Journal of Infrastructure, Policy and Development 2024, 8(8), 4972.
https://doi.org/10.24294/jipd.v8i8.4972

Article

Advancing financial analytics: Integrating XGBoost, LSTM, and Random Forest Algorithms for precision forecasting of corporate financial distress

Farida Titik Kristanti[1,*], Mochamad Yudha Febrianta[2], Dwi Fitrizal Salim[2], Hosam Alden Riyadh[1,3], Yoga Sagama[2], Baligh Ali Hasan Beshr[3]

[1] Department of Accounting, School of Economics and Business Telkom University, Bandung 40257, Indonesia
[2] Department of Management, School of Economics and Business Telkom University, Bandung 40257, Indonesia
[3] Administrative Science Department, College of Administrative and Financial Science, Gulf University, Sanad 26489, Kingdom of Bahrain
*** Corresponding author:** Farida Titik Kristanti, faridatk@telkomuniversity.ac.id

CITATION

Kristanti FT, Febrianta MY, Salim DF, et al. (2024). Advancing financial analytics: Integrating XGBoost, LSTM, and Random Forest Algorithms for precision forecasting of corporate financial distress. Journal of Infrastructure, Policy and Development. 8(8): 4972. https://doi.org/10.24294/jipd.v8i8.4972

ARTICLE INFO

Received: 3 March 2024
Accepted: 26 April 2024
Available online: 7 August 2024

Abstract: This study thoroughly examined the use of different machine learning models to predict financial distress in Indonesian companies by utilizing the Financial Ratio dataset collected from the Indonesia Stock Exchange (IDX), which includes financial indicators from various companies across multiple industries spanning a decade. By partitioning the data into training and test sets and utilizing SMOTE and RUS approaches, the issue of class imbalances was effectively managed, guaranteeing the dependability and impartiality of the model's training and assessment. Creating first models was crucial in establishing a benchmark for performance measurements. Various models, including Decision Trees, XGBoost, Random Forest, LSTM, and Support Vector Machine (SVM) were assessed. The ensemble models, including XGBoost and Random Forest, showed better performance when combined with SMOTE. The findings of this research validate the efficacy of ensemble methods in forecasting financial distress. Specifically, the XGBClassifier and Random Forest Classifier demonstrate dependable and resilient performance. The feature importance analysis revealed the significance of financial indicators. Interest_coverage and operating_margin, for instance, were crucial for the predictive capabilities of the models. Both companies and regulators can utilize the findings of this investigation. To forecast financial distress, the XGB classifier and the Random Forest classifier could be employed. In addition, it is important for them to take into account the interest coverage ratio and operating margin ratio, as these finansial ratios play a critical role in assessing their performance. The findings of this research confirm the effectiveness of ensemble methods in financial distress prediction. The XGBClassifier and RandomForestClassifier demonstrate reliable and robust performance. Feature importance analysis highlights the significance of financial indicators, such as interest coverage ratio and operating margin ratio, which are crucial to the predictive ability of the models. These findings can be utilized by companies and regulators to predict financial distress.

Keywords: decision trees; financial distress; LSTM Random Forest; XGBoost
JEL Clasification: C15; C45; G32; G33; M41

1. Introduction

The pursuit of financial sustainability is a fundamental objective for several companies, as it serves as a crucial safeguard against the perils of bankruptcy, which can have far-reaching ramifications on both the economic and social fronts. Hence, the timely identification of challenges assumes paramount significance, as it enables the implementation of proactive interventions. According to Sharma and Mahajan (1980), the identification of problems enables the implementation of regular activities aimed at mitigating the costs associated with failure.

In the early stages of bankruptcy prediction research, Fitzpatrick (1932) employed financial ratios and rates as predictive indicators, without resorting to the utilization of statistical models. The statistical model introduced by Altman (1968) employing multiple discriminant analysis continues to be widely employed by practitioners for the purpose of forecasting corporate health issues. The logit model was initially formulated by Ohlson (1980), explained that in predicting there are 3 things that must be considered, firstly paying attention to all the information in the company's financial reports, secondly using large or large amounts of data to predict and also paying attention to stronger factors. To strengthen it in further research, it is necessary to add variables. The basic variables used include Size, TLTA, WCTA, CLCA, OENEG, NITA, FUTL, INTWO, CHIN. Selecting a large sample that is close to the population will result in bias, so the appropriate method for predicting this problem can be using the Probit model (Zmijewski 1984). Shumway (2001) proposed a hazard model that incorporates time factors for the purpose of predicting business financial hardship and consider basic variables such as Size, past returns. In Bonello et al.'s (2018) research on US state companies using 3–5 years of back data to be trained in predicting the level of corporate bankruptcy. the methods used include Decision Tree, the Naïve Bayes classifier and the Artificial Neural. The methods mentioned are part of the Machine Learning method that can predict the level of bankruptcy of companies, Machine Learning is highly dependent on the variables used to strengthen prediction models such as company profitability, liquidity, management efficiency, leverage, company size, and industry type.

Noviantoro and Huang (2021) Testing to predict the level of visitor behavior at e-commerce companies in Turkey by comparing a number of models such as Machine Learning Algorithms, Decision Tree, Random Forest, Neural Net, Deep Learning, Naïve Bayes, K-NN Classifier, Logistics Regression, and the results show that Random Forest get more precise results in predicting the level of behavior of ecommerce visitors in Turkey. So that machine learning methods can be used in the current era of digitalization to predict all problems, not only company finances, consumer behavior and others. Precise results depend on the variables used, and the length of time the data is used to train the model that has been created.

During the early 1990s, several researchers, including Altman et al. (1994), the ANN model is better than traditional models such as linear discriminant (LDA) or logit analysis in Italian companies, the model used was tested separately between retail and manufacturing companies. Kristanti et al. (2023) machine learning research has also been developed on the type of construction companies in Indonesia with the best ANN model of 25 inputs, 20 hidden layer neurons, and 1 best output model whose results predict that of the 17 research samples tested there are 6 companies experiencing distress and the rest are not distress. Added by Kristanti and Dhaniswara (2023), on companies in Indonesia who found a prediction model of the level of distressed by including several variables such as current ratio (CR), return on assets (ROA), debt to asset ratio (DAR), total asset turnover (TATO), and cash flow to debt ratio was tested with the larning machine method and logistic regression. The results obtained that the machine learning model is better than the model designed by logistic regression to predict the level of bankruptcy in Indonesia. Prediction of company distress levels was carried out in Slovakia using the CART algorithm which produces

a binomial decision tree where the results obtained by the model have a high level of accuracy with simpler tests, the sample used is also relatively short from 2016–2018 (Durica et al., 2021). The Random Forest model has been marketed by Zhong and Wang (2022) to manufacturing companies in China in 2016–2021. It is necessary to be aware that the Profitability variable has a very large contribution to predicting the level of company distress which is reflected in the increase in return on equity (Zhong and Wang, 2022). In research in China in 2007–2017 the Random Forest (RF) model was better than the Support Vector Machine (SVM), Decision Tree (DT), baggingDT, oblique random forest (obRF), Kernel ridge regression (KRR) and Bayes models. in DFDF data classification (Shen, Liu, et al., 2020). There are differences in results from previous research which prove that the machine learning method is better than traditional regression models, so this research aims to create and prove that the variables used in the research are able to produce more precise results for companies in Indonesia.

Prior studies employed statistical methodologies and machine learning techniques. The study employs machine learning techniques and utilizes a diverse range of ways to get optimal outcomes. The objective of this study is to employ machine learning techniques, specifically Decision Tree (DT), Random Forest (RF), Long Short-Term Memory (LSTM), and Vector Machine (VM), to forecast financial distress in companies that are publicly traded on the Indonesian Stock Exchange. The utilization of financial ratios facilitates the formulation of prognostications. Furthermore, this study will also evaluate the performance and conduct feature selection analysis for each model. The findings of this study are anticipated to enhance understanding, specifically in the realm of financial management, particularly in the domain of predicting financial instability through the utilization of diverse machine learning techniques. The study findings are anticipated to offer further insights for enterprises and authorities, particularly in Indonesia, on the optimal machine learning model and the key ratio that significantly impacts their performance.

The structure of this paper's subsequent sections is as follows: First, the authors identify the research background by using the examination of the theoretical framework employed in this investigation in Section 2. In Section 3, the approach is discussed, while Part 4 is dedicated to presenting the outcome and facilitating subsequent discussion. This paper will conduct a comprehensive analysis, formulate a conclusion, and then present recommendations.

2. Literature review

Financial distress prediction is a critical domain in finance, and machine learning algorithms have increasingly become pivotal tools for enhancing prediction accuracy. This literature review explores the effectiveness of various machine learning algorithms, such as Decision Trees (DT), Random Forest (RF), Long Short-Term Memory (LSTM), and Support Vector Machine (SVM), in forecasting financial distress (Ling and Cai, 2022; Liu, Li, et al., 2022). These studies reveal that both traditional and novel machine learning techniques are effective at predicting when a company will encounter financial difficulties, with methodologies such as the advanced AWOA-DL method achieving remarkable accuracies in distress prediction

(Elhoseny et al., 2022).

Recent research highlights the superiority of machine learning models over traditional models like the Z-Score, particularly in predicting financial distress among Chinese A-listed construction firms (Rahman and Zhu, 2024). This suggests a broader applicability of these machine learning models across different markets and industries. Furthermore, innovative machine learning approaches, such as wolf pack-optimized long-term and short-term memory neural networks, have proven effective in crisis prediction, adding a new dimension to predictive analytics (Ling and Cai, 2022).

Notably, discussions encompass the trade-off between interpretability and performance in machine learning algorithms, underscoring the importance of balancing model complexity with explanatory power. Moreover, the application of tree-based gradient boosting models exemplifies efforts to enhance both modeling and explanation in distress prediction (Liu, Li, et al., 2022).

Expanding further into the fintech sector, Halteh et al. (2024) employ Artificial Neural Networks (ANNs) to forecast financial distress among FinTech unicorns. Their findings, which highlight the importance of financial ratios like return on capital, current ratio, quick ratio, and debt-to-equity ratio as significant predictors of financial distress within FinTech unicorns.

Moreover, studies continue to uncover specific financial ratios that serve as significant predictors of distress, such as return on capital employed, cash flows to total liability, and debt to equity ratio (Sehgal et al., 2021). Additionally, investigations into diverse factors influencing financial distress, such as corporate social responsibility, indicate a growing complexity in the factors that predictive models must consider (Song, 2023). This comprehensive analysis underscores the multifaceted nature of financial distress prediction and the crucial role played by advanced machine learning techniques in navigating this complex field.

