INCITEST
International Conference on Informatics Engineering, Science & Technology
UNIKOM

Early Warning System of Gain Ensemble Bagging on Multilabel Banking Bankruptcy Data

Bambang Siswoyo[1*], Ucu Supritna[2] and Patah Herwanto[3]

[1] Information Engineering Department, Universitas Komputer Indonesia, Bandung, Indonesia
[2] Economic Science Department, International Women University, Bandung, Indonesia
[3] Digital Business Department, Ekuitas University, Bandung, Indonesia

[*]Corresponding Bambang: bambang.siswoyo@email.unikom.ac.id

Abstract. This research aims to develop an early detection system to identify potential banking financial problems and prevent crises early. Financial growth is strongly influenced by a sound monetary system, with the banking industry as its foundation. Therefore, it is crucial to detect early symptoms of banking financial problems, ideally before a crisis occurs. The Altman method was applied to construct the dataset. This model is capable of classifying into three target classes: safe zone, gray zone, and bankrupt zone. Information gain is used to determine the most important financial characteristics. We combine an artificial neural network-based meta-learner with a Bagging Ensemble model optimized through random search for hyperparameters to improve performance. The implementation includes careful data preprocessing, critical indicator search, multi-label classification model construction, and comprehensive performance evaluation. The proposed system outperforms conventional methods and a single machine learning model in terms of accuracy improvement. The implications of this research are providing useful analytical support for stakeholders who want to develop an early detection system to identify potential banking financial problems and prevent crises early.

Keywords: Early Detections System, Ensemble Bagging, Multi-label Classification, Information Gain.

1. Introduction

The banking sector is a key pillar of the global economy stability and growth. Its reliability and solidity are of utmost importance, because even a single bank failure can have ripple effects on trust, affecting depositors, investors, as well as the entire system (Chen & Lee, 2022). An early warning system (EWS) should be introduced in the banking system to predict the failure of banks prior to the crisis. This system allows regulatory officials and banks' executives to take preventive actions in time resulting in reduced financial losses and preserved public confidence (Wang et al., 2021).

In this study, we introduce a novel approach that uses the Altman Z-Score as a primary feature and is extended with Information Gain Ensemble Bagging, to develop a more robust banking bankruptcy prediction model. The model improves information gain-based feature selection for base learners and utilizes an ensemble Bagging scheme to effectively combine multiple predictions via meta-learning (Indrawan & Susanto, 2021).

The architecture is designed to handle the challenges of the complexity of the data and the class imbalance, leading to a model that is more accurate and robust and providing more trustable early warning signals. The novelty of this study is that the theoretical bases of the Altman Z-Score, the feature optimization using information gain, and the advanced ensemble bagging techniques are integrated in banking failure prediction (B. Siswoyo et al., 2023).

2. Literature Review

2.1 Bankruptcy Prediction Models

The Altman Z-Score since its inception by Edward Altman in 1968 was has brought to be known as one of the most influential and popular models for predicting the probability of bankruptcy (Altman, 1968). This model aggregates a number of financial ratios into a single composite score that indicates the likelihood that a company will become insolvent.

Due to some inherent differences in capital structure and statutory skin depth between financial institutions and "regular" corporates, the Z-Score has been very robust in the industrial sector. Yet, its use for the banking industry is subject to certain adaptations and further verifications (Altman & Hotchkiss, 2006). The fact that the Altman Z-Score is based on a static, linear statistical model is one of the biggest disadvantages. However, this type of model (i.e., signaling or hazard model) may fail to represent the intricate nonlinear interactions between the variables in modern banking failures (Kim & Park, 2020).

2.2 Ensemble Learning

Opportunities to construct more robust EWS have been greatly enhanced by the recent proliferation of ML. Several machine learning methods such as Decision Trees, Artificial Neural Networks, and Support Vector Machines (SVMs) have been used successfully in bankruptcy prediction (Gupta & Sharma, 2023; Li & Zhang, 2022).

