TIN: Terapan Informatika Nusantara
Vol 6, No 11, April 2026, page 2106-2117
ISSN 2722-7987 (Media Online)
Website https://ejurnal.seminar-id.com/index.php/tin
DOI 10.47065/tin.v6i11.9589

SMOTE-Based Oversampling for Imbalanced Digital Fraud Risk Classification

Ika Nur Laily Fitriana[1,*], Fonda Leviany[2], Kurnia Sari Kasmiarno[3], Mohammad Okky Mabruri[4]

[1] Faculty of Science and Technology, Statistics Study Program, Universitas Terbuka, Tangerang Selatan, Indonesia
[2] Faculty of Science and Technology, Data Science Study Program, Universitas Terbuka, Tangerang Selatan, Indonesia
[3] Faculty of Economics and Business, Islamic Economics Study Program, Universitas Terbuka, Tangerang Selatan, Indonesia
[4] Faculty of Industrial Technology and System Engineering, Department of Engineering Physics, Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia
Email: [1,*]ika.nur@ecampus.ut.ac.id, [2]fonda.leviany@ecampus.ut.ac.id, [3]kurnia.sari@ecampus.ut.ac.id, [4]okkymabrur@gmail.com
Correspondence Author Email: ika.nur@ecampus.ut.ac.id

Abstract–Digital fraud risk among university students is an important issue, yet classification using survey-based indicators is complicated by class imbalance. This study examined whether Synthetic Minority Over Sampling Technique (SMOTE) improves Digital Fraud Risk classification among Universitas Terbuka students. This research used primary survey data from 498 respondents and modeled using five predictors representing financial literacy, digital financial literacy, monthly gross income, age, and job tenure. The evaluated models were Gaussian Naive Bayes, Random Forest, calibrated linear Support Vector Machine (SVM), Radial Basis Function SVM, and XGBoost. The performance of model was evaluated using confusion matrix, accuracy, balanced accuracy, precision, recall, F1 score, ROC-AUC, PR-AUC, MCC and Kappa. This research revealed that without oversampling, some models showed higher nominal accuracy but zero recall for High risk. It means that accuracy is insufficient for model selection under imbalance. In contrast, SMOTE increased recall for the High risk class across all models and improved PR AUC in several cases. The SMOTE based Random Forest achieved the highest test PR AUC (0.415), whereas the SMOTE based RBF SVM achieved the highest recall (0.659). Diagnostic analyses for the selected SMOTE based Random Forest provided evidence of non-random predictive signal, although overall discriminative performance remained moderate.

Keywords: Digital Fraud Risk; Imbalanced Classification; SMOTE; Survey-Based Prediction; Machine Learning

1. INTRODUCTION

The rapid diffusion of digital financial services has expanded access to payments, savings, and credit, but it has also increased exposure to fraudulent schemes conducted through online channels. Recent reviews show that machine-learning applications in fraud detection have grown rapidly, particularly in transaction level settings such as credit card fraud detection and financial anomaly detection (Ali et al., 2022; Sulaiman et al., 2022). However, fraud vulnerability is not determined solely by transactional patterns. Individual characteristics such as financial literacy, digital financial literacy, risk perception, and socioeconomic conditions may also shape exposure to fraudulent offers and the ability to recognize or avoid them.

This broader perspective is particularly relevant for student populations. University students are active users of digital financial services, yet they may differ substantially in financial knowledge, digital capability, age, income, and employment experience. In this regard, Universitas Terbuka (UT) students provide a meaningful case study. As an open and distance-learning university, UT operates through digital academic and administrative systems, so its students are routinely embedded in online environments and are likely to engage with digital payments, mobile banking, online marketplaces, and app based financial services in everyday life. At the same time, UT students are socially heterogeneous, spanning a wide range of ages, employment situations, and income levels. This combination makes UT students a relevant population for examining how survey based indicators of financial capability relate to digital fraud risk.

Prior studies support the importance of this perspective. Digital financial literacy has been shown to be associated with financial well-being and with the ability to protect against digital fraud (Choung et al., 2023). Other studies indicate that financial literacy and overconfidence can influence fraud victimization risk (Xiao et al., 2022). In addition, classical machine-learning methods have already been used to model fraud victimization in survey-based contexts (Lokanan & Liu, 2021). Nevertheless, these studies do not establish whether a compact survey instrument can reliably classify digital fraud risk in the specific context of UT students.

From a methodological standpoint, the problem also belongs to the literature on learning from imbalanced data. When the minority class is substantively important, conventional model training often becomes biased toward the majority class and may fail to detect rare but meaningful cases (Carvalho et al., 2025; Gao et al., 2026). In this context, SMOTE remains widely used because it creates synthetic minority observations and can help classifiers learn decision boundaries that are less dominated by the majority class. Recent studies also report that SMOTE can improve minority-class detection performance in applied imbalanced-learning settings, especially when the main objective is to improve identification of rare cases rather than to maximize nominal accuracy alone (Khalid et al., 2024; Sayegh et al., 2024).

However, the advantage of SMOTE is still context-dependent, and its effectiveness has not been established for compact survey data on digital fraud risk among Universitas Terbuka students. Therefore, the purpose of this study was to compare the performance of SMOTE and non-SMOTE approaches on survey data related to the digital fraud risk of UT students. Using the same predictors, train test split, tuning procedure, and classification models. The classification model used were Random Forest, GNB, support vector machines with radial and calibrated linear decision functions,

TIN: Terapan Informatika Nusantara
Vol 6, No 11, April 2026, page 2106-2117
ISSN 2722-7987 (Media Online)
Website https://ejurnal.seminar-id.com/index.php/tin
DOI 10.47065/tin.v6i11.9589

and XGBoost. This study examined whether SMOTE provides a more effective strategy than non-SMOTE training for detecting students in the high risk category.

2. RESEARCH METHODOLOGY

2.1 Basic Research Framework

This study used a quantitative cross-sectional design using primary data obtained from survey data collected from Universitas Terbuka students. The analytic sample consisted of 498 respondents. The research location was Universitas Terbuka, and the study focused on whether survey-based indicators of financial capability could classify digital fraud risk. The dependent variable was Digital Fraud Risk, operationalized as a binary outcome with two categories: Low and High. The independent variables captured key dimensions of financial capability and respondent characteristics, including financial literacy, digital financial literacy, monthly gross income, age, and job tenure. The observed class distribution consisted of 352 respondents in the Low-risk class and 146 respondents in the High-risk class. This distribution indicates a class imbalance that can potentially bias supervised learning models toward the majority class and reduce sensitivity to the minority class. Therefore, the study compared baseline model training on the original data with a SMOTE-based oversampling strategy to assess whether resampling improves classification performance for the minority class. Table 1 summarizes the variables used in the analysis.

Table 1. Research Variables

Variables	Type of Data
Digital Fraud Risk	Categoric : Low, High
Financial Literacy Score	Numeric
Digital Financial Literacy Score	Numeric
Monthly Gross Income (Million IDR)	Numeric
Age (Years)	Numeric
Job Tenure (Months)	Numeric

2.2 Research Stages

The stages of this research were organized in Figure 1.