Despite the advancements in financial distress prediction methodologies and their expanding applicability across various sectors, there remains a significant gap in targeted research within specific emerging markets, notably Indonesia. The unique economic dynamics and regulatory environments in Indonesia necessitate tailored analytical models that can accurately predict financial distress in this context. This study aims to fill this critical gap by integrating sophisticated machine learning algorithms such as XGBoost, LSTM, and Random Forest to develop a robust predictive framework suited to the Indonesian market. The need for this research is underscored by the increasing complexity of financial markets and the crucial role that precise, reliable financial distress predictions play in ensuring the stability and health of companies within these markets. By focusing on Indonesia, this research not only contributes to the broader field of financial analytics but also provides actionable insights that can significantly benefit Indonesian regulators and companies in mitigating financial risks.

2.1. Decision trees

Financial distress prediction within companies constitutes a pivotal area of research, with a notable shift towards the adoption of machine learning techniques, particularly Decision Trees (DT). Numerous studies have underscored the

effectiveness of machine learning models in this regard. Sehgal et al. (2021) specifically highlighted the superiority of machine learning-based distress prediction models over traditional time series models, focusing on developing tailored models within the Indian corporate landscape. Elhoseny et al. (2022) further emphasized the significance of machine learning models, particularly deep learning-based approaches, in predicting financial distress, marking a notable shift in the financial risk assessment paradigm.

Moreover, the literature elucidates the diverse applications and methodologies within the field, including feature selection, dataset considerations, and algorithm performance evaluation (Long et al., 2022). Notably, Song (2023) addressed the challenges of class-imbalanced datasets and introduced techniques like SMOTE to enhance prediction accuracy. Additionally, Ryll and Seidens (2019) explored non-linearities in time series data for financial market prediction, indicating the continuous evolution and integration of advanced methodologies in financial distress prediction.

The collective body of research underscores a growing trend towards leveraging advanced computational methods to enhance predictive accuracy and efficiency in financial risk assessment, exemplifying the evolving landscape of financial distress prediction (Ryll and Seidens, 2019). This integration of advanced methodologies and exploration of diverse factors not only enriches the predictive capabilities but also contributes significantly to proactive risk management and strategic decision-making within companies.

2.2. Random forest

Financial distress prediction within corporate entities has emerged as a critical focus in financial research, particularly accentuated by the adoption of machine learning techniques, notably Random Forest (RF). This interest is underscored by a wealth of studies dedicated to developing and deploying statistical and machine learning models for this purpose, aiming to enhance predictive accuracy amidst the complexities of financial landscapes (Gregová et al., 2020; Tron et al., 2022). Notably, these models have demonstrated superiority over traditional time series methods, especially in contexts marked by heightened corporate financial distress (Sehgal et al., 2021). Central to these endeavors is the identification of multifaceted factors—financial, managerial, and textual—that collectively contribute to the manifestation of financial distress characteristics within companies.

The exploration of various machine learning algorithms, such as support vector machines and sparse algorithms, has been instrumental in crafting robust financial distress prediction models, with hybrid approaches showcasing notable efficacy (Shen and Chen, 2022). Moreover, advancements in feature selection methodologies, facilitated by genetic algorithms, have significantly augmented the predictive capabilities of these models, particularly evident in the Chinese listed company context (Song, 2023).

In tandem, recent research has emphasized the pivotal role of machine learning in enhancing financial forecasting endeavors, transcending the realm of distress prediction to encompass broader financial decision-making processes (Hota et al., 2020). Such endeavors have not only outstripped conventional linear techniques in

forecasting accuracy but have also showcased promising applications in predicting stock market indices, volatility, and optimizing inventory management (Karathanasopoulos and Osman, 2019; Nasution et al., 2022; Ramos-Pérez et al., 2019). Amidst this backdrop, the comparative analyses delineated in recent literature underscore the significance of selecting appropriate models and discerning specific financial indicators to refine the accuracy of distress predictions, thus elucidating the invaluable role of machine learning in fortifying financial decision-making processes and risk management endeavors (Gregová et al., 2020; Tron et al., 2022).

2.3. XGBoost

Financial distress prediction, a pivotal domain within finance, is integral for anticipating and navigating potential economic challenges encountered by companies. This area has witnessed a surge in interest, notably with the ascension of machine learning techniques, particularly XGBoost, renowned for its predictive prowess (Tissaoui et al., 2022). Demonstrating superior accuracy over traditional methods, these machine learning models, including SVM, deep learning, and ensemble methods, have emerged as formidable tools in forecasting financial distress (Ayuni et al., 2022; El-Bannany et al., 2020).

Critical to enhancing predictive performance is meticulous feature selection, facilitated by methodologies like genetic algorithms, and the integration of diverse data sources spanning financial, textual, and social responsibility domains (Song, 2023). Comparative analyses underscore the efficacy of XGBoost in corporate financial distress prediction, notably outperforming traditional techniques like linear regression and ensemble methods (Tissaoui et al., 2022). Further accentuating the significance of model selection and evaluation, researchers have scrutinized the performance of XGBoost against alternatives like random forest and support vector machines, illuminating the nuanced strengths and limitations of each approach (Lai et al., 2023).

Complementary research delves into the application of diverse methodologies, including SVM, ANN, Cox Proportional Hazard model, and Altman Z-Score method, in predicting financial distress across various sectors (Kristanti et al., 2023). Integrating these methodologies underscores the importance of amalgamating different paradigms to fortify the accuracy and reliability of financial distress forecasts, ultimately empowering proactive risk management and strategic decision-making within companies. These endeavors collectively underscore the evolving landscape of financial distress prediction, driven by the integration of advanced computational techniques and traditional financial models.

2.4. Long short-term memory

Financial distress prediction within companies has emerged as a pivotal area of research, attracting substantial attention in recent years. A notable facet of this research involves the utilization of machine learning techniques, with particular emphasis on Long Short-Term Memory (LSTM) models, renowned for their adeptness in forecasting financial distress. This trend is underscored by studies exploring LSTM's applicability across various financial forecasting tasks, spanning stock market

prediction, forex forecasting, and energy consumption prediction (Sheng and Ma, 2022).

In parallel, a surge in interest has been observed regarding the application of machine learning algorithms, including LSTM, for predicting financial distress within corporate entities (Yousaf et al., 2021). While the specific contexts may vary, these studies collectively underscore the growing trend of leveraging advanced computational techniques to enhance distress prediction accuracy. Highlighted the significance of financial ratios in distress prediction, while Elhoseny et al. (2022) demonstrated the effectiveness of the AWOA-DL method. Moreover, the research by Yousaf et al. (2021) extends this exploration to evaluate the impact of board diversity on predicting financial distress, thereby enriching the discourse surrounding machine learning's effectiveness in this domain.

The integration of LSTM models within financial distress prediction frameworks offers a promising avenue for enhancing decision-making processes in the financial domain. LSTM's capability to capture long-term dependencies and nonlinear patterns in time series data renders it a valuable tool for forecasting financial indicators. As evidenced by the studies reviewed, LSTM demonstrates effectiveness across diverse financial forecasting tasks, thus emphasizing its potential for improving decision-making processes in the financial realm (Hájek and Munk, 2023; Sirisha et al., 2022). Collectively, these studies underscore the burgeoning interest in leveraging advanced computational techniques, particularly LSTM, to enhance the accuracy of financial distress prediction within companies.

2.5. Support vector machine

Financial distress prediction, a cornerstone of corporate finance research, has witnessed burgeoning interest, particularly with the advent of machine learning methodologies. Studies abound in leveraging Support Vector Machine (SVM), showcasing its versatility across various sectors such as manufacturing, real estate, infrastructure, transportation, and banking (Ayuni et al., 2022). This methodological diversity underscores SVM's robustness in analyzing financial data, providing invaluable insights into the likelihood of companies encountering financial adversities.

Expanding the repertoire, Artificial Neural Networks (ANN) have emerged as potent alternatives for financial distress prediction, offering nuanced approaches to pattern recognition and analysis (Kristanti et al., 2023). Noteworthy efforts have been directed towards refining predictive models through feature selection methods and sampling techniques, aiming to bolster their performance and reliability (Vu et al., 2019).

Crucially, financial distress prediction models encompass a gamut of financial ratios and indicators, encompassing liquidity, profitability, leverage, and company size, as pivotal factors (Kholisoh and Dwiarti, 2020). The comprehensive evaluation further extends to encompassing both internal and external factors, emphasizing their profound impact on a company's financial stability (Ye et al., 2020).

Parallelly, a rich tapestry of statistical models, including the Altman Z-score model and logistic regression, has been instrumental in prognosticating financial distress, leveraging historical financial data and key performance indicators

(Khamisah et al., 2021). Complementary efforts have delved into integrating disparate models and techniques to furnish a holistic assessment of financial risk (Dai et al., 2022).

Simultaneously, recent research underscores the pivotal role of machine learning algorithms, particularly SVM, in fortifying financial distress prediction endeavors (Abdullah et al., 2023). These investigations reveal SVM's prowess in navigating various sectors and its comparative advantage over traditional linear techniques, aligning with the overarching endeavor of facilitating proactive risk management and informed decision-making processes. Through a harmonized synthesis of machine learning innovations and traditional financial frameworks, researchers strive to forge predictive models that not only elucidate the dynamics of financial distress but also empower stakeholders with actionable insights across diverse industries.

3. Methodology

The present study aims to perform a comparative analysis of four distinct machine learning algorithms to predict financial distress. The algorithms under consideration include Random Forest and XGBoost, which are both ensemble tree-based methods renowned for their accuracy and robustness in handling various types of data. The Long Short-Term Memory (LSTM) network, a specialized form of recurrent neural network, is chosen for its proficiency in managing time-series data, an attribute particularly pertinent to the financial domain. Support Vector Machine (SVM) is also included in this study due to its capability as a boundary-based classifier, which is adept at finding the optimal separation between different classes.

Variable Y is a binary value of 1 or 0 which is the value of each company's Earnings Before Interest and Taxes (EBIT) in years 1 and −1. EBIT values in years −1 and 1 indicate that the company fails to earn a profit for the company which will impact the company's ability to pay interest, debt, and taxes in the current year. These conditions will make the company continue to be in a distress position, if these conditions are experienced continuously by the company, it will be difficult for the company to rise and will fall into a position of bankruptcy. Therefore, companies with this model can find the right strategy to improve their financial distress.

Given the propensity for class imbalance within financial datasets, where instances of distress may be significantly outnumbered by normal cases, two sampling techniques are employed to ensure a balanced representation of classes. Random Under-Sampling (RUS) is utilized to mitigate class imbalance by randomly discarding instances from the majority class, thus aligning the class distribution more closely with that of the minority class. Conversely, the Synthetic Minority Over-sampling Technique (SMOTE) is deployed to augment the minority class through the generation of synthetic samples, thereby enriching the dataset without the loss of information.

The efficacy of the selected algorithms, when applied in conjunction with the RUS and SMOTE sampling techniques, is evaluated across a suite of metrics. Accuracy serves as the primary indicator of overall performance, reflecting the proportion of true results among the total number of cases examined. The Area Under the Receiver Operating Characteristic curve (AUC ROC), for both training and testing datasets, provides insight into the true positive rate relative to the false positive rate,

offering a measure of the model's ability to distinguish between the classes. The AUC for ROC probabilities further elucidates the model's discriminative capacity. Additionally, the AUC Precision-Recall metric is scrutinized, particularly for its relevance in imbalanced datasets, as it reflects the model's aptitude in identifying the minority class.