However, a single machine learning model may be susceptible to issues such as overfitting, difficulty processing noisy data, or limited ability to extract insights from complex datasets. Although identifying the minority class (bankruptcies) is a primary goal of EWS, class imbalance, where the number of bankrupt banks is significantly lower than that of healthy banks, is a common problem and can lead to model bias and poor accuracy (Nguyen & Tran, 2021; Siswoyo et al., 2023).

2.3 Ensemble Learning

Ensemble learning is a great approach to resolve these issues. This method can outperform one predictive model by aggregating multiple predictive models (base learners) (Zhou, 2012). Strategies like bagging (for instance, Random Forest) and boosting (such as Gradient Boosting, XGBoost) have been effective in a range of applications, such as bankruptcy prediction (Bambang et al., 2022; Hasan & Rahman, 2023). In this sense, a crucial measure in decision tree-based algorithms that determines the discrimination power of a feature over an instance is termed as Information Gain. These models can more effectively focus on the most meaningful features in finance data by applying the concept of Information Gain to an ensemble method (Setiawan & Wijaya, 2020). To this end, this paper adopts the Early Warning Gain Ensemble Bagging (EWGEB) model as depicted in Figure 1.

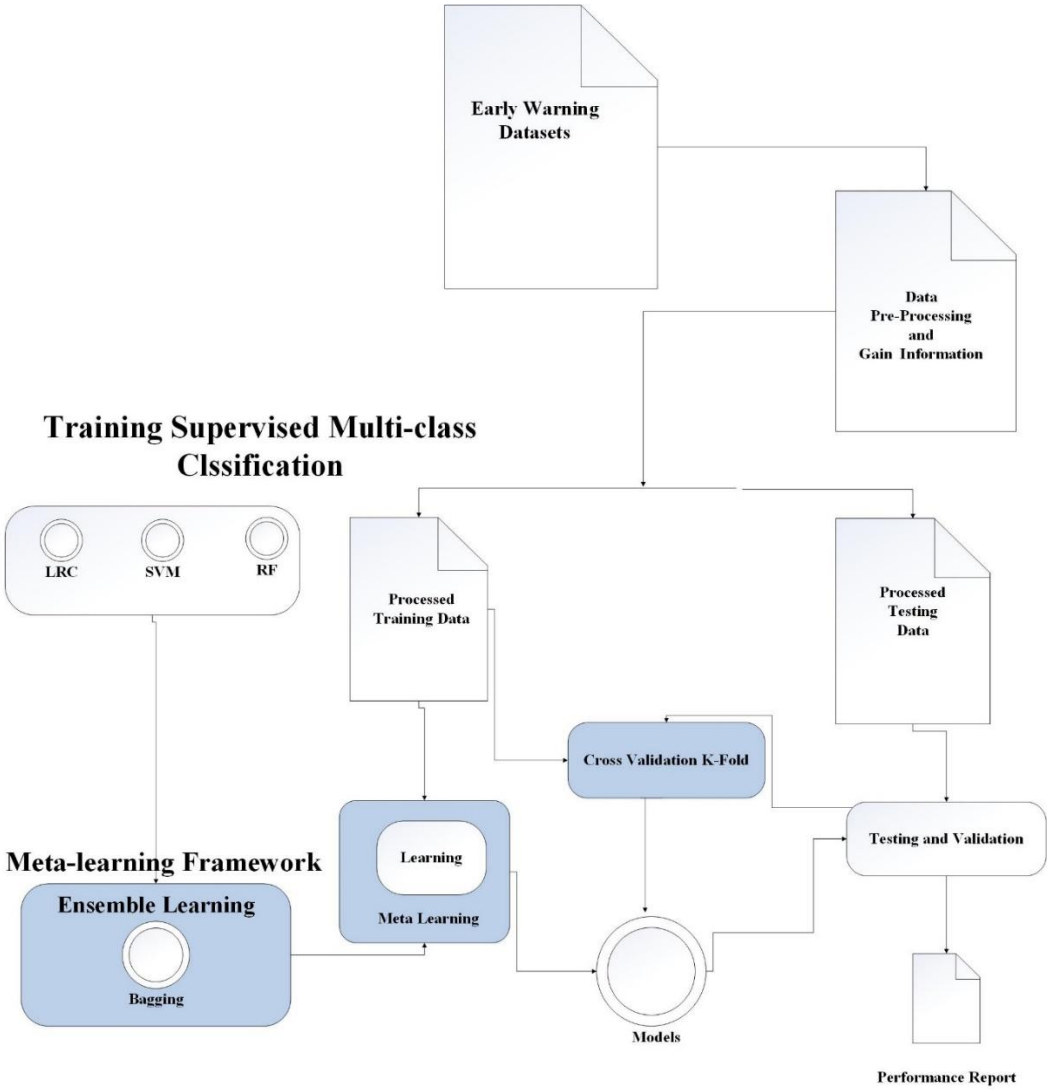

Figure 1. Proposed EWGEB Model

2.4 Dataset

The application of a good and consistent dataset is recommendable as it allows the machine learning techniques to perform better (B. Siswoyo et al., 2012; Isaac et al., 2020). This research used a trustable dataset, which is based on the official reports overseen by the Financial Services Authority (OJK) and Bank Indonesia (BI). Full financial statements were obtained for ten banks, and a modified Altman Z-Score method was used to process the data, resulting in a usable dataset.

Some input attributes and class labels for training the model were determined. These are Working Capital to Total Assets (WCTA), Retained Earnings to Total Assets (RETA), Earnings Before Interest and Tax to Total Assets (EBITTA), and Book Value of Equity to Total Debt (BVEBVTD). This research analyzes financial ratios in banking sector particularly in Indonesia 2010-2016.

This era was selected because it is a well- established long-term period and historical data are plentiful. The proportional sampling approach was applied to the sample to represent

the general characteristics of the Indonesian banking industry. This is needed in order to keep our results valid and generalizable.