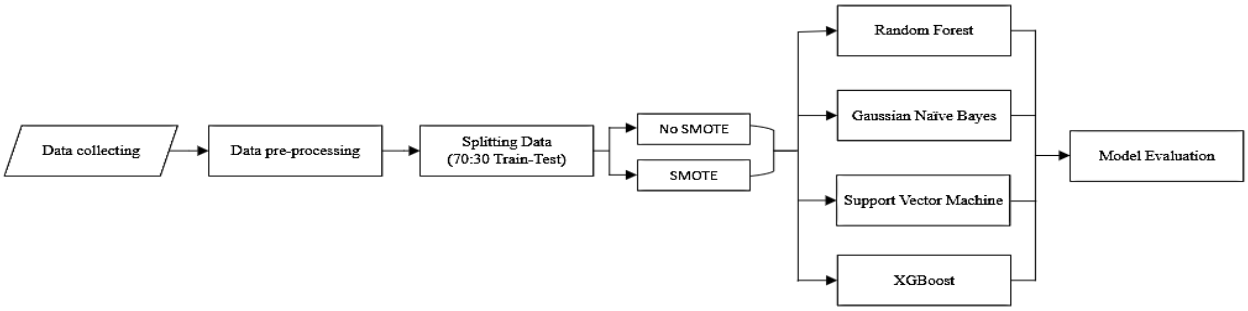


Figure 1. Research Flowchart

This study follows a systematic machine learning workflow consisting of several stages. The procedure began with data collection, in which primary survey data were obtained from Universitas Terbuka students who were employed and or reported earning income. The raw survey responses were then subjected to data preprocessing to ensure data quality and analytical readiness, including screening for completeness and consistency, correcting invalid entries, and preparing variables for modeling.

Following preprocessing, the dataset was partitioned using a stratified 70:30 train test split to preserve the original class proportions in both subsets. Two parallel training strategies were subsequently implemented to enable a direct and controlled assessment of oversampling. In the first strategy, models were trained on the original training data without resampling. In the second strategy, SMOTE was applied exclusively to the training set to mitigate class imbalance through the generation of synthetic minority class observations. The test set was retained in its original form under both strategies to prevent information leakage and to ensure that performance estimates reflect the naturally imbalanced evaluation condition.

Model development was conducted using machine learning algorithms, namely Random Forest, GNB, RBF SVM, Calibrated Linear SVM, and XGBoost, with hyperparameter tuning performed during training. Finally, all trained models were assessed on the testing set using standard classification metrics and confusion matrix analysis. This pipeline supports a transparent comparison between no SMOTE and SMOTE based training and ensures that model performance is evaluated in a systematic and reproducible manner consistent with best practice in imbalanced classification.

TIN: Terapan Informatika Nusantara
Vol 6, No 11, April 2026, page 2106-2117
ISSN 2722-7987 (Media Online)
Website https://ejurnal.seminar-id.com/index.php/tin
DOI 10.47065/tin.v6i11.9589

2.3 Machine Learning

Recent studies on digital fraud risk detection increasingly rely on supervised machine learning algorithms that can model complex and nonlinear relationships in high-dimensional with good classification performance (Leviany, Kasmiarno, and Fitriana., 2025). Machine learning models such as Random Forest, GNB, Support Vector Machine, and XGBoost are often benchmarked together, with performance evaluated using metrics such as accuracy, precision, recall, and F1 Score.

2.3.1 Random Forest

Random Forest (RF) is an ensemble learning method that constructs multiple decision trees on bootstrapped samples and aggregates their outputs, typically via majority voting for classification tasks (Bhaduri et al., 2025; Breiman, 2001; Pantic et al., 2025). RF reduces the correlation among decision trees by randomly selecting both samples and features. First, it draws a sample of data from the original training set, and then it randomly selects a subset of features to build each decision tree. This double randomization lowers the dependence between trees, helps prevent overfitting, and improves the model's accuracy (Salman, Kalakech, & Steiti 2024). The formula for predicting the input data using RF is as follows.

$$\hat{y} = majority\_vote(h_1(x), h_2(x), \dots, h_k(x)) \quad (1)$$

The formula shows that the final prediction $\hat{y}$ in RF is obtained by majority vote among the individual tree predictions $h_1(x), h_2(x), \dots, h_k(x)$. In other words, the class chosen by most trees becomes the model's output for input.

2.3.2 Gaussian Naïve Bayes

Gaussian Naïve Bayes (GNB) is a probabilistic classifier based on Bayes' theorem that assumes conditional independence among features and models continuous attributes with Gaussian distributions. Despite its simplifying independence assumption, GNB is attractive in fraud-related applications because of its fast training and prediction time, which makes it suitable as a baseline or as part of larger fraud detection pipelines. The formula for Gaussian Naïve Bayes is as follows (Saputra, 'Alauddin, & Azizan 2025):

$$P(x_i|y) = \frac{1}{\sqrt{2\pi\sigma^2 y}} exp\left(-\frac{(x_i-\mu_y)^2}{2\sigma_y^2}\right) \quad (2)$$

This formula calculates the probability $P(x_i \mid y)$ that feature x_i belongs to class y, assuming a normal distribution with mean μ_y and variance σ_y^2. Specifically, the probability density peaks at the mean and decays exponentially with the squared distance from it, normalized by the Gaussian constant.

2.3.3 Support Vector Machine (SVM)

SVM classifier lies in its ability to identify an optimized decision boundary representing the largest separation (maximum margin) between classes (Guido, Ferrisi, Lofaro, & Conforti, 2024). The goal of SVM is to find a good hyperplane in the input space to separate the two classes. The hyperplane is found to measure the margin between the two classes by finding their maximum point. Based on the kernel used, the SVM is divided into two types: linear SVM and non-linear SVM. In this paper, non-linear SVM using Radial Basis Function (RBF) kernel used to classify data that cannot be separated linearly. The RBF feature space has an infinite number of dimensions determined by its parameters. Therefore, when iterated, it can produce a unique linear solution, making it the best classification process. The RBF kernel equation is below:

$$\mathrm{K}(x, z) = exp[-\gamma \| x - z \|^2] \quad (3)$$

This formula defines the RBF kenel $\mathrm{K}(x, z)$, commonly used in Support Vector Machines. It computes the similarity between input vectors $\boldsymbol{x}$ and $\boldsymbol{z}$, where $\gamma > 0$ controls the kernel width and $\| x - z \|^2$ is the squared Euclidean distance. Higher similarity (closer to 1) occurs when $\boldsymbol{x}$ and $\boldsymbol{z}$ are nearby in feature space, enabling SVMs to capture nonlinear decision boundaries.

2.3.4 XGBoost

XGBoost is built on decision trees, a widely used supervised learning approach introduced by Quinlan (1986) for classification and regression tasks. XGBoost builds models iteratively, with each new model aiming to improve the prediction errors of the previous model. The loss function optimized in boosting is as shown in equation:

$$\mathcal{L}^{(t)} = \sum_{i=1}^{n} l\left(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)\right) + \Omega(f_t) \quad (4)$$

This formulation describes the optimization goal at iteration t. The first term is the summation of the loss function $l(\cdot)$ over all n training samples, which measures the discrepancy between the true target value y_i and the updated prediction obtained by adding the new model $f_t(x_i)$ to the previous prediction $\hat{y}_i^{(t-1)}$. The second term, $\Omega(f_t)$,

TIN: Terapan Informatika Nusantara
Vol 6, No 11, April 2026, page 2106-2117
ISSN 2722-7987 (Media Online)
Website https://ejurnal.seminar-id.com/index.php/tin
DOI 10.47065/tin.v6i11.9589

is a regularization component that penalizes the complexity of the newly added function f_t. By minimizing this objective, the learning process aims to improve predictive accuracy while controlling model complexity to prevent overfitting.

