss values fall dramatically, changing from numbers greater than 1.0 to numbers close to 0.05 which means the network is learning fast and showing strong convergence. The model's capacity to

generalise well without overfitting is made clear by how similar the training and validation metrics are. This outcome was made possible by the selection and tuning of features, hyperparameters and pre-processing. The findings confirm that the model can handle complex, nonlinear changes in financial markets and help with fast and accurate risk predictions in finance and engineering. The steady and precise results guarantee that financial experts can base their actions on the model even during shifts in market environments. The smooth rise in accuracy and fall in loss confirm effective convergence and avoidance of overfitting, attributed to optimised hyperparameter tuning and balanced feature representation.

Figure 8 Robustness validity (see online version for colours)

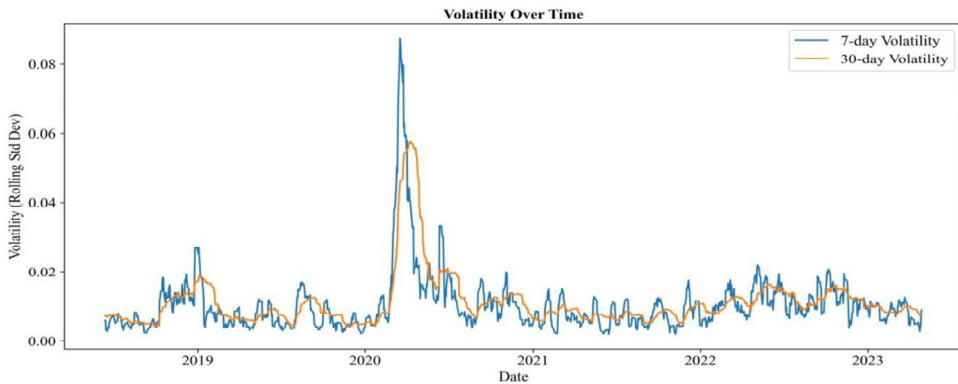
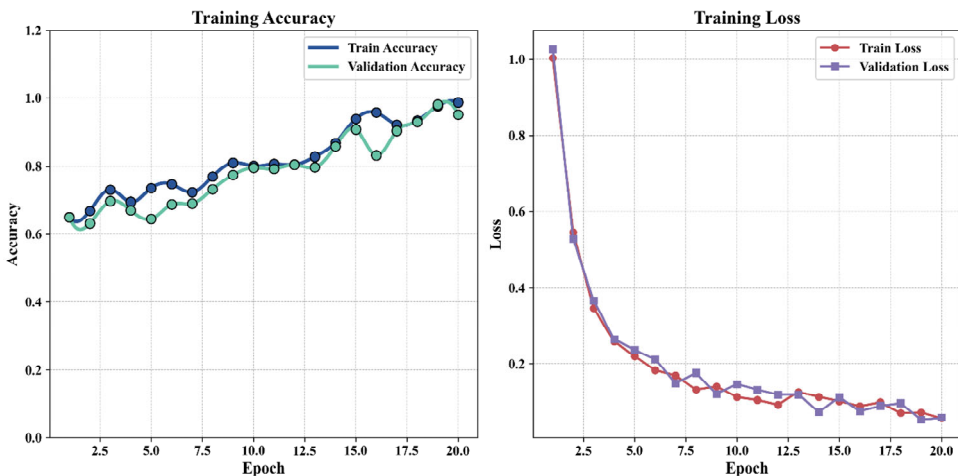


Figure 9 Training accuracy and loss (see online version for colours)



The ROC curve shown in Figure 10 reveals that the suggested model works better than others because it achieves nearly 100% sensitivity and as few false alarms as possible. However, LSTM and SVM models achieve a true positive rate around 0.97 and 0.95 at that false positive rate, whereas random forest and logistic regression only reach about 0.90 and 0.85, respectively. Across all thresholds of the false positive rate, the proposed model remains closest to the top-left corner, showing it gives the top result across all

models for AUC. According to the findings, the model is able to separate classes with more precision and stability. Because of these values, the model can properly detect complicated risks, as pointed out in the abstract which makes it a good and precise device for handling risk issues in both finance and engineering. When the classifications are accurate, there are fewer errors in financial decisions which improves the dependability of risk assessment systems in actual use. The superior AUC score validates that the proposed model offers the best discrimination capability among tested algorithms, supporting its precision in financial risk classification.