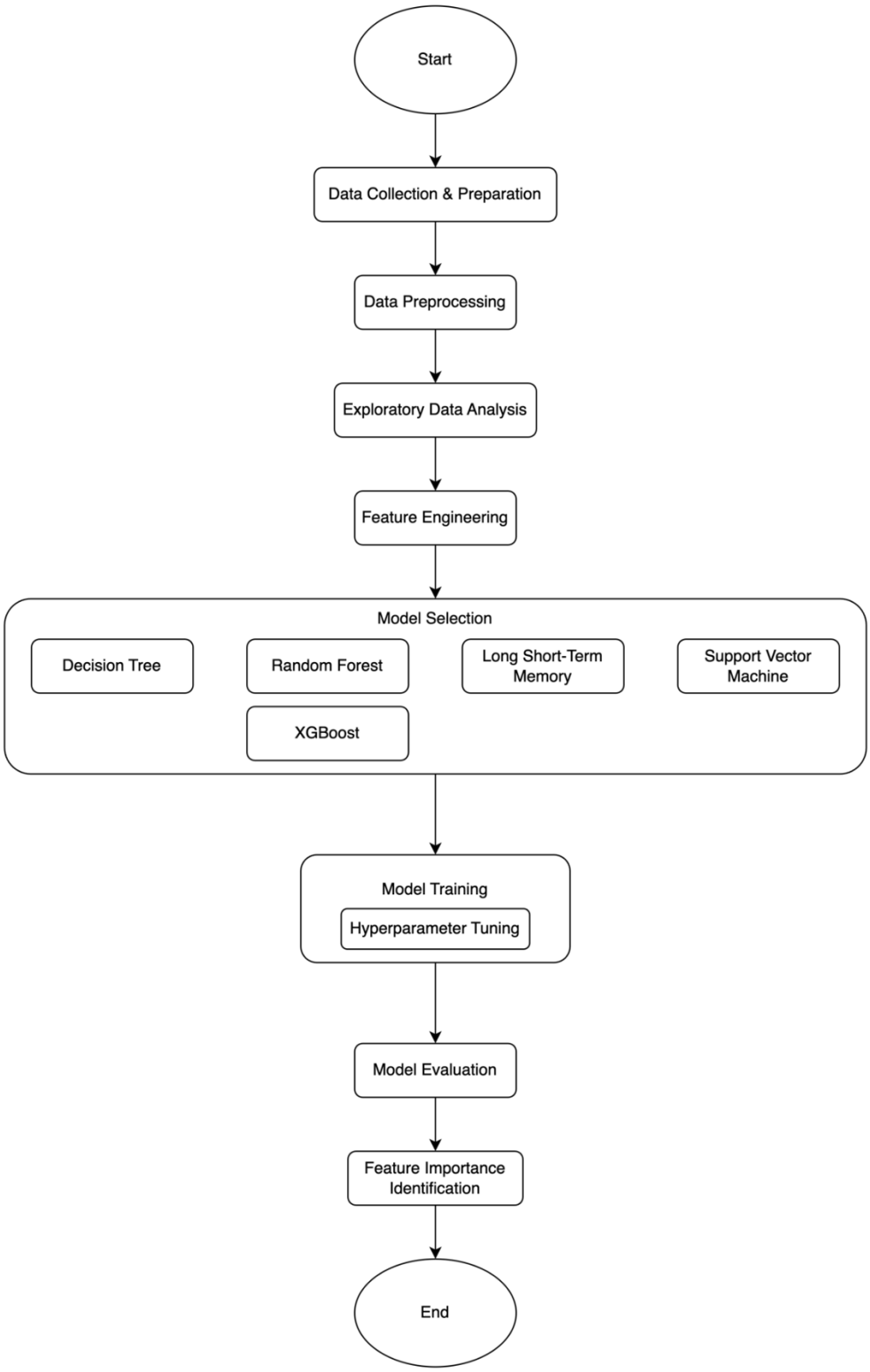


Figure 1. Process illustration of the study.

Precision, the F1 Score, and Recall are also calculated, with the F1 Score providing a harmonic mean of Precision and Recall, thereby presenting a balanced view of the model's precision in class prediction and its sensitivity in identifying all relevant instances. Through this comprehensive evaluation, the study aspires to delineate the strengths and limitations of each algorithm in the context of financial distress prediction, contributing valuable insights to the field of financial risk management. **Figure 1** depicts the assessment procedure for the machine learning model.

3.1. Data collection and preparation

For this study, a comprehensive dataset spanning ten years from 2013 to 2022 was utilized, comprising data from 437 companies across multiple industries in Indonesia. The dataset utilized for this research comprises data sourced from the Indonesia Stock Exchange (IDX), which serves as the primary repository of company information and financial data pertinent to securities trading within the Indonesian market (Febrianta, 2024). The dataset encompasses various financial indicators from companies, focusing on the classification of financial distress—a crucial element reflecting a company's financial health. The dataset is a comprehensive version where all variables of the ratio of financial reports are utilized in **Table 1**.

Table 1. Ratio variables.

No	Ratio Variables	No	Ratio Variables
X1	ROA(C) Before tax, interest, and depreciation	X36	Quick asset/current liabilities
X2	ROA(A)% after tax	X37	cash / current liability
X3	ROA(B) after tax, before interest & depreciation	X38	current liability to assets
X4	Operating margin	X39	operating funds to liability
X5	Realized gross profit margin	X40	Inventory/working capital
X6	Operating profit ratio	X41	Inventory/current liability
X7	Net interest rate before tax	X42	current liability / sales
X8	Net interest rate after tax	X43	working capital/equity
X9	Operating expense ratio	X44	current liability/equity
X10	Cash flow ratio	X45	long-term liability to current assets
X11	Interest rate	X46	Total income/total expense
X12	Tax rate (A)	X47	Total expense/assets
X13	Net value per share (A)	X48	Liquid assets turnover rate
X14	Cash flow per share	X49	Quick asset turnover rate
X15	Operating profit per share	X50	working capital turnover rate
X16	Net profit per share before tax	X51	Cash turnover
X17	Operating profit growth rate	X52	Cash flow to sales
X18	After-tax net profit growth rate	X53	Fixed assets to assets
X19	Growth rate of total assets	X54	Current liability to liability
X20	Net worth growth rate	X55	Current liability to equity
X21	Total return on assets growth rate	X56	Equity to long-term liability
X22	Current ratio	X57	Cash flow to total assets

Table 1. (*Continued*).

No	Ratio Variables	No	Ratio Variables
X23	Quick ratio	X58	Cash flow to liability
X24	Total debt/total net worth	X59	CFO to assets
X25	Debt ratio %	X60	Cash flow to equity
X26	Net worth/assets	X61	current liabilities to current assets
X27	Total asset turnover	X62	Liability-assets
X28	Turnover of accounts receivable	X63	Net Income to total assets
X29	Inventory turnover rate (times)	X64	Gross profit to sales
X30	Fixed asset turnover	X65	Net Income to stockholder's Equity
X31	Net worth turnover rate (times)	X66	Liability to equity
X32	working capital to total assets	X67	Interest coverage ratio (Interest expense to EBIT)
X33	Quick asset/total asset	X68	Net income
X34	current assets/total assets	X69	Equity to liability
X35	cash/total assets		

The dataset utilized in this study encompasses a comprehensive array of financial indicators, offering a thorough insight into the financial position of companies. It comprises a total of 72 features, each representing a distinct financial metric or identifier. Among the key features included in the dataset are: the "class" variable, serving as the target variable indicating the financial distress status of companies, with values of 0 representing "No" and 1 representing "Yes"; "sektor," which specifies the sector classification of the company; "company_code," denoting the unique stock code assigned to each company; and "year," indicating the reporting year of the financial data. These features collectively provide a multifaceted perspective for analyzing the financial health and distress status of companies within the dataset.

Financial metrics like roa_after_tax, operating_margin, cash_flow_ratio, growthrate_of_total_assets, net_worth_growth_rate, current_ratio, total_debt_total_net_worth, debt_ratio, and several others.

Data preprocessing was a critical step to ensure data quality and readiness for analysis. The dataset was meticulously cleaned to address inconsistencies, replacing problematic values with NaN to standardize missing data representation. A selective type conversion process was applied, converting relevant columns to a floating-point format to facilitate numerical computations, while preserving the original formats of categorical and identifying columns such as "class", "sector", "company_code", and "year". This preserved the integrity and usability of the dataset for machine learning purposes.

Data Cleaning:

In the data cleaning process, we identify and handle various forms of "noise" in our data, which can interfere with the analysis. We replace these inconsistencies, specifically values such as "#DIV/0!", "#Value!", "#VALUE!", "#REF!", and "-", with NaN (Not a Number) to standardize the representation of missing or undefined data.

Type Conversion:

During type conversion, we meticulously examine all columns that require a data

type adjustment. We convert their data types to float, facilitating numerical operations and computations on these columns. However, we intentionally exclude specific columns like “class”, “sector”, “company_code”, and “year” from this conversion process because they serve different purposes, such as categorization or identification, and are more suitable in their original format. This careful modification ensures that the dataset is primed for more complex procedures like analysis and machine learning model training, with data types being consistent and appropriate for such operations.

3.2. Synthesis and reporting

The findings from the study were systematically documented, highlighting the effectiveness of the models and the criticality of feature selection. The performance of various models, especially in precision-recall aspects, was discussed, emphasizing the need for models to maintain precision at high recall levels. The results were uploaded to the Neptune platform for tracking and future reference, ensuring a comprehensive and accessible record of the research outcomes.

In conclusion, the methodology chapter provides a detailed account of the systematic approach taken in this study to predict financial distress using machine learning models. From initial data preparation to complex model evaluation, the research methodology was designed to ensure the development of reliable and interpretable predictive models, with a keen focus on the practical application and generalizability of the results.

An extensive exploratory data analysis (EDA) phase followed, where the dataset was prepared for machine learning. Feature reduction and encoding were conducted, with non-essential features removed and categorical data transformed via one-hot encoding. The data was then restructured to improve readability and to ensure order, particularly placing the ‘class’ column prominently for easy reference. The dataset was divided into training and testing sets in an 80/20 split, employing stratified sampling to ensure representative class distribution, given the imbalanced nature of the data.

3.2.1. Feature reduction and encoding

We streamline our dataset by removing less critical features like ‘kode_perusahaan’, considering potential data discrepancies with new companies. We apply one-hot encoding to the ‘year’ and ‘sektor’ columns to convert categorical data into a format suitable for machine learning models.

3.2.2. Data reorganization and duplication removal

We restructure our DataFrame to position the ‘class’ column first, followed by the newly encoded ‘year’ and ‘sektor’ columns, and finally, the original features. This step enhances readability and order. We also ensure the removal of any duplicate columns inadvertently created during the encoding process.