2.4 SMOTE

In many research datasets, the class labels are often imbalanced, meaning that the proportion of samples in each class is not evenly distributed. Therefore, an appropriate solution is required to address the class imbalance problem using a data-level approach. This study focuses on handling imbalanced data through data-level techniques, which modify the dataset by rebalancing the minority and majority classes (Hairani, Widiyaningtyas, & Prasetya, 2024). We propose the techniques for addressing class imbalance using oversampling approach named the Synthetic Minority Over-sampling Technique (SMOTE). SMOTE generates new synthetic data by performing linear interpolation between minority class samples based on their k-nearest neighbors (Elreedy, Atiya, & Kamalov, 2024). Previous studies have demonstrated the use of average SMOTE has been shown to improve the performance of predictive models (Kivrak et al., 2024; Malhotra & Lata, 2022; Wibowo & Fatichah, 2021).

3. RESULTS AND DISCUSSION

3.1 Data Exploration

Before proceeding to further analysis, the descriptive statistics of all predictor variables are presented to provide an initial overview of the characteristics of the data. Table 2 shows that the respondents were generally characterized by relatively high financial literacy, with a mean score of 5.920 and a median of 6.000, while digital financial literacy was moderate to high, with a mean of 3.884 and a median of 4.000. The average age was 24.317 years, with a median of 23 years, indicating that the sample was concentrated in early adulthood, which is consistent with the profile of university students. Monthly gross income and job tenure displayed much wider dispersion than the literacy variables, as reflected in their large standard deviations and broad ranges. This pattern suggests substantial socioeconomic heterogeneity within the student sample. In analytical terms, the descriptive statistics indicate that the dataset combines moderate class imbalance with meaningful variation in socioeconomic and literacy-related predictors, which supports the use of comparative modeling to assess whether oversampling improves detection of the minority high-risk group.

Table 2. Descriptive statistics of predictor variables

Variables	Mean	SD	Median	Min	Max
Financial Literacy Score	5.920	1.026	6.000	1.000	7.000
Digital Financial Literacy Score	3.884	0.992	4.000	0.000	5.000
Monthly Gross Income (Million IDR)	3,850,032.765	3,529,863.319	3,000,000.000	200,000.000	30,000,000.000
Age (Years)	24.317	4.825	23.000	18.000	43.000
Job Tenure (Months)	34.171	40.087	24.000	1.000	288.000

The survey data collected from Universitas Terbuka (UT) students who are employed or earn income show that the low Digital Fraud Risk class substantially dominates the other class, indicating a clear majority–minority class distribution. Specifically, 70.68% of respondents (n = 352) are classified as low risk, while 29.32% (n = 146) are classified as high risk, as shown in Figure 2. The majority class (Low) is roughly 2.4× the minority class (High). This distribution demonstrates a marked class imbalance in the target variable. Without explicit mitigation, some models can achieve reasonable accuracy while producing poor recall for the High class, limiting their value for risk screening. Such an imbalance can bias machine-learning models toward the majority class and degrade the detection performance for the minority (high-risk) class; therefore, data balancing techniques are required prior to model development to improve minority-class learnability and achieve more reliable classification outcomes.

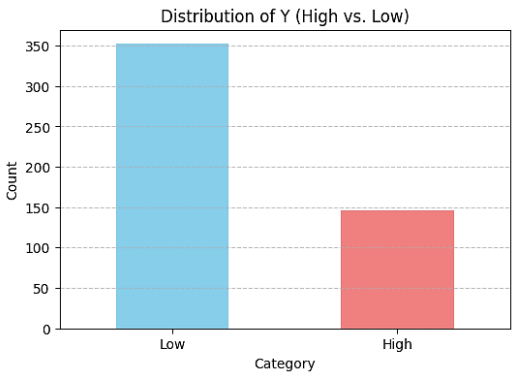


Figure 2. Distribution of Category in Digital Fraud Risk

TIN: Terapan Informatika Nusantara
Vol 6, No 11, April 2026, page 2106-2117
ISSN 2722-7987 (Media Online)
Website https://ejurnal.seminar-id.com/index.php/tin
DOI 10.47065/tin.v6i11.9589

After splitting the dataset into training and testing sets using a stratified 70:30 ratio to preserve the original class proportions, the training set comprised 246 observations in the Low Digital Fraud Risk class and 102 observations in the High Digital Fraud Risk class. As shown in Figure 3, applying SMOTE effectively balanced the training data by generating synthetic samples for the minority class, resulting in an equal number of observations in each class (246 per class). This balancing step enables the model to learn Digital Fraud Risk patterns more effectively and mitigates bias toward the majority class. Notably, SMOTE was applied only to the training set, while the testing set retained its original class distribution. This approach ensures that model evaluation remains objective and reflects the true imbalance present in real-world data.

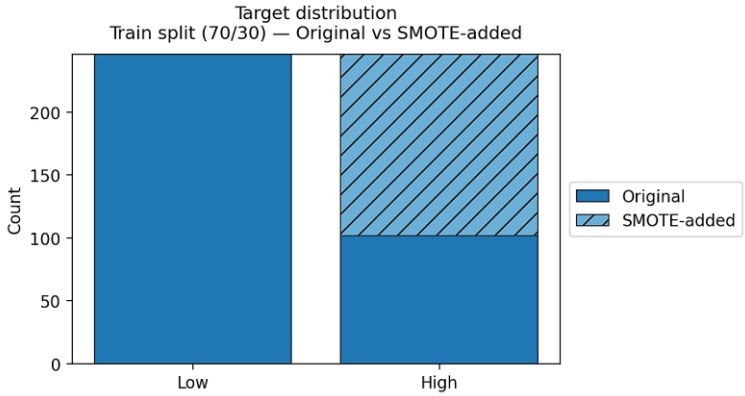


Figure 3. Target Distribution Original vs SMOTE in Training Data

3.2 Classification Result Using Gaussian Naïve Bayes (GNB)

The confusion matrix provides an interpretable summary of how each classifier performed on the testing data. Figure 4 provides the confusion matrix result using GNB. GNB correctly classified 97 Low cases and misclassified 9 Low cases as High. For the High class, it correctly identified only 5 cases while misclassifying 39 as Low. This error profile indicates limited sensitivity to High-risk observations and substantial under detection of the minority class. After SMOTE, the model identified 21 High cases correctly and reduced High class false negatives to 23. However, false positives increased, with 31 Low digital fraud risk misclassified as High, and true negatives decreasing to 75. In substantive terms, SMOTE shifted GNB from a majority-oriented classifier toward a more risk sensitive classifier that better captures High risk observations, although it increased the number of Low digital fraud risk flagged as High.