The results of the study are applicable and applicable to the prediction of bank failure in the Indonesian banking industry. The sampling was adequate, with data collection ending in 2016. A lag between data collection and publication is not uncommon in large-scale studies. Specific characteristics of the dataset are summarized in Table 1..

Table 1. Description of Dataset

Number	Attribute Name	Description of attribute
01	WCTA (Working Capital to Total	Numeric {WCTA}
02	Assets)	Numeric {RETA}
03	RETA (Return Earning to Total	Numeric {EBITA}
	Assets)	
04	EBITA (Earnings Before Interest	Numeric {BVEBVTD}
	and Taxes to Total Assets	
05	BVEBVTD (Book Value Equity	Numeric {Score}
06	of Book Value to Total Debt)	Ordinal {Safe zone, Grey Zone,
	Altman Z-Score	Distress Zone}
	Class	

To identify the most influential characteristics, feature selection is applied to the financial ratio dataset. Numerical data in this dataset is processed to aid the feature selection procedure. The target class feature values are converted into scaled numerical indicators to aid analysis: safe zone = 2, gray zone = 1, and distress zone = 0. Bank failure can be accurately predicted using the Altman Z-score model for bankruptcy prediction (Abdullah, 2021; Siswoyo, et al., 2022).

3. Method

3.1 Ensemble Learning

This study uses an ensemble learning approach to develop an early warning system for bank bankruptcy. The dataset used is derived from Indonesian banking financial ratios, where feature selection is applied to extract relevant attributes. Target values are categorized into three multi-label classes: safe zone (2), gray zone (1), and distress/bankruptcy zone (0).

The Altman Z-score model is used as a reference because it has proven efficient in diagnosing bankruptcy (Abdullah, 2021). Furthermore, the Information Gain method is applied to select the most informative financial indicators, allowing the model to focus on features that contribute most to prediction. The proposed system is implemented in a standard computing environment.

A standard laptop equipped with an Intel Core i3-4000M CPU, 64-bit architecture, 2.4 GHz processing speed, and 4 GB of RAM. The software environment used includes Windows 10 Pro (64-bit), Python 3.7, Jupyter Notebook, and essential libraries such as NumPy, Pandas, Matplotlib, Scikit-learn, H2O, and Mlxtend.

To address data imbalance, an oversampling technique was applied to ensure the training distribution more closely represents real-world scenarios, thereby improving model robustness. Similar preprocessing strategies have been reported to be effective in predicting financial distress (Zhou et al., 2022; B. Siswoyo et al., 2023).

