**Article Type:** Research Paper

# Ensemble learning with imbalanced data handling in the early detection of capital markets

Putri Auliana Rifqi Mukhlashin[1]*, Anwar Fitrianto[1], Agus M Soleh[1], and Wan Zuki Azman Wan Muhamad[2]

**Abstract**

**Research aims:** This study aims to create an early detection model to predict events in the Indonesian capital market.

**Design/Methodology/Approach:** A quantitative study comparing ensemble learning models with imbalanced data handling detected early capital market events. This study used five ensemble learning models—Random Forest, ExtraTrees, CatBoost, XGBoost, and LightGBM—to detect early events in the Indonesian capital market by handling imbalanced data, such as under sampling (RUS), oversampling (SMOTE, SMOTE-Broder, ADASYN), and over-under sampling (SMOTE-Tomek, SMOTE-ENN), weighted (class weight). Global and regional stock markets, commodities, exchange rates, technical indicators, sectoral indices, JCI leaders, MSCI, net buys of foreign stocks, national securities, and national share ownership all predicted the lowest return of Crisis Management Protocol (CMP) binary responses.

**Research findings:** Hyperparameters and thresholds were tuned to produce the optimum model. The best model had the highest G-mean. ExtraTrees with SMOTE-ENN predicted the highest number of one-day events, with a G-Mean of 96.88%. LightGBM with SMOTE handling best predicted five-day events with an 89.21% G-Mean. With a G-Mean of 89.49%, CatBoost with SMOTE-Border handling was the best for a 15-day event. In addition, LightGBM with SMOTE-Tomek handling and 68.02% G-Mean was best for 30-day events. Further, performance evaluation scores decreased with increased prediction time.

**Theoretical contribution/Originality:** This work relates more imbalance handling methods and ensemble learning to capital market early detection cases. Practitioner/Policy implication: Capital markets can indicate economic stability. Maintaining capital market efficacy and economic value requires a system to detect pressure.

**Research limitation/Implication:** This study used ensemble learning models to predict capital market events 1, 5, 15, and 30 days ahead, assuming Indonesian working days. The model's forecast results are expected to be utilized to monitor the capital market and take precautions.

**Keywords:** Capital Market; Early Detection; Ensemble Learning; Imbalance Class; Risk Event

**AFFILIATION:**
[1] Department of Statistics, Faculty of Mathematics and Science, IPB University, West Java, Indonesia

[2] Institute of Engineering Mathematics, Universiti Malaysia Perlis, Arau, Malaysia

***CORRESPONDENCE:**
putriaulianarifqi@gmail.com

**DOI:** 10.18196/jai.v24i2.17970

**CITATION:**
Mukhlashin, P. A. R., Fitrianto, A., Soleh, A. M., & Muhamad, W. Z. A. W. (2023). Ensemble learning with imbalanced data handling in the early detection of capital markets. *Journal of Accounting and Investment, 24*(2), 600-617.

**ARTICLE HISTORY**
**Received:**
20 Feb 2023
**Revised:**
17 Mar 2023
**Accepted:**
19 Mar 2023



**JAI Website**:

## Introduction

Imbalanced data can be handled in various approaches, including under sampling and oversampling. While under sampling is done by reducing the

majority class randomly (Wang & Liu, 2021), oversampling is accomplished by randomly duplicating the minority class (Putri & Dhini, 2019).

The oversampling approach includes SMOTE, SMOTE-Borderline, and ADASYN. The Synthetic Minority Over-Sampling Technique (SMOTE) creates new samples from the minority class by taking the nearest minority class samples. SMOTE-Borderline generates new data along the line between the minority class and its nearest neighbors. Besides, Adaptive Synthetic Sampling (ADASYN) uses a distribution weight for instances in the minority class based on the model's difficulty level of learning data. Faris et al. (2020) showed that SMOTE improves the prediction strength in the geometric mean (G-Mean) and type II errors. However, both under sampling and oversampling techniques are limited (Rahardja et al., 2023). Under sampling can remove important parts of the majority class, while oversampling can cause overfitting in the model. Sir and Soepranoto (2022) found that SMOTE-ENN (Edited Nearest Neighbor), included in the over-under sampling approach, is the best resampling technique. Indrawati (2021) also uncovered that SMOTE-ENN improved the accuracy performance of SVM by 2%–23%. Another handling approach is class weight, which gives more weight to the minority class. This method efficiently handles class imbalance (Asundi et al., n.d.).

Ensemble models combine multiple machine learning models to improve prediction performance and accuracy (Mishraz et al., 2021). They are typically used in classification problems and are better than single models, such as logistic regression, SVM, and neural networks, because the prediction errors of a single model are not always made by another model (Lutfiani et al., 2023). Examples of ensemble models include bagging (bootstrap aggregating) and boosting methods. These methods combine predictions from multiple single models through voting or weighted voting (Mishraz et al., 2021).

Several models within the bagging method yield good evaluation results, such as Extremely Randomized Trees (ExtraTrees) and Random Forest (RF). Bagging and boosting models outperform the others in predicting bank failure, with the RF model having the highest area under the ROC curve (AUC) value at 93% (Liu et al., 2021). Thakkar and Chaudhari (2021) found that RF produced the best results, with an accuracy of 90.4%. According to research by Islam et al. (2019), Extremely Randomized Trees (ExtraTrees) generated comparatively better results with a ROC-AUC value of 94.2%. ExtraTrees also exhibited promising precision performance—more than one-period prediction (Aini et al., 2023). Meanwhile, models that fall under the boosting method include Categorical Boosting (CatBoost), Extreme Gradient Boosting (XGBoost), and Light Gradient Boosting (LightGBM). Research by Aly et al. (2022), using a four-period class imbalanced dataset, showed that CatBoost with oversampling SMOTE provides a high accuracy result in predicting bankruptcy in Poland, with an AUC value in each period of more than 90% (Santoso et al., 2022). Carmona et al. (2019) revealed that XGBoost provided the best result with an accuracy of 95%. With its ability to prevent overfitting and create predictions that can be applied generally, XGBoost capability increases accuracy in bank failure prediction. In the study, XGBoost outperforms both the conventional method (Logistic Regression) and the modern machine learning approach (Random Forest). A study by Wang et al. (2022) also demonstrated that LightGBM was better than Decision

Tree (DT), K-Nearest Neighbor, and RF in predicting bankruptcy with an F1-Score of 87.63%.

On the other side, the capital market is a platform that brings together parties that need capital and investors. The capital market can be an indicator of a country's economic stability and needs to be monitored to provide optimal benefits. In this regard, early warning systems can help reduce negative impacts on the capital market. Pressure on the capital market can also be caused by various factors, such as technical indicators, macroeconomics, portfolio management, regional or global integration, monetary policy, the global financial crisis, and behavioral economics factors such as public mood, news, risk aversion, and consumer confidence (Hermawan et al., 2023). Specifically, in machine learning modeling, pressure events are rare occurrences called class imbalance problems. Imbalanced class classification is where the majority class distribution has a larger proportion than the minority class. The class ratio is divided into three categories: mild when the minority class proportion is 20% to 40%, moderate when it is 1% to 20%, and extreme when it is less than 1% of the total data available (Google Developers, 2021).

Regarding the Chinese capital market early warning system, the RF algorithm has the highest accuracy for identifying bond market crises using under sampling and double-sampling methods, with 96.08 and 90.2%, respectively (Zhang & Chen, 2022). In developing an early warning system for predicting stock market crises in China based on market indicators and mixed frequency investor sentiments, the Artificial Neural Network (ANN)-based model demonstrates more stable performance, with an accuracy range of 98% to 99% (Pramono et al., 2022). For other models, such as Support Vector Machine (SVM), Decision Tree (DT), Random Forest (RF), Gradient Boosting (GBDT), K-Nearest Neighbor (KNN), and Logistic Regression (LR), the prediction accuracy is greatly influenced by different stock markets (Lu et al., 2021). On early warning, Wang and Liu (2021) showed that a long short-term memory (LSTM) network produces satisfying performance with a test-set accuracy of 96.4% and an average of 2.8 days of forewarning. Cross-validation, back-testing, and a reality check demonstrate the model's reliability and practical value in real-time decision-making.