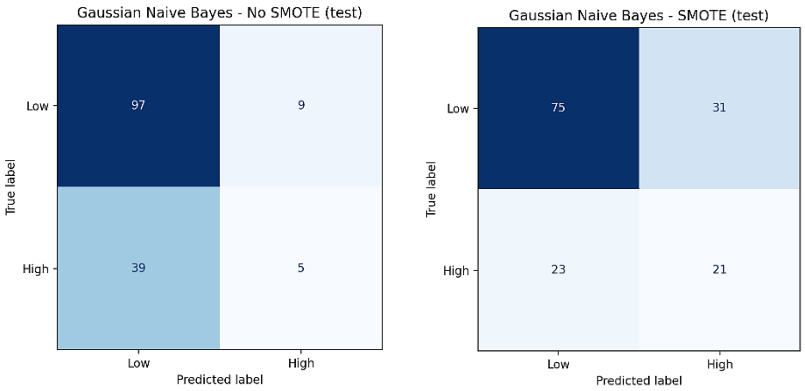


Figure 4. Confusion Matrix of GNB No SMOTE vs SMOTE

3.2 Classification Result Using Random Forest

In the no SMOTE strategy in Figure 5, Random Forest correctly classified 99 Low digital fraud risk and misclassified 7 Low digital fraud risk as High. For the High class, it correctly identified only 3 digital fraud risks, while 41 High were misclassified as Low. This pattern demonstrates that the model largely learned the majority class structure and rarely produced positive predictions for the minority class. In the SMOTE condition, the Random Forest confusion matrix is consistent with a marked improvement in High class detection. Based on the test set totals and the performance metrics reported previously, the SMOTE based Random Forest corresponds to 16 correctly identified High digital fraud risk and 28 High digital fraud risk misclassified as Low, while classifying 84 Low digital fraud risk correctly and misclassifying 22 Low digital fraud risk as High. This profile indicates a substantially lower false negative burden than the no SMOTE model, with a moderate increase in false positives. Relative to the other SMOTE based models, Random Forest provides a more balanced allocation of errors between sensitivity and specificity, which aligns with its favorable precision recall performance.

TIN: Terapan Informatika Nusantara
Vol 6, No 11, April 2026, page 2106-2117
ISSN 2722-7987 (Media Online)
Website https://ejurnal.seminar-id.com/index.php/tin
DOI 10.47065/tin.v6i11.9589

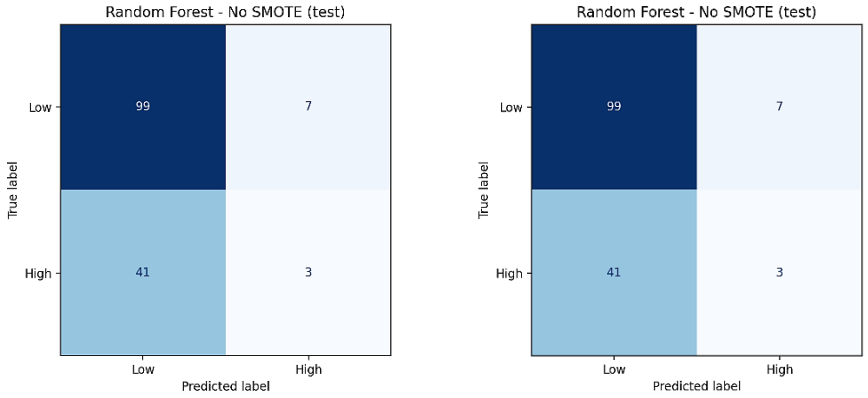


Figure 5. Confusion Matrix of Random Forest No SMOTE vs SMOTE

3.3 Classification Result Using SVM

The confusion matrices in Figure 6 reveal a pronounced imbalance effect for both SVM variants in the no SMOTE condition. The calibrated linear SVM predicted only the Low Digital Fraud Risk class, yielding 106 true negatives and 44 false negatives, with zero true positives for High Digital Fraud Risk. The RBF SVM exhibited the same behavior. These outcomes illustrate that, under severe imbalance and a default decision threshold, margin-based classifiers can collapse into a majority class prediction regime, producing high nominal accuracy while failing entirely on minority detection.

After SMOTE, both SVM variants recovered meaningful High Digital Fraud Risk detection. The calibrated linear SVM correctly identified 23 High Digital Fraud Risk observations and misclassified 21 High Digital Fraud Risk observations as Low, while producing 36 false positives and 70 true negatives. The RBF SVM produced the highest sensitivity among the shown matrices, correctly identifying 29 High Digital Fraud Risk observations and misclassifying 15 as Low. This improved sensitivity came with the largest false positive burden, with 47 Low Digital Fraud Risk observations misclassified as High and 59 Low Digital Fraud Risk observations correctly classified. Thus, the SMOTE based RBF SVM is best characterized as highly sensitive for screening purposes, whereas the SMOTE based calibrated linear SVM provides a more moderate balance between sensitivity and false alarms.

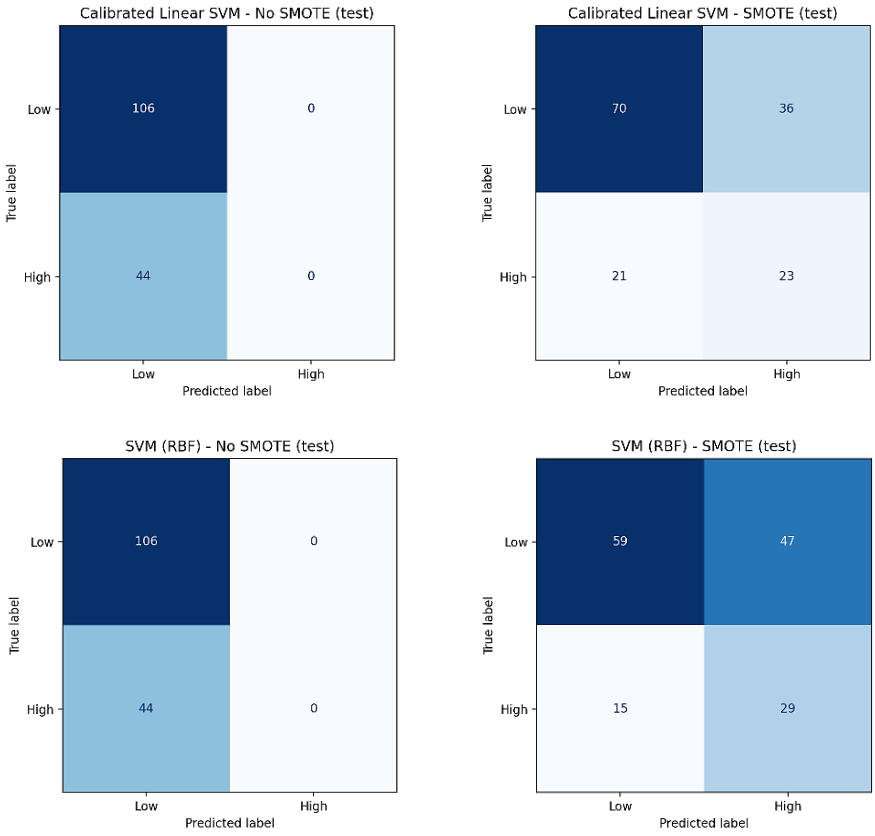


Figure 6. Confusion Matrix of Calibrated Linear SVM and RBF SVM No SMOTE vs SMOTE

TIN: Terapan Informatika Nusantara
Vol 6, No 11, April 2026, page 2106-2117
ISSN 2722-7987 (Media Online)
Website https://ejurnal.seminar-id.com/index.php/tin
DOI 10.47065/tin.v6i11.9589

3.4 Classification Result Using XGBoost

The confusion matrix using XGBoost can be seen in Figure 7. XGBoost in no SMOTE condition also failed to identify High Digital Fraud Risk, producing 105 true negatives, 1 false positive, 44 false negatives, and zero true positives. This again reflects majority class dominance and the practical limitation of relying on nominal accuracy under imbalance. After SMOTE, XGBoost correctly identified 15 High Digital Fraud Risk observations and misclassified 29 as Low, while misclassifying 30 Low Digital Fraud Risk observations as High and correctly classifying 76 Low Digital Fraud Risk observations.