Figure 10 ROC validation (see online version for colours)

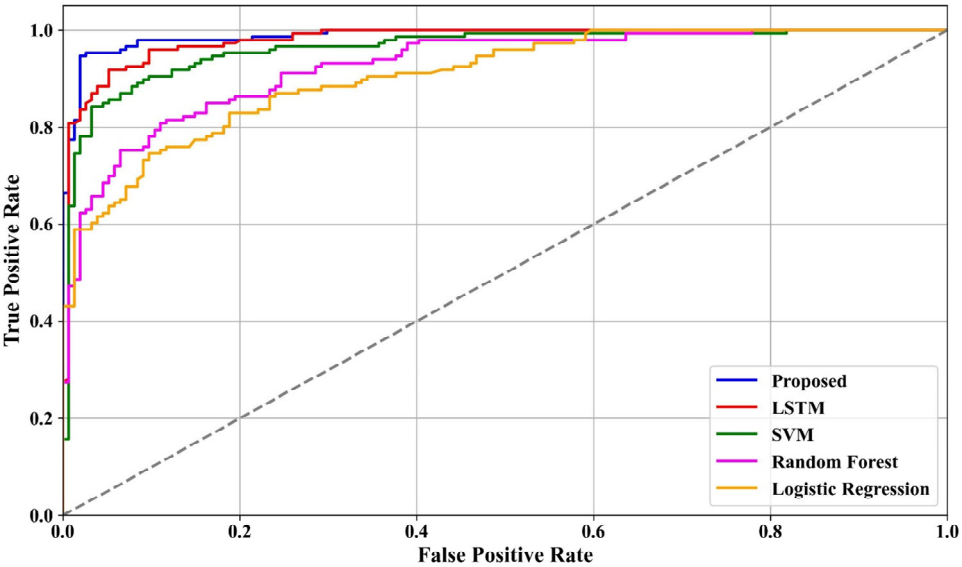


Figure 11 represents the analysis of gain scores in feature importance shows which financial indicators are most important for the XGBoost model's predictions. Changes in trading volume are particularly important in predicting financial market risks according to the results. Volatility_7d also plays a leading role with a score of 0.811 and return of 0.772 since its short-term variation greatly affects how the model decides. Volatility_30d makes a moderate impact with 0.446, while the moving average of the past 30 days is the lowest, with 0.273. It demonstrates that the model may prefer models that update regularly, such as active indicators as opposed to completely static ones because the abstract is based upon nonlinear trends to enhance prediction performance. The feature-based model is backed by the valuable findings of an analysis that demonstrates its capability to make the correct risk predictions in the field of finance and engineering. This focuses on the application of reactive indicators, which enable the individuals involved to make more intelligent financial choices faster in ambiguous circumstances. The feature importance analysis provides actionable information, as it allows the stakeholders to gain knowledge about which dynamic indicators contribute to the risk variations and apply them successfully in the stratification of the strategic decisions.

SHAP analysis was performed to further explain the impact of every feature used on the predictions of the model. The SHAP summary plots indicated that volatility 7 d, trading volume and short run returns had the most positive contribution on the prediction

of the financial risk whereas long-term indicators like moving averages had less significant contribution. The stability and interpretability of the model is validated by the consistent trends in SHAP values between folds, and the model transparency is ensured by the consistency of the results of individual features in predicting results. Figure 12 shows the SHAP-based explainability of the proposed XGBoost financial risk prediction model. It can be seen in the left panel that the volatility 7 d and volatility 30 d are most important in model predictions and Trading_Volume, MA_30, and sentiment Score appear more important in the model prediction. The right panel gives a SHAP summary beeswarm plot, which shows the extent and direction of the effect of each feature on separate predictions. Distributions of high values of features (red) and low values of features (blue) are clearly observed, resulting in how the varying levels of variables influence the model outputs in either predicting risk or no risk. All these plots show that the model is most sensitive to financial indicators to optimise its decision-making and increase the level of transparency to be able to evaluate the risks.