Using machine learning models, research on class imbalances for the Early Warning System (EWS) or early detection in the capital markets is still rare, especially in Indonesia. This study, therefore, aims to use ensemble learning models for the early detection of events in the capital market in Indonesia with the handling of imbalanced data for predictions 1 day, 5 days, 15 days, and 30 days in advance, assuming working days in Indonesia are Monday through Friday. This research is anticipated to contribute knowledge about handling imbalances from multiple data periods with ensemble learning in situations involving the early detection of capital markets in Indonesia or other similar mattresses. In practice, this research can be used to monitor and detect events in the Indonesian capital market to assist with decision-making (Putri et al., 2023).

## Literature Review and Hypotheses Development

The Early Warning System (EWS) in the field of the economic crisis has been developed for a long time. Many works of literature use statistical models as the EWS. However, with the advancement of technology, new modeling types known as ensemble learning have emerged (Bintoro et al., 2023). Many studies have been conducted to investigate whether these ensemble learning models can accurately predict crises (Candra et al., 2023).

Ensemble learning can avoid overfitting and bias in the model because it creates multiple machine learning algorithms to determine the optimal value by majority voting. In the case of the early warning system for bank failures, ensemble learning has the greatest accuracy compared to all other SMOTE-based methods for converting imbalanced to balanced data (Shrivastava et al., 2020). Gnip and Drotár (2019) have compared several ensemble machine learning methods applied to a recently acquired dataset of small and medium-sized enterprises in the Slovak Republic. In certain instances, the highest obtained prediction accuracy of the proposed classification models, measured by geometric mean, is nearly 100% (Hariguna et al., 2022).

A study (Bluwstein et al., 2021) using macro-financial data from 17 countries from 1870 to 2016 approached the machine learning model in the stock market crisis study, where it applied one of the ensemble models, Extremely Randomized Trees (Extra Trees) and produced the most accurate results. Tölö (2020), in predicting systemic financial crises one to five years ahead, which includes the crisis dates and annual macroeconomic series of 17 countries over the period 1870-2016, found that machine learning models can be significantly improved by using Long-Short Term Memory (RNN-LSTM) and Gated Recurrent Unit (RNN-GRU) neural nets (Marlina et al., 2023). In another study, Coffinet and Kien (2019) focused on detecting rare events in banking crisis cases, using data collected from 32 European and non-European countries from 2010 to 2017 (Pratama & Wijaya, 2023). The random forest method was found to be the best approach for computing an indicator for the probability of a banking crisis (Zanubiya et al., 2023). In the case of bank insolvencies with non-failed, failed, and assisted entity data from the Federal Deposit Insurance Corporation (FDIC) database from 2008 to 2014, the Random Forest model provided the best results by considering the flow of data and had more stable and consistent performance in all test samples (Petropoulos et al., 2020). The Random Forest, a modern classification tree ensemble technique, provides the best results with an accuracy of 90.4%, and includes global credit and real estate variables as predictors (Thakkar & Chaudhari, 2021).

Furthermore, the objective of Carmona et al. (2019) was to anticipate bank failure in the United States financial system (Mahardika & Irawan, 2022). Expanded Gradient Boosting was used for empirical analysis (Sipahutar et al., 2020). This method evolved from previous boosting methods such as AdaBoost and boosted classification trees (Widiastuti et al., 2023). The XGBoost algorithm's applicability and ability are to increase the accuracy of bank failure prediction (Kosasi et al., 2022). In addition, an early warning system predicts banking crises to identify excessive credit growth and aggregate leverage. Junyu (2020) showed that the accuracy of XGBoost has reached approximately 80%, providing a

novel method for predicting financial crises. Lu et al. (2021) have also proven that, among single classifiers, the combination of XGBoost and random under sampling outperforms the random forest in predicting the probability of healthy enterprises. The probability of financial crisis enterprises in the t-period is maintained at 92.86%, the misjudgment rate of normal enterprises is reduced to approximately 12%, and the overall prediction accuracy is enhanced through a straightforward integration of the Random Forest and XGBoost (Tussa'diah & Kartika, 2023).

## Research Method

### Data

This study used data from the Financial Services Authority. The predictor variables used were all numeric variables that fell into several categories, including Global Regional Stock Market, Commodities and Exchange Rates, Technical Indicators, Sector Indices, IHSG Leaders, Morgan Stanley Capital Indonesia (MSCI), Foreign Net Buy/Sell Stocks, and Government Securities (SBN) and SBN ownership, with a total of 2424 variables. The response variable employed was the event of the lowest Crisis Management Protocol (CMP) return. If the value is less than minus 5%, it is "Pressure," while if the value is greater than or equal to minus 5%, it is "Normal" (Soesilo and Tinggi, 2021). The time range of the data was from January 2010 to November 2020. A list of variables used can be found at bit.ly/thesisvariables. Afterward, this study compared ensemble learning models with imbalanced data handling to further predict the Indonesian capital market. In this regard, ensemble learning creates a more accurate and complete strong model by combining weak classifier models. Bagging and boosting are two ways to combine poor classifiers with strong ones. The bagging method generates base classifiers in parallel, while the boosting method successively generates them and influences later classifiers.

### Analysis Process

This study took the following steps: data preprocessing, feature selection, data division, handling class imbalance, ensemble learning modeling, model improvement, and evaluation. Each stage had subprocesses involved. The detailed flow is shown in Figure 1.

### Data Preprocessing

This process includes data cleaning, Exploratory Data Analysis (EDA), and feature engineering. Data were cleaned to remove or modify irrelevant, duplicate, and unformatted data. In EDA and Feature Engineering, the addition of the return variable from each predictor variable, which is a daily price index, and the conversion of the daily price index predictor variable into stock return value was carried out. $P_t$ is the stock price at period t, and $P_{(t-1)}$ is the stock price at the previous period t-1.