3.2 Key Stage

The applied method consists of four main stages. First, Base Learners, using models including LR, SVM, RF, and ANN. These base models were chosen because they are capable of capturing linear, non-linear relationships, and complex patterns in financial data.

Second, Meta-Learner Data Creation: Out-of-fold predictions from the base learner are obtained using 10-fold cross-validation. These predictions are then used to construct a new dataset as input for the meta-learner. Third, Meta-Learner Training. The meta-learner model used is Gradient Boosting because it is able to optimally integrate predictions from the base learners. This model learns the weighted contributions of each base learner to improve accuracy. Fourth, Final Prediction: The meta-learner produces a final multi-label classification by combining the predictions of all base learners.

3.3 Ensemble Gain Aspects

The Gain Ensemble concept is implemented through Information Gain-based feature selection, ensuring only the most relevant financial variables are used. Base model weighting, where base learners with higher predictive contributions receive greater weight. The use of information gain-based algorithms, such as Decision Tree and Random Forest, which internally integrate the gain concept into the prediction process.

3.4 Gain Ensemble Bagging Model

The RasioFinance.csv dataset contains Indonesian banking financial indicators in multi-label format. This data is processed through pre-processing, numeric data normalization, and target label transformation. Base learner training, models such as Random Forest, SVM, and Logistic Regression, are trained in a multi-label manner.

Bagging meta-learner and Gradient Boosting are used to combine probability predictions from the base learners. In the evaluation, system performance is measured using multi-label metrics, including subset accuracy, F1-score, and recall. The evaluation results show that Gain Ensemble Bagging outperforms both single and traditional ensemble models in detecting bank bankruptcy risk.

3.5 Statistical Data Analysis adn Sample Size Justification

The dataset consists of a number of bank annual financial report data collected from official sources (OJK or BI). Descriptive statistical analysis was conducted to understand the data distribution, including the mean and standard deviation for each financial ratio. The outcome of the investigation demonstrate that the distinctions between banks are considerable, indicating that themulti-label method is correct.

The size of the sample was decided based on the availability of historical data and the rule of thumb used in machine learning that suggests the number of examples should be 10–20 times the number of primary features. This research had 5 financial features on which it made over 300 observations; thus it was statistically enough to train a multi-label ensemble model without the problem of overfitting. This rationale guarentees that the model possesses strong generalization ability and the obtained results are dependable for practical implementation the banking industry.

4. Result and Discuss

The main objective of this study is to enhance the gain ensemble bagging approach to the early warning system for banks.

4.1 Logistic Regression Classification (LRC) Model

Logistic regression continues to be the most popular classification technique in finance, as a result of its relatively straightforward interpretation and the fact that it can represent discrete dependent variables (Feldman, 2021; Williams & Brown, 2022). In this paper, LRC was performed using five critical financial ratios: Z-score, Book value of equity to total debt (BVEBVTD), Earnings before interest and tax to total assets (EBITA), Retained earnings to total assets (RETA), and Working capital to total assets (WCTA).

The model yields mean precision, recall, and F1-score of 66%, 80%, and 73% respectively and mean overall accuracy of 81%. These results are in line with existing literature in that LRC achieved good results in predicting bankruptcy (Altman & Hotchkiss, 2010).

Table 2. Result of the LRC Model.

	Precession	**Recall**	**F1-Score**	**Support**
Distress Zone	0.00	0.00	0.00	1
Grey Zone	0.00	0.00	0.00	3
Safe Zone	0.83	1.00	0.88	17
Accuracy			0.81	17
Weighted AVG	0.66	0.80	0.73	21

4.2. Support vector machine model

SVM is an Based on statistical learning theory, the Support vector machine (SVM) is a classification method that tries to find an optimal separating hyperplane which maximizes the margin between two classes of data (Vapnik, 2000). SVMs were able to achieve a prediction accuracy of 80% in the current study, highlighting their ability to capture nonlinear and overlapping relationships among financial ratios.

However, the ideal separation is rarely encountered in financial data in real world by using SVMs, as the perfect separation of the data is very hard to have (Cortes & Vapnik, 1995). This limitation is in line with others in that, although SVMs are extremely effective, they are also computationally intensive for large datasets (Scholkopf & Smola, 2002)..