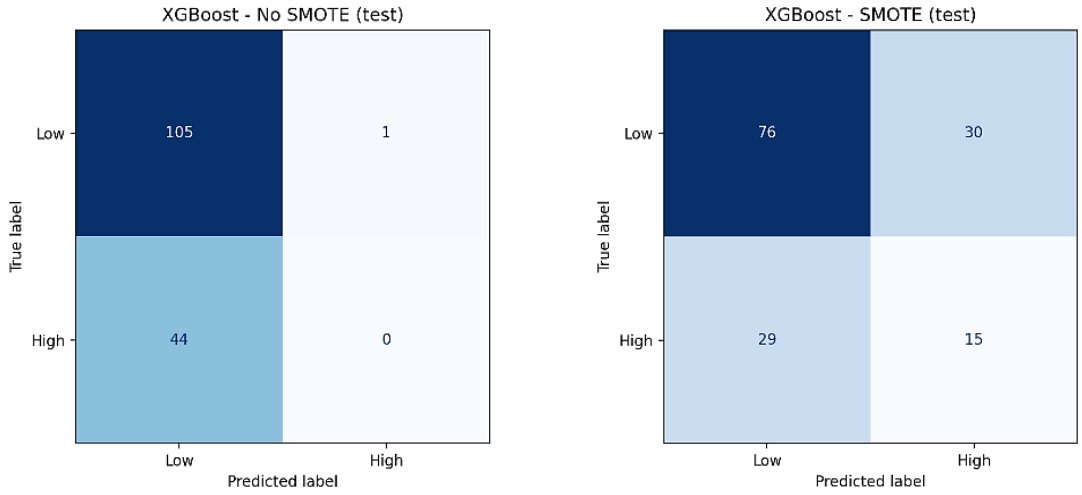


Figure 7. Confusion Matrix of XGBoost No SMOTE vs SMOTE

3.5 Comparative Summary

Across all evaluated algorithms, SMOTE consistently increased the number of High Digital Fraud Risk observations detected on the test set and reduced the number of High Digital Fraud Risk observations incorrectly predicted as Low. The primary cost was an increase in Low Digital Fraud Risk observations predicted as High, which is an expected consequence of shifting the decision boundary to improve minority sensitivity. From an applied perspective, model choice should therefore be driven by the operational objective. If the task is early screening where missed High Digital Fraud Risk observations are particularly costly, the SMOTE based RBF SVM provides the strongest sensitivity. If the objective is a more balanced performance profile that improves minority detection while limiting false alarms, the SMOTE based Random Forest offers a more practically stable compromise, consistent with the comparative metrics reported earlier.

Table 3. Performance Evaluation of No SMOTE and SMOTE of All Models

Model	Accuracy		Balance Accuracy		Precision		Recall		F1 Score	
	No-SMOTE	SMOTE	No-SMOTE	SMOTE	No-SMOTE	SMOTE	No-SMOTE	SMOTE	No-SMOTE	SMOTE
Random Forest	0.680	0.667	0.501	0.578	0.300	0.421	0.068	0.364	0.111	0.390
Gaussian Naive Bayes	0.680	0.640	0.514	0.592	0.357	0.404	0.114	0.477	0.172	0.438
SVM (RBF)	0.707	0.587	0.500	0.608	0.000	0.382	0.000	0.659	0.000	0.483
Calibrated Linear SVM	0.707	0.620	0.500	0.592	0.000	0.390	0.000	0.523	0.000	0.447
XGBoost	0.700	0.607	0.495	0.529	0.000	0.333	0.000	0.341	0.000	0.337

Table 3. Performance Evaluation of No SMOTE and SMOTE of All Models (Continued)

Model	ROC-AUC		PR-AUC		MCC		Kappa	
	No-SMOTE	SMOTE	No-SMOTE	SMOTE	No-SMOTE	SMOTE	No-SMOTE	SMOTE
Random Forest	0.549	0.608	0.369	0.415	0.004	0.163	0.003	0.163
Gaussian Naive Bayes	0.617	0.614	0.370	0.369	0.045	0.177	0.036	0.175

TIN: Terapan Informatika Nusantara
Vol 6, No 11, April 2026, page 2106-2117
ISSN 2722-7987 (Media Online)
Website https://ejurnal.seminar-id.com/index.php/tin
DOI 10.47065/tin.v6i11.9589

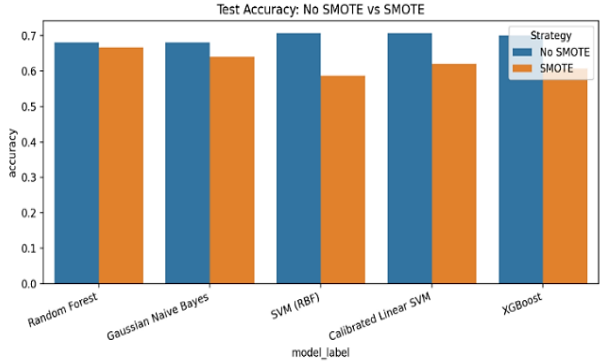
Model	ROC-AUC		PR-AUC		MCC		Kappa	
	No-SMOTE	SMOTE	No-SMOTE	SMOTE	No-SMOTE	SMOTE	No-SMOTE	SMOTE
SVM (RBF)	0.416	0.617	0.283	0.363	0.000	0.196	0.000	0.178
Calibrated Linear SVM	0.630	0.614	0.388	0.384	0.000	0.171	0.000	0.167
XGBoost	0.533	0.532	0.313	0.330	-0.053	0.058	-0.013	0.058

From Table 4, the test results clarify the practical effect of oversampling. No SMOTE produced the highest nominal accuracy in the calibrated linear SVM and RBF SVM (both 0.707), and the highest raw test ROC-AUC in the no-SMOTE calibrated linear SVM (0.630). However, both No-SMOTE SVM variants failed to identify any High cases at the default threshold, yielding zero recall, zero F1-score, and zero precision for that class. In contrast, SMOTE increased recall across all five models and improved balanced accuracy, precision, and F1-score in every model. The largest recall gain was observed in the RBF SVM, rising from 0.000 to 0.659, together with balanced accuracy increasing from 0.500 to 0.608, precision from 0.000 to 0.382, F1-score from 0.000 to 0.483, PR AUC from 0.283 to 0.363, and MCC from 0.000 to 0.196. For Random Forest, SMOTE improved balanced accuracy from 0.501 to 0.578, precision from 0.300 to 0.421, recall from 0.068 to 0.364, F1-score from 0.111 to 0.390, ROC-AUC from 0.549 to 0.608, PR AUC from 0.369 to 0.415, MCC from 0.004 to 0.163, and Cohen's Kappa from 0.003 to 0.163, while accuracy decreased only slightly from 0.680 to 0.667. More broadly, Table 4 shows that SMOTE shifted balanced summaries such as balanced accuracy, F1-score, MCC, and Cohen's Kappa in a more favorable direction across the five evaluated models. Within the SMOTE strategy, Random Forest achieved the highest test PR AUC, whereas the SMOTE-based RBF SVM achieved the highest recall.