$$\frac{P_t - P_{t-1}}{P_{t-1}} \quad \text{.....(1)}$$

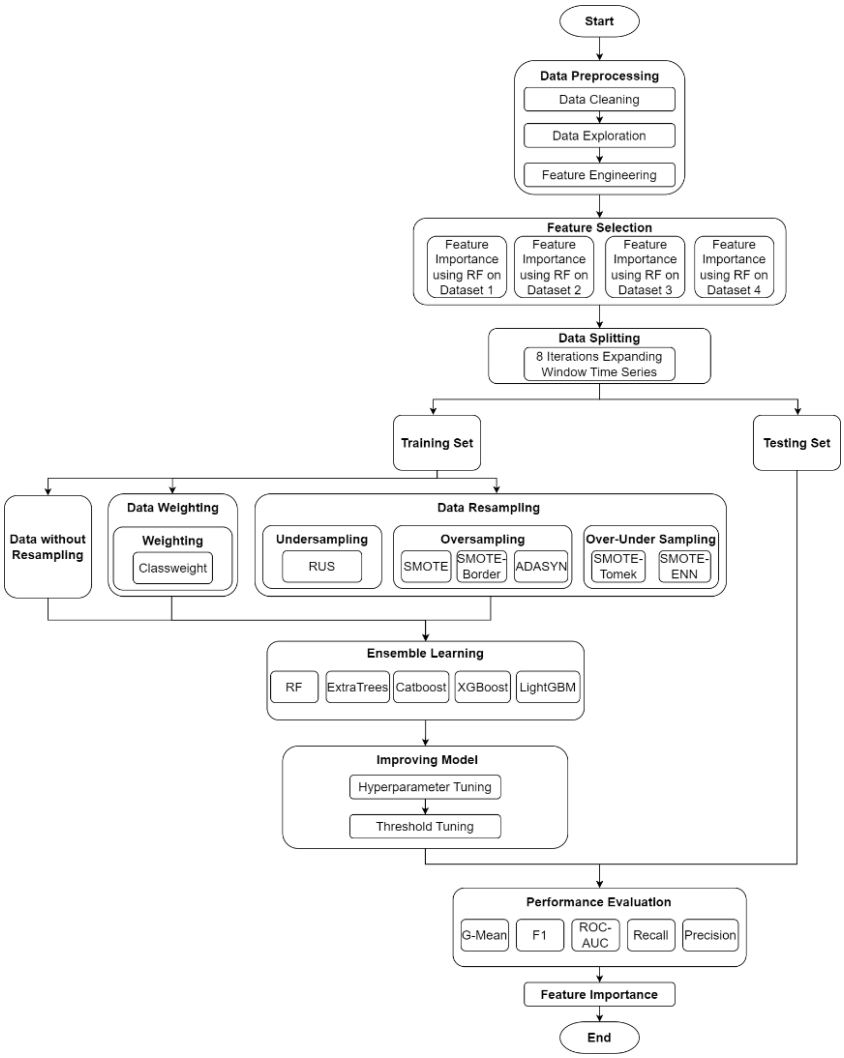



**Figure 1** Flow Diagram

In addition, other variables were added by creating moving average (MA) values. For example, the previously used ihsg_daily variable, the IHSG daily price index, was replaced with the return value ihsg_daily_return, and the moving average value was added, for example, ihsg_daily_return_movavg_five, or the five-day moving average of its daily return data. The MA value in this research was added to MA(k), with k = 5, 10, 15, and 30. Subsequently, data exploration was performed to identify the data patterns before and after each predictor variable's "Pressure" event. This pattern served as the basis for determining the change of the original variable into the time lag variable, such as in the example ihsg_daily_return_lag_i with i = 1, 2, 3, 4, and 5 to predict 1 day ahead. The original variable was replaced with the time lag variable (see Table 1) to avoid information leakage during the machine learning modeling.

**Table 1** Creating a time lag variable

| Period | X | x lag 1 | x lag 2 | … | x lag i | flag |
|---|---|---|---|---|---|---|
| 1 | -2.86 | NA | NA | … | NA | Normal |
| 2 | -1.72 | -2.86 | NA | … | NA | Pressure |
| 3 | 0.56 | -1.72 | -2.86 | … | NA | Normal |

X column is not utilized in the modeling process.

**Feature Selection**

Many initial variables were the basis for the feature selection process. Feature selection was performed using the Random Forest algorithm to remove weakly influencing variables and improve accuracy and classification performance (Chen et al., 2020). A feature selection was performed for each scenario, a prediction for 1 day, 5 days, 15 days, and 30 days ahead.

**Data Splitting**

The data division was carried out using the expanding window time series method (see Figure 2). This approach is called forward-chaining cross-validation (Vien et al., 2021). The data division was done annually by dividing the data into eight parts. However, 2014 and 2017 were not used as testing data due to the absence of "Pressure" events in those years.

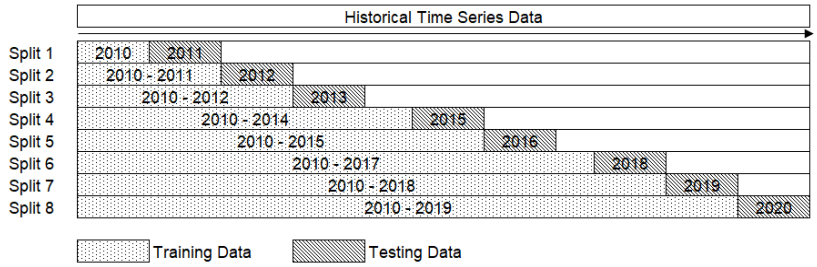



**Figure 2** Expanding Window Time Series

**Handling Class Imbalance Issue**

Based on Table 2, the proportion of "Pressure" events is 0.017, or 44 days out of 2620. It can be categorized as an extreme to moderate imbalance (Google Developers, 2021). Handling imbalanced class techniques are divided into four categories: under sampling (RUS), oversampling (ROS, SMOTE, SMOTE-Borderline, ADASYN), over-undersampling (SMOTE-Tomek, SMOTE-ENN), and weighting (class weight).

**Ensemble Learning Modeling**

The modeling was carried out on the training data using five models: ExtraTrees (Alfian et al., 2022), RF (Speiser et al., 2019), CatBoost (Jabeur et al., 2021), XGBoost (Qiu et al., 2021), and LightGBM (Sun et al., 2020). Each model was approached with and without the handling of class imbalance. It was also done for each prediction scenario.

**Table 2** The proportion of pressure event labels based on the year

| Year | Pressure Event |
|---|---|
| 2010 | 1.63% |
| 2011 | 4.45% |
| 2012 | 0.41% |
| 2013 | 3.20% |
| 2014 | 0.00% |
| 2015 | 0.81% |
| 2016 | 0.40% |
| 2017 | 0.00% |
| 2018 | 1.24% |
| 2019 | 0.41% |
| 2020 | 6.16% |

### Hyperparameter Optimization and Threshold

The hyperparameter optimization technique was performed using Optuna. Optuna is effectively used on the XGBoost model (Srinivas & Katarya, 2022). A total of 30 trials were performed to determine the best hyperparameter value of the model with imbalance class handling, such as the number of estimators, which is the number of decision trees, and max depth, which is the maximum depth of the decision tree. Threshold tuning was also performed using Optuna to obtain the optimal value based on the G-Mean.

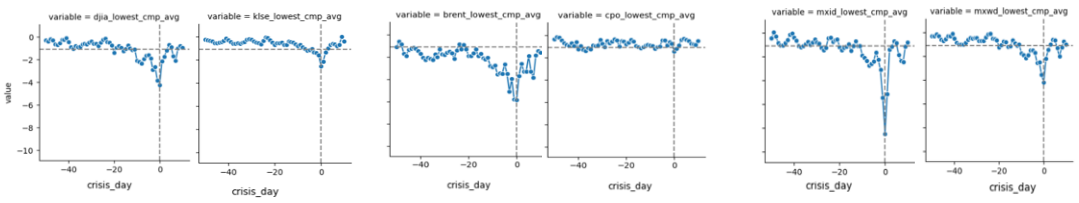
### Model Evaluation

This research aimed to predict the "Pressure" event for prevention and "Normal" for market capitalization monitoring. The evaluation was performed on the test data using the G-Mean value, maximizing both true positive and true negative while keeping both relatively balanced. The highest G-Mean value on the modeling algorithm is the basis for choosing the best model.

## Results and Discussion

### Data Exploration

The general pattern of the predictor variables with the pressure event was seen in this process. From the following Figure 3, to see the occurrence of pressure on the next day, on average, the predictor variables decreased in the previous 4 to 6 days. Then, the predictor variables decreased to see if there was pressure in the next five days, as seen in the previous 10 to 15 days. Meanwhile, to see pressure events in the next 15 days, the predictor variables decreased over the past 22 to 25 days. In addition, to predict the next 30 days, no pattern differed significantly from the predictor variable. Data exploration depicted that the predictor variables could detect pressure events before they happened.




(a) (b) (c)

**Figure 3** (a) Pattern of Global and Regional Market Indices: DJIA and KLSE; (b) Pattern of Commodity Prices: Brent and CPO; and (c) Pattern of MSCI: MXID and MXWD

**Feature Selection**

Feature selection was (see Figure 4) made to obtain variables that strongly influenced predicting pressure events in the capital market. In addition, it is expected to improve the performance of the classification algorithm in terms of computation. Fifty predictor variables were selected from the feature importance Random Forest model.