3.2.3. Dataset splitting

We proceed to divide the dataset into two subsets: a training set and a test set. The training set is used to train our model, while the test set is reserved and used to evaluate the model’s performance on unseen data. This is a critical step in machine learning practice to assess the model’s ability to generalize and not just memorize the training data.

We utilize a stratified sampling technique during this splitting to ensure that the training and test subsets have similar proportions of class labels as the original dataset. This is particularly important in cases of an imbalanced dataset where one class significantly outnumbers the other(s). Stratification aims to preserve the original class distribution in both training and test sets, thereby providing a more representative and fair ground for training and subsequently evaluating the model.

In our case, we allocate 80% of the data to the training set and the remaining 20% to the test set, a typical ratio that offers a balanced compromise between having enough data to learn from and enough to evaluate and test the robustness of the model. The use of a random state seed (e.g., 42) ensures reproducibility; the same data points will be allocated to the training and test sets each time the code is executed, which is essential for consistent results across multiple runs or users.

3.2.4. Data exploration

A crucial aspect of this exploration is understanding the distribution of our numerical data, specifically through the calculation of skewness values. Skewness provides us with insight into the symmetry, or lack thereof, in the distribution of our data points. Identifying high skewness values helps in pinpointing features that may require transformation to approximate normal distributions, potentially improving the performance of subsequent modeling.

To better interpret the skewness metrics, we adopt a visual approach. By dividing our data features into manageable chunks, we create horizontal bar plots displaying the skewness value for each variable. This visual representation is effective for quickly identifying variables with extreme skewness values, either positive or negative, that might warrant further investigation or transformation. The graphical approach complements the quantitative analysis, offering an intuitive understanding of data distribution characteristics.

3.2.5. Handling missing values

To address these missing values, we utilize an imputer in our preprocessing stage that substitutes these absent values with a minor constant of 0.0001. We opt for this small constant, as a zero value could convey significant information in our dataset, such as the nonexistence of a certain feature. This method is applied across both our training and testing datasets, confirming the absence of any missing values post-implementation.

3.2.6. Handling class imbalance with RUS and SMOTE

Recognizing the potential imbalance in our dataset, we implement Random Under Sampling (RUS) and Synthetic Minority Over-sampling Technique (SMOTE) to rectify it. Both methods aim to equalize the number of instances for each class, but while RUS does this by reducing the majority class, SMOTE generates synthetic data for the minority. We visualize the class distribution before and after the application of RUS and SMOTE using count plots, which are saved as images for reference. This exhaustive exploration and preprocessing regimen set the stage for accurate, bias-free machine learning models downstream.

3.3. Model development

The study established a set of baseline models to serve as a reference for the performance of more complex models developed later in the research. These models included Decision Trees, Random Forests, Long Short-Term Memory (LSTM) networks and Support Vector Machines. Each model was trained on both SMOTE and RUS processed datasets, with performance meticulously logged for comparison on metrics such as the F1 score and training time.

When we conduct a machine learning experiment, it's vital to establish a starting point or a reference to measure the impact of changes we make. This starting point is called a "baseline model". The primary purpose of using a baseline model is to have an initial measure of performance, which allows us to understand how subsequent changes, such as hyperparameter tuning, improve or deteriorate the model's performance.

In our approach, we commence by establishing various machine learning "baseline" models. These models are foundational and serve as our starting point for further experimentation. Specifically, the models that we choose for this baseline phase include Decision Trees, Support Vector Machines, Random Forests, and Long Short-Term Memory (LSTM) networks, among others. We then train these baseline models on two distinct datasets: one that has undergone the SMOTE and another that has been subjected to RUS. For each dataset type—SMOTE and RUS—we perform a series of operations. Firstly, we train the models and subsequently evaluate them. By doing so, we can monitor the performance of these models in real-time. Post training, we organize and update a log that meticulously captures the outcomes of our training sessions. This log, a structured data format, is instrumental in tracking the models' performance, especially metrics such as the F1 score, and the time taken for training.

The next phase involves delving deeper, as we embark on a rigorous process of hyperparameter tuning to unearth the optimal configuration that can further enhance our model's performance. By combining various hyperparameters, we will navigate through the vast search space, trying to find that perfect blend that offers us superior results.

3.4. Model evaluation and validation

Model evaluation was not confined to basic accuracy but also encompassed ROC-AUC, precision, recall, and the F1 score, among other metrics. Evaluations were conducted on both the training and testing datasets to verify the models' capacity to generalize. ROC curves and Precision-Recall curves were generated, alongside confusion matrices, to provide deeper insight into model accuracy and error trade-offs. The study further investigated feature importance across models to interpret predictive behaviors.

The results from hyperparameter tuning and feature importance measurement were synthesized to understand the strengths and limitations of each model.

4. Result and discussion

The dataset that we use encompasses various financial indicators from companies, focusing on the classification of financial distress—a crucial element reflecting a

company's financial health. To improve the accuracy of the financial distress prediction model, the study collected 10 years of data from 2013 to 2022, including financial indicators such as liquidity ratios, profitability ratios, and solvency ratios for each company. In the realm of machine learning, establishing a "baseline model" is a critical initial step for any experimental approach. This model serves as a reference point, providing an initial performance metric against which the impact of subsequent modifications, such as hyperparameter adjustments, can be measured as in **Table 2**.

Table 2. Descriptive analysis of variable.

Variable Description	Mean	Median	Maximum	Minimum	Std.Dev.	Observations
AFTER_TAX_NET_PROFIT_GROWTH_RATE	−1.401.201	−0.007336	2.380.367	−51290.79	7.875.562	4.370
CASH___CURRENT_LIABILITY	2.073.225	0.210331	1.808.950	0.000001	3.060.749	4.370
CASH___TOTAL_ASSETS	0.586662	0.066096	8.796.298	0.000001	1.739.055	4.370
CASH_FLOW_PER_SHARE	9.325.604	1.179.676	60968.85	−5.465.633	1.070.687	4.370
CASH_FLOW_RATIO	1.57×10^{10}	0.076155	4.74×10^{12}	−1.545.481	1.80×10^{11}	4.370
CASH_FLOW_TO_EQUITY	0.248911	0.126029	2.045.130	0.000001	0.653364	4.370
CASH_FLOW_TO_LIABILITY	0.174819	0.076015	1.545.481	−1.103.141	3.571.399	4.370
CASH_FLOW_TO_SALES	0.553243	0.127619	3.513.412	0.000001	6.645.606	4.370
CASH_FLOW_TO_TOTAL_ASSETS	0.110887	0.055285	1.138.276	0.000001	1.732.003	4.370
CASH_TURNOVER	−2.441.078	0.201571	6.377.365	−3.856.912	1.294.311	4.370
CFO_TO_ASSETS	0.088048	0.043392	1.138.276	−1.965.502	1.735.353	4.370
CURRENT_ASSETS_TOTAL_ASSETS	0.467460	0.432888	1.000.000	0.000001	0.285333	4.370
CURRENT_LIABILITIES_TO_CURRENT_ASSETS	4.208.026	0.694406	41964.05	0.000001	1.122.982	4.370
CURRENT_LIABILITY___SALES	3.488.946	0.477163	1.184.197	−1.889.230	3.497.179	4.370
CURRENT_LIABILITY_EQUITY	1.244.592	0.471364	3.672.056	−1.860.558	7.881.438	4.370
CURRENT_LIABILITY_TO_ASSETS	1.374.667	0.272107	3.183.894	0.000001	4.870.098	4.370
CURRENT_LIABILITY_TO_EQUITY	2.593.348	0.520400	1.988.911	−1.860.558	3.526.683	4.370
CURRENT_LIABILITY_TO_LIABILITY	0.668331	0.666154	3.689.123	−2.525.453	0.931818	4.370
CURRENT_RATIO	9.049.922	1.382.343	4.562.098	0.000001	9.473.631	4.370
DEBT_RATIO__	3.109.284	0.516846	3.461.972	0.000001	8.858.385	4.370
EQUITY_TO_LIABILITY	1.175.850	0.934036	9.782.454	−0.999717	2.302.825	4.370
EQUITY_TO_LONG_TERM_LIABILITY	2.586.568	3.462.907	905496.3	−1.650.641	13731.64	4.370
FIXED_ASSET_TURNOVER	4.184.365	1.173.536	4.694.564	−2.361.911	7.173.037	4.370
FIXED_ASSETS_TO_ASSETS	0.515089	0.551715	1.000.000	0.000001	0.277983	4.370
GROSS_PROFIT_TO_SALES	0.158153	0.110056	3.931.334	−1.712.579	9.086.417	4.370
GROWTH_RATE_OF_TOTAL_ASSETS	1.213.237	0.057403	3.981.966	−1.000.000	6.066.726	4.370
INTEREST_COVERAGE_RATIO__INTEREST_EXPENSE_TO_EBIT_	6.906.733	0.043007	27406.29	−1.615.582	4.151.445	4.370

Table 2. (*Continued*).