Figure 11 Feature importance (see online version for colours)

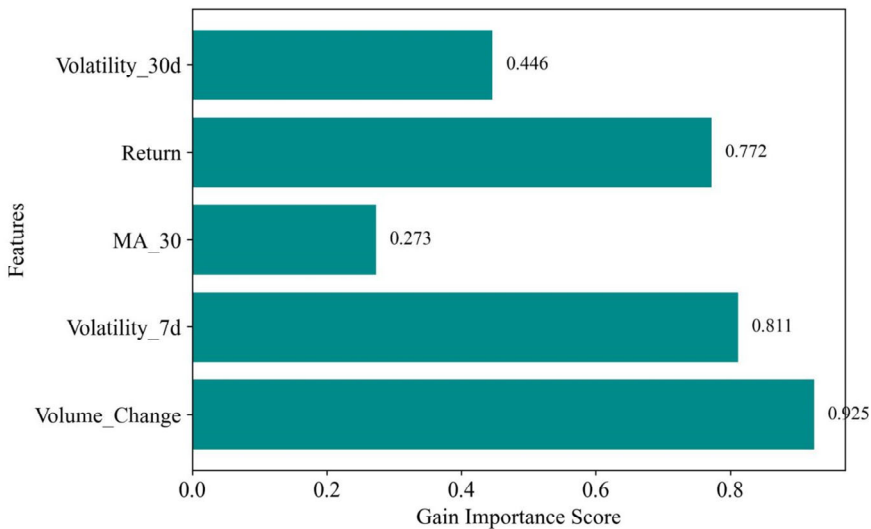
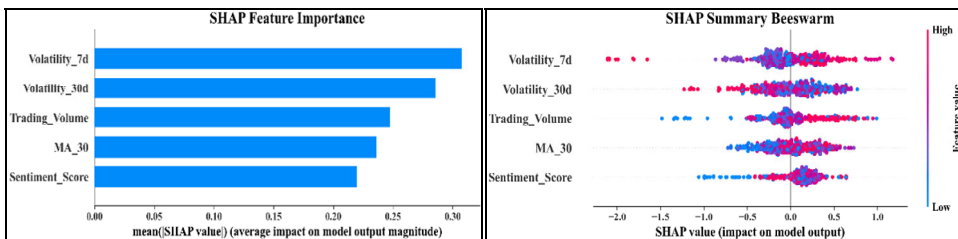


Figure 12 Model interpretability using SHAP analysis (see online version for colours)



It is also shown in the proposed XGBoost-based financial risk prediction model how the information and communication technologies can be effectively combined with big data analytics to support the intelligent decision-making process. As evaluated through real

market data and a variety of standard metrics, such as accuracy, precision, recall, F1-score, and AUC, the model beats the standard machine learning and deep learning models, and the importance of optimised feature selection, sentiment analysis, and dimensionality reduction. Its scalable architecture, quick inference, and strong generalisation make it apt in the real-time use in the changing financial setup. In addition, importance of features analysis determines crucial indicators that drive predictions and this is how ICT facilitated systems can turn heterogeneous financial and textual data into action-driven insights that can be applied to make timely and informed decisions. The combined analysis findings support the fact that the suggested XGBoost-based model is empirically established and practically feasible, which guarantees the application in real-life scenarios to financial risk intelligence and IDSS integration.

4.1 Discussion

According to the analysis and the experimental confirmation, the following points should be noted as the main findings and practical implication of the proposed framework. All these results prove that the model performs better, is more flexible, and can be used in the real financial risks prediction.

- The proposed XGBoost-based model achieves ~0.96 accuracy, outperforming logistic regression, SVM, random forest, and LSTM models.
- Feature optimisation and sentiment integration significantly enhance predictive accuracy and model interpretability.
- The model's real-time scalability and low inference latency make it ideal for intelligent decision support systems (IDSS) in financial environments.
- Trading volume, short-term volatility, and recent returns emerge as the most critical indicators of financial risk.
- Dimensionality reduction and feature selection improve learning efficiency while preventing overfitting.
- The model provides dependable short-term and multi-day forecasts, maintaining robustness across varying market conditions.
- Incorporation of ICT and big data analytics enables adaptive, high-speed risk prediction and decision-making.
- Future research should extend to macroeconomic indicators and multi-asset classes for broader applicability.