In the prediction of pressure events for the next 1 day, the variables with a high impact on the occurrence of pressure on the capital market were the IHSG, sectoral indices (financial, manufacturing), leading stocks (LQ45, Kompas) and the MSCI index (MXID) which were the highest 50 variables or 2.1% of all variables, based on RF prediction 1 day ahead feature importance covering 23.23% importance of the model. Then, the mining sector, indices (PCOMP, OMX), and exchange rates (YEN, YUAN) for the next 5 days covered 12.70%. Furthermore, for the next 15 days, foreign investors' SBN holdings, the exchange rate (YEN), the mining sector, cons good, global/regional stock index returns (SZCOMP, IBEX, TOPIX), and commodity prices (BRENT) covered 10.77%. Then, for the next 30 days, foreign investors' SBN holdings, exchange rates (YEN, USD), leading stocks (Kompas), the IHSG in the agribusiness sector, finance, and the MSCI index (MIXD) covered 16.03%.

**Ensemble Model Results**

The results of prediction modeling for the next 1-day (see Table 3) showed that the RF model with SMOTE-ENN treatment produced a G-Mean value of 0.9628 with a max depth of 2 and several decision trees of 42 at a threshold of 0.5393, in which this handling was better than other RF models. ExtraTrees with SMOTE-ENN handling produced the best performance value with a G-Mean of 0.9668, max depth of 4 and number of decision trees of 82 at a threshold of 0.5141. The CatBoost model with SMOTE-Border handling produced the highest G-Mean value of 0.9533 at max depth 4, and the number of decision trees was 710, the threshold of 0.4656. Likewise, with CatBoost, the XGBoost model handled SMOTE-Border with a max depth value of 4 and several decision trees of 86 and, at an optimal threshold of 0.5852, produced the highest G-Mean value of 0.9324. Whereas in LightGBM, SMOTE-ENN produced a good performance as seen from G-Mean 0.9486 with optimal threshold and hyperparameters of 0.5525 with a max depth of 1 and a total of 22 decision trees. In general, the ExtraTrees model with SMOTE-ENN handling was the best algorithm in predicting events in the capital market 1 day ahead with a higher

G-Mean value of around 2% - 5% compared to other models at the same treatment and 1% - 3% compared to other best algorithms. Some models produced low G-Mean (<50%), i.e., without handling, RUS on LightGBM models, ROS on CatBoost and XGBoost models, class weight on CatBoost, XGboost, and LightGBM models.

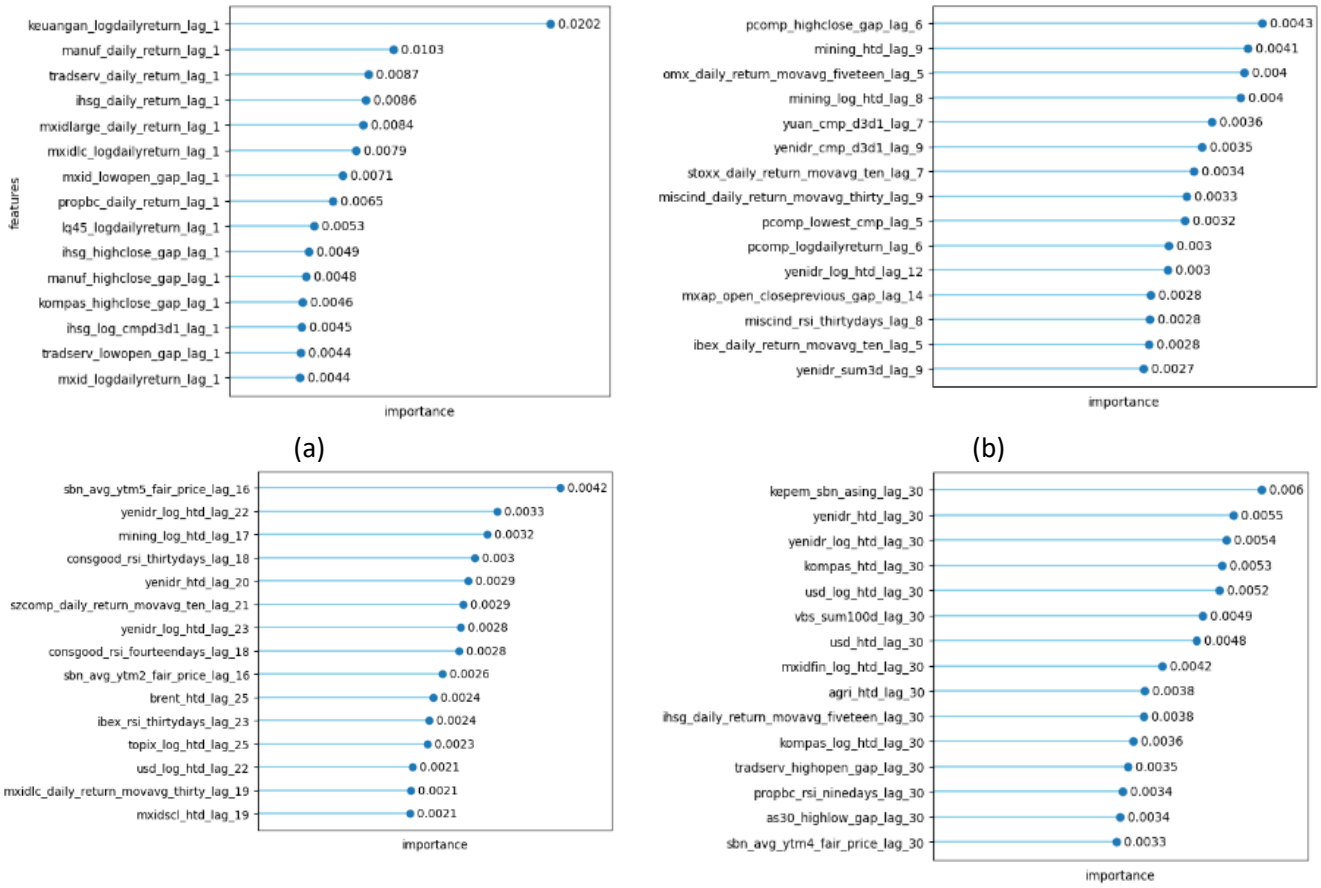



**Figure 4** Feature Selection Results (a) Prediction for the Next 1-Day, (b) Prediction for the Next 5 Days, (c) Prediction for the Next 15 Days, and (d) Prediction for the Next 30 Days

Furthermore, for predicting events for the next 5 days (see Table 3), the RF model with SMOTE-Border handling produced the best value compared to other RF models with a G-Mean value of 0.8529 at max depth 4, and the number of decision trees was 562 at a threshold of 0.6384. The ExtraTrees-SMOTE-Border algorithm produced the highest G-Mean of 0.8614 with a max depth of 4 and 82 decision trees at a threshold of 0.5141. The CatBoost model with SMOTE handling produced the highest G-Mean value of 0.8543 at max depth 4, and the number of decision trees was 541 at a threshold of 0.6955. The XGBoost model with SMOTE-Tomek handling produced the highest G-Mean value of 0.8812 in the optimal hyperparameter combination max depth of 4 and the number of decision trees of 881 at a threshold of 0.6312. The LightGBM model with SMOTE handling

with a max depth of 4 and several decision trees of 991 and at a threshold of 0.7590 produced the highest G-Mean value of 0.8921. Viewed as a whole, for predictions in the next 5 days, LightGBM with SMOTE treatment produced the highest G-Mean value. Compared to other models with SMOTE handling, the LightGBM-SMOTE algorithm showed a higher performance value, with a higher G-Mean value of about 1% - 14% higher than the same treatment and 1% - 4% compared to the other best algorithms. Some treatments produced a low G-Mean (0 - 50%), i.e., without overall handling producing a bad G-Mean value of less than 30%, RUS and class weight less than 40%, and ROS less than 50%.