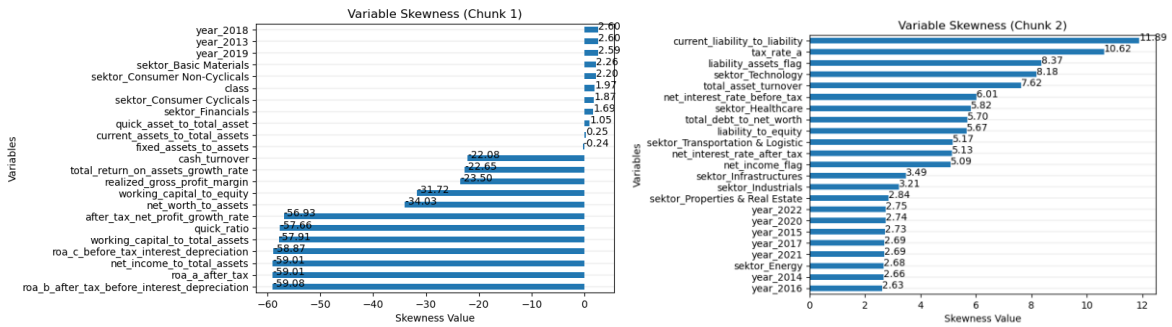
Variable Description	Mean	Median	Maximum	Minimum	Std.Dev.	Observations
INTEREST_RATE	2.53×10^{11}	2.47×10^{10}	6.85×10^{13}	-7.36×10^{10}	1.40×10^{12}	4.370
INVENTORY_CURRENT_LIABILITY	0.720310	0.242931	1.664.576	−9.212.581	3.890.647	4.370
INVENTORY_TURNOVER_RATE__TIMES_	5.300.869	3.423.849	90603.18	0.000001	1.452.914	4.369
INVENTORY_WORKING_CAPITAL	1.874.709	0.127797	2.138.035	−1.664.474	3.614.352	4.370
LIABILITY_ASSETS__FLAG	7.39×10^{12}	1.51×10^{12}	4.48×10^{14}	0.000001	2.30×10^{13}	4.370
LIABILITY_TO_EQUITY	1.758.278	0.950433	3.705.741	−2.708.534	9.934.413	4.370
LIQUID_ASSETS_TURNOVER_RATE	4.032.774	1.387.595	40844.91	−4.759.532	1.064.399	4.370
LONG_TERM_LIABILITY_TO_CURRENT_ASSETS	2.372.600	0.290651	28734.35	−7.414.321	6.928.108	4.370
NET_INCOME__FLAG	6.84×10^{11}	5.22×10^{10}	5.88×10^{13}	-5.93×10^{13}	3.45×10^{12}	4.370
NET_INCOME_TO_STOCKHOLDER_S_EQUITY	1.043.465	1.539.014	23678.07	−2.292.681	5.938.467	4.370
NET_INCOME_TO_TOTAL_ASSETS	0.503996	0.020270	3.612.443	−1.396.863	5.859.892	4.370
NET_INTEREST_RATE_AFTER_TAX	6.84×10^{11}	5.22×10^{10}	5.88×10^{13}	-5.93×10^{13}	3.45×10^{12}	4.370
NET_INTEREST_RATE_BEFORE_TAX	1.00×10^{12}	7.58×10^{10}	7.04×10^{13}	-6.47×10^{13}	4.67×10^{12}	4.370
NET_PROFIT_PER_SHARE_BEFORE_TAX	1.434.129	1.759.467	33921.11	−2.025.719	9.271.322	4.370
NET_VALUE_PER_SHARE__A_	7.349.136	2.406.975	296312.2	−27899.55	5.105.552	4.370
NET_WORTH_ASSETS	−2.109.519	0.483004	0.999898	−3.460.972	8.858.384	4.370
NET_WORTH_GROWTH_RATE	0.735231	0.063306	1.822.220	−2.930.838	2.923.707	4.370
NET_WORTH_TURNOVER_RATE__TIMES_	1.188.869	0.899266	24135.42	−12298.22	5.348.876	4.370
OPERATING_EXPENSE_RATIO	1.465.316	0.623091	5.883.192	−3.831.337	1.639.116	4.370
OPERATING_FUNDS_TO_LIABILITY	1.301.007	0.286272	4.560.157	−5.848.174	1.124.885	4.370
OPERATING_MARGIN	0.095220	0.083510	3.931.334	−1.712.579	9.121.754	4.370
OPERATING_PROFIT_GROWTH_RATE	5.267.447	−0.061657	21023.07	−6.269.538	3.197.557	4.370
OPERATING_PROFIT_PER_SHARE	1.901.653	3.146.730	36390.11	−1.995.829	1.025.852	4.370
OPERATING_PROFIT_RATIO	0.183225	0.097490	3.931.334	−1.712.579	9.103.815	4.370
QUICK_ASSET_CURRENT_LIABILITIES	5.840.507	0.811269	4.562.098	−0.658876	8.333.158	4.370
QUICK_ASSET_TOTAL_ASSET	0.287858	0.239216	0.998125	−0.260805	0.211717	4.370
QUICK_ASSET_TURNOVER_RATE	3.911.958	1.039.825	4.694.564	−2.361.911	7.169.879	4.370
QUICK_RATIO	−5.843.372	0.862515	4.562.098	−54173.14	8.246.093	4.370
REALIZED_GROSS_PROFIT_MARGIN	0.288054	0.277068	6.699.502	−5.700.093	1.397.105	4.370
ROA_A___AFTER_TAX	0.504864	0.021050	3.612.443	−1.396.863	5.859.891	4.370

Table 2. (*Continued*).

Variable Description	Mean	Median	Maximum	Minimum	Std.Dev.	Observations
ROA_B__AFTER_TAX__BEFORE_INTEREST__BEFORE_DEPRECIATION	0.554148	0.039896	3.612.443	−1.391.164	5.856.417	4.370
ROA_C__BEFORE_TAX__BEFORE_INTEREST__BEFORE_DEPRECIATION	0.628795	0.017586	3.657.774	−9.336.895	5.711.430	4.370
TAX_RATE__A_	3.08×10^{11}	2.56×10^{10}	3.12×10^{13}	0.000001	1.23×10^{12}	4.370
TOTAL_ASSET_TURNOVER	0.931011	0.509011	3.651.683	−1.657.770	1.451.701	4.370
TOTAL_DEBT_TOTAL_NET_WORTH	1.719.871	0.949081	3.705.741	−2.708.534	9.914.485	4.370
TOTAL_EXPENSE_ASSETS	0.836095	0.423860	4.004.248	−0.004573	6.180.066	4.370
TOTAL_INCOME_TOTAL_EXPENSE	1.446.112	1.083.125	9.754.333	−4.464.932	1.505.343	4.370
TOTAL_RETURN_ON_ASSETS_GROWTH_RATE	−1.672.307	−0.098730	3.032.513	−5.462.346	1.379.827	4.370
TURNOVER_OF_ACCOUNTS_RECEIVABLE	8.970.593	6.344.989	3468875.	−1.688.682	52567.19	4.370
WORKING_CAPITAL_EQUITY	0.293659	0.337934	8.613.119	−3.095.847	5.902.774	4.370
WORKING_CAPITAL_TO_TOTAL_ASSETS	−0.818243	0.136359	0.997619	−3.183.253	4.863.255	4.370
WORKING_CAPITAL_TURNOVER_RATE	0.493384	0.049079	9.289.595	−1.431.938	1.618.066	4.370
Y(Dependent)	0.148970	0	1	0	0.356100	4.370

Note: Data Processed, Variable Dependent 69 Ratio and 1 Variable Independent biner 1and 0.

Our visual analysis of skewness metrics reveals valuable insights. We have organized our dataset features into distinct groups and generated horizontal bar plots to illustrate the skewness values associated with each variable. This graphical representation serves as a practical tool for promptly identifying variables exhibiting significant positive or negative skewness, which may necessitate additional scrutiny or transformation. The integration of graphical visualization complements our quantitative analysis, facilitating a more intuitive comprehension of the distribution characteristics inherent in the data.



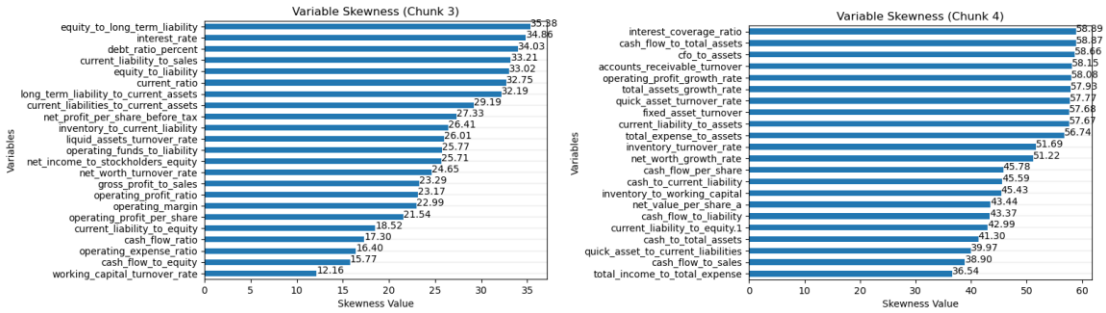


Figure 2. Data distribution characteristics.

In the dataset, most variables exhibit significant skewness, except for two: 'current_assets_to_total_assets' and 'fixed_assets_to_assets,' with skewness values of 0.252856 and −0.243993, respectively. These values are within the accepted range of normal skewness (−0.5 to 0.5), indicating a more symmetrical distribution around the mean showed on **Figure 2**.

1) 'current_assets_to_total_assets': A skewness of 0.252856 suggests that this feature is slightly right-skewed, but not significantly. This feature likely follows a distribution that is relatively symmetrical, meaning most companies in the dataset have a balance between current assets and total assets. This balance might be indicative of healthy financial practices, where companies maintain sufficient liquid assets relative to their size.
2) 'fixed_assets_to_assets': This feature, with a skewness of −0.243993, is slightly left-skewed. However, since the skewness is relatively low, the distribution of companies' fixed assets relative to total assets is fairly symmetrical. The slight left skew might indicate that there are a few companies with exceptionally high fixed assets compared to total assets.

The essence of a baseline model lies in its role as a benchmark, enabling researchers to discern whether further tweaks lead to improvements or degradations in performance. Our methodology begins with the creation of various machine learning baseline models. These foundational models act as the bedrock from which our experimentation expands. Our selection of baseline models encompasses a range of algorithms, including Decision Trees, Support Vector Machines, Random Forests, and Long Short-Term Memory (LSTM) networks, among others.

We adopt a dual dataset training approach, applying these models to datasets processed with Synthetic Minority Over-sampling Technique (SMOTE) and Random Under Sampling (RUS), respectively. Each baseline model undergoes a comprehensive evaluation process on both the SMOTE and RUS datasets. This involves not only training the models but also a real-time assessment of their performance. Post-training, we diligently record the results in a structured log, which is pivotal for monitoring and comparing the performance metrics of each model, with a keen focus on critical measures such as the F1 score and training duration.

Following the baseline assessments, we delve into the intricate phase of hyperparameter tuning. This process is exhaustive and aimed at discovering the optimal configurations to elevate the model's efficacy. Through a systematic exploration of the hyperparameter space, we seek that elusive combination that promises to yield superior performance, thereby fine-tuning our models to approach

their theoretical best. In order to maximize the performance of our models, we engage in a thorough process of hyperparameter tuning. The refinement of machine learning models through hyperparameter tuning is a nuanced process that significantly enhances model performance.

Our research delves into this domain by initializing an LSTM model, a variant of Recurrent Neural Networks renowned for their proficiency with sequential data. The model architecture includes an LSTM layer with 50 units and 'relu' activation, along with a Dense output layer with a 'sigmoid' activation function, aligning with the needs of binary classification. Model evaluation extends beyond basic accuracy, incorporating metrics like ROC-AUC, precision, recall, and the F1 score. Evaluations are executed on both training and test sets to ensure model robustness and generalizability. ROC curves, Precision-Recall curves, and confusion matrices offer deeper insight into model accuracy and the balance between different error types.

The classification reports generated post-evaluation detail precision, recall, and F1-scores for each class, saved in HTML format and uploaded to the Neptune platform for enhanced accessibility and longitudinal tracking. The results of hyperparameter experimentation are meticulously recorded, converting complex data into an interpretable format, highlighting the most effective hyperparameters identified. We unify these results in a comprehensive report via Neptune, ensuring that the latest metrics from retuned models are accurately reflected, replacing outdated data.

Our investigation also encompasses the measurement of feature importance across various models. Inherent attributes in models like Random Forest and XGBoost provide immediate insights, while for others, such as linear-kernel SVCs, we derive importance through alternative methods. Visual representations like bar plots are utilized for clearer communication of these importances. For an in-depth understanding of predictions, particularly in complex models like LSTM, we utilize SHapley Additive exPlanations (SHAP) values. These offer a transparent explanation of each feature's influence on the predictions. In addition, permutation importance is applied to determine the significance of features within LSTM networks, overcoming the challenge posed by their unique input shape requirements.