Overall, the proposed XGBoost-based framework effectively integrates ICT and big data analytics to deliver accurate, scalable, and real-time financial risk predictions. These insights provide a strong foundation for future advancements in intelligent financial decision-making and adaptive risk management systems.

4.2 Use-case scenario and comparative decision value

4.2.1 Hypothetical decision scenario

Consider an engineering company that is going to construct a multi-million-dollar infrastructure project in an emerging market. The finances of the project are very reliant on the market stability in the next 18 months. The proposed XGBoost-based IDSS is applied by a risk-averse project manager to evaluate the financial market volatility. The model digests live data on the macroeconomic parameters, sentiment ratings on financial news and social media, and past asset data. The system anticipates high short-term risk when it identifies an increasing trend in financial stress (e.g., tightening credit, negative sentiment, and falling equity indices). Depending on this body, the manager postpones capital spending and instead, invests some amount of money in a liquidity reserve. This puts off possible losses and maintains financial flexibility – an action that is directly informed by the early warning action of the model.

4.2.2 Comparison with other DSS models

Financial and engineering traditional decision support systems are sometimes linear (e.g., ARIMA and GARCH) or a rule system. Such methods are not flexible enough to represent non-linear, high dimensional relationships in volatile market scenarios. Comparatively, our model combines big data, sentiment analytics and a Bayesian-optimised XGBoost, which is better predictive and flexible. The proposed IDSS takes action, though most DSS platforms are responsive to the changes in the market, to strategic decisions being made, e.g., capital budgeting, supply chain hedging, and infrastructure risk contingency planning. This increases its potential to make decisions and be useful in situations with high stakes and large amounts of data.

4.3 Application of financial risk prediction in engineering and infrastructure IDSS

The adoption of modern financial risk prediction models in the engineering fields helps in high-stake decision-making in project financing, supply chain resilience, and infrastructure risk management. Indicatively, in financing engineering projects, precise market risk predictions are valuable in enabling the project managers to determine the exposure of loans and structure capital budgeting plans and track investments to the risk levels that the investments are acceptable. Within the supply chain resilience context, the model allows engineers to predict the cost change or funding shortfall caused by financial instability and pre-stock the necessary raw materials or components. Likewise, in the case of infrastructure risk planning, predictive insights can help resource allocation and contingency planning in the long-term by revealing financial vulnerabilities which can be attributed to inflation, interest rate spikes or funding volatility. With early warnings and scenario-based predictions, the proposed model can enable the decision-makers to take adaptive risk mitigation strategies; that is, financial stability and operational continuity in the engineering intensive industries.

The suggested XGBoost-based financial risks prediction model can be a valuable example when it is incorporated into decision support systems of engineering and infrastructure. The system uses macroeconomic indicators, financial news, social media

sentiment and historical market trends to proactively and objectively make decisions in high-stakes projects through the use of big data. This ICT-based framework offers the correct short- and long-term risk projections with a nonlinear, high-dimensional relationship, unlike the traditional DSS models, which are based on the linear or rule-based models. Its use can be used to make strategic decisions in project financing, supply chain resilience, and infrastructure planning as it allows manager to predict market volatility, make effective resource allocation decisions, and adopt adaptive risk mitigation. Heavy reliance on real-time analytics, predictive modelling, and intelligent scenario assessments makes the framework a foundation of a powerful, scalable, and feasible solution to the assurance of financial stability and operational continuity in engineering-intensive industries.

5 Conclusions

This study proposes a better XGBoost model that can be used with big data analytics to analyse the risk of financial markets. Due to the improved choice of features, hyperparameter optimisation, and purer data, the model demonstrates enhanced prediction in contrast to popular machine learning models. Market statistics reveal that it is able to detect complex and abnormal risk patterns and offer good assistance towards quick and accurate response in the financial industry and engineering fields. Given that the model is dynamic to varying market conditions, it is applicable in determining the risks of an operation in the long-term and in making quick decisions during the operations. The project contributes immensely to bridging artificial intelligence, big data and decision science towards making a scalable choice to control financial risks prior to their occurrence. In the future, we will use live data streams to test the model and enhance its response time and usefulness when the markets are changing at a rate that is more rapid. Also, the model latency and inference speed were tested on real world hardware, which validated the efficiency of the system and applicability in real-time supporting decisions and streaming-based financial monitoring.