The RF-SMOTE-ENN algorithm, with an optimal threshold of 0.5453, max depth of 4, and several decision trees of 897, was the best model for predicting events for the next 15 days, with the highest G-Mean of 0.7889 (see Table 3). The ExtraTrees model with SMOTE handling produced the highest G-Mean of 0.7573 with a max depth of 4, the number of decision trees was 60, and the threshold was 0.5285. The CatBoost model with SMOTE-Border handling produced the highest G-Mean value of 0.8949, optimal at a max depth of 4, number of decision trees of 798, and threshold of 0.3413. The XGBoost-SMOTE-Border algorithm revealed the best results with a G-Mean value of 0.8336 obtained after tuning with a threshold of 0.4006, a max depth of 2, and several decision trees of 490. In the LightGBM model, handling SMOTE-ENN produced the best G-Mean value of 0.8245 at the optimal max depth of 4, the optimal number of trees of 895, and the threshold of 0.7426. In general, the CatBoost-SMOTE-Border model was the best for predicting capital market events for the next 15 days. Compared with other SMOTE-Border models, it produced higher G-Mean values of 6% - 28% and 1% - 4% compared to the other best algorithms. Models that did not handle the imbalance class or use other handling methods (weight class, RUS, ROS) produced low G-Mean values (<30%).

Predicting the pressure events in the capital market earlier can help take more effective preventive measures. However, the predicted results for the next 30 days (see Table 3) were not good enough compared to predictions for other shorter periods. The RF model with SMOTE-Tomek handling produced the highest G-Mean value of 0.5968 at the threshold of 0.6007, max depth of 4, and the number of trees 32. The ExtraTrees-ADASYN algorithm produced the highest G-Mean value of 0.4162 at the threshold of 0.5255, max depth of 4, and the number of decision trees of 45. In the CatBoost model, the SMOTE-ENN treatment produced the highest G-Mean value of 0.6468 at a max depth of 4, a number of decision trees of 463, and a threshold of 0.4795. The XGBoost model with SMOTE handling produced the highest G-Mean value of 0.6632 at a max depth of 4, and the number of decision trees was 646. The LightGBM model with SMOTE-Tomek handling at a threshold of 0.6318 with a max depth of 3 and several decision trees of 224 produced the highest G-Mean value of 0.6802. Generally, compared to the SMOTE-Tomek handling model, LightGBM uncovered a higher G-Mean value of about 5% - 48% and 2% - 26% compared to the other best algorithms. Meanwhile, models with imbalanced class weight, RUS, and ROS handling produced G-Mean values below 50%.

**Table 3** Evaluation results of prediction model (in percent)

| Scenario | Model | G-Mean value from Imbalanced Class Handling Results | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | None | RUS | ROS | ADASYN | SMOTE | SMOTE Border | SMOTE Tomek | SMOTE ENN | *Class Weight* |
| 1 day ahead | RF | 0.6628 | 0.9051 | 0.9109 | 0.9078 | 0.891 | 0.9391 | 0.9026 | 0.9628[a] | 0.6712 |
| | ExtraTrees | 0.7672 | 0.8518 | 0.912 | 0.9411 | 0.939 | 0.9551 | 0.9446 | **0.9668[b]** | 0.7871 |
| | CatBoost | 0.7147 | 0.9132 | 0.4171 | 0.9182 | 0.8887 | 0.9533[c] | 0.8846 | 0.9386 | 0.4578 |
| | XGBoost | 0.4252 | 0.7537 | 0.2729 | 0.9018 | 0.8938 | 0.9324[d] | 0.8940 | 0.9158 | 0.2667 |
| | LightGBM | 0.1925 | 0.4514 | 0.7814 | 0.8942 | 0.901 | 0.9445 | 0.9047 | 0.9486[e] | 0.3154 |
| 5 days ahead | RF | 0.2108 | 0.3824 | 0.4026 | 0.8288 | 0.8124 | 0.8529[a] | 0.8121 | 0.8137 | 0.3522 |
| | ExtraTrees | 0.2324 | 0.1939 | 0.2236 | 0.7555 | 0.7563 | 0.8614[b] | 0.7543 | 0.7713 | 0.325 |
| | CatBoost | 0.1096 | 0.389 | 0.0961 | 0.8426 | 0.8543[c] | 0.8314 | 0.8009 | 0.8282 | 0.0914 |
| | XGBoost | 0.2507 | 0.2213 | 0.1237 | 0.8762 | 0.8795 | 0.8677 | 0.8812[d] | 0.8721 | 0 |
| | LightGBM | 0.0625 | 0.1537 | 0 | 0.8456 | **0.8921[e]** | 0.8811 | 0.8872 | 0.8777 | 0 |
| 15 days ahead | RF | 0.0375 | 0.2042 | 0 | 0.7583 | 0.7589 | 0.6155 | 0.767 | 0.7889[a] | 0.2387 |
| | ExtraTrees | 0 | 0.2861 | 0.1832 | 0.7024 | 0.7573[b] | 0.7545 | 0.7429 | 0.7279 | 0.2554 |
| | CatBoost | 0.0437 | 0.1803 | 0.0452 | 0.7929 | 0.822 | **0.8949[c]** | 0.8128 | 0.8234 | 0.0968 |
| | XGBoost | 0.104 | 0.2627 | 0.0453 | 0.8017 | 0.7926 | 0.8336[d] | 0.7979 | 0.8023 | 0 |
| | LightGBM | 0 | 0.2100 | 0 | 0.8034 | 0.8022 | 0.8228 | 0.8025 | 0.8245[e] | 0 |
| 30 days ahead | RF | 0.1462 | 0.1783 | 0 | 0.3929 | 0.4912 | 0.198 | 0.5968[a] | 0.3657 | 0.4564 |
| | ExtraTrees | 0 | 0.205 | 0.1041 | 0.4162[b] | 0.2546 | 0.2406 | 0.2046 | 0.3057 | 0.1227 |
| | CatBoost | 0.0422 | 0.4616 | 0.0388 | 0.4249 | 0.5919 | 0.589 | 0.5823 | 0.6484[c] | 0.0985 |
| | XGBoost | 0.0707 | 0.3512 | 0.0766 | 0.6316 | 0.6632[d] | 0.6237 | 0.6329 | 0.6263 | 0 |
| | LightGBM | 0.0371 | 0.2747 | 0.052 | 0.5215 | 0.6363 | 0.6583 | **0.6802[e]** | 0.648 | 0 |

The bold value indicates the highest evaluation value among all models.
[a] The highest evaluation value for the Random Forest (RF) model
[b] The highest evaluation value for the ExtraTrees model
[c] The highest evaluation value for the CatBoost model
[d] The highest evaluation value for the XGBoost model
[e] The highest evaluation value for the LightGBM model

The best model for every scenario yielded a different handling (see Figure 5). Not only did it have a high geometric mean value, but it also yielded high values in other metrics, such as the F1 score, indicating that it was correctly identifying most of the pressure events in the dataset and making accurate predictions, recall, or the proportion of actual pressure events that were correctly identified as pressure events by the model, and the precision value, denoting that most of the time when the model predicted "pressure," it actually was "pressure," and the AUC-ROC closed. For the 1-day prediction, the ExtraTrees model with SMOTE-ENN handling was the best with all evaluation metrics greater than 95%. For a 5-day prediction, the LightGBM model with SMOTE handling had the highest rate of "pressure" event capture (91.02%) and the strongest precision (92.75%). It also produced the best model in LightGBM, with an average score of about 80% in instances of imbalanced data and similar things (Wang et al., 2022), where the length of historical data used was around 10 years, and cross-validation was utilized to avoid overfitting. However, the hyperparameter tuning procedure in their research employed a grid search, while this study used OPTUNA. For the 15-day prediction, the CatBoost model with SMOTE-Border handling was the best, where all metrics were still satisfactory, with more than 80% for all metrics. Similar to previous studies (Aly et al., 2022), the best class imbalance treatment employed SMOTE-based oversampling, and both feature selection and parameter tuning were carried out to obtain the best model. However, (Aly et al., 2022) study differs in that cross-validation was not employed, while this study used cross-validation expanding window time series. Then, for the 30-day prediction, the LightGBM

model with SMOTE-Tomek handling was the best, with all evaluation metrics greater than 60%. As can be seen, the evaluation's effectiveness degrades as the prediction grows even further. Generally, this study aligns with the results of Indrawati (2021) and (Shrivastava et al., 2020), which found that SMOTE-Based oversampling is an effective method for handling class imbalance.