In models like Logistic Regression and Decision Trees, feature importance is assessed through the absolute values of coefficients and intrinsic metrics, respectively, with visualizations aiding in the comprehensive interpretation of these importances. This multifaceted approach to hyperparameter tuning and feature importance evaluation ensures the development of highly tuned and interpretable machine learning models. In the rigorous analysis of hyperparameter tuning across various machine learning models, the XGBClassifier demonstrates notable efficacy when integrated with Synthetic Minority Over-sampling Technique (SMOTE), achieving an approximate accuracy of 91.88% and an AUC-ROC score of around 87.61%. This level of performance, characterized by a harmonious balance between precision and recall, indicates its strong capacity for effectively classifying both classes within the dataset.

The RandomForest Classifier, applied in conjunction with both SMOTE and Random Under Sampling (RUS), also exhibits competitive performance. It maintains high accuracy and AUC-ROC scores, with a notable distinction: while the accuracy of RandomForest with RUS is marginally lower than its SMOTE counterpart, its AUC-

ROC is marginally higher. This suggests that the RandomForest with RUS may excel in the ranking of predictions, a crucial factor in certain decision-making contexts. Sequential models and Support Vector Classifier (SVC), consistent with earlier performance on the dataset, present lower metrics in accuracy, AUC-ROC, and other critical measures in comparison to ensemble models such as XGBoost and RandomForest. The protracted training times of these models further diminish their efficiency, indicating potential constraints in their application where time is a significant factor.

Feature importance measurement

The elucidation of feature importance within machine learning models is a pivotal component for the interpretation of model predictions, particularly after rigorous hyperparameter tuning. This is exemplified in our analysis of the dataset, where XGBClassifier and RandomForestClassifier have demonstrated noteworthy performance, especially when paired with Synthetic Minority Over-sampling Technique (SMOTE) in **Table 3**. The ascendancy of these models is largely attributable to a comprehensive set of features that enable precise predictions regarding financial distress.

Table 3. Feature selection (RUS).

No.	Decision Tree	XGBoost	Random Forest	LSTM	SVM
1	NA	Operating Profit per Share	Operating Margin	Fixed Assets to Assets	NA
2	NA	Operating Margin	Operating Profit per Share	Current Assets/Total Assets	NA
3	NA	Income to Total Expenses	Income to Total Expenses	Income to Total Expenses	NA
4	NA	Operating Profit Ratio	Net Income Flag	Operating Margin	NA
5	NA	Operating Profit Growth Rate	Operating Profit Ratio	Gross Profit to Sales	NA

An in-depth examination of feature importance within the XGBoost model reveals ‘operating_profit_per_share’ as the most significant feature, with a preeminent importance value. This prominence signifies the variable’s critical role in shaping the predictive model’s decisions, thereby suggesting that operating profit per share is a central determinant in assessing financial health. In a similar vein, other features such as ‘operating_margin’ and ‘interest_coverage’ also emerge as key influencers within the XGBoost model, as evidenced by their notable importance values. These variables’ substantial impact on the model’s decision-making process underlines their relevance in predicting financial distress. Conversely, specific sectors like ‘Consumer_Cyclical’ and ‘Technology’ are positioned lower on the importance chart, indicating their relatively marginal contribution to the model’s predictions. This disparity in feature importance highlights the varying degrees of influence that different variables possess in the assessment of financial distress in **Table 4**.

Table 4. Feature selection (SMOTE).

No.	Decision Tree	XGBoost	Random Forest	LSTM	SVM
1	NA	Operating Profit per Share	Operating Margin	Fixed Assets to Assets	NA
2	NA	Operating Margin	Operating Profit per Share	Income to Total Expenses	NA
3	NA	Interest Coverage Ratio	Income to Total Expense	Current Assets/Total Assets	NA
4	NA	Income to Total Expenses	Tax Interest Depreciation	Operating Profit per Share	NA
5	NA	Operating Profit Ratio	Net Income Flag	Realized gross profit margin	NA

Turning our attention to the RandomForest model, 'operating_margin' stands out as the paramount feature, suggesting that this metric is of utmost significance in the model's predictive accuracy. Furthermore, variables such as 'income_to_total_expense' and 'interest_coverage' carry considerable weight in the model's determinations, asserting themselves as critical components for prediction. However, akin to the observations in the XGBoost model, the RandomForest model also assigns lower importance to certain features, including specific year labels and sector indicators. This indicates a consensus between the two models regarding the lesser predictive power of these variables.

In summation, the analysis of feature importance within XGBClassifier and RandomForestClassifier models provides invaluable insights into the variables that are most indicative of financial distress. By recognizing the features with the most significant impact on predictions, we can not only improve our understanding of the models' internal mechanisms but also refine the selection of variables for future model training, ensuring that the predictive models we develop are both interpretable and reliable in **Table 5**.

Table 5. Model comparison.

Parameter	Random Forest		XGBoost		LSTM		SVM	
	RUS	SMOTE	RUS	SMOTE	RUS	SMOTE	RUS	SMOTE
Accuracy	0.957	0.974	0.966	0.989	0.851	0.923	0.87	0.903
AUC ROC Train	0.997	0.995	0.989	0.999	0.977	0.999	0.953	0.976
AUC ROC Test	0.968	0.981	0.977	0.981	0.859	0.85	0.866	0.835
AUC ROC Prob Train	1	1	1	1	0.996	1	0.981	0.989
AUC ROC Prob Test	0.994	0.997	0.997	0.999	0.899	0.915	0.927	0.92
AUC Precision Recall	0.962	0.978	0.982	0.994	0.619	0.751	0.656	0.68
Precision	0.889	0.926	0.908	0.975	0.737	0.848	0.756	0.803
F1 Score	0.922	0.951	0.938	0.978	0.771	0.849	0.791	0.818
Recall	0.968	0.981	0.977	0.981	0.859	0.85	0.866	0.835

Source: data processed.

The outcomes of the comparative study reveal that the integration of XGBoost with the Synthetic Minority Over-sampling Technique (SMOTE) outshines other algorithmic combinations in predicting financial distress, as evidenced by an exemplary accuracy of 0.989. This finding underscores the effectiveness of XGBoost when coupled with an advanced sampling method that counters class imbalance, thereby enhancing its predictive precision. In stark contrast, the Long Short-Term

Memory (LSTM) network, in alliance with Random Under-Sampling (RUS), registers the lowest accuracy, suggesting its limited utility in this specific predictive task.

When examining the Area Under the Receiver Operating Characteristic (AUC ROC) across both training and testing phases, all models display commendable scores on training data, with certain algorithms achieving the pinnacle of performance with scores of 1.0. This phenomenon highlights the models' capabilities in differentiating between classes during the learning process. In the testing phase, the models—XGBoost and Random Forest, both paired with RUS and SMOTE—showcase an exceptional ability to generalize, as reflected in their high AUC ROC scores exceeding 0.980.

All Models Hyperparameter Tuning Results XGBClassifier with SMOTE appears to be the most effective model, showing the highest accuracy (approx. 98.86%) and AUC-ROC test score (approx. 98.06%). Its precision, recall, and F1 score are also impressive, indicating a well-balanced model in terms of identifying both classes of the target variable. RandomForestClassifier with SMOTE also performs well but trails slightly in accuracy and AUC-ROC test score compared to XGBClassifier. However, it displays a slightly higher AUC for precision-recall, which might make it more suitable if the cost of false positives is high.

Models trained with RUS (Random Under-Sampling) generally have lower accuracy than their SMOTE counterparts, which indicates that SMOTE might be more effective for this dataset in handling class imbalance. Sequential models (presumably neural networks) have significantly longer training times and lower performance metrics across the board, suggesting that, for this dataset, traditional machine learning models outperform them. SVC (Support Vector Classifier) with both SMOTE and RUS performs the poorest among the tried models in terms of accuracy and AUC-ROC. The long training time with SMOTE and less satisfactory scores might make it less desirable for this specific problem.

Feature Importance Measurement Following the results of hyperparameter tuning, we've discerned the significant performance of XGBClassifier and RandomForestClassifier models, particularly when SMOTE is utilized. These models' superior efficacy on the "Ratio" dataset underscores the vitality of an extensive feature set for precise financial distress predictions. As we further delve into the intricacies of these models, understanding feature importance becomes paramount. This not only sheds light on the variables steering the models' decisions but also gives a clearer picture of the key financial metrics influencing the probability of financial distress. Here's a detailed assessment of feature importance across the two datasets:

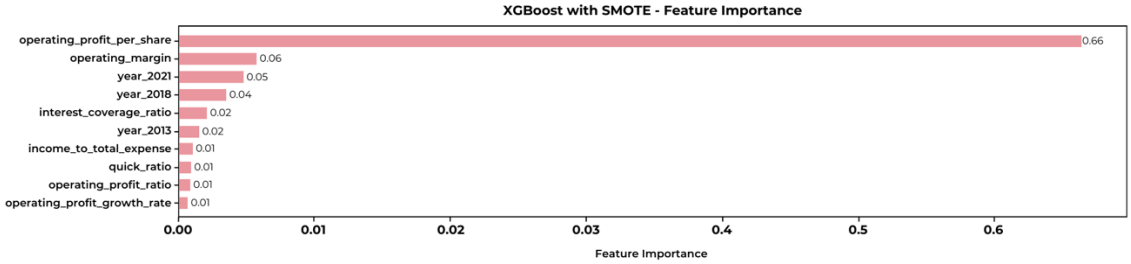


Figure 3. Feature importance XGBoost with SMOTE.

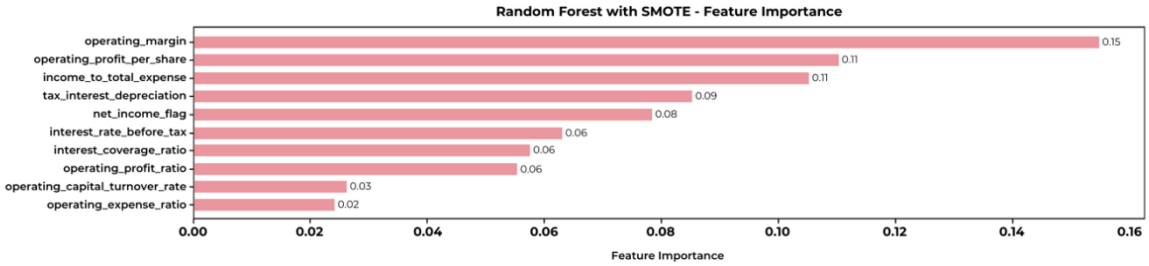


Figure 4. Feature importance Random Forest with SMOTE.