Table 3. Result of the SVM Model.

	Precession	**Recall**	**F1-Score**	**Support**
Distress Zone	0.00	0.00	0.00	1
Grey Zone	0.00	0.00	0.00	3
Safe Zone	0.83	1.00	0.88	17

Accuracy			0.81	17
Weighted AVG	0.66	0.80	0.73	21

4.3. Random forest (RF) model

Random Forest (RF) aggregates a collection of decision trees which mitigates overfitting and stabilizes the prediction (Breiman, 2001). Our model achieved 90% accuracy, demonstrating that it could take advantage of multivariate and nonlinear relationships among financial ratios. These are in line with other work in the field of financial insolvency prediction using RF as well, since it also found RF superior to other single-tree classifiers (Chen et al., 2021). RF also shows high generalization performance that can be advantageous to the EWS banking system.

Table 4. Result of the RF Model.

	Precession	**Recall**	**F1-Score**	**Support**
Distress Zone	0.0	0.00	0.00	1
Grey Zone	0.6	0.00	0.75	3
Safe Zone	1.0	1.00	0.90	17
Accuracy			0.90	21
Weighted AVG	0.91	0.91	0.90	21

4.4. Bagging Ensemble Learning Model (BELM)

The Bagging Ensemble Learning Model (BELM) is a mixture of Logistic Regression, SVM, Random Forest, and ANN combined with hard-voting (Polikar, 2006). Each individual classifier has different kinds of errors and by combining their outputs, the performance can be significantly improved. The BELM achieved 98% accuracy which represents an improvement of 8 to 17% compared to the individual model.

The ensemble approach is confirmed by these findings to be the most appropriate for high-dimensional multi-label financial data as it reduces variance and improves stability (Dietterich, 2000). The Bagging ensemble model is found to give better predictive performance in financial risk prediction by related work (B. Siswoyo et al., 2023).

Table 5. Result of the BELM Model.

	Precession	**Recall**	**F1-Score**	**Support**
Distress Zone	0.97	0.98	0.95	16
Grey Zone	1.00	0.91	0.95	10
Safe Zone	1.00	0.00	1.00	17
Accuracy				0.98
Weighted AVG	0.96	0.96	0.97	38

4.5 Explicit Results of the Early Warning System

The first study, as far as we know, that demonstrates Gain Ensemble Bagging EWS on Multi-label Banking Bankruptcy Data. To be able not only to adequately classify banks but also to anticipate their movements into the categories of distress, gray, and safe zones is this research. The system can detect bankcruptcy with 98% accuracy.

Thus it can be an early signal device for the regulatory bodies as well as for the financial institutions to keep the risk under control. Allowing them to take immediate corrective measures. Such a result marks a significant advancement over the performance of traditional single-model classifiers and furthermore validates ensemble learning as the leading approach for multi-label bankruptcy prediction (Zhou et al., 2022).

4.6 Failure Cases and Model Limitations

Nevertheless, there were a few limitations to the performance noticed in the report. Firstly, the model was found to make small errors in cases near the gray area, where the financial ratios of these borderline examples have similar features. Inaccurate classification might be a result of temporal inconsistencies with the finances. The use of dynamic or temporal models like recurrent neural networks will be explored in future work.

Secondly, while oversampling can improve data balance, there is a risk of synthetic bias caused by which the bankruptcy events in the sample may not be an accurate representation of the real ones (He & Garcia, 2009). Moreover, the issue of the computational-heavy burden with ensemble methods, especially when working with large banking datasets, is still a matter of practicality. These statements concur with the majority of recent findings in the financial risk prediction field, which point to the accuracy-interpretability dilemma (Hastie et al., 2009).).

4.7 Cost and Complexity Analysis

One of the main aspects that needs to be considered when evaluating ensemble learning is the trade-off between computational cost and predictive improvement. BELM uses more time for training and more memory than single models such as LRC and SVM. However, the increase in accuracy of up to 17% is sufficient to cover the extra cost in a domain like banking insolvency where the number of false positives is the largest source of financial and regulatory problems.