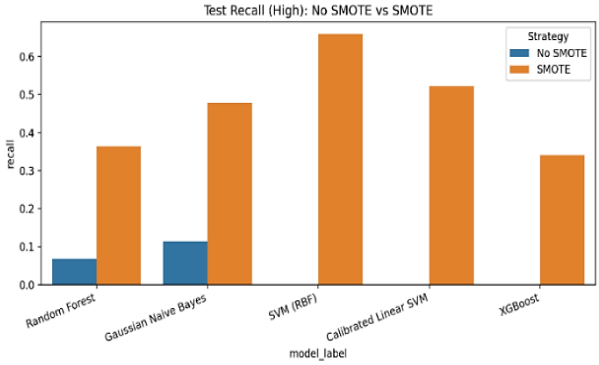
Figure 8 provides a compact visual comparison of test set performance under the no SMOTE and SMOTE across the five tuned classifiers. As seen in Figure 8, the effect of SMOTE is primarily expressed as a shift in minority class detection, rather than a uniform improvement across all metrics. In accuracy performance (Figure 8a), models trained without SMOTE exhibit consistently higher nominal accuracy than their SMOTE counterparts. This pattern is expected under class imbalance because accuracy is strongly influenced by performance on the majority of Low Digital Fraud Risk class. The modest reduction in accuracy after SMOTE therefore reflects a deliberate redistribution of classification errors, where improving sensitivity to the minority High Digital Fraud Risk class necessarily increases the number of Low observations predicted as High. This tradeoff is most clearly visible in the recall for the High Digital Fraud Risk class. Across all five models, SMOTE yields a substantial increase in recall relative to the no SMOTE condition. In several baseline models, recall (Figure 8b) is near zero, indicating that the classifier rarely, or never, predicts the High class at the default threshold. After SMOTE, recall improves markedly for every algorithm, with the largest gain observed for the RBF SVM. This confirms that oversampling effectively counteracts the majority class dominance observed in the baseline setting and enables the classifiers to recognize minority class patterns that were underrepresented during training.

The ROC AUC (Figure 8c) shows a more heterogeneous response. SMOTE improves ROC AUC for Random Forest and RBF SVM, while changes are comparatively small for GNB, calibrated linear SVM, and XGBoost. This is consistent with the interpretation that oversampling primarily improves class specific sensitivity and decision boundary formation but does not necessarily translate into uniformly higher-ranking performance across all models.

The PR AUC (Figure 8d) provides the most application relevant summary under imbalance, as it emphasizes performance on the minority positive class. Here, SMOTE yields improvements for several models, most notably Random Forest and RBF SVM, whereas changes are marginal for GNB and calibrated linear SVM. Within the SMOTE strategy, Random Forest achieves the strongest PR AUC, indicating the most favorable precision recall trade off when ranking High Digital Fraud Risk observations. At the same time, the model with the highest recall under SMOTE is the RBF SVM, highlighting that the most sensitive classifier is not necessarily the best in precision weighted evaluation.

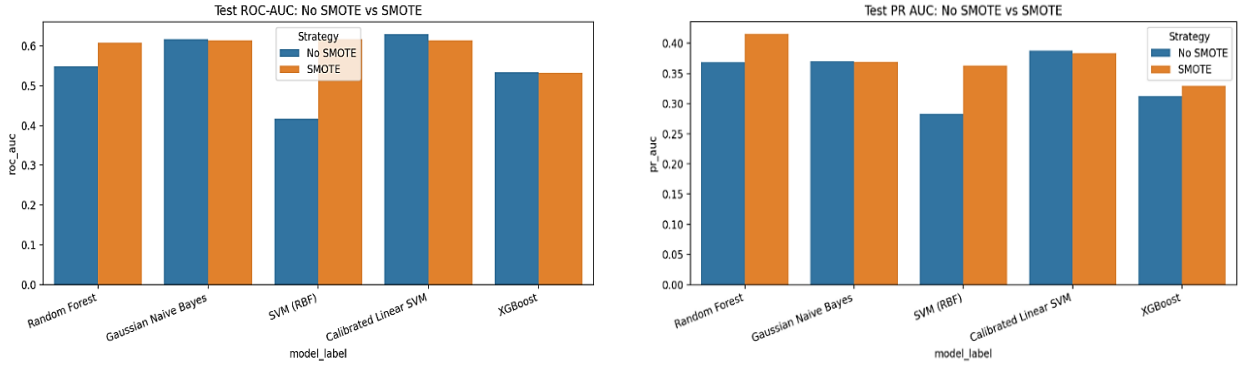


(a)



(b)

TIN: Terapan Informatika Nusantara
Vol 6, No 11, April 2026, page 2106-2117
ISSN 2722-7987 (Media Online)
Website https://ejurnal.seminar-id.com/index.php/tin
DOI 10.47065/tin.v6i11.9589



(c) (d)

Figure 8. Comparison (a) Accuracy, (b) Recall, (c) ROC-AUC and (d) PR AUC between no-SMOTE and SMOTE strategies across models in test dataset

3.6 Diagnostic Analysis

Two complementary diagnostics were performed for the selected SMOTE-based Random Forest, because this model provided the strongest precision-recall performance within the SMOTE strategy and therefore best represented the paper's SMOTE-focused comparison. Table 4 summarizes the main diagnostic results used to evaluate whether the selected model captured non-random signal and whether additional training data might improve performance.

Table 4. Summary of diagnostic results for the selected SMOTE-based Random Forest

Diagnostic measure	Value
Observed test ROC–AUC	0.608
Observed test PR–AUC	0.415
Permutation p-value (ROC–AUC)	0.020
Permutation p-value (PR–AUC)	0.020
Learning-curve mean ROC–AUC at 20% training fraction	0.536
Learning-curve mean ROC–AUC at 100% training fraction	0.608
Learning-curve mean PR–AUC at 20% training fraction	0.336
Learning-curve mean PR–AUC at 100% training fraction	0.415

A permutation test was used to benchmark the observed performance against a null distribution obtained by randomly shuffling the outcome labels and refitting the model. The observed test ROC–AUC (0.608) and PR–AUC (0.415) exceeded nearly all values under the label-shuffled baseline. Specifically, only 2 of 100 permutations matched or exceeded the observed ROC–AUC, and 2 of 100 matched or exceeded the observed PR–AUC, yielding empirical p-values of 0.020 for both metrics (Table 4). These results provide evidence that the model captures predictive structure beyond random label assignment. However, the absolute magnitude of ROC–AUC and PR–AUC remains moderate, and the permutation evidence should therefore be interpreted as supporting the presence of real signal rather than implying strong discriminative performance.

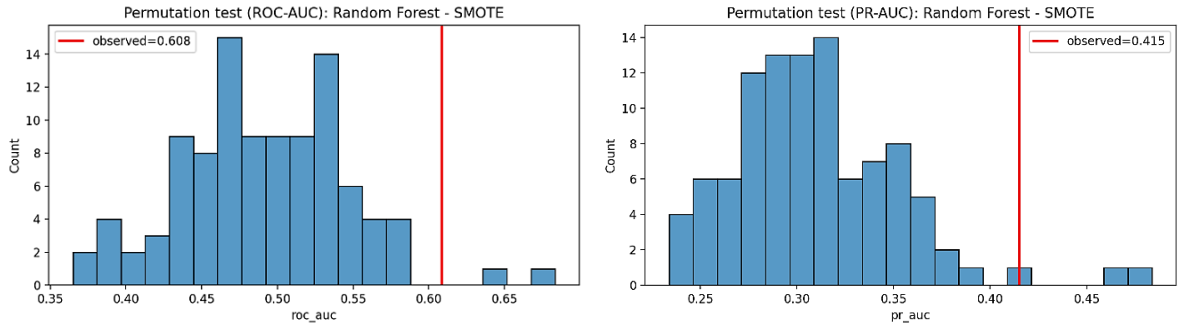


Figure 9. Permutation test for ROC–AUC and PR-AUC in the selected SMOTE-based Random Forest

As shown in Figure 9, the observed ROC–AUC falls in the extreme right tail of the permutation distribution, consistent with the empirical p-value reported in Table 4. Figure 9 supports the same conclusion for PR–AUC that the observed value remains above almost all permuted outcomes. It means that performance is unlikely to be attributable to chance alone.