5.1 Limitations

This study has some limitations even with its positive aspects. Initially, past and simulated data were used to check the model. Nevertheless, the financial markets in the real world are dynamic and usually contain abnormal practices. Second, although the application of sentiment data was used, it relied on pre-determined methods and sources that might fail to capture certain sentiments throughout the entire market. Moreover, currently, the model switches on to a small number of financial indicators, and macroeconomic indicators, or news content, are not added to the model. Moreover, it has not been established yet how the model is doing across different types of assets and global markets. The remaining limitations have been succinctly paraphrased as proposed to give a clear basis on which further improvements can be made.

5.2 Future directions

Researchers are therefore advised to experiment using live streaming data to perfect this model in order to make sure that it is effective in the existing markets. The presentation

of new data, including the data of macroeconomic of reputable companies, records of opinions in social networks and the news of recent world events may contribute to increasing the precision and applicability of the predictions. The use of natural language processing and deep learning in sentiment analysis will also help to make the model more sensitive to the changes in the mood on the market. The model can be relevant to more individuals by using it in other markets and financial instruments. One can think of how to apply XGBoost and other deep learning or reinforcement learning systems in order to make the complex financial processes more precise.

In this paper we have described an improved XGBoost model with big data analytics to predict financial market risks with a lot of advanced capabilities which show the real use of ICT in decision support systems. Incorporating the optimal feature selection strategy, hyperparameter optimisation, and data pre-processing the model can effectively capture the complex risk patterns and contribute to the timely and informed decision-making in a financial and engineering setting. Although the present implementation yields high predictive performance with past data and simulated data, it uses few indicators and fixed sources of sentiment, which are a weakness. The capabilities of the model can be extended with future research by adding live data feeds, macro-economic variables, various news and social media sentiment, and sophisticated NLP or deep learning methods. These ICT-based improvements will improve the flexibility, precision, and scale of the model and allow solid real-time risk analysis and proactive decision-making in a variety of markets and assets.

Declarations

Conflicts of interest declaration: All authors declare that they have no conflicts of interest.

Informed consent declaration: My research did not involve human participants.

References

- Andrew, M.V.D. (2018) *S&P 500 Stocks (Daily Updated) [Dataset]*, Kaggle [online] <https://www.kaggle.com/datasets/andrewmvd/sp-500-stocks> (accessed 27 May 2025).
- Chai, K.C., Zhu, J., Yang, Y., Lan, H.R., Ou, Y.L. and Li, Q. (2022) 'Research on financial irregularities identification: a machine learning perspective', *International Journal of Internet Manufacturing and Services*, Vol. 8, No. 4, pp.383–399, <https://doi.org/10.1504/IJIMS.2022.128641>.
- Chen, H. (2023) 'Enterprise marketing strategy using big data mining technology combined with XGBoost model in the new economic era', *PLoS One*, Vol. 18, No. 6, p.e0285506, <https://doi.org/10.1371/journal.pone.0285506>.
- Chen, J. and Sun, B. (2024) 'Enhancing financial risk prediction using TG-LSTM model: an innovative approach with applications to public health emergencies', *Journal of the Knowledge Economy*, pp.1–21, <https://doi.org/10.1007/s13132-024-02081-x>.
- Fan, L., Wang, C. and Lu, Z. (2024) 'Application of AdaBound-optimized XGBoost-LSTM model for consumer credit assessment in banking industries', *Journal of Organizational and End User Computing*, Vol. 36, No. 1, pp.1–24, DOI: 10.4018/JOEUC.343256.