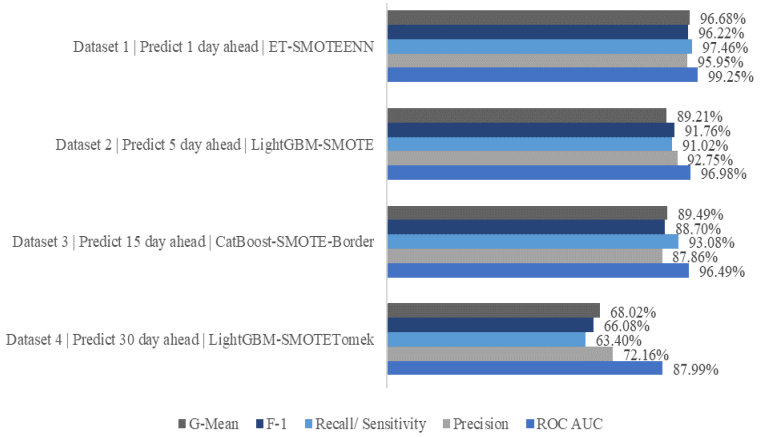



**Figure 5** Summary of the Best Model

**Feature Importance**

Based on the best algorithm ExtraTrees-SMOTE-ENN, it can be seen that the variables at a time lag 1 had the most impact, which can be said that predicting the pressure event 1 day ahead was influenced by the event 1 day before (see Figure 6). The IHSG, industrial sector, trade, manufacturing, and MSCI Indonesia Index impact predicting the 1 day ahead event. The LightGBM-SMOTE algorithm showed that the most influential variables for predicting market events 5 days ahead were in the range of 5-14 days, namely indices and foreign stock exchanges such as the return of China's Yuan 8 days, the Philippine stock exchange 6 days, the Australian stock exchange 6 days, the return of Thailand's index 6 days ago, SBN, and the consumer goods industry sector. Based on the best algorithm CatBoost-SMOTEBorder, which predicts 15 days ahead, the most influential factors were in the range of 16-25 days ago, which include SBN on days 16 and 20, the USD index on day 22, Japan TOPIX on day 25, the XAUD sector which is the gold/silver sector, Brent oil on days 19 and 25, trade, industry, utility, and transportation sectors with significant impact. Based on the best algorithm LightGBM-SMOTETomek, it can be seen that the moving average of foreign net buy/sell shares, the USD index, Shenzen, Yuan, the moving average MSCI, and some sectors such as basic industry, agriculture, and manufacturing, Brent oil, SBN, and the CSI300 and S&P 500 indices 30 days before are important factors in predicting events in the market 30 days ahead.

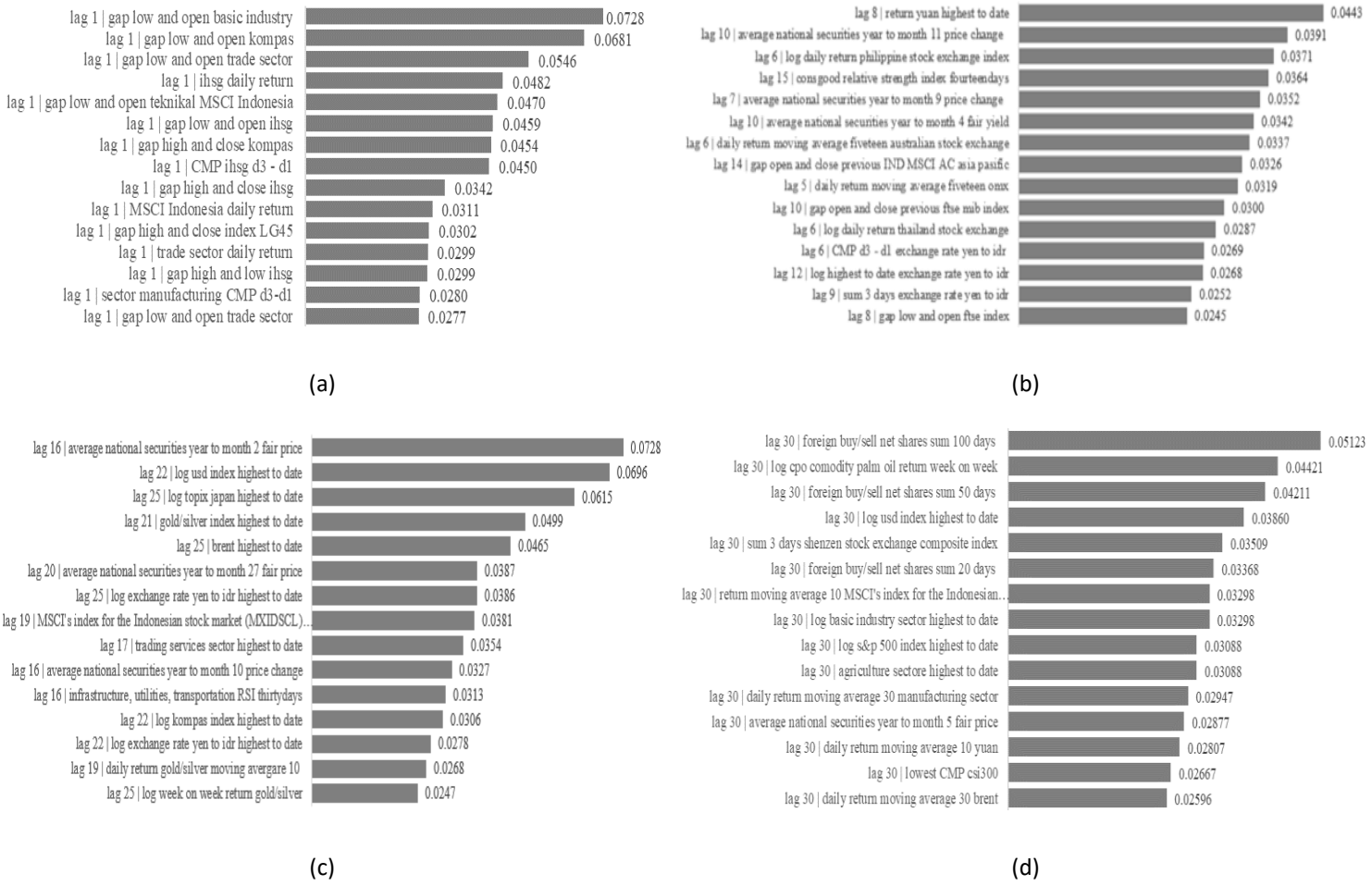



**Figure 6.** Feature Importance Prediction Model (a) 1 Day Ahead; (b) 5 Days Ahead; (c) 15 Days Ahead; and (d) 30 Days Ahead

## Conclusion

This study used ensemble learning modeling with and without handling the imbalance class to detect events in the Indonesian capital market. There are differences in the best algorithm for each scenario. For a 1-day prediction, the ExtraTrees model with SMOTE-ENN handling had the highest G-Mean value of 0.9668. Then, for a 5-day prediction, the LightGBM algorithm with SMOTE handling had a G-Mean value of 0.8921. For a 15-day prediction, the CatBoost algorithm with SMOTE-Border handling had a G-Mean value of 0.8949. Furthermore, for a 30-day prediction, the LightGBM algorithm with SMOTE-Tomek handling had a G-Mean value of 0.6802. In conclusion, the further the prediction, the weaker the model's performance. In this study, effective methods for handling the class imbalance problem in machine learning models were oversampling techniques (SMOTE, SMOTEBorder) and over-under sampling (SMOTE-ENN, SMOTE-Border). On the other hand, random under sampling (RUS), random oversampling (ROS), and class weight

were less effective methods for handling the class imbalance problem in machine learning models in this study.