TIN: Terapan Informatika Nusantara
Vol 6, No 11, April 2026, page 2106-2117
ISSN 2722-7987 (Media Online)
Website https://ejurnal.seminar-id.com/index.php/tin
DOI 10.47065/tin.v6i11.9589

The learning-curve analysis showed gradual improvement as the effective training fraction increased. Mean ROC-AUC rose from 0.536 at the 20% fraction to 0.608 at the full training fraction, while mean PR AUC increased from 0.336 to 0.415 (Figure 10). These gains are directionally favorable and consistent with the expectation that larger training sets can improve model stability. However, the increments are not large, which suggests that additional observations alone may not be sufficient to produce strong discrimination unless richer predictors are also introduced. In this sense, the learning-curve pattern is reasonable for a survey-based study. It supports the validity of the modeling approach while also making clear that the present feature set imposes substantive limits on predictive performance.

Figure 10 shows that ROC-AUC improved as the effective training fraction increased, although the slope became flatter at higher fractions. Figure 9 shows a similar pattern for PR AUC, with gains from smaller to larger training fractions but no evidence of a sharp late-stage increase. Read together with Table 4, Figures 10 suggest that additional data could improve performance, but likely not enough to overcome the present limitations of the feature set on their own.

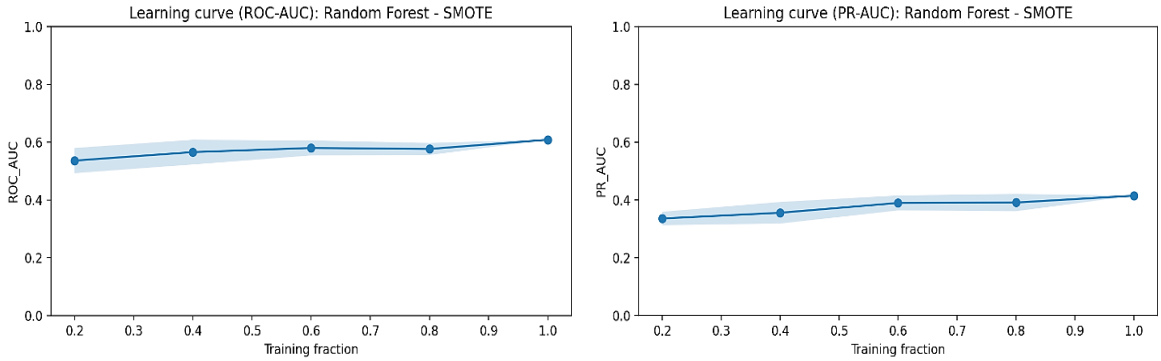


Figure 10. Learning curve for ROC-AUC and PR-AUC in the selected SMOTE-based Random Forest

3.7 Discussion

The main contribution of this study lies in the explicit comparison between no-SMOTE and SMOTE strategies for classifying Low and High digital fraud risk from UT students. Across the reported results, the comparison consistently shows that the choice of sampling strategy materially affected model behavior. In particular, SMOTE improved the model's ability to identify the minority class, whereas several baseline models appeared favorable only when performance was assessed mainly using overall accuracy.

This finding is important because the no-SMOTE models could achieve relatively high accuracy while still failing to identify high-risk digital fraud risk. The clearest examples were the no-SMOTE RBF SVM and calibrated linear SVM, both of which reached 0.707 accuracy but produced zero recall for the high-risk class. In imbalanced risk classification, this pattern reflects the well-known "accuracy paradox": a model can perform well on aggregate while failing on the digital fraud risk of greatest practical concern. Accordingly, these findings reinforce the methodological position that accuracy should not be treated as the primary criterion for model selection in imbalanced classification settings, where the operational objective is to reduce missed detections among minority high-risk digital fraud risk.

Within this comparative framework, SMOTE demonstrated its strongest advantage in minority-class sensitivity. Recall increased in all five evaluated models after oversampling, and PR AUC also improved in several cases, especially for Random Forest, the RBF SVM, and XGBoost. The SMOTE-based Random Forest provided the most balanced overall result for the study objective, improving recall from 0.068 to 0.364, ROC-AUC from 0.549 to 0.608, and PR AUC from 0.369 to 0.415, while accuracy declined only slightly from 0.680 to 0.667. Collectively, these results indicate that, for the present survey dataset, SMOTE is particularly useful when the practical goal is improved identification of students in the High Digital Fraud Risk category, rather than maximizing nominal accuracy.

This result is consistent with recent studies reporting that SMOTE-based oversampling is especially valuable when the main objective is stronger minority-class detection rather than maximization of nominal accuracy alone. Chen et al., (2024) note that recent imbalanced-learning research continues to show the usefulness of oversampling for improving minority-class performance, while Khalid et al., (2024) and Sayegh et al., (2024) report that SMOTE-supported modeling improved the detection of minority or attack classes in applied classification settings. Thus, the present findings align with the broader literature show that the practical benefit of SMOTE is most visible in recall-oriented or minority-focused evaluation.

At the same time, the results also show that SMOTE should be interpreted as a targeted improvement rather than a universal solution. Some no-SMOTE models retained higher accuracy, and the no-SMOTE calibrated linear SVM achieved the highest raw test ROC-AUC. For GNB and the calibrated linear SVM, PR AUC changed only marginally after oversampling. Accordingly, the evidence from this study supports a conditional conclusion: SMOTE is preferable when emphasis is placed on minority detection, but its benefit is less decisive when the evaluation focus is limited to overall accuracy or isolated ranking metrics.


TIN: Terapan Informatika Nusantara
Vol 6, No 11, April 2026, page 2106-2117
ISSN 2722-7987 (Media Online)
Website https://ejurnal.seminar-id.com/index.php/tin
DOI 10.47065/tin.v6i11.9589

The broader metric set reported in Table 4 further clarifies this conclusion. PR AUC is especially relevant in imbalanced settings because it focuses attention on performance for the minority class, while MCC and Cohen's Kappa provide more balanced summaries than accuracy alone. When these metrics are considered together, the SMOTE-based Random Forest emerges as the most practical model in the present study, not because it is perfect, but because it offers the strongest balance between minority detection and overall predictive performance.

From a substantive perspective, the results indicate that the survey-based predictors contain meaningful signal related to Digital Fraud Risk. The permutation test results indicate that the selected model captured non-random signal, and the learning-curve results suggest that the available data already support stable, if still moderate, predictive patterns. This is an important finding for survey-based fraud-risk research because it shows that financial literacy, digital financial literacy, age, income, and job tenure can function as relevant indicators of vulnerability even without the richer behavioral detail typically available in transaction-level datasets.

Nevertheless, the remaining performance gap should be interpreted carefully. Rather than implying failure, it more plausibly reflects the practical limits of prediction from a compact survey-based design. The train-test differences in the tree-based models also suggest that richer predictors and broader samples would likely improve generalization. The train–test performance differences observed in the tree-based models also suggest that broader samples and richer predictors may improve generalization. Thus, the study contributes both (i) a methodological insight: SMOTE improves minority-case detection in this setting; and (ii) a substantive insight: survey-based student characteristics contain usable, though incomplete, information for Digital Fraud Risk classification.