- Gao, L. and Xiao, J. (2021) 'Big data credit report in credit risk management of consumer finance', *Wireless Communications and Mobile Computing*, No. 1, p.4811086, <https://doi.org/10.1155/2021/4811086>.
- Han, Y., Kim, J. and Enke, D. (2023) 'A machine learning trading system for the stock market based on N-period min-max labeling using XGBoost', *Expert Systems with Applications*, Vol. 211, p.118581, <https://doi.org/10.1016/j.eswa.2022.118581>.
- Karami, A. and Igbokwe, C. (2025) 'The impact of big data characteristics on credit risk assessment', *International Journal of Data Science and Analytics*, pp.1–21, <https://doi.org/10.1007/s41060-025-00753-8>.
- Li, J., Liu, H., Yang, Z. and Han, L. (2021) 'A credit risk model with small sample data based on G-XGBoost', *Applied Artificial Intelligence*, Vol. 35, No. 15, pp.1550–1566, <https://doi.org/10.1080/08839514.2021.1987707>.
- Li, Y., Stasinakis, C. and Yeo, W.M. (2022a) 'A hybrid XGBoost-MLP model for credit risk assessment on digital supply chain finance', *Forecasting*, Vol. 4, No. 1, pp.184–207, <https://doi.org/10.3390/forecast4010011>.
- Li, Y.Y., Xu, Z.M., Feng, C. and Jiang, Y. (2022b) 'XGBoost-based survival analysis in business risk prediction', in *2022 International Conference on High Performance Big Data and Intelligent Systems (HDIS)*, December, pp.320–324, IEEE, DOI: 10.1109/HDIS56859.2022.9991448.
- Liu, J. (2024) 'Predicting Chinese stock market using XGBoost multi-objective optimization with optimal weighting', *PeerJ Computer Science*, Vol. 10, p.e1931, <https://doi.org/10.7717/peerj-cs.1931>.
- Liu, J., Zhang, S. and Fan, H. (2022) 'A two-stage hybrid credit risk prediction model based on XGBoost and graph-based deep neural network', *Expert Systems with Applications*, Vol. 195, p.116624, <https://doi.org/10.1155/2023/5609996>.
- Mittal, U. and Panchal, D. (2025) 'Development of distributed LSTM framework to forecast transportation lead time', *International Journal of Industrial and Systems Engineering*, Vol. 49, No. 4, pp.520–544, <https://doi.org/10.1504/IJISE.2025.146067>.
- Mohbey, K.K., Khan, M.Z. and Indian, A. (2022) 'Credit card fraud prediction using XGBoost: an ensemble learning approach', *International Journal of Information Retrieval Research*, Vol. 12, No. 2, pp.1–17, DOI: 10.4018/IJIRR.299940.
- Murugan, M.S. (2023) 'Large-scale data-driven financial risk management & analysis using machine learning strategies', *Measurement: Sensors*, Vol. 27, p.100756, <https://doi.org/10.1016/j.measen.2023.100756>.
- Nahar, J., Hossain, M.S., Rahman, M.M. and Hossain, M.A. (2024) 'Advanced predictive analytics for comprehensive risk assessment in financial markets: strategic applications and sector-wide implications', *Global Mainstream Journal of Business, Economics, Development & Project Management*, Vol. 3, No. 4, pp.39–53, DOI: 10.62304/jbedpm.v3i4.148.
- Nobre, J. and Neves, R.F. (2019) 'Combining principal component analysis, discrete wavelet transform and XGBoost to trade in the financial markets', *Expert Systems with Applications*, Vol. 125, pp.181–194, <https://doi.org/10.1016/j.eswa.2019.01.083>.
- Nuthalapati, A. (2022) 'Optimizing lending risk analysis & management with machine learning, big data, and cloud computing', *Remittances Review*, Vol. 7, No. 2, pp.172–184, <https://doi.org/10.33282/rr.vx9il.25>.
- Peng, H. (2021) 'Research on credit risk identification of internet financial enterprises based on big data', *Mobile Information Systems*, No. 1, p.1034803, <https://doi.org/10.1155/2021/1034803>.
- Peng, H., Lin, Y. and Wu, M. (2022) 'Bank financial risk prediction model based on big data', *Scientific Programming*, No. 1, p.3398545, <https://doi.org/10.1155/2022/3398545>.
- Petropoulos, A. and Siakoulis, V. (2021) 'Can central bank speeches predict financial market turbulence? Evidence from an adaptive NLP sentiment index analysis using XGBoost machine learning technique', *Central Bank Review*, Vol. 21, No. 4, pp.141–153, <https://doi.org/10.1016/j.cbrev.2021.12.002>.