Based on the study's findings, it is hoped it will be able to contribute knowledge about how to handle imbalances from multiple data points with ensemble learning in scenarios involving early detection of the Indonesian capital market or other similar situations in the application of machine learning, which has previously rarely been studied and can be further explored. It is also expected that this research can be utilized to monitor and detect events in the Indonesian capital market to assist in decision-making because the metrics evaluation findings show excellent performance in detecting events in the capital market. In this study, the ensemble learning model was only utilized for bagging and boosting purposes; there was no comprehensive description of the ensemble learning model's role in predicting how events would affect the capital market. This study also has limitations regarding the newest data. Therefore, for future research, the best existing models can be combined with other ensemble learning methods, such as stacking, and the feature importance of the best model can be explained with more advanced interpretations, such as LIME (Local Interpretable Model Agnostic Explanation).

## References

Aini, Q., Manongga, D., Rahardja, U., Sembiring, I., & Efendy, R. (2023). Innovation and Key Benefits of Business Models in Blockchain Companies. *Blockchain Frontier Technology*, *2*(2), 24-35. https://doi.org/10.34306/bfront.v2i2.161

Alfian, G., Syafrudin, M., Fahrurrozi, I., Fitriyani, N. L., Atmaji, F. T. D., Widodo, T., ... & Rhee, J. (2022). Predicting breast cancer from risk factors using SVM and extra-trees-based feature selection method. *Computers*, *11*(9), 136. https://doi.org/10.3390/computers11090136

Aly, S., Alfonse, M., & Salem, A. B. M. (2022). Intelligent Model for Enhancing the Bankruptcy Prediction with Imbalanced Data Using Oversampling and CatBoost. *International Journal of Intelligent Computing and Information Sciences*, *22*(3), 92-108. https://doi.org/10.21608/ijicis.2022.105654.1138

Asundi, R. V., Prakash, R., & Kumar, K. (n.d.). *Class Weight technique for Handling Class Imbalance*.

Bintoro, B. P. K., Lutfiani, N. and Julianingsih, D. (2023) 'Analysis of the Effect of Service Quality on Company Reputation on Purchase Decisions for Professional Recruitment Services', APTISI Transactions on Management (ATM), 7(1), pp. 35–41. https://doi.org/10.33050/atm.v7i1.1736

Bluwstein, K., Buckmann, M., Joseph, A., Kapadia, S., & Simsek, Ö. (2021). Credit growth, the yield curve and financial crisis prediction: Evidence from a machine learning approach. ECB Working Paper No. 2021/2614. http://dx.doi.org/10.2139/ssrn.3969562

Candra, O., Chammam, A., Rahardja, U., Ramirez-Coronel, A. A., Al-Jaleel, A. A., Al-Kharsan, I. H., ... & Rezai, M. M. (2023). Optimal Participation of the Renewable Energy in Microgrids with Load Management Strategy. *Environmental and Climate Technologies*, *27*(1), 56-66. https://doi.org/10.2478/rtuect-2023-0005

Carmona, P., Climent, F., & Momparler, A. (2019). Predicting failure in the US banking sector: An extreme gradient boosting approach. *International Review of Economics & Finance*, *61*, 304-323. https://doi.org/10.1016/j.iref.2018.03.008

Chen, R. C., Dewi, C., Huang, S. W., & Caraka, R. E. (2020). Selecting critical features for data classification based on machine learning methods. *Journal of Big Data*, *7*(1), 52. https://doi.org/10.1186/s40537-020-00327-4

Coffinet, J., & Kien, J. N. (2019). Detection of rare events: A machine learning toolkit with an application to banking crises. *The Journal of Finance and Data Science*, *5*(4), 183-207. https://doi.org/10.1016/j.jfds.2020.04.001

Faris, H., Abukhurma, R., Almanaseer, W., Saadeh, M., Mora, A. M., Castillo, P. A., & Aljarah, I. (2020). Improving financial bankruptcy prediction in a highly imbalanced class distribution using oversampling and ensemble learning: a case from the Spanish market. *Progress in Artificial Intelligence*, *9*, 31-53. https://doi.org/10.1007/s13748-019-00197-9

Gnip, P., & Drotár, P. (2019, September). Ensemble methods for strongly imbalanced data: bankruptcy prediction. In *2019 IEEE 17th International Symposium on Intelligent Systems and Informatics (SISY)*, 155-160. IEEE.

Google Developers. (2021). *Machine Learning.* https://developers.google.com/machine-learning/data-prep/construct/sampling-splitting/imbalanced-data

Hariguna, T., Rahardja, U., & Sarmini. (2022). The Role of E-Government Ambidexterity as the Impact of Current Technology and Public Value: An Empirical Study. *Informatics*, *9*(3), 67. https://doi.org/10.3390/informatics9030067

Hermawan, A., Sunaryo, W., & Hardhienata, S. (2023). Optimal Solution for OCB Improvement Through Strengthening of Servant Leadership, Creativity, and Empowerment. *Aptisi Transactions on Technopreneurship (ATT)*, *5*(1Sp), 11-21. https://doi.org/10.34306/att.v5i1Sp.307

Indrawati, A. (2021) 'Penerapan Teknik Kombinasi Oversampling Dan Undersampling Untuk Mengatasi Permasalahan Imbalanced Dataset', JIKO (Jurnal Informatika dan Komputer), 4(1), pp. 38–43. https://doi.org/10.33387/jiko.v4i1.2561

Islam, S. R., Eberle, W., Ghafoor, S. K., Bundy, S. C., Talbert, D. A., & Siraj, A. (2019). Investigating bankruptcy prediction models in the presence of extreme class imbalance and multiple stages of economy. *arXiv preprint arXiv:1911.09858*.

Jabeur, S. B., Gharib, C., Mefteh-Wali, S., & Arfi, W. B. (2021). CatBoost model and artificial intelligence techniques for corporate failure prediction. *Technological Forecasting and Social Change*, *166*, 120658. https://doi.org/10.1016/j.techfore.2021.120658

Junyu, H. (2020, August). Prediction of Financial Crisis Based on Machine Learning. In *2020 The 4th International Conference on Business and Information Management* , 71-75. https://doi.org/10.1145/3418653.3418674

Kosasi, S., Yuliani, I. D. A. E., & Rahardja, U. (2022, February). Boosting e-service quality of online product businesses through it leadership. In *2022 International Conference on Science and Technology (ICOSTECH)*, 1-10. IEEE. 10.1109/ICOSTECH54296.2022.9829036

Liu, Q., Wang, C., Zhang, P., & Zheng, K. (2021). Detecting stock market manipulation via machine learning: evidence from China Securities Regulatory Commission punishment cases. *International Review of Financial Analysis*, *78*, 101887. https://doi.org/10.1016/j.irfa.2021.101887

Lu, S., Liu, C. and Chen, Z. (2021) 'Predicting stock market crisis via market indicators and mixed frequency investor sentiments', Expert Systems with Applications. Elsevier, 186, p. 115844. https://doi.org/10.1016/j.eswa.2021.115844