4. CONCLUSIONS

This study compared tuned classifiers trained with and without SMOTE for Low-High digital fraud-risk classification using survey data from Universitas Terbuka students. The results reveal that the main effect of SMOTE was not universal improvement on every metric, but a more reliable ability to detect the minority high-risk class. In contrast, several no-SMOTE models retained higher nominal accuracy while failing to identify any high-risk, which confirms that accuracy alone is not sufficient for evaluating imbalanced fraud-risk classification. Within the SMOTE strategy, Random Forest provided the most balanced practical performance by achieving the highest test PR AUC, while the SMOTE-based RBF SVM achieved the highest recall. Overall, the findings indicate that the survey variables contain usable and non-random information about digital fraud risk, even though the resulting class separation remains moderate. Accordingly, in this survey-based context, SMOTE is most appropriate when the operational objective prioritizes improved identification of High-risk students rather than the maximization of nominal accuracy. Future research should extend this work by incorporating richer and more behaviorally proximal predictors, expanding the sample size, and adopting repeated validation schemes to improve the robustness of performance estimates. In addition, threshold optimization and cost-sensitive learning may further enhance minority-class detection and strengthen model generalizability in real-world deployments.

REFERENCES

Ali, A., Abd Razak, S., Othman, S. H., Eisa, T. A. E., Al-Dhaqm, A., Nasser, M., Elhassan, T., Elshafie, H., & Saif, A. (2022). Financial Fraud Detection Based on Machine Learning: A Systematic Literature Review. *Applied Sciences*, *12*(19), 9637. https://doi.org/10.3390/app12199637

Bhaduri, D., Toth, D., & Holan, S. H. (2025). A Review of Tree-Based Methods for Analyzing Survey Data. *WIREs Computational Statistics*, *17*(1). https://doi.org/10.1002/wics.70010

Breiman, L. (2001). *Random Forests*. *45*, 5–32.

Carvalho, M., Pinho, A. J., & Brás, S. (2025). Resampling approaches to handle class imbalance: a review from a data perspective. *Journal of Big Data*, *12*(1), 71. https://doi.org/10.1186/s40537-025-01119-4

Chen, W., Yang, K., Yu, Z., Shi, Y., & Chen, C. L. P. (2024). A survey on imbalanced learning: latest research, applications and future directions. *Artificial Intelligence Review*, *57*(6), 137. https://doi.org/10.1007/s10462-024-10759-6

Choung, Y., Chatterjee, S., & Pak, T.-Y. (2023). Digital financial literacy and financial well-being. *Finance Research Letters*, *58*, 104438. https://doi.org/10.1016/j.frl.2023.104438

Elreedy, D., Atiya, A. F., & Kamalov, F. (2024). A theoretical distribution analysis of synthetic minority oversampling technique (SMOTE) for imbalanced learning. *Machine Learning*, *113*(7), 4903–4923. https://doi.org/10.1007/s10994-022-06296-4

Gao, X., Xie, D., Zhang, Y., Wang, Z., Chen, C., He, C., Yin, H., & Zhang, W. (2026). A comprehensive survey on imbalanced data learning. *Frontiers of Computer Science*, *20*(11), 2011622. https://doi.org/10.1007/s11704-025-50274-7

Guido, R., Ferrisi, S., Lofaro, D., & Conforti, D. (2024). An Overview on the Advancements of Support Vector Machine Models in Healthcare Applications: A Review. *Information*, *15*(4), 235. https://doi.org/10.3390/info15040235

TIN: Terapan Informatika Nusantara
Vol 6, No 11, April 2026, page 2106-2117
ISSN 2722-7987 (Media Online)
Website https://ejurnal.seminar-id.com/index.php/tin
DOI 10.47065/tin.v6i11.9589

Hairani, H., Widiyaningtyas, T., & Dwi Prasetya, D. (2024). Addressing Class Imbalance of Health Data: A Systematic Literature Review on Modified Synthetic Minority Oversampling Technique (SMOTE) Strategies. *JOIV : International Journal on Informatics Visualization*, *8*(3), 1310. https://doi.org/10.62527/joiv.8.3.2283

Khalid, A. R., Owoh, N., Uthmani, O., Ashawa, M., Osamor, J., & Adejoh, J. (2024). Enhancing Credit Card Fraud Detection: An Ensemble Machine Learning Approach. *Big Data and Cognitive Computing*, *8*(1), 6. https://doi.org/10.3390/bdcc8010006

Kivrak, M., Avci, U., Uzun, H., & Ardic, C. (2024). The Impact of the SMOTE Method on Machine Learning and Ensemble Learning Performance Results in Addressing Class Imbalance in Data Used for Predicting Total Testosterone Deficiency in Type 2 Diabetes Patients. *Diagnostics*, *14*(23), 2634. https://doi.org/10.3390/diagnostics14232634

Leviany, F., Kasmiarno, K. S., & Fitriana, I. N. L. (2025). Predicting Digital Fraud Risk Using Support Vector Machine Classifier A Case Study Of Universitas Terbuka Students. *Proceeding of The International Seminar on Business, Economics, Social Science and Technology (ISBEST)*, 54–60. https://doi.org/10.33830/isbest.v5i1.7407

Lokanan, M., & Liu, S. (2021). Predicting Fraud Victimization Using Classical Machine Learning. *Entropy*, *23*(3), 300. https://doi.org/10.3390/e23030300

Malhotra, R., & Lata, K. (2022). Handling class imbalance problem in software maintainability prediction: an empirical investigation. *Frontiers of Computer Science*, *16*(4), 164205. https://doi.org/10.1007/s11704-021-0127-0

Pantic, I. V., Paunovic Pantic, J., Valjarevic, S., Corridon, P. R., & Topalovic, N. (2025). Artificial intelligence – based approaches based on random forest algorithm for signal analysis: Potential applications in detection of chemico - biological interactions. *Chemico-Biological Interactions*, *418*, 111624. https://doi.org/10.1016/j.cbi.2025.111624

Salman, H. A., Kalakech, A., & Steiti, A. (2024). Random Forest Algorithm Overview. *Babylonian Journal of Machine Learning*, *2024*, 69–79. https://doi.org/10.58496/BJML/2024/007

Saputra, D., 'Alauddin, A. A. F., & Azizan, M. (2025). Comparative Analysis of Gaussian Naïve Bayes and Categorical Naïve Bayes Algorithms with Laplace Smoothing in COVID-19 Detection. *Jurnal Ilmu Komputer Dan Informatika*, *5*(1), 69–78. https://doi.org/10.54082/jiki.286

Sayegh, H. R., Dong, W., & Al-madani, A. M. (2024). Enhanced Intrusion Detection with LSTM-Based Model, Feature Selection, and SMOTE for Imbalanced Data. *Applied Sciences*, *14*(2), 479. https://doi.org/10.3390/app14020479

Sulaiman, B. R., Schetinin, V., & Sant, P. (2022). Review of Machine Learning Approach on Credit Card Fraud Detection. *Human-Centric Intelligent Systems*, *2*(1–2), 55–68. https://doi.org/10.1007/s44230-022-00004-0

Wibowo, P., & Fatichah, C. (2021). An in-depth performance analysis of the oversampling techniques for high-class imbalanced dataset. *Register: Jurnal Ilmiah Teknologi Sistem Informasi*, *7*(1), 63. https://doi.org/10.26594/register.v7i1.2206

Xiao, X., Li, X., & Zhou, Y. (2022). Financial literacy overconfidence and investment fraud victimization. *Economics Letters*, *212*, 110308. https://doi.org/10.1016/j.econlet.2022.110308