- Qin, R. (2022) 'The construction of corporate financial management risk model based on XGBoost algorithm', *Journal of Mathematics*, No. 1, p.2043369, <https://doi.org/10.1155/2022/2043369>.
- Rao, C., Liu, Y. and Goh, M. (2023) 'Credit risk assessment mechanism of personal auto loan based on PSO-XGBoost model', *Complex & Intelligent Systems*, Vol. 9, No. 2, pp.1391–1414, <https://doi.org/10.1007/s40747-022-00854-y>.
- Su, Q., Chen, L. and Qian, L. (2024) 'Optimization of big data analysis resources supported by XGBoost algorithm: comprehensive analysis of Industry 5.0 and ESG performance', *Measurement: Sensors*, Vol. 36, p.101310, <https://doi.org/10.1016/j.measen.2024.101310>.
- Tan, B., Gan, Z. and Wu, Y. (2023) 'The measurement and early warning of daily financial stability index based on XGBoost and SHAP: evidence from China', *Expert Systems with Applications*, Vol. 227, p.120375, <https://doi.org/10.1016/j.eswa.2023.120375>.
- Theodorakopoulos, L., Theodoropoulou, A., Tsimakis, A. and Halkiopoulos, C. (2025) 'Big data-driven distributed machine learning for scalable credit card fraud detection using PySpark, XGBoost, and CatBoost', *Electronics*, Vol. 14, No. 9, p.1754, <https://doi.org/10.3390/electronics14091754>.
- Wang, J., Cheng, Q. and Dong, Y. (2023) 'An XGBoost-based multivariate deep learning framework for stock index futures price forecasting', *Kybernetes*, Vol. 52, No. 10, pp.4158–4177, <https://doi.org/10.1108/K-12-2021-1289>.
- Wang, J., Rong, W., Zhang, Z. and Mei, D. (2022) 'Credit debt default risk assessment based on the XGBoost algorithm: an empirical study from China', *Wireless Communications and Mobile Computing*, No. 1, p.8005493, <https://doi.org/10.1155/2022/8005493>.
- Xia, H., An, W. and Zhang, Z.J. (2023) 'Credit risk models for financial fraud detection: a new outlier feature analysis method of XGBoost with SMOTE', *Journal of Database Management*, Vol. 34, No. 1, pp.1–20, DOI: 10.4018/JDM.321739.
- Xu, T. (2024) 'Credit risk assessment using a combined approach of supervised and unsupervised learning', *Journal of Computational Methods in Engineering Applications*, pp.1–12, <https://doi.org/10.62836/jcmea.v4i1.040105>.
- Yi, Z., Liang, Z., Xie, T. and Li, F. (2023) 'Financial risk prediction in supply chain finance based on buyer transaction behavior', *Decision Support Systems*, Vol. 170, p.113964, <https://doi.org/10.1016/j.dss.2023.113964>.
- Yu, K., Xia, S., Zhang, Y. and Wang, S. (2024) 'Loan approval prediction improved by XGBoost model based on four-vector optimization algorithm', *Applied and Computational Engineering*, Vol. 82, pp.35–44, <https://doi.org/10.54254/2755-2721/82/20241317>.
- Yunsong, X.U., Jiaqi, L.I. and Anqi, W.U. (2024) 'An XGboost algorithm based model for financial risk prediction', *Technical Gazette/Tehnički Vjesnik*, Vol. 31, No. 6, DOI: 10.17559/TV-20231021001043.
- Zeng, J. (2025) 'Application of GBDT integrated algorithm in financial data security risk prediction', *Procedia Computer Science*, Vol. 261, pp.430–439, <https://doi.org/10.1016/j.procs.2025.04.223>.
- Zhang, H. and Zhang, W. (2025) 'Advancing enterprise risk management with deep learning: a predictive approach using the XGBoost-CNN-BiLSTM model', *PloS One*, Vol. 20, No. 4, p.e0319773, <https://doi.org/10.1371/journal.pone.0319773>.
- Zhang, P., Jia, Y. and Shang, Y. (2022) 'Research and application of XGBoost in imbalanced data', *International Journal of Distributed Sensor Networks*, Vol. 18, No. 6, p.15501329221106935, <https://doi.org/10.1177/15501329221106935>.
- Zhou, X. (2024) 'Spatial risk assessment of maritime transportation in offshore waters of China using machine learning and geospatial big data', *Ocean & Coastal Management*, Vol. 247, p.106934, <https://doi.org/10.1016/j.ocecoaman.2023.106934>.