Lutfiani, N., Wijono, S., Rahardja, U., Iriani, A., Aini, Q., & Septian, R. A. D. (2023). A Bibliometric Study: Recommendation based on Artificial Intelligence for iLearning Education. *Aptisi Transactions on Technopreneurship (ATT)*, *5*(2), 112-119. https://doi.org/10.34306/att.v5i2.279

Mahardika, R., & Irawan, F. (2022). The Impact Of Thin Capitalization Rules On Tax Avoidance In Indonesia. *JURNAL PAJAK INDONESIA (Indonesian Tax Review)*, *6*(2S), 651-662. https://doi.org/10.31092/jpi.v6i2S.1972

Marlina, E., Putri, A. A. and Suriyanti, L. H. (2023) 'Determinants of strategic management accounting implementation in Higher Education Institutions (HEIs) in Indonesia', Journal of Accounting and Investment, 24(2), pp. 306–322. https://doi.org/10.18196/jai.v24i2.16562

Mishraz, N., Ashok, S., & Tandon, D. (2021). Predicting Financial Distress in the Indian Banking Sector: A Comparative Study Between the Logistic Regression, LDA and ANN Models. *Global Business Review*, 09721509211026785.. https://doi.org/10.1177/09721509211026785

Petropoulos, A., Siakoulis, V., Stavroulakis, E., & Vlachogiannakis, N. E. (2020). Predicting bank insolvencies using machine learning techniques. *International Journal of Forecasting*, *36*(3), 1092-1113. https://doi.org/10.1016/j.ijforecast.2019.11.005

Pramono, . E. S. ., Rudianto, D. ., Siboro, F. ., Abdul Baqi , M. P. ., & Julianingsih, D. (2022). Analysis Investor Index Indonesia with Capital Asset Pricing Model (CAPM). *Aptisi Transactions on Technopreneurship (ATT)*, *4*(1), 35–46. https://doi.org/10.34306/att.v4i1.218

Pratama, A., & Wijaya, A. (2023). Implementasi Sistem Good Corporate Governance Pada Perangkat Lunak Berbasis Website PT. Pusaka Bumi Transportasi. *Technomedia Journal*, *7*(3), 340-353. https://doi.org/10.33050/tmj.v7i3.1917

Putri, H. R. and Dhini, A. (2019) 'Prediction of financial distress: Analyzing the industry performance in stock exchange market using data mining', in 2019 16th International Conference on Service Systems and Service Management (ICSSSM). IEEE, pp. 1–5. https://doi.org/10.1109/ICSSSM.2019.8887824

Putri, R. L., Hidayat, S., Wahyono, E., & Rahmawati, L. (2023). Big Data and Strengthening MSMEs After the Covid-19 Pandemic (Development Studies on Batik MSMEs in East Java). *IAIC Transactions on Sustainable Digital Innovation (ITSDI)*, *4*(2), 83-100. https://doi.org/10.34306/itsdi.v4i2.574

Qiu, Y., Zhou, J., Khandelwal, M., Yang, H., Yang, P., & Li, C. (2021). Performance evaluation of hybrid WOA-XGBoost, GWO-XGBoost and BO-XGBoost models to predict blast-induced ground vibration. *Engineering with Computers*, 1-18. https://doi.org/10.1007/s00366-021-01393-9

Rahardja, U. et al. (2023) 'Implementation of Tensor Flow in Air Quality Monitoring Based on Artificial Intelligence', International Journal of Artificial Intelligence Research, 6(1).

Santoso, R. E., Prawiyogi, A. G., Rahardja, U., Oganda, F. P., & Khofifah, N. (2022). Penggunaan dan Manfaat Big Data dalam Konten Digital. *ADI Bisnis Digital Interdisiplin Jurnal*, *3*(2), 88-91. https://doi.org/10.34306/abdi.v3i2.836

Shrivastava, S., Jeyanthi, P. M., & Singh, S. (2020). Failure prediction of Indian Banks using SMOTE, Lasso regression, bagging and boosting. *Cogent Economics & Finance*, *8*(1), 1729569.. https://doi.org/10.1080/23322039.2020.1729569

Sipahutar, R. J. et al. (2020) 'Drivers and Barriers to IT Service Management Adoption in Indonesian Start-up Based on the Diffusion of Innovation Theory', in 2020 Fifth International Conference on Informatics and Computing (ICIC). IEEE, pp. 1–8. 10.1109/ICIC50835.2020.9288556

Sir, Y. A. and Soepranoto, A. H. H. (2022) 'Pendekatan Resampling Data Untuk Menangani Masalah Ketidakseimbangan Kelas', J-ICON: Jurnal Komputer dan Informatika, 10(1), pp. 31–38. https://doi.org/10.35508/jicon.v10i1.6554

Soesilo, T. H., & Tinggi, M. M. P. (2021). *Analisis pengembangan sistem informasi gaji pegawai (sigap) menggunakan soft system methodology (Studi pada Biro Keuangan Universitas Brawijaya)*. Universitas Brawijaya.

Speiser, J. L., Miller, M. E., Tooze, J., & Ip, E. (2019). A comparison of random forest variable selection methods for classification prediction modeling. *Expert systems with applications*, *134*, 93-101. https://doi.org/10.1016/j.eswa.2019.05.028

Srinivas, P., & Katarya, R. (2022). hyOPTXg: OPTUNA hyper-parameter optimization framework for predicting cardiovascular disease using XGBoost. *Biomedical Signal Processing and Control*, *73*, 103456. https://doi.org/10.1016/j.bspc.2021.103456

Sun, X., Liu, M., & Sima, Z. (2020). A novel cryptocurrency price trend forecasting model based on LightGBM. *Finance Research Letters*, *32*, 101084. https://doi.org/10.1016/j.frl.2018.12.032

Thakkar, A., & Chaudhari, K. (2021). Fusion in stock market prediction: a decade survey on the necessity, recent developments, and potential future directions. *Information Fusion*, *65*, 95-107. https://doi.org/10.1016/j.inffus.2020.08.019

Tölö, E. (2020). Predicting systemic financial crises with recurrent neural networks. *Journal of Financial Stability*, *49*, 100746. https://doi.org/10.1016/j.jfs.2020.100746

Tussa'diah, H., & Kartika, N. Y. (2023). Critical Discourse Analysis on Linguistic Ideology of The Netizens Comments. *ADI Journal on Recent Innovation*, *4*(2), 110-121. https://doi.org/10.34306/ajri.v4i2.838

Vien, B. S., Wong, L., Kuen, T., Rose, L. F., & Chiu, W. K. (2021). A Machine Learning Approach for Anaerobic Reactor Performance Prediction Using Long Short-Term Memory Recurrent Neural Network. *Struct. Health Monit. 8apwshm*, *18*, 61.

Wang, D. N., Li, L., & Zhao, D. (2022). Corporate finance risk prediction based on LightGBM. *Information Sciences*, *602*, 259-268. https://doi.org/10.1016/j.ins.2022.04.058

Wang, H., & Liu, X. (2021). Undersampling bankruptcy prediction: Taiwan bankruptcy data. *Plos one*, *16*(7), e0254030. https://doi.org/10.1371/journal.pone.0254030

Widiastuti, T., Karsa, K., & Juliane, C. (2023). Evaluasi Tingkat Kepuasan Mahasiswa Terhadap Pelayanan Akademik Menggunakan Metode Klasifikasi Algoritma C4. 5. *Technomedia Journal*, *7*(3), 364-380. https://doi.org/10.33050/tmj.v7i3.1932

Zanubiya, J., Meria, L., & Juliansah, M. A. D. (2023). Increasing Consumers with Satisfaction Application based Digital Marketing Strategies. *Startupreneur Bisnis Digital (SABDA Journal)*, *2*(1), 12-21.

Zhang, Z. and Chen, Y. (2022) 'Tail risk early warning system for capital markets based on machine learning algorithms', Computational Economics. Springer, 60(3), pp. 901–923. https://doi.org/10.1007/s10614-021-10171-0