XGBoost:

1) The most significant feature in the XGBoost model in **Figures 3 and 4** seems to be 'operating_profit_per_share', with a dominant importance value. This suggests that operating profit per share is a major determinant in the prediction process.
2) Following closely are features like 'operating_margin' and 'interest_coverage', both of which have noticeable importance values, indicating their significance in the model's decision-making.
3) Features towards the end of the chart, such as 'Consumer_Cyclical' and 'Technology' sectors, have a lower importance value, suggesting that they have a minor role in the prediction process.

Random Forest:

1) In the Random Forest model, 'operating_margin' tops the chart, indicating it's the most influential feature.
2) Features like 'income_to_total_expense' and 'interest_coverage' are also of high importance, making them crucial predictors for the model.
3) Much like the XGBoost model, some features towards the bottom, such as specific year labels and sector indicators, have reduced importance in the Random Forest's prediction process.

The AUC ROC Probability metrics, which gauge the models' discriminative prowess, resonate with the aforementioned AUC ROC results, presenting XGBoost and Random Forest as the superior classifiers. Furthermore, within the purview of imbalanced datasets, the AUC Precision-Recall emerges as a pivotal metric. Herein, XGBoost augmented with SMOTE achieves a remarkable score of 0.994, affirming its efficacy in distinguishing the minority class amidst a predominant majority. Conversely, LSTM generally lags in this respect, suggesting a potential misalignment with the data's imbalanced nature.

A granular analysis of Precision and Recall places XGBoost with SMOTE at the forefront once more, with the highest precision indicating a substantial rate of true positive identifications. Equally impressive is its recall, on par with Random Forest with SMOTE, both capturing the lion's share of actual positive cases. The F1 Score, which harmonizes precision and recall, further cements the standing of XGBoost with SMOTE as the forerunner, with a score of 0.978, symbolizing an optimal balance between identifying true positives and the entirety of positive instances.

The discourse on these results illuminates the superior performance of tree-based methods, especially XGBoost in combination with SMOTE, across a multitude of evaluative metrics. The success of these methods can be ascribed to their intrinsic

capacity to unravel complex, non-linear relationships inherent in financial data. The underperformance of LSTM could potentially be attributed to its intricate structure and a propensity for overfitting, which may not align with non-temporal data patterns. The Support Vector Machine (SVM), despite its resilience to outliers and adaptability in high-dimensional spaces, does not exhibit competitive edge in this particular predictive modeling challenge. This study, therefore, posits the XGBoost algorithm, especially when applied with SMOTE, as a robust predictive tool in the domain of financial distress forecasting.

There are several studies that try to develop prediction models to predict the level of company distress such as in the research of Altman et al (2022) they examined MSME companies in Croatia to create a prediction model for default by combining a number of variables such as a number of traditional financial ratios, MSME company payment behavior variables. The prediction model is getting better with the addition of several additional variables such as management changes, employee changes and tenure of employees. Testing with the help of Least Absolute Shrinkage and Selection Operator (LASSO) and then continued with the use of machine learning. Barboza et al. (2017) found that the Random Forest model is better for predicting distress than the traditional prediction model in a sample of companies in North America. There are several studies that use a number of variables to strengthen in modeling financial distress can be read in the research of Carton and Hofer (2006), Cano et al. (2015), Tsai (2014), Thi Vu et al. (2019).

The use of machine learning to create models to predict the level of bankruptcy of companies in the digital era, researchers are competing to create models by adding a number of variables from various sources from previous research and adding new variables to strengthen research results indicated by an increase in the accuracy of predictions on the models created. Modeling to predict bankruptcy/distress can help external and internal parties. Internal parties such as company management in making future company strategies. Internal parties such as investors, creditors, and regulators, investors consider before investing in a company by calculating from existing models and finding results with high accuracy.

From these results, investors determine whether the company matches the criteria of investors. Furthermore, the creditor where this party is a company or bank that will channel loan funds to the company, it is very necessary to have very accurate prediction results to ensure that the company that will be channeled loan funds will be able to pay its debt in the future, this aims to minimize the level of risk of default. The last is the regulator where this party is responsible for providing input to companies that are distressed or threatened with bankruptcy, if the company cannot get out of the bankruptcy or distress zone then the regulator can revoke the business license of the company. Cultrera (2020) explains that MSMEs need assistance such as fragmentation procedures from regulators to get out of the financial crisis so that companies can operate to strengthen the economy in a country.

There is a difference in Citterio (2024) research in that banking sector institutions were the first targets to have an impact during the crisis in 2008-2009, so there is a need for early warning for banks to prevent the threat of crisis early. Therefore, it is necessary for banks to pay attention to non-financial variables such as ESG (Environmental, Social and Governance) variables to create a financial distress model

for banking institutions. The use of non-financial variables was also suggested by Huang et al. (2024) for input such as social responsibility reports, then to make predictions, researchers usually use binary codes 1 and 0 so that in the future they can be more detailed, such as medium, medium and high for the percentage of bankruptcy.

5. Conclusion

This study has thoroughly examined the use of different machine learning models to predict financial distress in Indonesian companies by utilizing the Financial Ratio dataset collected from the Indonesia Stock Exchange (IDX), which includes financial indicators from various companies across multiple industries spanning a decade.

This study has rigorously investigated the application of various machine learning models to predict financial distress using the Financial Ratio dataset, encompassing a decade's worth of financial indicators from a diversity of companies. The research has traversed through several phases, including preprocessing, exploratory analysis, baseline model establishment, hyperparameter tuning, and evaluation of feature importance, culminating in the nuanced understanding and prediction of financial distress. The study's findings from the hyperparameter tuning and feature importance measurements highlighted the XGBClassifier and RandomForestClassifier as standout performers, particularly when integrated with SMOTE, which improved the models' accuracy, AUC-ROC, precision, and recall. The feature importance analysis underscored the significance of certain financial indicators, such as interest coverage ratio and operating margin ratio, which were instrumental in the models' prediction capabilities.

ROC and Precision-Recall curves provided a deeper understanding of the models' diagnostic abilities, revealing the ensemble methods' superiority in balancing sensitivity and specificity. In conclusion, this study affirms the effectiveness of ensemble methods in financial distress prediction, with XGBClassifier and Random ForestClassifier offering robust and reliable performance. The comprehensive approach to model selection, preprocessing, and hyperparameter tuning has resulted in models that are not only highly accurate but also have the potential to generalize across different datasets. The findings of this study contribute significantly to the body of knowledge in financial analytics and offer a framework for developing predictive models that can be tailored to various financial contexts. The insights gained from feature importance and model evaluation serve as a guide for future research, emphasizing the critical role of precise feature selection and model tuning in the pursuit of predictive excellence.

The limitation in this study is the incorporation of all sectors to be studied so that the results can be said to be biased, for further research, it is possible to use the model of the research variables to be tested again on a similar group of companies in order to get more precise results and in accordance with the criteria of each sector.

Author contributions: Conceptualization, FTK and MYF; methodology, DFS; investigation, HAR; writing—original draft preparation, YS; writing—review and editing, YS; supervision, FTK; project administration, BAHB. All authors have read and agreed to the published version of the manuscript.

Funding: This paper is funded by PPM-PTM Grants of the Ministry of Education, Culture, Research and Technology of 2023 (03/SP2H/RT-MONO/LL4/2023).

Conflict of interest: The authors declare no conflict of interest.

References

Abdullah, M., Chowdhury, M., Uddin, A., & Moudud-Ul-Huq, S. (2023). Forecasting nonperforming loans using machine learning. Journal of Forecasting, 42(7), 1664-1689. https://doi.org/10.1002/for.2977

Altman, E. I. (1968). Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. The Journal of Finance, 23(4), 589–609. https://doi.org/10.2307/2978933

Altman, E. I., Balzano, M., Giannozzi, A., & Srhoj, S. (2022). Revisiting SME default predictors: The omega score. Journal of Small Business Management, 61(6), 2383–2417. https://doi.org/10.1080/00472778.2022.2135718

Altman, E. I., Marco, G., & Varetto, F. (1994). Corporate distress diagnosis: Comparisons using linear discriminant analysis and neural networks (the Italian experience). Journal of Banking and Finance, 18(3), 505–529. https://doi.org/10.1016/0378-4266(94)90007-8

Ayuni, N., Lasmini, N., & Putrawan, A. (2022). Support vector machine (svm) as financial distress model prediction in property and real estate companies., 397-402. https://doi.org/10.2991/978-2-494069-83-1_72

Barboza, F., Kimura, H., & Altman, E. (2017). Machine learning models and bankruptcy prediction. Expert Systems with Applications, 83, 405–417. https://doi.org/10.1016/j.eswa.2017.04.006

Bonello, J., Bredart, X., & Vella, V. (2018). Machine learning models for predicting financial distress. Journal of Research in Economics, 2(2), 174–185. https://doi.org/10.24954/jore.2018.22

Cano-Rodríguez, M., Sánchez-Alegría, S., & Arenas-Torres, P. (2015). The influence of Auditor's opinion and Auditor's reputation on the cost of debt: Evidence from private Spanish Firmsla Influencia de la Opinión de Auditorãa y la reputación del auditor en el Coste de la Deuda: Evidencia en las empresas españolas no cotizadas (Spanish). Spanish Journal of Finance and Accounting / Revista Española de Financiación y Contabilidad, 45(1), 32–62. https://doi.org/10.1080/02102412.2015.1111096

Carton, R. B., & Hofer, C. W. (2006). Measuring organizational performance: Metrics for entrepreneurship and strategic management research. Edward Elgar Publishing.

Citterio, A. (2024). Bank Failure Prediction Models: Review and outlook. Socio-Economic Planning Sciences, 92, 1–26. https://doi.org/10.1016/j.seps.2024.101818

Cultrera, L. (2020). Evaluation of bankruptcy prevention tools: Evidences from COSME programme. Economics Bulletin, 40(2), 978–988.

Dai, C., Liu, J., & Zhou, T. (2022). Forecasting cash for companies-the case of CATL. In: Proceedings of the 2022 2nd International Conference on Economic Development and Business Culture (ICEDBC 2022). pp. 687-696. https://doi.org/10.2991/978-94-6463-036-7_102

Durica, M., Frnda, J., & Svabova, L. (2021). Financial distress prediction in Slovakia: An application of the CART algorithm. Journal of International Studies, 14(1), 201–215. https://doi.org/10.14254/2071-8330.2021/14-1/14

El-Bannany, M., Sreedharan, M., & Khedr, A. (2020). A robust deep learning model for financial distress prediction. International Journal of Advanced Computer Science and Applications, 11(2). https://doi.org/10.14569/ijacsa.2020.0110222

Elhoseny, M., Metawa, N., Sztanó, G., & El-Hasnony, I. (2022). Deep learning-based model for financial distress prediction. Annals of Operations Research. https://doi.org/10.1007/s10479-022-04766-5

Febrianta, Mochamad Yudha. (2024). Financial Report Data of 437 Company in Indonesia. Telkom University Dataverse. https://doi.org/10.34820/FK2/ZT2PEC

Fitzpatrick, P. J. (1932). A Comparison of Ratios of Successful Industrial Enterprises with Those of Failed Firms. Certified Public Accountant, 12, 598-729.