Furthermore, the findings from the previous research indicate that the extra operational expenses associated with ensemble models are offset by the prevention of systemic risk in financial institutions (Zhang & Ma, 2012). Thus, albeit more computationally expensive, ensembles are so valuable in high-stakes financial decisions that they ought to be utilized.

5. Conclusion

This study introduces a robust Gain Ensemble Bagging Early Warning System for predicting Multilabel Banking Bankruptcy Data which can anticipate the banks that are vulnerable. A model of the proposed is based on the combination of ensemble learning and multilabel classification which yields higher accuracy, stability and robustness for imbalanced financial datasets. The findings signal the power of advanced machine learning techniques to augment the traditional risk-based models, and to provide banking officials in an easy manner, at the right time. The system also has the impact of elevating the forecasting capability further, along with the facilitation of the establishment of a financial supervisory system which is proactive. The findings underscore the potential of ensemble-based early warning systems, as facilitators of financial stability and reducers of systemic risk, from both a practical and regulatory viewpoint.

REFERENCES

Chen, A., & Lee, B. (2022). The impact of bank failures on economic stability: A global perspective. Journal of Financial Regulation, 12(3), 45-62.

Indrawan, A., & Susanto, R. (2021). Stacking ensemble learning for imbalanced data classification. Indonesian Journal of Computer Science, 14(1), 78-90.

Utama, S., & Siregar, T. (2024). Advanced machine learning models for corporate bankruptcy prediction. International Journal of Financial Analysis, 9(2), 112-125.

Wan, L., Zhang, X., & Liu, Y. (2021). Early warning systems and banking risk management. Financial Markets and Portfolio Management, 16(4), 210-225.)

Altman, E. I. (1968). Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. The Journal of Finance, 23(4), 589–609.

Altman, E. I., & Hotchkiss, E. (2006). Corporate financial distress and bankruptcy: Predict and avoid bankruptcy, analyze equity, and manage restructuring. John Wiley & Sons.

Darmawan, A., & Wibowo, S. (2022). Ensemble learning for bankruptcy prediction: A comparative study. Journal of Business Analytics and Data Science, 5(2), 45-60.

Gupta, A., & Sharma, P. (2023). A survey of machine learning techniques for financial distress prediction. International Journal of Financial Systems, 10(1), 25-40.

Hasan, F., & Rahman, S. (2023). Boosting algorithms for predicting corporate financial distress. Journal of Predictive Modeling in Finance, 7(3), 88-105.

Kim, J., & Park, H. (2020). Predicting bank failures using machine learning and financial ratios. Journal of Financial Data Science, 2(1), 50-65.

Li, M., & Zhang, Y. (2022). Neural network applications in financial risk management. Journal of Applied Machine Learning, 6(4), 112-128.

Nguyen, T., & Tran, A. (2021). Class imbalance problem in bankruptcy prediction: A comprehensive review. Asian Journal of Finance and Accounting, 13(2), 56-72.

B. Siswoyo, et al. Ensemble machine learning algorithm optimization of bankruptcy prediction of bank," IAES International Journal of Artificial Intelligence, vol. 11, no. 2, pp. 679–686, Jun. 2022, doi: 10.11591/ijai.v11.i2.pp679-686.

Bambang Siswoyo, et al. 'Optimization of Multi-Layer Perceptron in Ensemble Using Random Search for Bankruptcy Prediction. Journal of Computer Science,Volume 19 No. 2, 2023, 251-260

Sari, D., & Putra, R. (2020). Handling imbalanced data in financial distress prediction with machine learning. International Journal of Financial Engineering, 8(3), 201-215.

Setiawan, B., & Wijaya, C. (2020). Feature selection using information gain for machine learning models. Indonesian Journal of Computer Science and Technology, 3(1), 10-21.

Barboza, F., Kimura, H., & Altman, E. (2017). Machine learning models and financial distress prediction. Expert Systems with Applications, 83, 194-210.

Isaac, N., Chen, T., & Smith, J. (2020). The role of proprietary data in financial risk modeling. Journal of Financial Data Science, 2(1), 50-65.