Gregová, E., Valášková, K., Adamko, P., et al. (2020). Predicting financial distress of slovak enterprises: comparison of selected traditional and learning algorithms methods. Sustainability, 12(10), 3954. https://doi.org/10.3390/su12103954

Hájek, P., & Munk, M. (2023). Speech emotion recognition and text sentiment analysis for financial distress prediction. Neural Computing and Applications, 35(29), 21463-21477. https://doi.org/10.1007/s00521-023-08470-8

Halteh, K., Alkhouri, R., Ziadat, S., & Haddad, F. (2024). Fintech Unicorns Forecaster: An AI Approach For Financial Distress Prediction. Migration Letters, 21(S4 (2024)), 942–954.

He, J., Hao, Y., & Wang, X. (2021). An interpretable aid decision-making model for flag state control ship detention based on smote and XGBoost. Journal of Marine Science and Engineering, 9(2), 156. https://doi.org/10.3390/jmse9020156

Hong, A., Gao, M., Gao, Q., & Peng, X. (2022). Non-stationary financial time series forecasting based on meta-learning. Electronics Letters, 59(1). https://doi.org/10.1049/ell2.12681

Hota, S., Jena, S., Gupta, B., & Mishra, D. (2020). An empirical comparative analysis of nav forecasting using machine learning techniques., 565-572. https://doi.org/10.1007/978-981-15-6202-0_58

Huang, Y., Wang, Z., & Jiang, C. (2024). Diagnosis with incomplete multi-view data: A variational deep financial distress prediction method. Technological Forecasting and Social Change, 201, 1–12. https://doi.org/10.1016/j.techfore.2024.123269

Karathanasopoulos, A., & Osman, M. (2019). Forecasting the dubai financial market with a combination of momentum effect with a deep belief network. Journal of Forecasting, 38(4), 346-353. https://doi.org/10.1002/for.2560

Khademolqorani, S., Zeinal Hamadani, A., & Mokhatab Rafiei, F. (2015). A hybrid analysis approach to improve financial distress forecasting: Empirical evidence from Iran. Mathematical Problems in Engineering, 2015, 1–9. https://doi.org/10.1155/2015/178197

Khamisah, N., Listya, A., & Saputri, N. (2021). Does financial distress has an effects on audit report lag? (study on manufacturing companies listed in indonesia stock exchange). Akuntabilitas, 15(1), 19-34. https://doi.org/10.29259/ja.v15i1.13058

Kholisoh, S., & Dwiarti, R. (2020). The analysis of fundamental variables and macro economic variables in predicting financial distress. Management Analysis Journal, 9(1), 81-90. https://doi.org/10.15294/maj.v9i1.36395

Kristanti, F. T., & Dhaniswara, V. (2023). The accuracy of artificial neural networks and logit models in predicting the companies' financial distress. Journal of Technology Management and Innovation, 18(3), 42–50. https://doi.org/10.4067/s0718-27242023000300042

Kristanti, F. T., Safriza, Z., & Salim, D. F. (2023). Are Indonesian construction companies financially distressed? A prediction using artificial neural networks. Investment Management and Financial Innovations, 20(2), 41–52. https://doi.org/10.21511/imfi.20(2).2023.04

Lai, J., Lin, Y., Lin, H., et al. (2023). Tree-based machine learning models with optuna in predicting impedance values for circuit analysis. Micromachines, 14(2), 265. https://doi.org/10.3390/mi14020265

Le, X.-H., Ho, H. V., Lee, G., & Jung, S. (2019). Application of long short-term memory (LSTM) neural network for flood forecasting. Water, 11(7), 1387. https://doi.org/10.3390/w11071387

Ling, T., & Cai, Y. (2022). Financial crisis prediction based on long-term and short-term memory neural network. Wireless Communications and Mobile Computing, 2022, 1-8. https://doi.org/10.1155/2022/5728470

Liu, J., Li, C., Ouyang, P., et al. (2022). Interpreting the prediction results of the tree-based gradient boosting models for financial distress prediction with an explainable machine learning approach. Journal of Forecasting, 42(5), 1112-1137. https://doi.org/10.1002/for.2931

Liu, Y., Song, C., Tian, Z., & Shen, W. (2023). Identification of high-risk patients for postoperative myocardial injury after CME using Machine Learning: A 10-year Multicenter Retrospective Study. International Journal of General Medicine, 16, 1251–1264. https://doi.org/10.2147/ijgm.s409363

Long, X., Kampouridis, M., & Jarchi, D. (2022). An in-depth investigation of genetic programming and nine other machine learning algorithms in a financial forecasting problem. https://doi.org/10.1109/cec55065.2022.9870351

Nasution, A., Matondang, N., & Ishak, A. (2022). Inventory optimization model design with machine learning approach in feed mill company. Jurnal Sistem Teknik Industri, 24(2), 254-272. https://doi.org/10.32734/jsti.v24i2.8637

Noviantoro, T., & Huang, J.-P. (2021). Applying Data Mining Techniques To Investigate Online Shopper Purchase Intention Based On Clickstream Data. Review of Business, Accounting & Finance Volume, 1(2), 130–159.

Ohlson, J. A. (1980). Financial ratios and the probabilistic prediction of bankruptcy. Journal of Accounting Research, 18(1), 109–131. https://doi.org/10.2307/2490395

Paule-Vianez, J., Gutiérrez-Fernández, M., & Coca-Pérez, J. L. (2019). Prediction of financial distress in the Spanish banking system. Applied Economic Analysis, 28(82), 69–87. https://doi.org/10.1108/aea-10-2019-0039

Probst, P., Boulesteix, A.-L., & Bischl, B. (2019). Tunability: Importance of Hyperparameters of Machine Learning Algorithms. Journal of Machine Learning Research, 20, 1–32. https://doi.org/ https://doi.org/10.48550/arXiv.1802.09596

Rahman, M. J., & Zhu, H. (2024). Predicting financial distress using machine learning approaches: Evidence China. Journal of Contemporary Accounting & Economics, 20(1), 100403. https://doi.org/10.1016/j.jcae.2024.100403

Ramos-Pérez, E., Alonso-González, P., & Núñez-Velázquez, J. (2019). Forecasting volatility with a stacked model based on a hybridized artificial neural network. Expert Systems With Applications, 129, 1-9. https://doi.org/10.1016/j.eswa.2019.03.046

Ryll, L., & Seidens, S. (2019). Evaluating the performance of machine learning algorithms in financial market forecasting: a comprehensive survey. https://doi.org/10.48550/arxiv.1906.07786

Salehi, M., Mousavi Shiri, M., & Bolandraftar Pasikhani, M. (2016). Predicting corporate financial distress using data mining techniques. International Journal of Law and Management, 58(2), 216–230. https://doi.org/10.1108/ijlma-06-2015-0028

Sehgal, S., Mishra, R. K., Deisting, F., & Vashisht, R. (2021). On the determinants and prediction of corporate financial distress in India. Managerial Finance, 47(10), 1428–1447. https://doi.org/10.1108/mf-06-2020-0332

Sharma, S., & Mahajan, V. (1980). Early warning indicators of business failure. Journal of Marketing, 44(4), 80. https://doi.org/10.2307/1251234

Shen, F., Liu, Y., Wang, R., & Zhou, W. (2020). A dynamic financial distress forecast model with multiple forecast results under unbalanced data environment. Knowledge-Based Systems, 192, 105365. https://doi.org/10.1016/j.knosys.2019.105365

Shen, Z., & Chen, S. (2022). Financial distress prediction: a hybrid tracking model approach. Asian Journal of Economics Business and Accounting, 185-192. https://doi.org/10.9734/ajeba/2022/v22i24906

Sheng, Y., & Ma, D. (2022). Stock index spot–futures arbitrage prediction using machine learning models. Entropy, 24(10), 1462. https://doi.org/10.3390/e24101462

Shumway, T. (2001). Forecasting bankruptcy more accurately: A simple hazard model. The Journal of Business, 74(1), 101–124. https://doi.org/10.1086/209665

Sirisha, U., Belavagi, M., & Attigeri, G. (2022). Profit prediction using arima, sarima and lstm models in time series forecasting: a comparison. IEEE Access, 10, 124715-124727. https://doi.org/10.1109/access.2022.3224938

Song, Y. (2023). Class-imbalanced financial distress prediction with machine learning: incorporating financial, management, textual, and social responsibility features into index system. Journal of Forecasting, 43(3), 593-614. https://doi.org/10.1002/for.3050

Thi Vu, L., Thi Vu, L., Thu Nguyen, N., et al. (2019). Feature selection methods and sampling techniques to financial distress prediction for Vietnamese listed companies. Investment Management and Financial Innovations, 16(1), 276–290. https://doi.org/10.21511/imfi.16(1).2019.22

Tissaoui, K., Zaghdoudi, T., Hakimi, A., & Nsaibi, M. (2022). Do gas price and uncertainty indices forecast crude oil prices? fresh evidence through XGBoost modeling. Computational Economics, 62(2), 663-687. https://doi.org/10.1007/s10614-022-10305-y

Tron, A., Dallocchio, M., Ferri, S., & Colantoni, F. (2022). Corporate governance and financial distress: lessons learned from an unconventional approach. Journal of Management & Governance, 27(2), 425-456. https://doi.org/10.1007/s10997-022-09643-8

Tsai, C.-F. (2014). Combining cluster analysis with classifier ensembles to predict financial distress. Information Fusion, 16, 46–58. https://doi.org/10.1016/j.inffus.2011.12.001

Vu, L., Vu, L., Nguyen, N., et al. (2019). Feature selection methods and sampling techniques to financial distress prediction for vietnamese listed companies. Investment Management and Financial Innovations, 16(1), 276-290. https://doi.org/10.21511/imfi.16(1).2019.22

Wang, L., Wang, X., Chen, A., et al. (2020). Prediction of type 2 diabetes risk and its effect evaluation based on the XGBOOST model. Healthcare, 8(3), 247. https://doi.org/10.3390/healthcare8030247

Ye, L., Li, Y., Zhang, H., & Kou, Y. (2020). Internal and external factors leading to corporate financial distress a case study of HuaYi brothers media group. Saudi Journal of Economics and Finance, 4(6), 281-286. https://doi.org/10.36348/sjef.2020.v04i06.014

Yousaf, U., Jebran, K., & Wang, M. (2021). Can board diversity predict the risk of financial distress? Corporate Governance, 21(4), 663-684. https://doi.org/10.1108/cg-06-2020-0252

Zhong, J., & Wang, Z. (2022). Artificial intelligence techniques for financial distress prediction. AIMS Mathematics, 7(12), 20891–20908. https://doi.org/10.3934/math.20221145

Zmijewski, M. E. (1984). Methodological issues related to the estimation of financial distress prediction models. Journal of Accounting Research, 22, 59–82. https://doi.org/10.2307/2490859